fMRI data analysis : statistics, information and dynamics

Bertrand Thirion

THÈSE

présentée à

INRIA SOPHIA-ANTIPOLIS

pour obtenir le titre de

DOCTEUR EN SCIENCES

École Doctorale Télécom Paris

Spécialité

Signal & Image

soutenue par

Bertrand Thirion

le 1^{er} octobre 2003

Titre

fMRI data analysis: statistics, information and dynamics

Analyse de données d' IRM fonctionnelle: statistiques, information et dynamique

Directeur de thèse : Olivier Faugeras

Remerciements

Je voudrais tout d'abord remercier les membres du jury de cette soutenance, les rapporteurs Jean-Baptiste Poline et Lars Kai Hansen, ainsi qu'Isabelle Bloch, Nicolas Ayache et Yves Burnod. Outre leur compétence, leur indulgence, et l'intérêt qu'ils ont accepté de porter à ce travail, c'est grâce à eux que cette thèse n'est pas simplement un rapport technique mais une *publication*.

Cette thèse est la conclusion de trois années de travail au sein du projet Robotvis/Odyssée. Je suis très heureux de ce temps passé à Sophia-Antipolis, et de tout ce que j'ai pu y apprendre.

Plus que ces choses, ce sont en fait les membres de l'équipe qui resteront présents à ma mémoire: ceux qui m'ont suggéré des voies et parfois mis sur la piste -Gerardo, Christophe, Thierry; ceux qui m'ont donné les coups de main rapides ou sérieux, et sans qui l'informatique ne ressemblerait qu'à un triste épouillage - Théo, Robert, Lionnel, Jean-Philippe ; ceux qui -avec autant de mérite- m'ont aidé au delà de l'informatique stricte -Marie-Cécile, Nicolas -qui fut en outre un voisin très sympathique, David ; ceux qui ont régulièrement partagé leurs économies et leurs discussions lors des pauses goûter et/ou autres apéritifs - en plus des précédents, Mickaël, Jan, Emmanuel, Maureen, Jacques, Fred, Florent, Christophe, Matthieu, Pierre et Nour; Diane, pour les petits-déjeuners, le canyoning et le ski de rando. Je dois une reconnaissance plus grande encore à Olivier qui, avec courage et rigueur m'a entraîné dans l'approche de ces disciplines nouvelles et m'a soutenu tout au long de cette thèse -et surtout à l'heure des choix pour le futur. Par ailleurs, je n'oublie pas le personnel de l'INRIA, dont l'efficacité et le dévouement offre des moyens remarquables et un cadre de travail bien agréable au quotidien. Avant de quitter ces murs, je salue aussi David, Fabien, Aubin et tous ceux avec qui j'ai partagé un peu de ma vie, au travail ou ailleurs.

Ma soutenance est aussi un enterrement, celle d'une vie d'étudiant, passionnante, heureuse et ... longue. Je remercie mes parents qui lui ont donné l'impulsion initiale, et m'ont permis de réaliser des choses que ni eux ni moi ne projetions. Graces leur soit rendues pour la confiance et les moyens qu'ils m'ont offerts généreusement. S'ils ne m'ont pas eu trop longtemps à charge, ils ont eu néanmoins le mérite de supporter - au sens français et anglais- un fils pas toujours commode.

Quant à ce dernier point, ils sont néanmoins devancés par Marie, qui fut la première et la plus exigeante lectrice du présent manuscrit, se faisant ainsi pardonner de m'avoir donné quelques distractions lors de ma première année de thèse, et un nouveau co-locataire un peu bruyant lors de la deuxième année.

A Marie,
A Irénée,
A ε',

Résumé

Dans cette thèse, nous discutons et proposons un certains nombre de méthodes pour l'analyse de données d'IRM -imagerie par résonance magnétique- fonctionnelle. L'IRM fonctionnelle est une modalité récente de l'exploration du cerveau: elle produit des séquences d'images reflétant l'activité métabolique locale, celle-ci reflétant l'activité neuronale. Nous nous intéressons tout d'abord à la modélisation des séries temporelles obtenues pour chaque voxel séparément, en faisant appel aux techniques de prédiction linéaire et au calcul de l'information des processus modélisés. Nous étudions ensuite différentes généralisations multivariées de ce modèle. Après avoir rappelé et discuté certaines techniques classiques (analyse en composantes indépendantes, regroupement), nous proposons successivement une approche linéaire fondée sur la théorie des systèmes à état et une approche non-linéaire fondée sur les décompositions à noyau. Le but commun de ces méthodes -qui peuvent se compléter- est de proposer des décompositions qui préservent au mieux la dynamique des données. Nous introduisons ensuite une approche nouvelle par réduction de la dimension des données; cette approche offre une représentation plus structurée et relativement agréable à visualiser. Nous montrons ses avantages par rapport aux techniques linéaires classiques. Enfin, nous décrivons une méthodologie d'analyse qui synthétise une grande partie de ce travail, et repose sur des hypothèses très souples. Nos résultats offrent ainsi une description globale des processus dynamiques qui sont mis en image lors des expériences d'IRM fonctionnelle.

Abstract

In this thesis, we discuss and propose several methods for functional MRI -magnetic resonance imaging- data analysis. Functional MRI is a recent modality for the study of brain function: it produces image sequences that reflect local brain metabolic activity, which in turn reflects neural activity. We first deal with the modeling of each voxel-based temporal pattern, using linear prediction techniques and estimating the information contained in the temporal processes. We then study different multivariate generalizations of this model. After recalling and discussing several classical methods (independent components analysis, clustering), we first propose a linear approach based on state-space modeling, and then a non-linear approach based on kernel decompositions. The common objective of these methods -that are nevertheless complementary- is to propose decompositions that preserve optimally the data dynamics. We then introduce a novel point of view based on dimension reduction of the dataset, which allows for a more structured representation and helps for visualization. We show its effectiveness with respect to classical linear decomposition techniques. Finally, we describe a methodology of analysis that synthesizes different parts of this work, and relies on soft hypotheses. Our results give a global description of the dynamical processes embedded in the data produced by functional MRI.

Contents

List of Figures

List of Tables

"Compter voir la vérité sortir de la pensée revient à confondre le boin de pensée avec l'appétit de savoir."
Hannah Arendt,
La vie de l'esprit I. La pensée.

" Arithmétique ! Algèbre ! Géométrie ! Trinité grandiose ! Triangle lumineux ! Celui qui ne vous a pas connues est un insensé ! "
Lautréamont,
Les chants de Maldoror. Chant deuxième.

Chapter 1

Introduction

"Compter voir la vérité sortir de la pensée revient à confondre le besoin de pensée avec l'appétit de savoir."
Hannah Arendt,
La vie de l'esprit I. La pensée.

" Arithmétique ! Algèbre ! Géométrie ! Trinité grandiose ! Triangle lumineux ! Celui qui ne vous a pas connues est un insensé ! "
Lautréamont,
Les Chants de Maldoror. Chant deuxième.

Dans cette introduction, nous allons expliquer la problématique que nous avons suivie dans ce travail. Supposant d'abord que le lecteur connaît l'IRM fonctionnelle et les possibilités qu'elle offre pour l'exploration du cerveau, nous commencerons par expliquer notre point de vue et notre démarche par rapport au vaste problème de l'analyse de données d'IRM fonctionnelle. Cela étant, nous dresserons une description de la suite de ce rapport. Enfin, nous essaierons de rendre compte des problèmes inexplorés au cours de notre travail.

1.1 Pourquoi parler de statistiques? De dynamique? D'information?

1.1.1 Estimation et inférence

Soulignons tout d'abord que l'analyse de données d'IRM fonctionnelle (IRMf) est motivée par les possibilités que cette modalité apporte dans la connaissance de l'activité du cerveau. Techniquement, on dira que les données d'IRMf permettent d'effectuer des inférences sur le fonctionnement de l'activité du cerveau. Ceci recouvre entre autres la

localisation des principales zones fonctionnelles, l'étude des co-occurences de signaux au sein du cortex, ou la structure temporelle de ces signaux. Mais si nous parlons d'inférence, c'est pour indiquer que ce travail est produit à partir d'hypothèses sur les mécanismes cognitifs, moteurs ou sensoriels étudiés. Avant d'être une modalité d'exploration, l'IRMf est une modalité de validation.

On comprend dès lors que les premiers efforts en analyse de données (disons depuis 1990-1993 [11]) aient porté sur l'établissement d'un cadre statistique rigoureux pour permettre l'inférence [173]. De fait, un standard et un logiciel d'analyse se sont imposés sous le vocable SPM (Statistical Parametric Mapping). Toute méthode d'inférence reposant sur des hypothèses précises concernant le signal, il a aussi fallu affiner progressivement la connaissance que l'on avait sur la structure du signal. C'est crucial, car des hypothèses inexactes rendent l'inférence inefficace. Ce problème a progressivement induit deux branches importantes dans le processus d'analyse [172]: l'exploration d'une part, l'estimation d'autre part (dans les deux cas, on peut également parler de détection, bien que ce moment n'ait pas précisément le même sens). L'exploration vise à mettre en évidence les structures essentielles du signal, tandis que l'estimation sert à trouver les paramètres pertinents pour capter les informations essentielles du signal et permettre l'inférence. Signalons d'ores et déjà que ces problèmes ne peuvent être réglés une fois pour toutes, étant donnés les différents modèles expérimentaux, la possibilité d'expérimenter sur différentes espèces animales, l'utilisation éventuelle d'agents de contraste, mais surtout la variabilité inter- et intra-individuelle.

La première possibilité à considérer pour résoudre les questions d'estimation est bien sûr d'inverser le processus connu de générations des images, de modéliser les sources d'artefacts et finalement d'identifier l'information. Cette démarche est toutefois peu répandue dans la communauté d'analyse, d'une part parce que les mécanismes physiologiques à l'origine du signal IRMf sont encore partiellement inconnus, d'autres part du fait de la multiplicité des sources vraisemblables du signal (tous les processus physiologiques du sujet étudié, la signature de la machine), enfin parce que la génération du signal est extrêmement complexe -autrement dit, le processus n'est pas *inversible*. Dans le cadre de notre travail, nous n'avions pas de possibilité simple d'aborder ces aspects, et avons donc préféré une approche a posteriori, fondée sur l'observation des images et la mise en évidence des causes vraisemblables. D'une certaine manière, nous avons donc ainsi considéré le système de stimulation et d'imagerie comme une machine inconnue dont seuls les données (*input*) et les sorties (*output*) étaient connues. Nous ne considérons donc pas les sources auxiliaires d'information (mesure du rythme cardiaque, du mouvement des yeux, *monitoring* divers).

Après ce premier développement, on conviendra qu'à ce jour les statistiques sont le moyen de produire un modèle à partir de données empiriques multiples. Dans notre cas, nous nous servirons de ces statistiques pour identifier les structures principales des données, non pour faire de l'inférence.

1.1.2 Modélisation multivariée, modélisation dynamique

Concrètement, les données d'IRMf constituent l'échantillonnage d'un certain volume à différents instants consécutifs. Il s'agit donc d'un ensemble à quatre dimensions (trois d'espace, une de temps). Il convient de s'arrêter un instant sur ce partage:

- Si on considère que les volumes sont recalés, chaque décours temporel décrit le signal au niveau d'un voxel (élément de volume élémentaire). En confrontant cette information à celle du paradigme expérimental, on pourra caractériser la réponse au niveau d'un voxel. Il s'ensuit que l'interprétation des données se fait essentiellement par examen des décours temporels. Ceci justifie à nos yeux l'introduction de la notion de dynamique, entendue comme l'étude des processus temporels. Plus prosaïquement, la dynamique consistera à considérer le problème embarrassant de la corrélation temporelle des données: une fois le signal d'intérêt retranché, celles-ci ressemblent fort peu à un bruit blanc; il faut donc prendre en compte leur structure temporelle complexe.

- Malheureusement, ce projet à peine esquissé se heurte à une double réalité: d'une part les données temporelles sont relativement courtes, une centaine, quelques centaines au plus d'échantillons par voxel, d'autre part un rapport signal à bruit relativement faible. La conjonction de ces deux faits empêche d'envisager une simple estimation et/ou exploration univariée -i.e. effectuée sur chaque voxel indépendamment.

Par ailleurs, la connaissance présente du cerveau confirme que le voxel est souvent trop petit pour caractériser une unité fonctionnelle du cortex; on a donc tout intérêt à effectuer le travail d'exploration et d'estimation sur des domaines un peu étendus, voire sur l'image entière. L'analyse multivariée est cet art qui consiste à profiter de la multiplicité des voxels pour bâtir un modèle fiable des phénomènes présents tout en mettant à jour des différences fonctionnelles d'une région à l'autre.

1.1.3 Qu'entend-on par *information* ?

L'établissement de modèles statistiques quantitatifs à partir de données empiriques se traduit assez naturellement en terme d'information, entendue au sens de la théorie de l'information: toute suite aléatoire, et plus généralement toute chaîne de caractères peut être vue comme une source d'information ayant des caractéristiques mesurables.

Un malentendu potentiel existe entre ce que le neurophysiologiste entendra par information, se référant au contenu des données, et ce que le modélisateur vise par ce terme, à savoir la caractérisation statistique d'un système complexe de données (on verra d'ailleurs que les notions d'information et de complexité sont très liées).

Nous essayerons toutefois de recouper ces deux points de vue: par exemple, on peut considérer que la quantité d'information contenue dans une série temporelle se définit par la possibilité de prédire sa valeur à des instants futurs- ce que nous considérons comme une propriété structurelle de cette série. Or précisément, si cette série est une fonction du paradigme expérimental, elle est parfaitement prédictible, c'est-à-dire qu'elle contient un maximum d'information, ou encore qu'elle est de complexité minimale. Finalement, l'emploi des concepts dérivés de la théorie de l'information se justifiera dès lors que l'on sera dans le cadre d'un modèle acceptable pour le neurophysiologiste. Aussi, lorsque nous tenterons de produire un modèle intelligible des données, nous ne manquerons pas d'inclure le paradigme expérimental, et nous réserverons tous nos commentaires à la part des données qui est expliquée par ce paradigme. Notons que l'on pourrait tenir un raisonnement très similaire sur le lien entre la structure spatiale très localisée des foyers d'activation et l'entropie réduite des cartes spatiales qui permettent de les décrire.

1.2 Détail de notre démarche

1.2.1 L'IRM fonctionnelle

L'IRM fonctionnelle (IRMf) est une modalité d'exploration du cerveau d'une très grande sophistication. Pour la comprendre, il faut rapprocher deux points de vue:

Point de vue physique Fondée sur le principe physique de la résonance magnétique nucléaire, elle permet de produire des images du cerveau d'une assez bonne précision (typiquement 3mm pour des images fonctionnelles) en un temps relativement bref (typiquement 3s pour un volume). Sans entrer dans le détail de la physique d'acquisition, nous rendons compte des phénomènes subtils qui permettent de produire des images à partir de la résonance des noyaux des molécules d'eau (voir 2.2).

Point de vue biologique Plus récemment, on a découvert que les images acquises étaient sensibles à l'activité métabolique locale. Cette activité entretient elle-même des liens assez complexes avec l'activité neuronale (synaptique) du cortex (voir 2.1). Ceci a fait de l'IRM un moyen de mesurer l'effet BOLD (blood oxygen-level dependent), et finalement de mesurer l'activité du cerveau dans les tâches motrices, cognitives ou sensorielles.

Pour achever ce tour d'horizon, introduisons deux concepts importants:

Le paradigme expérimental Il s'agit de la succession des états ou des stimulations qui caractérisent l'expérience du point de vue du sujet (voir 2.3.3). Il va de soi que l'interprétation de données se fait en fonction de ce paradigme. On conviendra de le représenter par un vecteur d'état (généralement binaire, mais des variations sont possibles).

Les pré-traitements L'exploitation des séquences d'images d'IRM fonctionnelle nécessite préalablement une séquence de pré-traitements: typiquement, une estimation du mouvement au cours de l'acquisition, si nécessaire une correction de ces mouvements, un recalage temporel des données pour compenser les décalages entre coupes d'acquisition, la normalisation spatiale des images, le lissage, l'élimination de signaux de basse fréquence (*detrending*). Nous faisons une critique rapide de ces méthodes et explicitons nos choix méthodologiques en 2.4.

1.2.2 Etat de l'art de l'analyse de données d'IRMf

Le deuxième chapitre vient classiquement exposer l'état de l'art sur notre sujet. Nous commençons par suggérer les difficultés techniques que l'analyse devra surmonter (section 3.1.1), puis nous explicitons la différence entre les méthodes inférentielles et les méthodes exploratoires (section 3.1.2). Reste ensuite à affronter la multiplicité -le foisonnement- des solutions existantes. Pour ne pas trop nous disperser, nous nous arrêtons sur la méthode standard, dite Statistical Parametric Mapping (SPM) (section 3.2.1), puis procédons par variations progressives autour de ce standard. Nous rappelons donc les contributions qui utilisent un modèle similaire, mais diffèrent sur des aspects techniques (section 3.2.2), puis les contributions qui s'appuient sur un modèle différent. Nous finissons par les méthodes multivariées (section 3.3), qui sont utilisées dans une approche exploratoire des données.

1.2.3 Modèle temporel de l'IRMf

Cela étant, nous nous penchons sur le problème essentiel de l'analyse de données en IRMf: la modélisation des effets temporels. Pour ce faire, nous revenons d'abord sur la phénoménologie de l'effet BOLD (section 4.1), et révisons les différentes modélisations proposées pour la réponse hémodynamique. Attendu que *i)* un certain nombre d'incertitudes subsistent sur la réponse "standard" *ii)* la variabilité semble être la règle, nous faisons le pari de la flexibilité pour nos modèles à venir. Reste ensuite à trouver un cadre général et cohérent pour l'étude du processus temporel qu'est un signal BOLD. Après quelques tâtonnements, nous adoptons (section 4.2) la décomposition de Wold comme modèle général. Un peu abusivement, nous identifions la composante déterministe de cette décomposition avec la fraction du signal liée au paradigme. Reste alors à mettre au point une procédure d'estimation. Rappelons encore que les séries temporelles sont courtes et bruitées; il faut donc paramétrer parcimonieusement le processus. Ce contrôle trouve une formalisation pratique avec la théorie de l'information, et notamment avec le principe dit *Minimum Description Length (MDL)* (section 4.3); ce dernier est en outre formellement identique au critère bayésien *BIC*, que nous développons dans l'appendice D et qui semble donner le modèle asymptotique le plus général de la complexité des modèles temporels.

Après cette première plongée dans la théorie de l'information, nous revenons (section 4.4) sur quelques problèmes généraux (normalité du bruit, linéarité de la réponse), mais repoussons les modèles trop généraux qui complexifient inutilement l'analyse, et réduisent l'efficacité des résultats.

1.2.4 Analyse multivariée et théorie de l'information

Une des difficultés de l'IRMf est que l'étude univariée (voxel par voxel) des données est à la fois difficile -à cause du faible rapport signal à bruit- insuffisante -car elle permet de constater des effets, non des interactions- et redondante - car les effets présents sont a priori nettement moins nombreux que les voxels. Une solution possible consiste à faire une estimation globale -au niveau de l'image- des effets présents: c'est ce en quoi consiste l'analyse multivariée. A certaines approches classiques mais limitées par des hypothèses sous-jacentes simplistes (analyse en composantes principales (PCA) ou coalescence (*clustering*)), nous préférons une rationalisation du problème par les concepts et la théorie de l'information. Nous appliquons ce point de vue à trois cas distincts:

Décomposition linéaire sous l'hypothèse gaussienne Nous revisitons en 5.1.4 l'analyse par corrélation canonique (CCA), qui est un modèle relativement simple et efficace, bien qu'évidemment limité par l'hypothèse gaussienne[1] . Il s'agit d'une adaptation de la PCA qui permet de prendre en compte certaines caractéristiques des données (corrélation spatiale ou temporelle, reproductibilité).

[1]La modélisation gaussienne est utile pour définir une hypothèse nulle sur le signal. Elle est donc fausse a priori lorsqu'il s'agit de rendre compte de la structure empirique des signaux, dans la mesure où l'on s'attend effectivement à des activations. Une mixture de deux gaussiennes au moins est alors nécessaire pour rendre compte des données.

Décomposition linéaire sous l'hypothèse non-gaussienne La décomposition linéaire des données sous l'hypothèse non-gaussienne conduit nécessairement au critère d'analyse en composantes indépendantes (ICA): c'est ce que nous montrons en partant d'une approche par maximum de vraisemblance (voir la section 5.2). Nous nous arrêtons un moment sur les techniques disponibles, étant donnée l'importance prise par l'ICA dans l'analyse de données d'IRM fonctionnelles. Nous rappelons aussi quelques limitations essentielles de l'ICA (communes à presque toutes les techniques multivariées): la non-considération des effets temporels (d'où des difficultés d'interprétation), et le choix du nombre de composantes à considérer dans la décomposition.

La quantification des données Parmi les méthodes non-linéaires, la quantification (ou *clustering*) semble bien adaptée pour décrire des données multimodales (au sens statistique). La méthode dite du goulot d'information (*information bottleneck*) fournit un cadre rigoureux et assez naturel pour le problème de la quantification des données (voir section 5.3). Comme elle ne peut s'appliquer qu'à des données de petite dimension, nous l'appliquons à des données pré-traitées, i.e. réduites à quelques coefficients associés à une matrice de dispersion. Nous montrons sur un exemple synthétique que cette méthode permet effectivement de séparer différents modes d'une distribution statistique des données.

1.2.5 Vers un modèle spatio-temporel des données d'IRMf

L'étape suivante consiste à réconcilier l'analyse temporelle et l'analyse multivariée. En l'espèce, il s'agit de modéliser le processus qui a généré les données, en prenant en compte la structure temporelle de ces données. Ceci est l'objet de l'analyse en composantes dynamiques (section 6.1.1). Cependant, encore une fois, il faut adapter l'analyse aux contraintes propres à l'IRMf: nous avons donc choisi une modélisation faisant intervenir un vecteur d'état (caché) qui décrit le processus générateur des données. Ce vecteur est lié aux données par une équation de mélange (*mixing*) ou d'observation. Se pose alors à nouveau la question de l'estimation des quantités impliquées dans le modèle. Nous décrivons d'abord une approche inspirée du filtrage de Kalman, mais utilisant une technique d'Expectation-Maximization (EM). Cependant, la bonne surprise vient d'un modèle linéaire assez simple (non itératif) qui permet d'estimer les mêmes quantités, avec une aussi bonne précision. Après avoir décrit cette solution en 6.2.2, nous nous penchons sur certains aspects techniques du problème (sections 6.2.2 et 6.2.3):

L'inclusion du paradigme expérimental Le modèle initial, purement auto-régressif, doit être adapté pour prendre en compte l'information du paradigme expérimental. Notons que le modèle résultant donne nécessairement une approximation médiocre de la réponse hémodynamique. Celle-ci aide pourtant à mieux caractériser les signaux d'activation dans le cas de l'analyse exploratoire.

L'estimation du rang du système Pour une fois, il est relativement facile de pouvoir estimer le rang du système, i.e. le nombre de composantes temporelles significativement structurées. Nous proposons pour cela un test par simulation et un test analytique, ce dernier reposant sur des hypothèses non validées; il convient tout de même dans la limite d'un petit nombre de composantes.

Une approche récursive Un problème de cette méthode est de reposer sur une réduction initiale des données par PCA. Nous montrons qu'en appliquant l'algorithme à des instants successifs de manière récursive, on peut utiliser plus de composantes dans la PCA initiale, limitant ainsi le risque de négliger des sources d'information présentes dans les données.

En fait la méthode des espaces-états peut s'appliquer à des cas très généraux (section 6.3): calcul du signal d'intérêt en un voxel à partir de données multi-session, estimation des principaux effets présents dans un ensemble de données, estimation de la réponse en un voxel en tenant compte des voisins...

1.2.6 Kernel PCA et mélange non linéaire

Nous avons décrit un certain nombre d'approches linéaires multivariées; il est tentant de regarder la possibilité d'analyses non-linéaires. L'intérêt d'icelles est de permettre des représentations redondantes (*overcomplete*) des données, permettant en particulier d'examiner finement une partie de l'espace des signaux. Techniquement, nous avons choisi la méthode d'analyse en composantes principales à noyau (*kernel PCA* ou *KPCA*) (section 7.1). L'avantage est que l'optimisation se fait par diagonalisation d'une matrice de covariance généralisée, évitant la minimisation d'un critère non convexe. Par ailleurs, selon le choix du noyau, la PCA apparaît comme un cas limite du modèle, permettant une interprétation claire de la non-linéarité. En somme, cette technique jouit d'un pouvoir descripteur bien supérieur aux techniques linéaires habituelles; toutefois, c'est au prix d'un coût élevé en calcul (la matrice de covariance a pour taille le nombre de voxels inclus), et le problème de la sélection du nombre de composantes significatives est encore plus compliqué que pour la PCA. Nous esquissons des solutions à ces problèmes en 7.2.

L'application de la méthode peut se faire sur les décours temporels d'un ensemble de données (section 7.3.1), ou en cascadant la méthode au modèle d'espace-état précédent. Dans ce cas, le choix d'un noyau polynomial simplifie la charge des calculs (section 7.3.2). Enfin, nous proposons une comparaison avec le modèle linéaire multivarié (section 7.3.3).

1.2.7 L'espace signal vu comme une variété: étude à partir des cartes Laplaciennes

On peut ne pas être satisfait d'une représentation redondante des données: en effet, une représentation enrichie offre sans doute plus de détails, mais risque de faire perdre de vue la structure d'ensemble des données. On cherche donc à réduire la dimension des données, toujours dans un cadre non-linéaire. Une approche géométrique (section 8.1) du problème consiste à considérer l'espace des signaux comme une variété à explorer. La géométrie nous dit alors que les fonctions propres de l'opérateur de Laplace-Beltrami constituent une carte, *optimale* au sens des moindres carrés, de cette variété. L'étape suivante consiste donc à implémenter ce modèle sur un graphe créé à partir des données, qui représente la variété. Un aspect sympathique du problème est que la matrice représentant les interactions est en fait très creuse du fait du caractère local de l'estimation, permettant le calcul sur des ensembles de données importants (section 8.2). Si la méthode ne donne pas de choix évident pour le nombre de composantes à considérer, il semble que dans bien des cas une représentation de petite dimension (deux ou trois) suffise.

Nous appliquons la méthode *i)* dans un cas purement exploratoire, en nous intéressant à la matrice des corrélations empiriques entre voxels (section 8.3.1) *ii)* dans un cas contraint où un modèle de filtre hémodynamique estimé est associé à chaque voxel -après pré-sélection (section 8.3.2). Il semble que la méthode donne une représentation concise des principaux effets présents dans les données. Malheureusement, elle n'explicite pas le *plongement* (*embedding*) qui va de la carte aux données.

1.2.8 Modèle intégré des données

Pour finir, nous proposons une synthèse de diverses idées développées dans ce travail. Dans un premier temps, nous revenons sur les modèles temporels en conciliant l'approche espace-état avec la théorie de l'information, afin d'obtenir un modèle hémodynamique associé à un petit nombre de séries temporelles (des voxels voisins, ou plusieurs sessions de données); voir la section 9.1. Nous obtenons un critère et une méthode d'estimation rapide.

Puis dans un second temps, nous bâtissons un modèle d'analyse multi-session recalées (qu'il s'agisse d'un ou de plusieurs sujets) en 9.2. L'idée est d'estimer un ensemble d'effets par session (typiquement, un espace de nuisance), un modèle hémodynamique par voxels (cf le paragraphe précédent); on parachève cette étude par la constitution d'une carte des signaux par la méthode Laplacienne. L'estimation de l'espace de nuisance se fait naturellement par la méthode d'espace-état (sans le paradigme expérimental). Celle-ci a pour intérêt de supprimer les composantes les plus corrélées du signal, dé-biaisant l'estimation au niveau des voxels - ce que nous vérifions en montrant que les coefficients d'auto-régression en chaque voxel reculent très significativement grâce à cette procédure. L'estimation de l'espace de nuisance et des signaux d'intérêt est donc conjointe, évitant le calcul redondant d'une structure d'auto-corrélation en chaque site.

Nos expériences montrent que l'on accroît ainsi la sensibilité dans la détection des signaux d'intérêt (section 9.3). On peut par ailleurs utiliser l'espace de nuisance dans une procédure d'inférence pour obtenir une estimation des amplitudes d'activation un peu plus sensible.

Enfin, nous terminons ce travail par quelques annexes techniques. Tout d'abord, l'appendice A décrit les données réelles et synthétiques qui nous ont servi dans nos expériences d'analyse. L'appendice B traite du seuillage statistique des cartes multivariées; nous y indiquons comment nous avons procédé en pratique dans ce travail. L'appendice D décrit la dérivation du critère d'information Bayésienne (BIC), que nous utilisons pour la juste paramétrisation des processus temporels. Elle s'appuie notamment sur l'appendice C qui définit les concepts de théorie de l'information utilisés dans cette thèse.

1.3 Quelques angles morts de ce travail

Nous faisons un point rapide sur certains aspects importants du problème qui ne sont pas abordés dans ce travail. Nous expliquons pourquoi ces impasses.

1.3.1 Les développements statistiques

Etant entendu que nous ne nous intéressons pas spécifiquement aux procédures d'inférence sur les données, nous ne faisons qu'un usage modéré des méthodes statistiques avancées. De fait, certaines parties, comme le chapitre 6, mériteraient des développements supplémentaires. Ce choix ne rend pas non plus justice aux travaux majeurs effectués dans ce domaine (théorie des champs gaussiens, taux de fausses découvertes *false discovery rate*). Nous nous contentons de réutiliser certaines recettes, avec quelques adaptations si nécessaire (voir l'appendice B). Remarquons en passant que l'utilisation d'une correction dite de Bonferroni induit des seuils un peu plus élevés pour le seuillage des cartes que l'absence de correction ou que l'utilisation du taux de fausse découvertes, mais que la différence n'est pas nécessairement flagrante.

1.3.2 L'information anatomique

Un autre aspect qui mériterait d'être développé est l'utilisation d'information ou de contraintes anatomiques dans le processus d'estimation. En effet, la corrélation spatiale est un aspect évident des données d'IRMf; l'ignorer est certainement très sous-optimal. Le lissage est une manière très médiocre de prendre en compte cette corrélation spatiale (en raison de la finesse du cortex, de l'ordre du voxel); en revanche, l'utilisation de parcellisation semble un moyen judicieux pour estimer localement la présence et la nature de signaux d'activation. Par exemple, on peut préférer une estimation locale du filtre hémodynamique à une estimation voxélique. En fait, le voxel est une unité de volume sans signification neurophysiologique, mal adaptée pour l'analyse anatomo-fonctionnelle.

Toutefois, cette approche suppose d'avoir recalé parfaitement les images anatomiques et fonctionnelles, d'avoir segmenté la matière grise sur les images anatomiques, et finalement d'avoir défini la parcellisation. Ayant choisi de nous intéresser uniquement à des problèmes fonctionnels, nous n'avons donc pas abordé cet aspect, pas plus que la fusion de données anatomo-fonctionnelles.

1.3.3 Les études multi-sujet

Enfin, nous n'avons pas explicitement travaillé sur la question des études multi-sujet. Celle-ci revêt une grande importance pour l'établissement et la validation de résultats en neurophysiologie. De fait, les analyses multi-sujet ont un intérêt pour l'inférence plutôt que pour l'exploration, et surtout, que pour l'estimation. En revanche, nous indiquons quelques pistes élaborées dans le cadre de l'analyse multi-session, qui s'étendent facilement au cas multi-sujet recalé (voir 9.3.2). Pour autant, le principal problème est le recalage lui-même, que nous n'abordons pas ici. Quant à l'étude éventuelle de données non recalées, elle repose exclusivement sur l'identification de signaux temporels, avec toute la fragilité que cela implique, étant donnés le bruit, l'impact de nuisance diverses sur le signal, et la brièveté des séries temporelles.

1.4 Publications associées à cette thèse

Les différentes pistes que nous avons explorées au cours de cette thèse nous ont permis d'effectuer plusieurs publications:

- Tout d'abord, dans le domaine de l'analyse univariée, nous nous sommes intéressé à la modélisation du signal par chaîne de Markov et à l'utilisation de l'information mutuelle pour la mesure des activations [205], [206].

- Nous nous sommes ensuite intéressé à des méthodes multivariées, abordées sous le thème de l'analyse en composantes dynamiques dans [207]. Il s'agissait alors d'une méthode d'analyse en composantes spatialement indépendantes contenant des éléments d'information temporelle.

- Nous avons repris le thème de l'analyse en composantes dynamiques, mais en le reformulant dans le cadre de modèles à état, et en introduisant une solution linéaire, dans [210].

- Parallèlement, nous avons abordé les modèles de mélange (*mixture*) non linéaire par l'analyse en composantes principales à noyau; cette approche est publiée en [208], et une comparaison avec le modèle linéaire multivarié est décrite en [211].

- Enfin, la méthode de clustering dite du goulot d'information (*information bottleneck*) fait également l'objet d'une publication [209]. Rappelons que cette méthode est une rationalisation par la théorie de l'information du problème du clustering de données expérimentales.

Ajoutons que, bien qu'inédite dans le cadre de l'IRMf, la méthode des plongements laplaciens présentée au chapitre 8 n'a pas encore fait l'objet d'une publication. Enfin, nous présentons au chapitre 5 un algorithme d'analyse en composantes indépendantes établi par Christophe Chef d'Hotel, non encore publié.

Part I

Modeling

Chapter 2

Functional MRI, a recent tool for the investigation of brain activity

This chapter outlines the importance of functional Magnetic Resonance Imaging (fMRI) studies in the understanding of brain function, and some fundamental features of this modality. The first section briefly exposes the physiological effects that underpin current image studies of brain activity. The second section deals with basic facts about Magnetic Resonance Imaging (MRI), and the third one outlines some particular aspects of functional MRI. In the fourth part, we describe the main usual preprocessing steps applied to fMRI data, i.e. the treatments that are performed prior to functional analysis of the data.

2.1 Investigating brain activity

2.1.1 Brain activity

A human cortex contains around 10^{12} neurons, the activity of which supports all the cognitive, sensory or motor processes of the body. Basically, neurons carry electrical information and exchange it through their synapses. This electrical information consists in the depolarization of the neuron membrane, which occurs through the exchange of ions - essentially potassium K^+ and sodium Na^+; this phenomenon is known as action potential. The information is exchanged at the level of synapses through the release of neurotransmitters that bind to receptor sites on post-synaptic terminals, yielding the next neuron to depolarization. Recovery from neuronal signaling requires uptake and repackaging of neurotransmitter and restoration of ionic gradients, all processes that consume Adenosin Triphosphate (ATP).

Important for us is that ATP consumption requires a continuous supply of glucose and oxygen; this supply is allowed by the Cerebral Blood Flow (CBF). As an illustration, the brain receives 15% of the total cardiac output of blood, and yet accounts for 2% of the body weight; in particular, the flow per gram tissue to gray matter is comparable to that in the heart muscle, the most energetic organ in the body [33]. Yet the brain has no reserve store of oxygen, and depends on continuous delivery by CBF.

As noticed very early [119], brain activity could be assessed by the measurement of regional CBF. Indeed, CBF increases substantially close to the areas of neural activation; moreover, this increase is graded according to the degree of activation [69]. Such effects

can be measured by Positron Emission Tomography (PET); but functional Magnetic Resonance Imaging (fMRI) measures another effect: the blood oxygen level dependent (BOLD) effect. Before turning to the latter, let us mention that the regulation of CBF is in itself a complicated phenomenon that involves several vasodilatory agents, like nitric oxyd (NO).

2.1.2 The BOLD effect

If there is an evidence that both flow and glucose metabolism increase substantially in activated areas of the brain, it has been established that the CBF change is not required to supply the glucose metabolism change; in fact oxygen metabolism -measured in terms of cerebral metabolic rate of oxygen, $CMRO_2$- increases much less than blood flow during brain activation. The result of this imbalance between CBF and $CMRO_2$ is a substantial drop in oxygen extraction and a corresponding drop in the deoxyhemoglobin content of the venous blood. The MR signal is sensitive to this change because deoxyhemoglobin (dHb) is paramagnetic, and the presence of dHb reduces the MR signal at rest (as will be explained next, the MR signal essentially depends on the total amount of deoxyhemoglobin within an image voxel). By contrast, activation induces a slight increase of the MR signal: this is known as the BOLD effect.

Note that the picture is in fact more complex: The BOLD signal change depends on the combined changes in CBF, $CMRO_2$ and Cerebral Blood Volume (CBV). Assuming that the CBV change is primarily on the venous side, increased CBV tends to increase local deoxyhemoglobin content simply because the volume of venous blood increases. This increase tends to counteract the effect of blood oxygenation, which decreases the deoxyhemoglobin content. The blood oxygenation effect dominates the blood volume effect (there is a net deoxyhemoglobin content decrease), so that the BOLD signal change is positive with activation.

Figure 2.1 represents a summary of the phenomena that occur during brain activation. However, the detailed explanation of these mechanisms remains unknown.

2.2 MRI: the principles

2.2.1 NMR

Nuclear Magnetic Resonance (NMR) is the basic physical phenomenon used for Magnetic Resonance Imaging (MRI). It concerns primarily hydrogen atoms present in the body (water molecules of the tissues). Since hydrogen nuclei have a single proton, they have a spin, which is an intrinsic angular moment. It can be associated with a magnetic dipole moment: each hydrogen nucleus behaves as a tiny magnet, with the north/south axis parallel to the spin axis. The sum of the moments of a sample of molecules is zero in the absence of a magnetic field; but in the presence of a magnetic field B_0, the spin axes of the protons precess around the direction of the magnetic field. The resulting magnetic moment of a sample is oriented in the direction of B_0. The frequency ν_0 of the precession is linearly related to the field by the gyromagnetic ratio γ, whose value depends on the nature of the nuclei.

$$\nu_0 = \gamma |B_0| \tag{2.1}$$

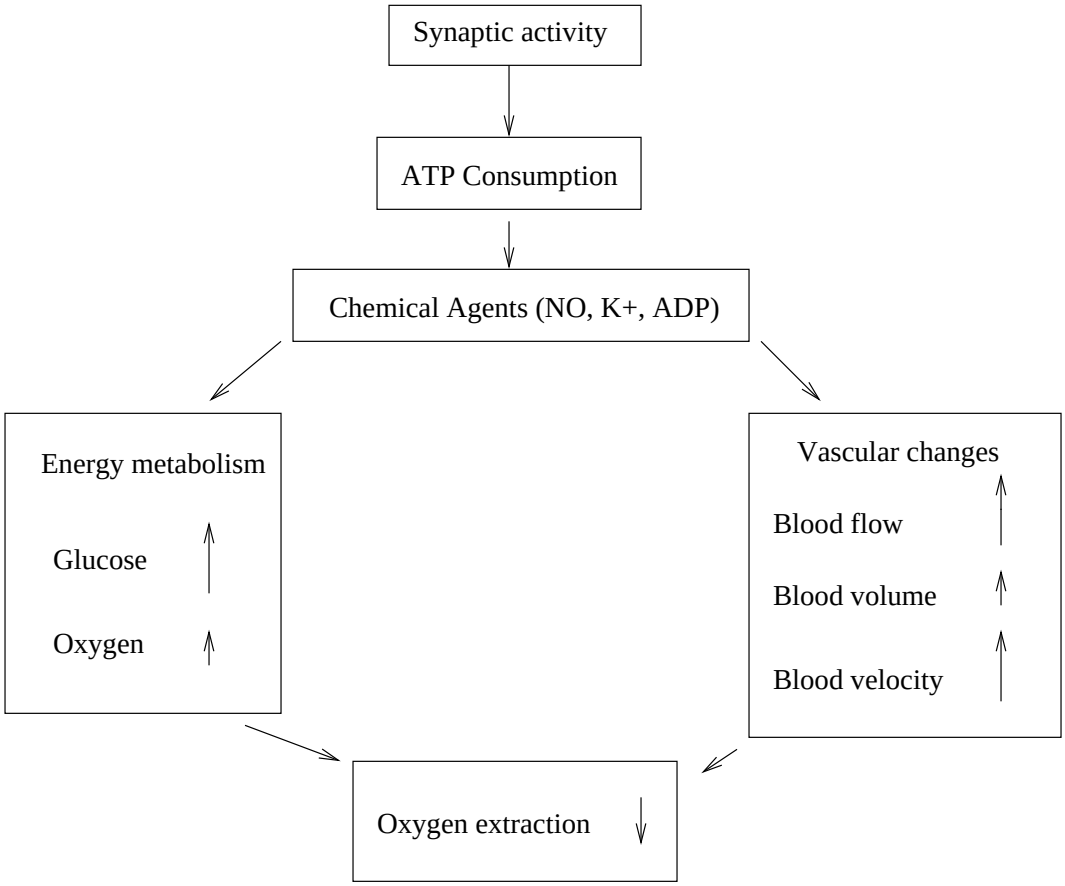


Figure 2.1: Physiological changes accompanying brain activation.
Functional neuroimaging is largely based on the metabolism and flow changes in the lower three blocks: the drop in oxygen extraction is the basis for the BOLD signal changes measured with fMRI, but the MR signal is potentially sensitive to blood flow, volume, and velocity as well. This figure is taken from [33].

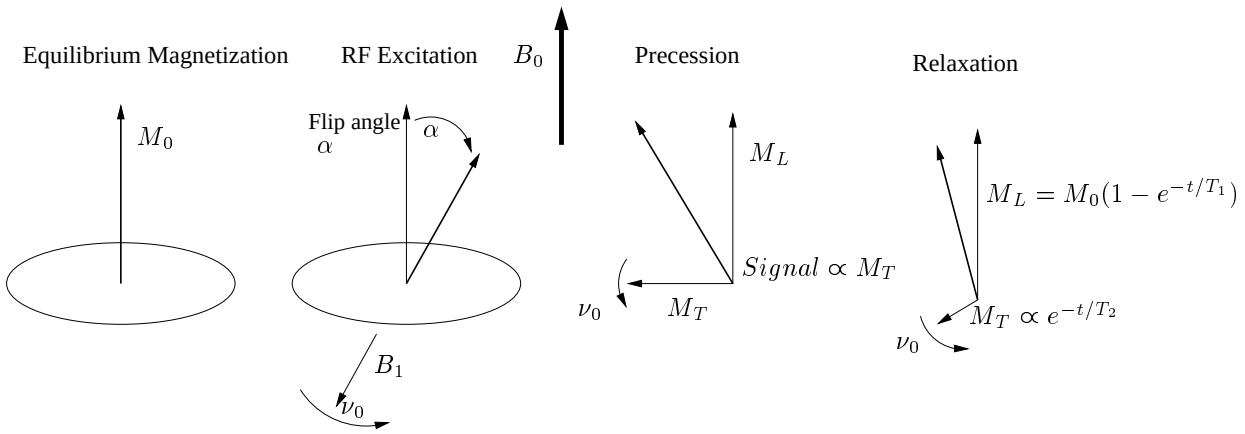


Figure 2.2: The basic physics of the NMR experiment.
In a magnetic field B_0, an equilibrium magnetization M_0 forms due to the alignment of nuclear dipoles (left). An RF pulse tips over M_0 creating a longitudinal component M_L and a transverse component M_T (middle). M_T precesses around the direction of B_0, generating a detectable MR signal. Over time M_T decays to zero with a relaxation time T_2 and M_L returns to M_0 with a relaxation time T_1(right). This picture is taken from [33].

Besides the precession of the nuclei, a second phenomenon is important to us: the relaxation of the nuclei. In the presence of a constant field B_0, the spin axes of the nuclei slowly tend to align with B_0, with a time constant called T_1.

The Radio Frequency pulse (RF pulse) technique consists in applying in addition to B_0 a transient field pulse B_1, orthogonal to B_0, rotating at the resonance frequency of the nuclei ν_0, and several orders of magnitude smaller. The resulting moment M of the item is flipped (usually by 30 or 90 degrees, according to the duration of the pulse); when B_1 is switched off, M precesses around B_0 and finally aligns with B_0: the transient transversal moment also called *Free Induction Decay* (FID)) cancels with a time constant T_2 while the longitudinal moment reaches its equilibrium with a time constant T_1. T_1 and T_2 depend on the environment, so that their local value can be used to discriminate between the tissues, proton density being a third signature to discriminate between tissues. The difference in the values of M are then measured by coils. An illustration of the phenomenon is given in figure 2.2.

The repetition of the basic RF pulse is also called a pulse sequence. Many kinds of pulse sequences are possible, and the sensitivity of the MR images to the different parameters can be adjusted by tuning the repetition time (TR) between consecutive pulses. Typical sequences may be of several kinds:

- The Gradient Echo Pulse Sequence simply consists of the repetition of the FID described previously. It is simply described by the value of the flip angle α and the repetition time (TR).
- The Spin Echo Pulse Sequence consists in applying a first 90 degrees pulse, then after a time TE/2 a 180 degrees pulse in the transverse plane; the effect of this

pulse is to refocus the signal whose phase has been quickly dispersed by local field inhomogeneities. Thus an echo of signal appears at time TE; the measurement is performed at this time; this echo can be repeated many times to sample the T_2 decay.

- The Inversion Recovery Pulse Sequence begins with a 180 degrees pulse and after a delay TI a 90 degrees pulse. It enhances the T_1 weighting of the image.

2.2.2 MRI

The pulse sequences produce a transient pattern of transverse magnetization across -and around- the brain. Magnetic Resonance Imaging (MRI) consists in imaging in three dimensions the distribution of the transverse magnetization within the brain. The principle is the following: the phase of the local signal is manipulated in such a way that the net signal traces out the spatial Fourier transform of the distribution of transverse magnetization.

The same coil is used for the transmission of the gradient and the reception of the signal; since a coil typically encompasses the body (the head of the subject) it measures a sum of the signals from each tissue in the head. Localization is based on relationship (2.1): the magnetic field B_0 is added with a gradient in the transverse direction. The selection of a particular frequency at the receiver part is then equivalent to the selection of a slice -a plane with a thickness of typically 1 to 10 mm- along the transverse direction (z). This procedure is called the *slice selection*.

Then, within each slice or plane spanned by the resulting directions (x and y), two gradients are applied during the relaxation. In the x direction, a negative gradient is applied after the RF pulse, and a positive one during acquisition, which creates a gradient echo during data acquisition, halfway through the second gradient pulse; the effect is that the precession frequency varies along the x axis, so that a Fourier transform of the signal gives its amplitude along the axis. This procedure is called *frequency encoding*.

Slice selection plus frequency encoding yield a two dimensional information. In the remaining y direction, a gradient field is applied for a short interval between the RF pulse and data acquisition; after cancellation of this field, the precession is at the uniform rate, but with a phase shift determined by the position on y. Repeating the frequency encoding many times with different phase shifts creates information on the y position. This third procedure is known as *phase encoding*. A summary of a basic imaging pulse sequence is given in figure 2.3.

Since functional imaging requires fast acquisition, some techniques have been developed for fast imaging, the most important being echo planar imaging (EPI); in this case, the gradient are oscillated very rapidly, so that sufficient gradient echoes are created to allow measurement of all the phase-encoding steps required for an image. The full data for a low-resolution image are acquired from the signal generated by one RF pulse. EPI requires strong gradients. An 2D image matrix of size 64×64 is acquired in about 30 to 100 ms (instead of a few minutes for a conventional T_2-weighted slice acquisition). After acquisition of the data in the frequency space, also known as $k-$space, the data is mapped into the 3D space by Fourier transform. The only information used in subsequent discussion is the magnitude of the (complex) signal at each voxel.

The acquisition of a 3D volume of data is the sequence of multiple slice acquisitions, that can be sequential (ordered in the z direction) or interleaved (pair slices acquired before

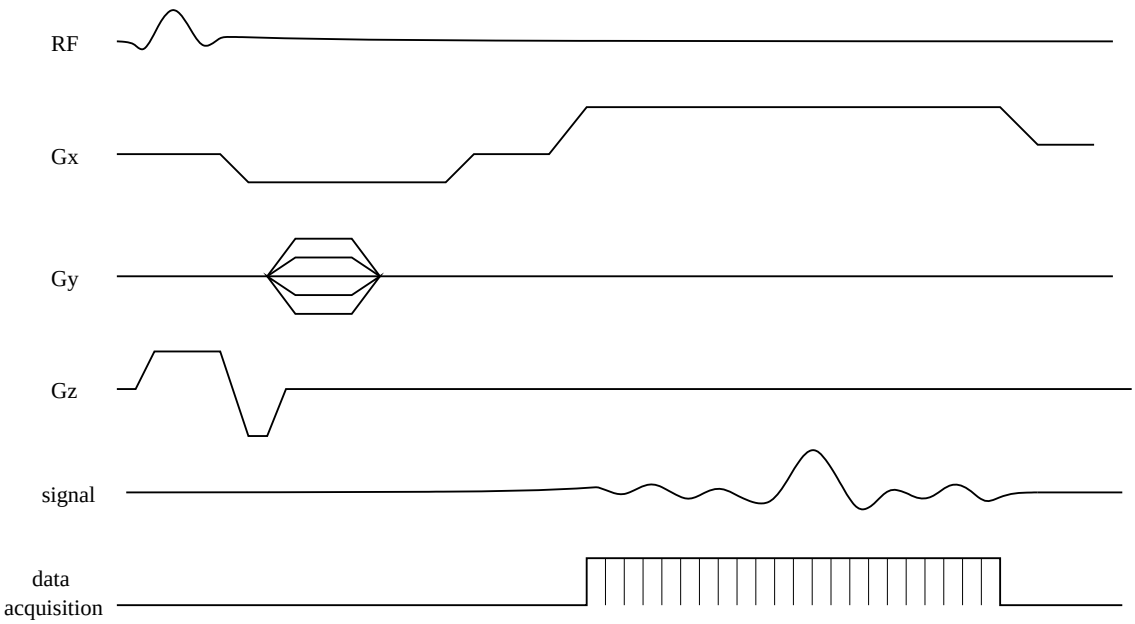


Figure 2.3: A basic imaging pulse sequence.
During the RF excitation pulse a gradient in z is applied (slice selection), and during read-out of the signal a gradient in x is applied (frequency encoding). Between these gradient pulses, a gradient pulse in y is applied, and the amplitude of this pulse is stepped through a different value each time the pulse is repeated (phase encoding). Typically 128 or 256 phase-encoding steps (repeats of the pulse sequence) are required to collect sufficient information to reconstruct an image. This figure is taken from [33].

odd slices). Direct 3D acquisition schemes have also been designed; they offer higher Signal to Noise Ratio (SNR), but they do not allow for fast imaging procedures.

2.3 functional MRI

We concentrate here on BOLD fMRI, although other functional imaging techniques are possible with fMRI (contrast agent methods, arterial spin labeling). This is because BOLD fMRI is more frequent than other methods.

2.3.1 Blood susceptibility depends on deoxyhemoglobin content

Functional MRI (fMRI) is a way of measuring the field distortion around the vessels due to deoxygenated blood. Indeed, while fully oxygenated blood has about the same susceptibility as other brain tissues, deoxyhemoglobin is paramagnetic and changes blood susceptibility. More precisely, the BOLD imaging is based on the following phenomena discovered by Ogawa et al. [168], [169]: in the normal human brain 40% of the oxygen delivered to the capillary bed in arterial blood is extracted and metabolized. There is thus a substantial amount of deoxyhemoglobin in the venous vessels, yielding an attenuation of the MR signal. When the brain is activated, the local flow increases substantially, but oxygen metabolism increases only in a small amount (see figure 2.1). As a result, the oxygen extraction is reduced, and the venous blood more oxygenated. The reduction of deoxyhemoglobin concentration leads to a signal increase (a few percents at 1.5T, 5-15% at 4T).

2.3.2 Mapping brain activity with BOLD signal changes

The prototype brain mapping experiment consists of alternating periods of stimulus task and control task (e.g. finger tapping and resting) [11]. The periods are typically 20-30 seconds or 10 times the repetition time (TR). The basic task alternance is repeated several times. Throughout these stimulus/control cycles dynamic EPI images are collected covering all or part of the brain, with relatively low resolution with respect to conventional anatomical images (e.g. $3 \times 3 \times 3$ mm^3 against $1 \times 1 \times 1$). Each image of the chosen slices are acquired in rapid succession, and after a time TR, this set of images is acquired again. This process results in a four dimensional dataset: three spatial dimensions plus time. Incidently, spatial distortions are not the same for anatomical and functional EPI images. This makes the coregistration of both modalities a difficult problem, since non-rigid deformations are involved.

Let us notice that the resulting data has neither a physical dimension nor an absolute scale, which would help for interpretation. An increase or decrease of signal can thus be described in terms of ratio of the mean level, or with its original absolute value.

2.3.3 The experimental design

The experimental paradigm is the sequence of events, stimuli or conditions that the subject undergoes during the scanning session. It ranges within one of the following families:

- The block design can be described in terms of state succession: during a number of scans, the subject is in a certain behavioral state, that changes for another period, and so on. When compatible with the studied function, this kind of design is usually considered as optimal for detection purposes, since it maximizes the signal to noise ratio (SNR) under certain hypotheses [140].

- Event-related designs cannot be described in terms of states, but rather as successions of stimuli onsets. They are especially useful for certain cognitive functions (e.g. language), and as a methodological purpose, for the precise characterization of the response (shape, delay, linearity) (e.g. [202]). Experiments based on adaptation [100] can be put in this category.

- Periodic designs are useful for some functions. This is in particular the case for retinotopy [196], [225] where the stimulation space (retinotopic polar angle, retinotopic eccentricity) is continuous.

- Parametric designs are relatively rare, and usually used for testing whether there is some proportionality between a parameter of the stimulation -e.g. the speed of a moving target, or the contrast of a flickering checkerboard- and the analyzed response [219].

Besides, the experimental paradigm may comprise one or many stimulation conditions, i.e. at least two conditions, including the baseline. In the first case, the question of interest is whether the stimulation elicits a response, while the second case yields more complex inferences; for example, one may study the brain areas involved in one of the tasks, or selectively activated by one experimental condition, or significantly more activated under one condition than under another one.

2.4 Preprocessing the data

Now we consider a four-dimensional dataset acquired under a given experimental paradigm. We call each volume of data acquired at a given TR an image; a set of sequentially acquired images is a run or a session. Many different sessions can be acquired for a given dataset, with repetition of the same experimental paradigm or not. Next, the same experiment can be replicated on many subjects to allow for neurophysiological inference.

2.4.1 Registration

Due to the motion of the subject during the experiment, the images have to be registered, so that a given voxel unambiguously represents a brain area for all the images. This involves two steps:

- Motion estimation: it is often assumed that the motion is rigid; this is only approximately true due to the intrinsic artifacts of EPI images, that induce non-rigid distortions between images even if the subject motion is rigid. Under this hypothesis, motion estimation boils down to the estimation of six parameters (3 translations, 3 rotations). The most current method consists in finding the rigid transformation

that minimizes the grey level difference between consecutive images; but this simple method has been shown to introduce artifacts, like spurious task-related motion estimates. For this reason, it is preferable to use more robust methods [70]. In particular, INRIAlign procedure reduces the influence of large intensity differences by weighting errors using a non-quadratic, slowly-increasing function. This is basically the principle of an M-estimator.

- Motion correction: according to Mangin et al. [70], this step should be performed only when the estimated motion is non-negligible with respect to the voxel size, since a reinterpolation of the data has ill-controlled effects on the data content. The usual method is a trilinear interpolation of the data that takes into account the motion estimates. We do not address the combination of motion with distortion.

Following motion estimation and correction, a step of spatial normalization can be performed; this consists of coregistering the functional images with a MR anatomical image of the same subject or with a template (this is of frequent use for multi-subjects studies), and then to interpolate the functional images into the template. Let us notice that in that case, the displacement field is considered as non-rigid, yielding a heavy computational load. Care should be taken when employing this procedure, because:

- The registration between images or different modalities and/or templates is very difficult. It requires non-rigid deformations (e.g. spatial stretching of the data). The effect on the resulting activation maps may be quite complex. Note that in the SPM 99 software, smoothing is then recommended "as a preprocessing step to suppress noise and effects due to residual differences in functional and gyral anatomy during inter-subject averaging". This is not very encouraging.

- This procedure dramatically increases the number of voxels of the dataset, which in turn increases the computation load for the analysis. This has a conservative effect on the corrected P-values obtained with a Bonferroni correction(see next chapter).

- Last, as any sub-sampling procedure, this simply increases the size of the images without adding any information. In fact, it is simpler to interpolate the final images (activation maps) into the template of interest instead of the raw functional images.

2.4.2 Smoothing

Spatial smoothing has become a standard routine for fMRI data analysis for two kinds of reasons [173], [127], [5]: *i)* The increase of the SNR *ii)* The interpretation of the images as gaussian random fields (under the null hypothesis that no activation pattern is present). Let us examine these two points:

- Smoothing the data spatially increases the SNR in the sense that it reduces the effect of the spatially uncorrelated noise with respect to a priori more structured signal of interest. This is of course at the expense of spatial precision (the spatial precision of functional images is quite coarse; in particular, the average grey matter width is not much greater than the typical voxel size, so that it is unavoidable that isotropic smoothing mixes tissues of different nature). The debate about spatial smoothing

yields a tradeoff between bias (the precision in activation localization) and variance (SNR gain by smoothing). However, it is clear that an optimal smoothing scheme is not the isotropic gaussian filter employed usually, but is brought by adaptative filters. An example of anatomically-based smoothing has been proposed in [5]. Other anatomically-informed procedures have also been proposed: parcellation [66] and anatomical basis functions [127].

- The second reason comes from the method adopted by the SPM software (see next chapter) to assess activation significance: it is assumed that the residual of the regression model can be treated as a gaussian random field with a certain smoothness; this is probably not true if one considers the original images, but becomes likely after the smoothing process.

It is important to consider that intrinsic spatial correlation is embedded in raw fMRI datasets; in particular, though many functional brain areas are largely sub-voxel, there is a consensus that reliable activation foci should encompass clustered voxels. Smoothing is then simply a means for canceling spatially high frequency noise in the data. Though this statement is correct, this intrinsic spatial correlation is certainly not isotropic. Moreover, there are other ways of taking spatial autocorrelations into account than data smoothing (see next chapter). Notice that the study of spatial correlations has been introduced as a particular way of processing the data, in particular with the help of scale-space theory (see [177] [47] and 5.1.4).

2.4.3 Removing global effects

This consists in subtracting to all voxel time courses the mean of each image. This preprocessing is aimed at removing some physiological effects that are assumed to be global over the image. However, this is at the risk of removing activation patterns, if the latter have an influence on the global mean. This is actually the case, and recent studies have enlightened the bias induced by this procedure [51].

2.4.4 Selecting voxels of interest

Signals of interest are expected only in the grey matter. It is thus tempting to ease the analysis by selecting only the anatomically relevant voxels for further analysis. More simply, one often uses a mask that keeps the brain voxels; a simple threshold on the T_2 averaged image of the sequence is often sufficient. More sophisticated methods can be employed (e.g. [120]), but they are integrated in the general framework of signal analysis. Some developments of this work will require such posterior selection.

2.4.5 Detrending

Looking at the temporal part of the dataset, we also observe the presence of intrinsic correlations. Although some methods use a smoothing in the temporal domain, this is not a systematic usage in fMRI data processing. In fact, the presence of temporal correlations is probably related to the intrinsic object of the measurement, in particular the presence of biological rhythms (respiratory, cardiac) that some authors propose to correct [121]. But there is more consensus on the presence of trends in the signal, which induces high temporal

correlation (these effects have typically low frequency) in the dataset. Their cancellation is thus important to enforce the stationarity hypothesis, which is fundamental in the analysis of the data. This is performed by removing the low frequencies of the signal [221], or by estimating adaptively the trend [147], or by fitting a wavelet basis [155]. A partial study of detrending methods is available in [203].

A practical method is the removal from each voxel-based signal of a fitted low frequency approximation:

$$x_{detrend}(t) = x(t) - (x * g)(t) \tag{2.2}$$

where g is e.g. a gaussian filter wider than the timing of the effects of interest. The computation of $x * g$ can be made quickly and efficiently through the use of recursive filters [52].

2.4.6 Temporal registration or slice timing

Another source of artifacts in the interpretation of the data is the fact that all the slices are not acquired at the same time, but at given fractions of the TR. This effect is problematic when the sequence of events is quick (of the order of the TR). In that case, it may be better to correct for this effect, that is to apply a kind of temporal registration between the slices, so that their acquisition can be considered as simultaneous for further study. This is currently done by preserving the spectrum of the signal obtained at different voxels and shifting their phase. This method has been popularized under the name of *slice timing correction*.

2.4.7 Preprocessing: what is essential ?

In this work, we are not particularly concerned by the analysis of pre-processing steps. Notice that this is currently studied with high-level procedures in several works [134] [51]. We simply state here which parts of the above mentioned procedures are essential to us:

- Motion estimation is essential, since it gives some insight on the behavior of the subject. Motion estimates that are correlated to the experimental paradigm indicate probably a bad estimation. We use the INRIAlign method (http://www-sop.inria.fr/epidaure/software/INRIAlign/).
- Motion correction should be applied as soon as motion amplitude is beyond a fraction of a voxel. Otherwise it can be bypassed.
- We analyze the functional images in their original format and warp only the resulting spatial maps towards the anatomical images. We use the method developed by [111] [41]. Since most of our data is based on single subject studies, we do not consider the problem of common registration.
- No global scaling is applied.
- We prefer to avoid spatial smoothing.
- Concerning block designs with long stimulation periods, the temporal registration can also be avoided.

- Voxel selection is applied by simple masking of the brain
- Detrending is performed through gaussian fitting -which can be done very quickly.

Chapter 3

fMRI data analysis: state of the art

This chapter summarizes the main ideas that have been proposed to analyze fMRI data. The basic distinction is between hypothesis-driven approaches, that perform a statistical validation of prior hypotheses and exploratory methods that give an account of the data. We will describe more precisely the Statistical Parametric Mapping (SPM) methodology, which is used in standard fMRI studies, then give a quicker account of different univariate and multivariate methods. The point is of course to discuss what each method assumes and proposes. However, there is no place for an exhaustive description of each method.

3.1 An overview

3.1.1 Position of the problem

fMRI data analysis consists in extracting relevant information from the spatio-temporal data produced by the experiment. In an engineering perspective, this means analyzing the response of a system given inputs and outputs (see figure 3.1). By *system* we mean the subject undergoing the experiment, who accomplishes -among other things- some sensory, motor or cognitive tasks, but also the measurement setting, including all the physical and computational steps that we briefly described in the first chapter and elementary preprocessing, e.g. motion correction ; by *input* we mean the information delivered to the subject in order for him to accomplish the experimental task ; in the remainder of this work we will refer to this input as the *experimental paradigm*; by *output* we mean the spatio-temporal collected dataset.

What is perhaps more problematic is the concept of relevant information: is it the response to the experimental paradigm -assuming that it can be defined unambiguously- or rather the amplitude of the response -which under some hypotheses is probably a more tractable concept- or the delay, or other characteristics (filter, linearity) of the response ? Or more generally, a synthetic representation of the dynamics of the dataset ? In spite of their naive formulations, these questions do not have plain answers, and all the authors have paid more attention to one or the other aspect. We will develop some parts of this discussion in the next subsection.

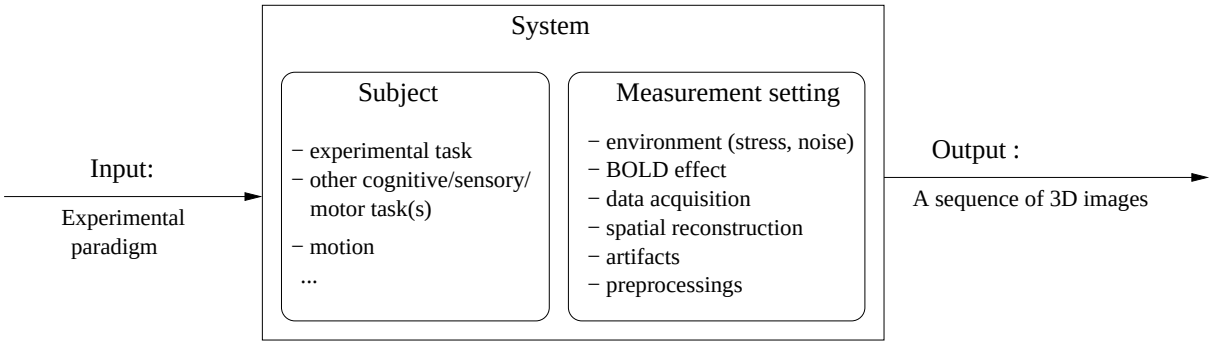


Figure 3.1: Schematic representation of the data generation
This includes the input - represented as the experimental paradigm, but in fact all the device that present information to the subject- the system itself -subject, measurement setting, basic processing- that produce the dataset. By this generation scheme, we want to point out that no simple relationship can give an account of the input/output relationship. Put more simply, the system is a kind of *black box* whose behavior can only be approximated.

Before going further, we emphasize that the intrinsic complexity of the system does not allow for simple interpretation (unlike engineering systems whose behavior is characterized by a transfer function). To be clear, let us recall some of the basic facts that prevent us from giving straightforward interpretation of the input/output relationship of the system (without listing gross artifacts sources, as defective stimulus presentation):

- The experiment takes place in a particular environment (in a tunnel, with strong acoustic noise).
- The subject is supposed to perform some motor/sensory/cognitive task, but we have little control on his real behavior; though this is the purpose of good experiment designs, the state of the subject does not reduce to what the experimental setting defines.
- The structure of the neural answer is not well understood (this is in fact the ultimate goal of neurobiology).
- Neither is the link between neural response and BOLD effect well understood.
- The respiratory and cardiac rhythms of the subject also appear in the BOLD response (with evident aliasing concerning the cardiac rhythm).
- Inhomogeneities in the magnetic field create some distortions in the EPI images.
- Subject motion is only partially corrected by standard methods ; the combination of subject motion with image distortion has complex effects.
- The machine itself has its own artifacts : signal drifts, thermal noise.

- The spatial reconstruction of the data has its own limitations (ghosting effects, introduction of spatial correlations).

This non-exhaustive list gives an idea of why the characterization of the system is far from simple, even if the first preprocessing described in the first chapter counteracts some of the nuisances. Let us also outline the effort of experimenters to measure the attention of the subject and monitor some of his natural rhythms in order to take this side information in the analysis.

As explained in section 2.3.3, the experimental paradigm can be of several kinds : block design, event-related design, continuously varying stimulation, parametric design; moreover, it may encompass one or several experimental conditions. Indeed, inference from neuroimaging data is usually based on subtractive logic. The concept of subtractions is highly associated with the idea of contrast in current methodology (see e.g. equation 3.10). In any case, one needs to define the input or excitatory variable of our system in a numerically practical way. The most frequent way is to define a matrix $P_c(t), c = 1..C,\ t = 1, .., T$, with one row for each experimental condition; typically $P_c(t) = 1/0$ if the subject undergoes/does not undergo condition c at time t. Then this vector can be treated as any quantity, for instance, it may be centered, normalized, whitened etc.

3.1.2 Hypothesis-driven and exploratory methods

fMRI data analysis has generated an abundant literature so that it is important to introduce proper distinctions for the understanding of the different methods. We propose here the characterization that is the most fundamental to us : the distinction between hypothesis-driven and exploratory methods.

Hypothesis-driven methods [173] postulate a certain form for the response to the experimental stimulation. This allows the parameterization of the response, and then the estimation of the model parameters ; this kind of methodology is followed by a statistical test that assesses the response estimation and concludes to the presence or the absence of an activation. More often, such methods are voxel-based (even if they have been generalized by spatially regularization with smoothing [5], Markov Random Fields [53], or parcellation [65]). Their potential weakness is their way of assuming a certain form of the response which may prove to be inadequate, thus yielding biased conclusions. On the other hand, these methods provide clear conclusions on a particular question. For example, they can answer the question "Is the time course of this data voxel correlated with the assumed response to the experimental paradigm?" with a clear response " Yes/No" associated with a probability value (P-value) of making mistakenly a positive statement. With some (still controversial) hypotheses on the data and noise structure, they can even answer a question like "Are there any clusters of connected activated (in the sense of being correlated with a known pattern) voxels in the image?", assorted with P-values. Even though considerable efforts have been done to retain flexible models in this framework (introduction of temporal correlation, response space of dimension greater than 1, introduction of nonlinearities or more generally deviations with respect to the assumed linear model (see in the next section)), they still cannot answer general questions like "What is the strongest pattern present in the dataset as a response to the experimental paradigm ?" or more generally "'What are the main spatio-temporal patterns present in this dataset ?".

This explains the necessity of introducing different methodologies that we will call exploratory [172]; in that case, the idea is to search for a generative model of the data, that is a model that shows which patterns appear in the dataset, and how these patterns are temporally/ spatially structured. This kind of approach is now rather multivariate, in the sense that it considers all voxels simultaneously. Its goal is thus to give an account of the data content, which is interesting if one wants to investigate cautiously which effect indeed appears in the data (given all the sources of confounds described earlier). However, such an approach does not give definitive answers to hypotheses concerning the dataset, since no question has been formulated in a closed form. Rather, the idea is to check by exploration what is the main response pattern present in the dataset, if some confounds can be identified (heart beat, respiration, motion, drifts). A perhaps more interesting -and more difficult- question is which patterns can be generalized from one dataset to the other, from one subject to the other.

Now we turn to a (non exhaustive) list of the most frequent methods, keeping in mind the basic distinction given above.

3.1.3 Taxonomy

An overview of the main existing methods is given on figure 3.2. Let us give a quick description before going to more details on the methods themselves. Hypothesis testing methods try by different ways to assess the presence of a given activation pattern. The methodological variations can be related to the estimation procedure used (temporal domain or frequency domain using Fourier transform), or to the statistical test employed to assess the presence of the activation. Sometimes, the differences are essentially a matter of vocabulary ; for example, Maximum Likelihood (ML) estimates are equivalent to the linear estimates given a regression scheme. Bayesian methods rather look for a maximum a posteriori (MAP) solution, given some priors on the response or parameters of interest. Moreover, some authors have used non-parametric statistics (Kolmogorov-Smirnov test, mutual information between the data and the experimental paradigm, conditional entropy rate of the data) for activation assessment.

On the exploratory part, we have a set of methods that extract information from the dataset, often without any prior knowledge of the experimental paradigm. The only way to process is then to use some structural properties that can discriminate between features of interest present in the data (decorrelation, independence, distance in a feature space, similarity, which yields respectively Principal Components Analysis (PCA), Independent Components Analysis (ICA), Clustering, and Self-organizing maps).

Some authors have tried to bridge the gap between the two families of methods. This is the case of multivariate linear models (MLM).

3.2 Univariate methods: review and discussion

In this section, we give some more details on the main univariate methods, with a special emphasis on the General Linear Model, which has been popularized with the Statistical Parametric Mapping (SPM) software. Of course, this description is by no way exhaustive.

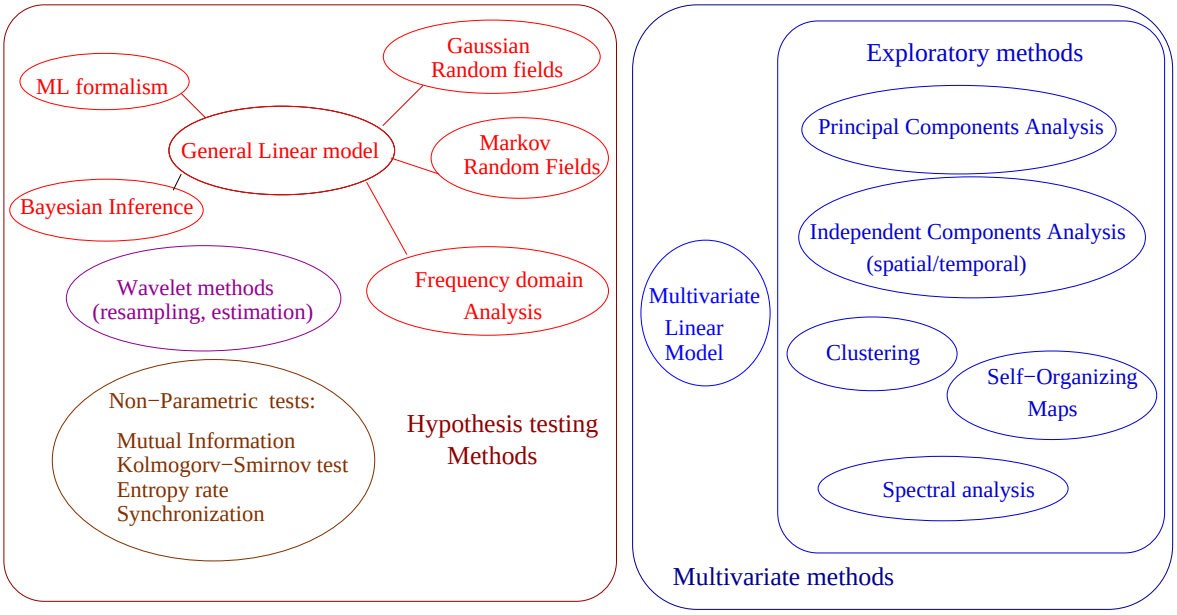


Figure 3.2: Overview of the main methods for fMRI data analysis.
They can be broadly divided into two categories: The hypothesis testing methods that primarily try to define which voxels can be said as activated given one signal model; these methods differ either by the signal estimation procedure or by the statistical method employed to assess the activation. On the right are the exploratory methods, that generally extract a set of meaningful patterns from the dataset. Some methods, as the multivariate linear model, try to play on both parts.

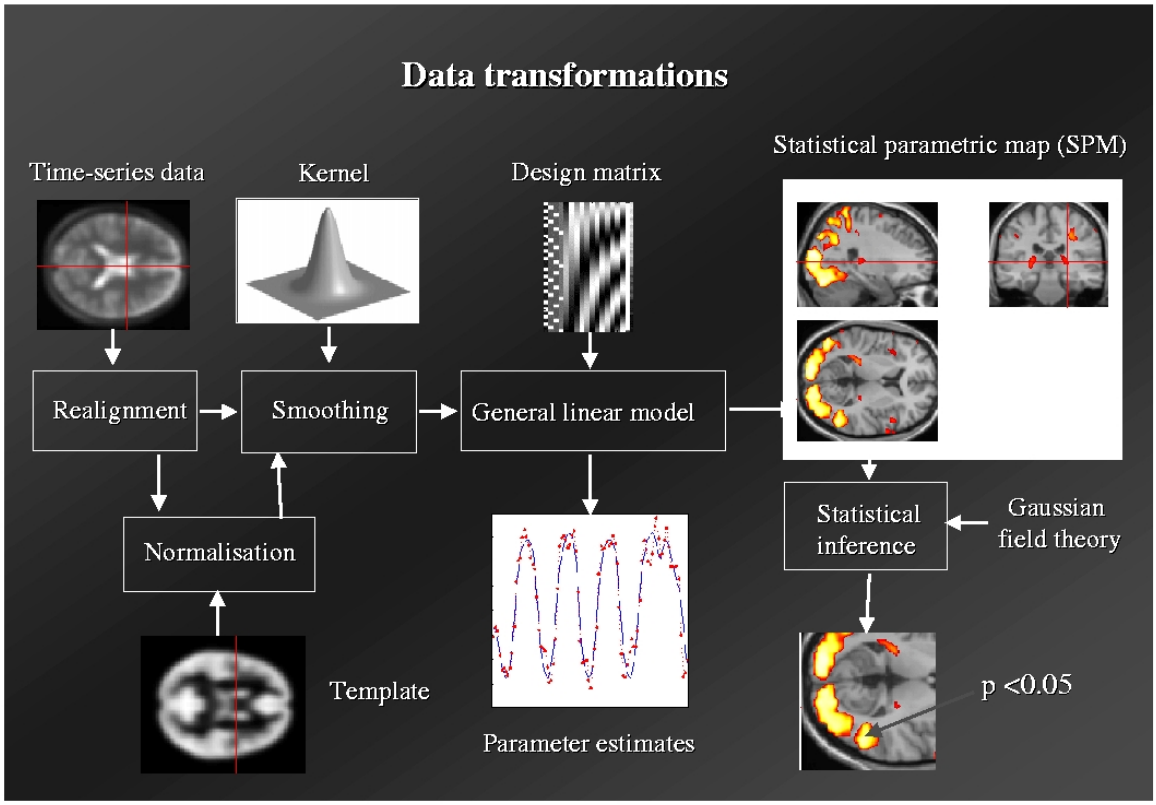


Figure 3.3: Typical data analysis performed with the SPM software.
This figure is borrowed from [77].

3.2.1 The general linear model (G.L.M.)

A complete description of the General Linear Model as implemented in the Statistical Parametric Mapping (SPM) software can be found in [77]. Here we sum up the steps of the model, which is important since almost all the fMRI literature refers to it. A general idea can be found in figure 3.3.

Let us denote X one session of a fMRI dataset, considered as a $N \times T$ matrix, where N is the number of voxels in the dataset (we do not consider here their spatial structure), and T the length of the time series. $X_n(t)$ will thus be the signal at voxel n and time t.

We assume that the subject undergoes different conditions of a given experimental paradigm. The conditions are defined by the time course of the effect (usually, a stimulation) $P_c(t), t = 1..T, c = 1..C$.

In a first approximation, the G.L.M. assumes that the response to each experimental tasks is proportional to a given time course, that can be defined through convolution of the stimulation $P_c(t)$ with a filter h known as the canonical hemodynamic response function (hrf). This is the consequence of two technical assumptions: *i)* the linearity of the response with respect to the stimulation, and *ii)* the time-invariance of the response

(a delay in the stimulation induces the same delay in the response); the linearity in the response is equivalent to a superposition principle : if two tasks are associated in the experiment, their effects in terms of response simply sum up. The filter h is assumed identical for all voxels. The most basic signal model is then

$$X_n(t) = b_0 + \sum_{c=1}^{C} b_c(n) h * P_c(t) + \epsilon_n(t) \tag{3.1}$$

Where b_0 is a constant of no interest, $b_c(n)$ the amplitude of the responses to the stimulus c and $\epsilon_n(t)$ a noise term. In the sequel, we assume that the signal is centered, so that $b_0 = 0$.

Let us denote $g_c(t) = h * P_c(t)$; one can specify other explicative variables for the analysis of the voxel time-course $X_n(t)$: head motion estimates during the session, or given functions of the time, like low frequency sinusoids, low order polynomials. More recently, the (first or second) temporal derivatives of the first components $g_c(t), c = 1..C$ have been used with the regressors of interest. We obtain thus a richer explicative model, modeled as a set of temporal regressors $G = g_r(t)$, with $r = 1..R, t = 1..T$, $R \geq C$, called the *design matrix*.

$$X_n(t) = \sum_{r=1}^{R} b_r(n) g_r(t) + \epsilon_n(t) \tag{3.2}$$

The corresponding vector of regressors $b(n) = (b_r(n))_{r=1..R}$ can be estimated by least squares estimation:

$$\hat{b}(n) = (G^T G)^{-1} G^T X(n) \tag{3.3}$$

with the dispersion matrix

$$\Lambda_b(n) = \sigma^2 (G^T G)^{-1} \tag{3.4}$$

where σ^2 is the residual variance. (It is implicit here that the design matrix is of full rank; pseudo inverses can be used in a general case, but full rank designs are recommendable). This technique yields the best linear unbiased estimator of $b(n)$ if one assumes normal, noise, and of $\Lambda_b(n)$ if additionally the noise is white.

Non-normality of the noise is difficult to control, but it is known that fMRI noise is correlated, at least temporally, so that three procedures can be performed: *i)* whiten the signal X together with the regression model G [182], [221], [223]; this procedure requires a careful estimation of the autocorrelation *ii)* estimate the noise covariance Σ [30], and replace equations (3.3) and (3.4) by

$$\hat{b}(n) = (G^T \Sigma^{-1} G)^{-1} G^T \Sigma^{-1} X_n. \tag{3.5}$$

$$\Lambda_b(n) = \sigma^2 (G^T \Sigma^{-1} G)^{-1} \tag{3.6}$$

at the risk of biasing the result if the estimate of Σ is poor, and *iii)* add more correlation than what is actually in the data, and derive a new noise covariance matrix [2]; neglecting initial correlations, one can also reduce the final correlation of the data to that induced by the analysis, or consider a simplified model [84]: equation (3.2) becomes

$$[k * X_n](t) = \sum_{r=1}^{R} b_r(n) [k * g_r](t) + [k * \epsilon_n](t) \tag{3.7}$$

where $k(t)$ is a low-pass filter and equation (3.3) becomes

$$\hat{b}(n) = (G^T K^T K G)^{-1} G^T K^T K X(n) \tag{3.8}$$

where K is the filter k written in a matrix form. The dispersion around the estimate of $\hat{b}(n)$ is given by the matrix

$$\Lambda_b(n) = \sigma^2 (G^T K^T K G)^{-1} G^T K^T K \Sigma K^T K G (G^T K^T K G)^{-1} \tag{3.9}$$

Both estimators (3.8) and (3.9) are biased; however it is argued in [84] that this bias is inferior to the bias induced by an improper whitening. More recently, an optimization of this procedure has been proposed in [40], based on an adaptative spline smoothing; the adaptation is allowed by a generalization error criterion (cross-validation). This results however in an heavier computational cost. It is shown that naive procedure (without correction) and uniform correction overestimate the ensuing tests. Whatever the method employed, these estimates can then be used to derive statistical maps. The statistics are related to the definition of *contrasts*; a contrast γ is a linear combination of the estimates $\hat{b}$; its interpretation is simple if one views the vector $\hat{b}(n)$ as the set responses at voxel n to the effects given by the matrix G. Contrasting them means that one checks whether the response to a given effect is stronger or weaker than to another effect. Once the linear combination and its dispersion (obtained from the dispersion matrix) are computed, one assesses the significance of the estimated response to the contrast by comparing it to its estimated dispersion:

$$\frac{\gamma \hat{b}(n)}{\sqrt{\gamma^T \Lambda_b(n) \gamma}} \sim t_d(n) \tag{3.10}$$

where t_d is the Student distribution with d degrees of freedom, which are derived from the design matrix G by standard methods [222]. The Student distribution is invocated here, because, under the null hypothesis (the contrast γ does not fit any particular feature of the time series $X_n(t)$), and assuming that the residual of the model is gaussian, then the quantity on the left side is actually distributed under a *Student* density. For practical application, the t scores are then converted into a normal variable z through standard procedures. The resulting map $z(n), n = 1..N$ is our first statistical map.

This map can be thresholded for a certain significance value (or P value): still under the null hypothesis, the probability that z is above a given threshold t (usually in absolute value) can be derived analytically. The inference consists in rejecting the null hypothesis given the statistical score.

One can also take a set of contrasts $\Gamma = \{\gamma_1, .., \gamma_I\}$ and derive a statistical score to assess the squared norm of the vector $\Gamma b(n)$ with respect to its dispersion:

$$\frac{(\Gamma \hat{b}(n))^T (\Gamma \hat{b}(n))}{\Gamma^T \Lambda_b(n) \Gamma} \frac{d_2}{d_1} \sim F_{d_1, d_2}(n) \tag{3.11}$$

where F_{d_1,d_2} is the Fisher distribution with d_1 and d_2 degrees of freedom; d_1 and d_2 are respectively the number of degrees of freedom from the numerator and the denominator, and are derived from standard procedures. The resulting map can be used similarly as a t map. An important difference with t maps is that the information about the sign of the effect (positive or negative) is lost.

The last topic that makes up the standard SPM analysis is the introduction of map-wise threshold; indeed, the P-value defined above is voxel-based, but we are interested in controlling the number of false positive voxels for a given map. The simplest way to do that is to assume that all the voxels are independent (in the sense that their associated signals can be viewed as statistically independent). Then the Bonferroni correction procedure applies: Let π be the probability that any voxel in the image has a z score above the threshold t under the null hypothesis, and ρ the same probability for one given voxel; then $\pi = \rho^N$. Given π and N, one straightforwardly derives ρ and then t. But this method suffers from several disadvantages:

- The method does not take into account the spatial correlations that violate the independence hypothesis.
- Consequently, an oversampling of the data increases the number of voxels N and thus the threshold t, whereas we still have the same map. The reason is that the increase in N is related to an increase in the spatial correlation.
- Last, the method does not take into account the spatial structure of supra-threshold voxels: a set of clustered voxels more likely represents an activation pattern than isolated voxels.

For these reasons, a more sophisticated framework has been developed. The idea is that, under the null hypothesis, the z-map is a gaussian random field of dimension D (2 or 3), in a volume V, with a given smoothness. It is possible to derive the expected Euler-Poincaré characteristic χ_t of this field once thresholded at the level t:

$$\mathbb{E}(\chi_t) = V|\Lambda|^{\frac{1}{2}}(2\pi)^{-\frac{D+1}{2}}H_D(t)e^{-\frac{t^2}{2}} \tag{3.12}$$

where Λ is the covariance matrix of the field (basically, its smoothness or inverse point spread function) and H_D the Hermite polynomial of degree D. Now, for a high value of t, one has

$$P(z_{max} \geq t) \approx P(\chi_t > 1) \approx 1 - e^{-E(\chi_t)} \approx E(\chi_t) \tag{3.13}$$

Equations (3.12) and (3.13) together give a new way of setting a threshold on a smooth gaussian random field. A concept of interest is the -dimensionless- notion of the number of resolution elements ($RESELS$): This number is given by

$$RESELS = \frac{V}{\prod_{i=1}^{D} FWHM_i}, \tag{3.14}$$

where $FWHM_i$ represents the full width at half maximum of the spatial filter associated with the map, in direction i, and related to the smoothness of the field by

$$V|\Lambda|^{\frac{1}{2}} = RESELS(4\log(2))^{\frac{D}{2}} \tag{3.15}$$

This method has been further improved with the introduction of the spatial extent of the activated areas [89], joint test on the height and size of supra-threshold clusters [178], and the robust estimation of smoothness in presence of activations [128]. Note that this approach is compatible with spatial smoothing of the data (more precisely, this model encourages smoothing, which makes the hypothesis of a stationary gaussian random field more credible).

3.2.2 Variations around the G.L.M.

A great part of the fMRI literature has concentrated on partial improvement of aspects of the linear model described above. While some authors have essentially tried to optimize the parameters of the GLM (especially the hemodynamic model [45], see the next chapter for more details), some others have refined the framework. We make here a non-exhaustive review of those contributions.

Variations in noise modeling

Nuisance space modeling: A important concern in fMRI analysis is the ability to find unbiased estimates of the regressors dispersion (see equation (3.4)). In many instances, it has been observed that confounds are present in the dataset; their main effect is an overestimate of the residual variance, then in the parameters dispersion, hence a loss in efficiency. The standard solution implemented in SPM is the removal of low frequencies by high pass filtering of the data [77, chapter 3]; some authors have proposed to estimate a nuisance subspace from the data [6]. The removal of the nuisance has the main effect of increasing the likelihood of the data (see further).

Autoregressive models: Given that the noise is temporally correlated, it seems necessary to take this into account in the dispersion estimate of the linear model. Friston et al. have thus introduced an AR(1) noise model in the GLM [84], but this model is uniform across the dataset, which is suboptimal [226]. Other authors use more general AR(p) models [124], [223], AR(1)+white noise model [182] [30] [181] or estimations from the frequency domain [123]. Complex models, including ARMA noise and trends, are possible, but the estimation of all the quantities involved requires the use of nonlinear optimization methods [141]. Wavelet resampling has also been proposed to assess the presence of activation with colored noise [29].

Some statistical refinements

Maximum Likelihood framework: The maximum likelihood (ML) framework is attractive for the problem of parameter estimation. Indeed, the least-squares estimation procedure in equations (3.3) or (3.5) yields also essentially maximum likelihood estimators of $b(n)$ under the gaussian noise hypothesis (respectively white or with known covariance). This idea has been generalized in [163] to deal with complex data.

Bayesian framework: Another statistical framework has received much attention for fMRI data analysis: the Bayesian framework. The main difference with the ML framework is the introduction of priors in the statistical model of the data; the inference is then made a posteriori, given the prior and the data fit. For example, in [124], the authors use Jeffrey's rule in order to have uninformative prior, hence no bias on the final posterior distribution.

Bayesian inference is now gaining popularity in the fMRI data analysis community; besides many Bayesian works on the hemodynamic response (see next chapter), a recent contribution is the introduction of hierarchical models in the case of *i)* multi-subject study, *ii)* inference when temporal correlations are considered. The introduction of Bayesian concepts is in fact a way to optimize the parameters (response amplitude, residual covariance)

in an expectation-maximization fashion [87] [81]. Bayesian formalism is a practical way to introduce some priors (e.g. spatial smoothness) in the estimation of effects, as has been done by many authors (see the paragraph on spatially regularized estimations).

Mixture model: Another way to assess the presence of activation in a set of voxels is to introduce a statistical distribution for the activated data, and to finally select the voxels that are more likely to be activated given that model. Unfortunately, this derivation can be done properly only for particular experimental paradigms, and still with restrictive hypotheses on the activation patterns [58]. A possible, still rarely explored way is to model the regressors distribution with a mixture of gaussians, with one mode representing the null hypothesis [15].

Non-parametric Permutation testing: For the specific goal of hypothesis testing, permutation tests are an alternative to the use of analytic functions as implemented in SPM [166]. Nevertheless, these methods are computationally demanding and ill-adapted for temporally or spatially correlated data [29]. This explains why their use is rare in fMRI data, or only with whitening procedures. However, permutation testing is of frequent use in the case of multi-subject studies, when two populations have to be contrasted.

Use of false discovery rate This technique refers to the specific problem of the correction necessary due to the multiple comparison problem. The standard solution has been the modeling of images as a gaussian random field. However, this requires very smooth data, hence intensive smoothing and blurring. Recently, a procedure has been proposed in [90] to overcome this problem: the idea is not to control the number of false positive above a given threshold, but to find a threshold for which a given fraction of subthreshold voxels will be false positive on average. Since this method works on the rate of false positive voxels rather than the number of false positive, it is not sensitive to the number of comparisons performed, unlike the Bonferroni correction. It can then be reliably used on unsmoothed dataset. However, a potential disadvantage of the method is that the threshold decreases when more voxels are *activated*, allowing for a weaker control of false positives.

Introduction of non-linearities

The linearity of the response with respect to the stimulation is controversial. It is rare that the stimulation is associated with an intensity parameter, so that the amplitude linearity is difficult to check (an exception is the case where the stimulation intensity is defined by a contrast or a frequency, e.g. for visual stimuli); but time invariance of the response poses more problems, essentially when dealing with event-related designs, in which case the succession of the stimulations can be very quick [219] [157]. This has incited Friston et al. to introduce a non-linear term as an additional term in the G.L.M. [83]. We review and discuss the question in section 4.1.

Analysis in the frequency domain

Some authors have introduced estimations in the frequency domain; this procedure has several advantages: first, the low frequencies of the noise are well separated from the remainder, which helps for signal identification, in particular for the hemodynamic response estimation [137]; second, the Fourier coefficients are approximately independent in the limit of infinitely long time series, allowing for unbiased hypothesis testing [147]. The evident disadvantage is that this procedure is fully correct only for periodic stimulations,

and does not help much in the case of multiple stimuli. However, this can be used for the detection of some physiological rhythms as in [158] [106].

Spatially regularized models

Another controversial point in fMRI data analysis is the use of smoothing. Indeed, some authors have advocated that it would be preferable to use spatial regularization of the noise estimate instead of smoothing both activation and noise [199] [181] [18] [19] [223] [123]. A potential limitation of these models is the adequacy of the spatial model, which is usually isotropic, whereas a more proper model should be based on the anatomical geometry of the cortex (at the resolution of fMRI, the cortex is essentially a folded surface).

Another way to regularize spatially fMRI data is the introduction of Markov random fields [53]. A related method is the spatial mixture model proposed in [108]. The inference based on Markov Random Fields can also be interpreted within the Bayesian framework, where the prior probability is given by the voxel information (time course), while the posterior probability takes into account the contextual information carried by the neighboring voxels. Last, the optimal solution of Markov Random Fields labeling is hard to obtain; for a two-classes problem, the normalized graph cuts methodology presented in [129] is an elegant and optimal solution. But one should recall that all these methods are obviously limited by the arbitrariness in the local interaction model.

3.2.3 Non-parametric methods

More marginally, some authors have worked on methods that demand less hypotheses, especially avoiding the formulation in terms of linear combination of effects (equations (3.1), (3.2)); but this is often at the expense of interpretability. Let us give a quick account of these more original works.

Use of the Kolmogorov-Smirnov test

The Kolmogorov-Smirnov (KS) test has been widely used for assessment of an activation without any assumption on the signal and noise distribution (see e.g. [184]). The idea is to compute the empirical distribution of the signal given an experimental condition, and then to test the identity of the obtained distributions. Rejection of the identity by the KS test is interpreted as an activation. However, the estimation of signal distribution is difficult for the short fMRI time series; thus the method is not efficient in general. It is furthermore very sensitive to non-stationarities in the signal [1], which is problematic for fMRI data. Last, a difference between two statistical distributions is easily interpretable if the difference consists in a shift of the mean, but not if the difference involves higher order moments of the distributions.

Use of information theoretic measures

A very similar approach has been proposed with the Kullback-Leibler divergence between the statistical distribution of the signal and the distributions of the latter once marginalized with respect to discrete values of the experimental paradigm; this amounts to computing the mutual information between the experimental paradigm and the signal [215] [129].

A nice feature of this approach is that it adapts well for more than two experimental conditions. But very similar critics as for the KS test can be addressed to this method, whose spatial maps are furthermore difficult to threshold [206]. A more adapted method has been proposed with the introduction of the entropy rate, instead of the entropy, to characterize the time courses, since it accounts for temporal information, and especially non-stationarities [64]. Some close ideas will be developed in section 4.3.

Phase synchronization

An original contribution to non-linear analysis of fMRI time series is the introduction of the concept of instantaneous phase in a signal, which is based on the Hilbert transform of the temporal signals. The instantaneous phase of the signal can then be compared with that of the assumed task-related response; a constant lag between the phases indicates a synchrony between the time course and the assumed effect, hence an activation [136]. There remains nonetheless the difficulty of the sensitivity of this technique with respect to noise and trends inevitably present in the data; moreover, the activation amplitude is not really quantified.

The idea of phase synchronization has also been used in [213], though with slight differences in the formulation. However, the algorithms proposed by the authors contains ad hoc steps, making it hardly generalizable; statistical inference is not straightforward -and often suboptimal- with such a complex modeling.

Markovian modeling

A property of Markov chains that makes them attractive for biological time series is that they implement a concept of causality: an effect (observation or experimental condition) at time t explicitly determines the observation at times $t+1$. An original application of Bayesian analysis has been the Markovian modeling of fMRI time series, where the prior is in fact the past values of the data [113]. A different use of Markov chain models has been proposed in [206], where the authors propose to simulate asymptotic signal distributions conditionally locked to conditions of the experimental paradigm. This is of course valuable only for block design experiments.

3.3 Multivariate methods: review and discussion

In this section, we develop the main multivariate methods that are used for exploratory analysis of the data. Once again, we have no assurance of giving a complete panel of the existing techniques. These methods are multivariate in the sense that all voxels are considered simultaneously. Principal and Independent Components Analysis are two methods that produce a generative model of the data; they are based on the assumption that the true model that generated the data is low dimensional, and that it can be recovered through a kind of *unmixing* technique. By contrast, clustering techniques and self-organizing maps are based on the assumption that the set of voxels can be split into different sets on which one effect is predominant.

3.3.1 Principal Components Analysis and related methods

The Principal Components Analysis (PCA) of fMRI data is often performed through a Singular Value Decomposition (SVD) technique, after centering of the dataset. The SVD simply decomposes the dataset into mutually orthogonal spatio-temporal components. Recalling that we consider a given dataset (a session) as a $N \times T$ matrix (N = number of voxels considered, T = length of the time series) which has been centered (the sum of each row is null), the SVD of X is

$$X = U\Sigma V^T; \tag{3.16}$$

where U and V are $N \times N$ and $T \times T$ orthogonal matrices, and Σ a $N \times T$ matrix with non-zero elements only on its diagonal. If one further requires that the diagonal elements of Σ are decreasing, the decomposition is unique up to change of sign of the columns of U and V; U and V respectively diagonalize XX^T and X^TX. The columns of U are interpreted as a set of images, and those of V as a set of signals. Performing a SVD of a raw dataset is a means to explore it which has rarely been employed; at least the data is corrected for some confounds(low frequency components) [80] [88] [4]. But a more common way to use it is to define a space of interest, which is analogous to a $K \times T$ design matrix of the experiment G, and to perform the SVD of XG^T. This technique has been further developed with different normalizations of the matrices X and G, yielding the Canonical Variate Analysis (CVA) method [79], Partial Least Squares (PLS) [151] and the Multivariate Linear Model (MLM) [224] [126]. The methods that use a design matrix are not really exploratory, since nothing is known outside the space of interest spanned by the rows of G.

Some authors have also tried to improve the basic setting with introducing spatial smoothness with Markov property [176].

We go more into the details of this technique and associated questions in chapter 4.

Nonlinear PCA: Let us also quickly mention the introduction of *non-linear PCA* by Friston et al. in [74] and [75]. In this work, Friston et al. introduced the concept of interaction of the spatial modes, and proposed a neural network architecture to estimate the deviation from the linear PCA. But this proposition has elicited little interest in the neuroimaging community, probably because of the difficult assessment of such results.

3.3.2 Independent Components Analysis

Let us recall that both the spatial and temporal components of the SVD are mutually orthogonal. By contrast, Independent Components Analysis (ICA) is devoted to the derivation of statistically independent components either in the spatial or in the temporal domain (but not both). This makes sense, since the independence of random variables is a much more constrained problem than their decorrelation. We will study this problem and its connection with information theory more thoroughly in chapter 4. Let us mention that ICA has much more often been used with a spatial independence criterion than with a temporal independence criterion, the most obvious reason being that there are much more voxels than time points to assess statistical independence.

The basic setting is the following (say, for spatial ICA):
The dataset X is a set of images $X(t)$ which are viewed as the *superposition* of independent images, which are called *sources* and noted S in the ICA language.

$$X^T = MS + E \tag{3.17}$$

The superposition is modeled by a mixing matrix M, and there is an additional noise term. In the most general setting, X, M, S and E are $N \times T$, $T \times K$, $K \times N$ and $T \times N$ matrices of rank T, $K < T$, K and $T - K$. K is the number of independent sources that have generated the observed data (it is an unknown parameter), and E is the residual noise [1]. The solution of the problem consists in estimating the matrices W and S so that

$$S = W(X - E)^T \tag{3.18}$$

The $K \times T$ matrix W is called the *unmixing* matrix and can be viewed as a generalized inverse of the mixing matrix M. Especially, the resulting matrix S can be interpreted as a set of *activation maps* associated with the different independent effects present within the dataset.

The solution of this problem involves

- The definition of an independence criterion(see chapter 4 and [57]), which determines a practical algorithm for the solution of (3.18).
- A separation between signal and noise, and, accordingly, an estimation of the rank K of the generative model.
- The interpretation of the spatial maps obtained from the algorithm [68].

This technique is now well established (see [153], [154], [35], [101], [54], [38]), and is probably the exploratory technique used most frequently for the study of fMRI data. Note that temporal ICA has also been proposed ([37], [171], [145]).

We can also notice that a link has been made between ICA and the general linear model in a *hybrid approach* [152]; this approach probably eases the interpretation of the ICA method.

Last, ICA -like PCA, though to a greater extent- can be seen as a way of denoising the data by separating the different effects present in the data (say, signal, confounds and noise) into different components [212].

CCA, an intermediate method: Canonical Correlation Analysis (CCA) is a means to detect the subcomponents of two multivariate datasets that are maximally correlated. It has been proposed as a way of investigating fMRI datasets in [72], with two possible applications: *i)* the derivation of the temporal components that are maximally correlated at lag 1, *ii)* the derivation of the most spatially smooth maps of the dataset. Speaking informally, this is an equivalent of ICA with the advantage of a non-iterative procedure (the criterion being bilinear after some normalization of the data). However, important issues such as dimension selection, are not solved with this method.

[1] The concept of noise is problematic here, since no specification is made neither on signal nor on noise. In fact, E is simply neglected in the hypothesis of noise free mixing, and is in practice the residual of the initial PCA performed for dimension reduction purposes.

3.3.3 Clustering

Clustering is another exploratory method based on the following statistical viewpoint: the dataset X is a set of N features (the temporal time courses) that belong to a given signal manifold or feature space $\mathcal{F}$. The distribution of the data in $\mathcal{F}$ can be modeled as a multimodal distribution; each mode will be characterized as a data cluster. The literature on fMRI data clustering deals with the following problems:

- The method that is used for deriving the final clusters: among others, C-means algorithm [10] , fuzzy C-means [12], [59], dynamical cluster analysis [13], deterministic annealing [220] have been proposed.

- The definition of the feature space $\mathcal{F}$, that is, of the metric that is used to quantify the similarity between time courses: This similarity can be measured by the Euclidean distance in the signal space of origin [220] or another distance based on correlations [60], or a Mahalanobis metric [96]. The choice of a correct metric is not obvious; an Euclidean metric can be a suboptimal choice [97] for high dimensional spaces $\mathcal{F}$.

- The quality of clustering results is difficult to assess. To solve these problems, authors have proposed some heuristics [60] [161], but these are not necessarily optimal; moreover they are used after convergence of the algorithms, or sometimes yield complex multistage strategies [59].

- This is related to the problem of the selection of the number of clusters [60]. It is intuitively clear that the choice of a given number of clusters corresponds to a certain bias/variance tradeoff, but this tradeoff is usually implicit.

These methods are efficient [12] from a computational point of view and can isolate interesting patterns in the data; this explains their relative success for the analysis of fMRI data. However, the solutions proposed to all the above problems are rather heuristic in nature. We address the question with a new clustering method in chapter 4. Additionally, clustering algorithms can spend a lot of efforts trying to isolate patterns of no interest; this is due to the absence of prior information in these methods. Recent works [67] suggest that introducing anatomical and functional information (i.e. a space of interest) improves both the generality and the precision of the method. Last, unlike PCA or ICA, they do not decompose the data into components, and thus do not benefit from the associated denoising effect.

3.3.4 Self-organizing maps

Self-organizing maps are a variant of the clustering method. The particularity of self-organizing maps is the use of a 2-dimensional map to represent the feature-space [62]. This map represents cluster centers, which are updated by taking into consideration randomly selected features of the dataset. The difference with respect to classical clustering methods is that the incorporation of a new feature into the nearest cluster has also an impact on the neighboring clusters, giving some consistency to the map. However this clustering method is quite technical, and requires the use of several non-interpretable parameters, so that the method is still of limited use for fMRI data analysis [165] [43].

3.3.5 Multivariate spectral analysis

Let us mention an original work on multivariate exploratory analysis of fMRI data based on the spectral representation of the data [162]. The authors present a way to derive spectral parameters (with confidence intervals) from the data using Wiener theory: these parameters are computed from each pair of voxels of the dataset: *i)* a coherence measure, that basically measures the synchronization between the voxel time courses at a frequency of interest, and *ii)* the phase lead, which measures the advance of one voxel over the other one. These quantities are original, and meaningful, but there is a real problem in analyzing the resulting $N \times N$ matrices in a systematic way. Moreover, this technique assumes that the information of interest is concentrated at a given frequency, which is only true for periodical experimental designs, and problematic if several experimental conditions alternate.

Beyond the methods

Let us make a final point on this chapter by noticing that the inflating corpus of existing analysis methods makes the choice of the proper methodology embarrassing. This has incited some authors to perform some meta analysis, i.e. some analyses that involve the parallel derivation of some of the above methods, each one being taken as a statistical signature of the data. This is notably the case of [104]; a toolbox has been created for practical use [105]. This approach has the advantage of combining the specific inputs of the included models. On the other hand, it may be computationally expensive and the combination of different statistical *signatures* is both difficult to perform and to interpret.

3.4 Conclusion

We consider that univariate analysis of fMRI data has received much attention and thus theoretically and practically sufficient solutions (though some particular aspects are still currently improved technically). We will thus rather work in the margins of this main approach, considering questions that are left behind by the standard GLM: what is actually the shape, amplitude or delay of the hemodynamic response associated with the different stimulations (chapter 4) ? What can we do from a strictly statistical point of view, i.e. without considering temporal modeling (multivariate approach, chapter 5) ? We will focus on the dynamics of the dataset, with either linear (chapter 6) or non-linear (chapter7) methods. In an exploratory (multivariate) spirit, we will also consider fMRI datasets as the embedding in a high dimensional space (the images) of a low dimensional system and study it with nonlinear methods (chapter 8). These different points of view will finally help us to propose an alternative analysis scheme that is data-driven, but with the ability to focus on signals of interest (chapter 9).

Chapter 4

Temporal modeling of fMRI data

In the study of fMRI data, the first concern is to build a temporal model of all phenomena that appear in the fMRI dataset. Though we do not pretend to solve definitively this issue, we make in this chapter a review of the main ideas that bridge the gap between our knowledge of the underlying biological and physical phenomena of the experiment, and the modeling of the empirical data. First, we summarize some common knowledge about the BOLD signal -which characterizes the signal of interest present in the dataset- and the fMRI measurements -which also include different, and sometimes problematic effects. Then we propose a general mathematical framework (the prediction theory) to embed this knowledge without introducing too many constraints of the nature of the signal. We make a first connection with information theory, which characterizes the information associated with stochastic processes; this allows for a generalization of the Maximum likelihood approach, where the structure of the representation is optimized with respect to a complexity criterion. Last we discuss some of the basic technical assumptions that underpin temporal modeling (noise normality, response linearity) and -briefly- the extension to nonlinear dynamical system fitting.

4.1 The input from biologists and experimenters

4.1.1 Phenomenological description of the BOLD effect

In figure 4.1, we reproduce the description of a typical signal given in [33]. The main features of this scheme are accepted by many authors, though with minor differences concerning the presence of an initial dip (which has not reached a consensus), and the post stimulation undershoot. Note that the description concerns more the shape of the stimulation than the values of the change -a few percent of signal change, depending on the stimulation, the area under consideration, the acquisition sequence and the scanner.

Let us recall that the BOLD effect is a quite complicated effect that depends on changes in cerebral blood flow (CBF), cerebral blood volume (CBV) and cerebral metabolic rate of oxygen ($CMRO_2$). We sum up these facts in figure 4.2. Then we go on with several points of interest concerning the BOLD effect.

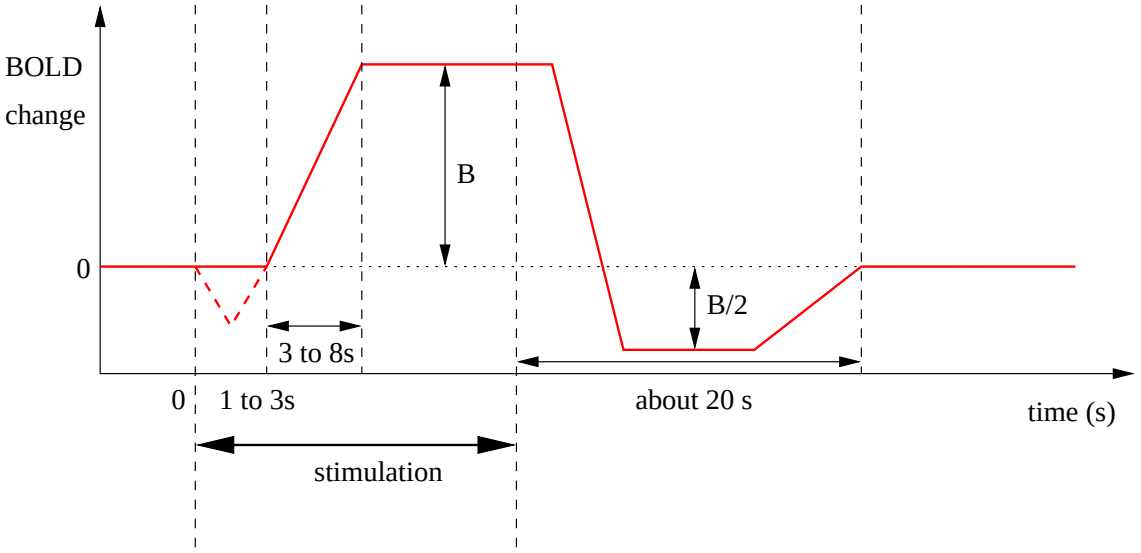


Figure 4.1: Schematic representation of the BOLD effect relative to a given stimulation. Note that the quantity involved is a percentage of signal change; since there is no prototypical value of change (which are typically in a range of 1% to 5%, depending on the task, the area of the response, the acquisition sequence) we have not precisely quantified the ordinate axis. The effect includes an initial delay of 1-3 s after the initialization of the stimulation, followed by a ramp of 3-8 s before plateau signal change is reached. After the end of the stimulus, the signal declines, and often undershoots the original baseline, with a value which is half of the peak. The undershoot takes around 20 s to resolve. A variant of the scheme is in the existence of an initial dip of 1-2 s at the beginning of the stimulation.

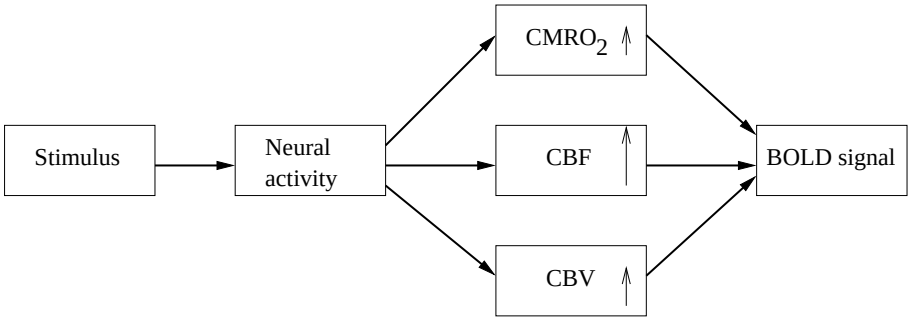


Figure 4.2: The chain of events leading to the BOLD signal.
The stimulus induces neural activity, which in counterparts triggers metabolic activity in the form of a large increase of cerebral blood flow (CBF), a moderate increase of cerebral blood volume (CBV) and a small increase of cerebral metabolic rate of oxygen ($CMRO_2$). The three effects result in the BOLD change.

Location of BOLD signal changes

Because the venous vessels undergo the largest changes in deoxyhemoglobin content, the largest BOLD signal probably arise around draining veins, which may be separated from the area of neuronal activation by as much as 1 cm or more. It has been shown [135] that the localization of BOLD and CBF changes do not always coincide. However, due to the small signal changes obtained usually, it does not seem possible to correct for such spatial bias.

The relationship between BOLD effect and neural activity

If it is commonly admitted that the BOLD effect actually reflects neuronal activity, a quantitative relationship between these two effects has not been established yet. At this point, let us simply mention here that neuronal activity can be quantified in at least two ways: *i)* the average rate of generation of action potentials, and *ii)* the average rate of neurotransmitter recycling within the region. The question of which quantity mainly triggers the BOLD signal is still under debate. However, an important contribution has been made in [142], which describes simultaneous recordings of neural signals -local field potentials and spiking activity- and fMRI responses from the visual cortex of monkeys. It was shown that BOLD activity is better explained, in the framework of linear systems theory, by local field potentials than by spiking activity. This suggests that the BOLD contrast mechanism reflects the input and intra-cortical processing of a given area rather than its spiking output.

The linearity of the BOLD response

Given the uncertainties concerning quantitative analysis of neuronal activity, the question has turned into the following: Can one establish a linear relationship between the stimulation timing and amplitude, and the final BOLD signal ? Mathematically, this question triggers the possibility of defining the BOLD response as linear time invariant filtering process (i.e. a convolution) of the stimulation time course. Several studies have been done experimentally comparing the response to brief stimuli to the response to longer stimuli, using visual [219], auditory [93] and motor [93] stimuli. The consistent result of these studies (see also section 4.1.2) is that the response is roughly linear, but that there is a definite nonlinear component: the response to a brief stimulus appears to be stronger than what would be expected given the response to a longer stimulus.
The nature of this nonlinearity may be related to either of the effects described in figure 4.2. First, the relationship between stimulation and neuronal response is certainly nonlinear, so that the main question is rather whether it is the only source of nonlinearity. While the relationship between neuronal activity and metabolic activity is often assumed linear, the relation of the latter with the BOLD effect can be affected by a ceiling effect, or by a different timing of CBF and CBV changes.
Let us conclude that it is not safe to assume uniquely a linear response, which is however the general hypothesis.

Dynamics of the BOLD response

Although most studies are mainly concerned by the difference between the basis signal level and the activation one, it is of interest to study the transient parts of the signal, at the beginning and end of the stimuli. Once again, these transiences may be either the effect of changes in neuronal activity and/or in the metabolic counterparts (combined change of CBV, CBF and $CMRO_2$). This difficult issue can nevertheless not be solved by the use of BOLD fMRI alone.
For example, the signal undershoot after stimulation is attributed to the following effect: the CBV remains elevated after stimulation while the CBF returns to baseline, so that the deoxyhemoglobin content remains elevated after flow has returned to a resting level. A biomechanical model called the balloon model has been proposed to explain this [31].
The initial dip is another important but controversial aspects of the BOLD response dynamics. First, the observation of this effect with BOLD fMRI is not systematic, and could be achieved only at high field strengths. Second, it has sometimes been interpreted as an initial increase in local deoxyhemoglobin prior to the larger and latter decrease, which suggests that it may map more tightly the areas of neural activity than the deoxyhemoglobin decrease, which causes the massive part of the BOLD effect [146]. But this hypothesis is not the unique one, leaving the question unanswered [32].

Let us end up this section with some remarks inspired by the phenomenological study of the BOLD response:

- Temporally, an activation signal is well characterized by a delayed increase.
- Spatially, the location of such increase is not necessarily accurate, but gives a reasonable approximation of the neuronal activation locus
- In a general approach, it is not safe to assume uniquely a linear response. This is however a frequent hypothesis, which is justified if the stimulus duration and inter stimulus interval are not too short.
- One should preserve some flexibility in the choice of the hemodynamic or impulse model in order to fit correctly the transient parts of the signal.

A perhaps more complex question, that we will not address, is the existence of brain networks including excitatory and inhibitory systems, which yields a more complex dynamical evolution of metabolic activity.

4.1.2 The hemodynamic response filter: history and discussion

The study of the hemodynamic response function (hrf) is one of the main efforts in fMRI data analysis literature. Schematically, the modeling and quantitative study of the hemodynamic response has progressed in several steps:

- the use of an optimal model within a parametric family
- the introduction and use of physiologically meaningful parameters
- the use of non-parametric models

- the study of the regional differences
- the study of the linearity

Let us give an account of these successive steps.

Parametric models

The first approach has been the use of heuristic model to describe the hemodynamic response. In fact, its purpose was to detect reliable activations by taking into account the delay in the response peak. The successive models were :

The model proposed in [82] was a Poisson filter, which is a particular choice within the gamma family. The filter can be parametered by a delay value, but the dispersion is also equal to the delay value, which is too restrictive.

A gaussian model has been proposed in [183]. It is discussed more thoroughly in the next paragraph.

Lange and Zeger [137] have introduced a gamma function, whose parameters had to be estimated in a computationally difficult manner.

A more physiologically plausible model has been retained for the hrf model of the SPM software, which is the difference between two gamma functions:

$$h(t) = \left(\frac{t}{d}\right)^a \exp\left(-\frac{t-d}{b}\right) - c\left(\frac{t}{d'}\right)^{a'} \exp\left(-\frac{t-d'}{b'}\right) \tag{4.1}$$

where $d = ab$ is the time to peak, $d' = a'b'$ is the time to undershoot, with $a = 6, a' = 12$ and $b = b' = 0.9s$; $c = 0.35$. This model has become the most frequent one (see e.g. [93]), since it models both activation, undershoot, and that both modes are not symmetrical.

Physiologically oriented parametric models

Among the parametric models used for the characterization of the hemodynamic response function (hrf), some of them were found preferable: for example, some particular parameterization of a gaussian kernel enables an interpretation of the estimated parameters, like in [132], [202]:

$$h(t, \beta) = \frac{\beta_0}{\beta_1\sqrt{2\pi}} \exp\left(-\frac{(t-\beta_2)^2}{2\beta_1^2}\right) + \beta_3 \tag{4.2}$$

with $\beta = (\beta_0, .., \beta_3)$; β_0 can be interpreted as the gain in the hrf, β_1 as the duration or dispersion of the hrf, β_2 as the delay of the hrf and β_3 as the baseline level. But a limitation of this model is the shape of the kernel, which is symmetrical around its peak, which is not confirmed by inspection of the hrf. The gaussian kernel shape also neglects the post-stimulus undershoot.
This model has been generalized in [133]. In particular, an asymmetric gaussian model has been introduced, but it induces a heavy computational load; another more physiological model was also introduced with inflow, outflow rates, but with still worse convergence.

The SPM hrf model can also be partially interpreted in terms of physiological parameters (equation 4.1).

A special emphasis has been put on the estimation of the delay parameter, which is of course more crucial in the analysis of the response dynamics. A first attempt had been

the introduction of the temporal derivative of the SPM standard hrf into the linear model, which was meant to account for and statistically assess the delay [78]. But the method has been shown not to allow for correct delay estimation [139]. In the latter work, an unbiased method is proposed for estimation of the delay.

Another delay estimation procedure is described in [192], which is based on the spectral study of the voxel time series. However, this study also showed the confounding effect of low frequency effects present in the data.

Use of non-parametric model

A more general approach to the hrf estimation is to introduce a finite impulse response (FIR) filter, and to optimize its coefficients given the data and some priors on the response. This approach has been pioneered in [98], and recast in a simplified framework in [149]. The estimation procedure is based on a Bayesian framework for the tuning of model hyperparameters; it has been generalized in [44], [150] and is probably close to optimal for estimation purposes. The computational load is heavy, but can be reduced with adequate approximations.

Note that Miezin et al [156] use a two-steps procedure, with first a FIR estimate followed by a fit to a 3-parameter delayed gamma function.

However, even when assuming that a best estimate of the hemodynamic response is derived, there remains some issues: the (spatial) non-stationarity of the hrf is taken into account in the model, but it would be interesting to allow for a study of the different hrf patterns across the dataset. Another concern is the non-linearity of the global response.

Regional differences

The regional differences for the response delay have been well characterized by the aforementioned models [98], [192], [139]. A widespread hypothesis is that the signal due to the veins is relatively late with respect to parenchymal signal, but there is also variability within the parenchyma itself. A perhaps more complete study has been presented in [156]. Let us recall some conclusions made by the author:

- The amplitude and time to peak of a hemodynamic response are quite reproducible for a subject and a stimulus, within a given region.
- The amplitude of the response depends on the trial presentation rate: closer stimulations induce weaker responses.
- The timing and amplitude of the responses differ between regions, so that no systematic relationship can be obtained.
- Across subjects, the amplitude of the response shows no significant correlation with timing of the response.
- Statistically significant timing differences can be observed between regions, allowing for specific analysis of those relative timings.

Finally, within-subject variability in terms of activation amplitude and timing has been described in [164] as high, and in some instances, comparable to between-subject variability. Of course, all these phenomena have been described in an empirical manner, and do not yield straightforward interpretation in terms of neural responses.

Response Linearity

Last, a strategic issue in the study of the hrf is that the actual response to a stimulation does probably not follow the simple linear convolution model. The consequence of this hypothesis is that the response derived from a quick stimulation does not help much for the case of long or closely presented stimuli. This specific issues has been studied in [49], [219], [157], [21] [116] [174]. In [219] it was shown that short stimulation periods and intensity yielded strong deviation from the otherwise acceptable linear model. This observation is also made in [93], with 4 s of inter stimulus interval being a reasonable limit of the linearity model. In [157], the study was done on both BOLD and CBF with a perfusion technique, and supported the idea of a linearity between neural activity -which is non-linearly related to the stimulation- and CBF, and a non-linearity (a saturation) between CBF and BOLD signal. In short, the BOLD response to short stimuli overpredicts the response to long stimuli. In [21], the authors further showed that the non-linearity in the response was not spatially stationary. So did Huettel and McCarthy in [116], who considered rather the non-linearity related to the refractory period after one first stimulation. Finally, in [174], the link with field strength is studied -high fields induce weaker non-linearities- and some evidence is given for a tissue (gray matter) specific nonlinearity, interpreted as a switch effect associated to activation; this also confirms the fact that deviations from linearity are significant for short stimulation only.
The link with standard regression models is yet not systematic, even if a connection has been made between the balloon model of Buxton et al. [31], and the explicit non-linear model from Friston et al. [83] in [85]. A perhaps more promising application has been given in [76], but its validation has still to be carried out.

In the remainder of this work, we will not be explicitly concerned with the precise hemodynamic response. Rather, keeping in mind all these difficulties in the use of fixed hemodynamic models, we will use flexible models to allow for correct signal detection given some simple hypotheses on the response.

4.1.3 Some sources of confounds in the acquisition

The noise that adds further to the signals from a particular voxel has two sources: random thermal noise and physiological fluctuations. The random thermal noise arises primarily from stray current signals in the receiver coil. This thermal noise is spread throughout the raw acquired data. The result in the reconstructed images can be accurately described as gaussian, independent from voxel to voxel, and with uniform variance. However, the variance of the signal over time is several times larger than what would be expected from thermal noise alone, and exhibits both spatial and temporal structure.

Physiological fluctuations include several effects: cardiac pulsations create a pressure wave that strongly affects the signal of flowing blood, but it also creates pulsations in CSF

and in the brain parenchyma itself. This motion creates non-uniform signal fluctuations. Usually, the cardiac rhythm is aliased in the data and can appear at low frequencies. The respiratory rhythm is easily observed due to the more adapted sampling time of fMRI data. There are also low frequencies resulting from either the scanner drift or slow physiological pulsations (vasomotion) [158].

It is well known that strongly autocorrelated noise (often called $\frac{1}{f}$ noise in the fMRI literature [221]) biases usual models and their statistical results; however this effect has been somehow minored in [148]. This kind of component is partially canceled by high-pass filtering (detrending) of the data [223].

4.1.4 Processing-related artifacts

Let us underline here another difficulty: if little is known or assumed about the neural activity associated with fMRI signals, the definition of the signals themselves is not unambiguous: it clearly depends on the preprocessing performed on the data. This effect is evident if we consider the spatial registration of the data, which is necessary to study the signal on a voxel basis, but it can also bias the data (see chapter 1, section 4).

Other effects, such as the temporal registration performed of the data to account for different acquisition timing are of paramount importance in the study of the response delay. Note that performing first the spatial or the temporal correction yields different datasets. Data smoothing is another controversial aspect of the problem, as global scaling or detrending (see chapter 1, section 4).

The optimization of preprocessing choices is not the topic of this work, but we insist that this question adds some relativity in the interpretation of experimental results. This is another incentive for not over-constraining the temporal model of the data.

4.2 A mathematical framework for the temporal model

Until the end of this chapter, we consider a time series $x(t)$, $t = 1, .., T$ e.g. a voxel time course, obtained from a fMRI dataset. We also assume, when necessary, that we have a description of the experiment, formulated as a paradigm. We ask two simple questions: First, does the time course contain information, and if yes, how can we measure it ? Second, How can we highlight the information of x that is related to the experimental paradigm ?

Before starting, let us point out the following question: if x is naturally a numerical variable, the experimental paradigm is most often defined as a sequence of events or epochs (that describe the subject state). The use of the latter in quantitative analysis requires the setting of a numerical variable. As most authors, we solve this problem by defining C experimental conditions -basically the different states/events of the experiment- and for each condition $c = 1..C$ a time course $P_c(t)$. For example $P_c(t) = 1$ means that the subject undergoes condition c at time t, while $P_c(t) = 0$ means that condition t was not realized at time t. The ensuing vector $P = P_c(t)$, $c = 1, .., C$, $t = 1, .., T$ will be treated as any numerical variable.

4.2.1 Stochastic processes and generative models

First, let us notice that the observation $x(t)$ is corrupted by noise -whatever the true nature of the noise. Consequently, it is advantageous to consider it as the realization of an underlying stochastic process $X(t)$. Analyzing a time series or a set of time series amounts then to the following inverse problem: *given the observation $x(t)$, try to identify the underlying stochastic process that generated it.* Such a task is possible only if we define which class of stochastic process the signal may belong to: we need some prior on the generative process of the data.

Considering all the previous work made on fMRI data, we will make the following choice:

- The part of the signal which is related to the paradigm is the result of the convolution of the latter with a finite impulse response (FIR) filter.
- The process that adds to this first component (it may be advisable to avoid the word *noise*, since it may reflect informative processes [22]) has an autoregressive structure.

Both priors suggest some way to deal with data analysis, respectively *projection* and *prediction*. Before going to these issues, let us clarify our mathematical framework. It is convenient to consider that the process $X(t), t \in 1, .., T$ (respectively $t \in \mathbb{Z}$) belongs to the Hilbert space of the finite (resp. infinite) processes of finite energy $\mathbb{R}^T$ (resp. l^2). Additionally, the process $X(t), t = 1, .., T$ will be assumed stationary; this means the following three properties:

$$\mathbb{E}(X(t)^2) < \infty \; \forall t \in [1, .., T] \tag{4.3}$$

$$\mathbb{E}(X(t)) = \mu \; (= 0) \; \forall t \in [1, .., T] \tag{4.4}$$

$$\mathbb{E}(X(r)X(t)) = \mathbb{E}(X(r+s)X(t+s)) \; \forall (r, t, s) \; / \; 1 < r, t < T - s \tag{4.5}$$

In other words, $X(t)$ has finite second order moments, a mean $\mu = 0$ and an autocovariance function $\gamma(u) = \mathbb{E}[X(t)X(t-u)]$. Note that none of these properties is in fact guaranteed for empirical data, but the assumption is necessary for the application of mathematical theory. However, detrending procedures (section 2.4.5) are explicitly designed for this hypothesis. Equations (4.3) to (4.5) concern only weak stationarity, and conditions on higher moments would be necessary for strict stationarity. However, concerning empirical data it is sufficient and more realistic to work on a weak stationarity hypothesis.

Projection: Assuming that the space of task-related activations has a closed structure - this is the case for *finite impulse response (FIR) functions* of the experimental stimulation- then noting this space S, the following decomposition holds:

$$X = \mathcal{P}_S X + Z, \tag{4.6}$$

where Z is simply defined as the residual of the equation, $\mathcal{P}_S$ is some projection operator, which can depend both on S and the assumed residual structure. The General Linear Model (section 3.2.1) is nothing but an implementation of this model. Noticing that projection can be seen as a minimization procedure

$\mathcal{P}_S x = \arg\min_{s\in S} \|x-s\|$, we see that this model requires two informations : the specification of S, and the specification of a norm (the ambient metric)$\|\|$.

Specification of S: it may include the canonical response to the paradigm, i.e. $S = span(h * P_c), c = 1..C$, but also some derivative of these patterns with respect to some parameters. Here, we use a more general FIR model $S = span(P_c(t-m)), c = 1..C, m = 1..M$. When the experimental paradigm reduces to a periodical stimulation of frequency ω, then $S = span(\sin(\omega t), \cos(\omega t))$ is an obvious choice.

Specification of the metric: this problem is related to the noise model ($z = x - s$), which is not known -this is the main weakness of this model. If z is assumed i.i.d gaussian, then the metric is the usual Euclidean one. If z has a known covariance structure (covariance matrix Λ) then the metric is the bilinear form associated to Λ^{-1}.

Prediction: The second approach is to consider that solving the inverse problem about the generative process amounts to finding a predictive model of the data. A typical model is autoregressive prediction, i.e. the definition of a predictor of $X(t+1)$ given its past values. In the Hilbertian framework, the predictor element of $\mathcal{H}_t = span(X(1), .., X(t))$ with minimum mean-square distance with $X(t+1)$ is noted $\mathcal{P}_{\mathcal{H}_t} X(t+1)$.

The solution of this problem is not difficult. Let us concentrate on the one step predictor

$$\hat{X}(t+1) = \mathcal{P}_{\mathcal{H}_t} X_{t+1},\ t > 0 \tag{4.7}$$

$\hat{X}(t+1) \in \mathcal{H}_t$ implies the existence of $\phi_{t1}, .., \phi_{tt}$ so that

$$\hat{X}(t+1) = \sum_{i=1}^{t} \phi_{ti} X(t+1-i), \tag{4.8}$$

where $\phi_{t1}, .., \phi_{tt}$ satisfy the prediction equation

$$\mathbb{E}(X(t+1)X(t+1-j)) = \mathbb{E}\left(\sum_{i=1}^{t} \phi_{ti} X(t+1-i)X(t+1-j)\right),\ j = 1, .., t \tag{4.9}$$

In terms of autocorrelation, this gives

$$\gamma(j) = \sum_{i=1}^{t} \phi_{ti}\gamma(i-j),\ j = 1, .., t, \tag{4.10}$$

i.e.

$$\Gamma_t \Phi_t = \gamma_t \tag{4.11}$$

where $\Gamma_t = [\gamma(i-j)]_{i,j=1..t}$, $\gamma_t = (\gamma(1), ..\gamma(t))$ and $\Phi_t = (\phi_{t1}, .., \phi_{tt})$. Equation (4.11) implies the existence of the solution, since matrix Γ_t is non singular under general conditions: *If $\gamma(0) > 0$ and $\gamma(h) > 0$ as $h \longrightarrow \infty$ then the covariance matrix Γ_t is non singular for every t.*

This simple setting, known as Yule-Walker equation, provides an excellent model for oscillating phenomena: for example, sinusoids are characterized by the equation $x(t+1) = 2\cos(\omega)x(t) - x(t-1)$. Moreover, the natural Euclidean metric can be used for the

estimation, since the residual is assumed white. But this approach clearly over-fits the data, since the problem has as many unknowns (ϕ_{Tt}) as data items ($\gamma(t)$).

Next we try to reconciliate prediction and projection approaches. Formally, the Wold decomposition together with the Kolmogorov formula give a framework for the identification of the generative process for X.

4.2.2 The Wold decomposition

The Wold decomposition of a process X defined for $t \in \mathbb{Z}$ gives a decomposition of the process into a deterministic process $V(t)$ and an uncorrelated stochastic process $U(t)$,

$$X(t) = U(t) + V(t) \tag{4.12}$$

First let us define $\sigma^2 = \mathbb{E}\Big[(\hat{X}(t+1) - X(t+1))^2\Big]$. Alternatively, σ can be defined by the Kolmogorov formula: let f the spectral density of the process X (defined a.e. on $[-\pi, \pi]$). Then

$$\sigma^2 = 2\pi \exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \log(f(\lambda))d\lambda\right) \tag{4.13}$$

This should be compared with the equivalent formula for the variance of X when considered as a random variable

$$v = \int_{-\pi}^{\pi} f(\lambda)d\lambda \tag{4.14}$$

$\sigma^2 < v$ always holds, and $\frac{\sigma^2}{v}$ is the proportion of the variance of the process that can be predicted given its past values.

Thus, if $\sigma^2 = 0$, the process is deterministic (i.e. fully predictable): $U(t) = 0$.

The other case (the unique case that makes sense for empirical data) is treated by the following theorem, stated in the hypothesis of an infinitely long time process (see [27] for details):
If $\sigma^2 > 0$, $X(t)$ can be expressed as

$$X(t) = \sum_{j=0}^{\infty} \psi_j Z(t-j) + V(t) \tag{4.15}$$

where
(i) $\psi_0 = 1$ *and* $\sum_{j=0}^{\infty} \psi_j^2 < \infty$,
(ii) $Z(t)$ *i.i.d.* $\sim \mathcal{N}(0, \sigma^2)$,
(iii) $Z(t) \in \mathcal{H}_t$, $\forall t \in \mathbb{Z}$,
(iv) $\mathbb{E}(Z(t)V(s)) = 0$, $\forall (s,t)$,
(v) $V(t) \in \bigcap_{n=-\infty}^{+\infty} \mathcal{H}_n$
(vi) $V(t)$ *is deterministic*

The stochastic process $U(t) = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ is thus defined as a $MA(\infty)$ process, with $Z(t)$ as innovation process.

Interestingly, the terms in (4.15) can be made explicit by letting

$$
\begin{aligned}
Z(t) &= X(t) - \mathcal{P}_{\mathcal{H}_{t-1}} X(t) && (4.16)\\
\psi_j &= \frac{1}{\sigma^2} < X(t), Z(t-j) > && (4.17)\\
V(t) &= X(t) - \sum_{j=0}^{\infty} \psi_j Z(t-j) && (4.18)
\end{aligned}
$$

However, these formula make sense, and in particular yield a non-trivial deterministic process $V(t)$ only in the case of infinitely long time series ($t \in \mathbb{Z}$), which is not accessible here.

A much similar conclusion can be drawn by looking at the spectral representation of the process: the spectral distribution of any stochastic process that satisfies the above hypotheses decomposes into two parts $f = f_U + f_V$, with f_U having an absolutely continuous density, while f_V is singular; f_U and f_V are the spectral densities of U and V respectively. Their definition naturally implies that only f_U is accessible from empirical (finite length) data.

However, the stochastic variance of the process can be at least approximated by empirical estimation of the process spectrum. In other words, we are able to quantify the intrinsic uncertainty of the process without more knowledge on its *content*! Moreover, we can consider some prior information in the definition of the deterministic component: the experimental paradigm, for instance. The next step relies then on the identification of $U(t)$ and $V(t)$ given our prior on the generative process.

4.2.3 Explicit prediction model

It seems that the analysis of the data in the frequency domain has several advantages in terms of predictability analysis. However, this is quite problematic: first, the estimation of the signal spectrum by FFT or related methods yields a poor estimate of the spectrum [27]. In practice, one has to use sophisticated methods to reduce the bias in the estimation of the spectrum [191]. Moreover, the study of the density does not consider important information as the phase of the Fourier coefficients. Last, unless the experimental paradigm is periodical, the interpretation of the spectrum in terms of task-related activity is problematic [147]. In particular, task-related responses to event related experiments have a spread spectral representation, and thus can be well described only in the temporal domain.

Therefore, we rather deal with an explicit model stated in the temporal domain as in the Wold decomposition. We simply identify the deterministic component $V(t)$ with task-related activity, the remainder being the stochastic component $U(t)$. This yields the following model

$$
\begin{aligned}
X(t) &= U(t) + V(t) && (4.19)\\
V(t) &= \sum_{m=0}^{M}\sum_{c=1}^{C} \beta_m \gamma_c P_c(t-m) && (4.20)\\
Z(t) &= U(t) - \sum_{l=1}^{L} \alpha_l Z(t-l) && (4.21)\\
\sigma^2 &= var(Z) && (4.22)\\
& && (4.23)
\end{aligned}
$$

This model is not exactly analogous to the Wold model, but it yields realistic estimation. Parameters $(\beta_0, .., \beta_M)$ represent the FIR *hemodynamic* model; $(\gamma_1, .., \gamma_c)$ represent the information of interest, i.e. the relative impact of each experimental condition on the process; $(\alpha_1, .., \alpha_L)$ are the autoregression parameters. Note that equations (4.19) and (4.21) are equivalent to

$$
\begin{aligned}
X(t) &= V(t) + A(t) + Z(t) && (4.24)\\
A(t) &= \sum_{l=1}^{L} \alpha_l Z(t-l) && (4.25)
\end{aligned}
$$

Maybe it is now clearer that processes $V(t)$, $A(t)$ and $Z(t)$ carry respectively information about the experimental paradigm, the past values of the process, and the deviation from prediction (innovation process). Next, we could generalize equation (4.20) by introducing a non-linearity

$$U(t) = F(\sum_{m=0}^{M}\sum_{c=1}^{C} \beta_m \gamma_c P_c(t-m)) \quad (4.26)$$

However, we do not consider this possibility here (see section 4.4.2 for some developments).

One may ask whether it would be simpler to consider $V(t)$ as a random signal, avoiding equations (4.21) or (4.25). But the distinction between these terms has several advantages:

- It allows for the description of potential nuisance signals in the dataset, opening the way towards identification.
- It unbiases the noise estimate: the true stochastic variance can be identified as $var(Z)$ instead of $var(V)$ (see the discussion about Kolmogorov formula in section 4.2.2).
- It allows for an exploratory approach if the experimental paradigm is not determined, or if one does not want to use this information (in particular if task related patterns are close to periodical or low frequency patterns).
- It makes the residual more *normal* (see section 4.4.1).

The next step is to estimate all the quantities involved in the model: (α_l), (β_m), (γ_c), but also -and this make the issue much more complex- L and M. For this reason, we will use an information theoretical framework. In fact, this stems quite naturally from a maximum likelihood approach; let us note $\theta = ((\alpha_l), (\beta_m), (\gamma_c))$ and q the number of independent parameters of the model $q = L + M + C - 1$; then the estimation of θ becomes a classical maximum likelihood problem:

$$\hat{\theta} = \arg\max_{\theta} \; p(X|\theta, P, \overleftarrow{X}, q) \tag{4.27}$$

where $\overleftarrow{X}$ is the past of X. Assuming that the model completely takes into account temporal correlations, so that $Z(t)$ is i.i.d., this right hand term factorizes

$$p(X|\theta, P, \overleftarrow{X}, q) = \prod_{t=1}^{T} p(X(t)|\theta, P, \overleftarrow{X}\,(t), q) \tag{4.28}$$

which incites us to consider the log-likelihood of the model

$$\mathcal{L}(X|\theta, P, \overleftarrow{X}, q) = \frac{1}{T} \sum_{t=1}^{T} \log\left(p(X(t)|\theta, P, \overleftarrow{X}\,(t), q)\right) \tag{4.29}$$

which we identify as the opposite of the conditional entropy

$$H(X|\theta, P, \overleftarrow{X}, q) = -\mathbb{E}[\log\left(p(X(t)|\theta, P, \overleftarrow{X}\,(t), q)\right)] \tag{4.30}$$

4.3 Prediction and information

4.3.1 Entropy rate of a stochastic process

Let us consider the stochastic process X. From the information theory point of view, it can be endowed with an entropy rate (see appendix C.1.4 and [48]). The formal definition of the entropy rate is

$$\eta(X) = \lim_{t \to \infty} \frac{1}{t} H(X(1), X(2), .., X(t)) \tag{4.31}$$

where $H(X(1), X(2), .., X(t))$ is the entropy of the joint distribution of $X(1), .., X(t)$. Let us recall that the definition of the entropy of a random variable Y of dimension d and probability density ρ is [1]

$$H(Y) = H(\rho) = -\int_{\mathbb{R}^d} \log \rho d\rho = \mathbb{E}_Y\left(-\log(\rho)\right) \tag{4.32}$$

The following property holds if the process X is stationary:

$$\eta(X) = \lim_{t \to \infty} H(X(t)|X(1), .., X(t-1)) \tag{4.33}$$

[1] In this work, we use the logarithm in basis e, by contrast with most of the information theoretic literature which has been developed in basis 2 . This choice has no impact on the theory.

This gives a more intuitive interpretation: the entropy rate can be interpreted as the additional randomness of the process at each time step, i.e. its non-predictable part. This quantity is also the limit of the increase of the extensive entropy in [20].

The entropy rate is easy to compute in at least two cases: for Markov processes and for gaussian processes. For biological data, the gaussian model will be more convenient. Let X be a gaussian stationary process. Let Γ be the covariance matrix of $X(1), .., X(t)$ (Γ is Toeplitz with $\Gamma_{ij} = \gamma(|i-j|)$, $\gamma(k)$ being the autocorrelation coefficient of order k). The joint entropy of the vector $X(1), ..X(t)$ is thus

$$H(X(1), .., X(t)) = \frac{1}{2} \log \left(det(2e\pi\Gamma) \right) \tag{4.34}$$

But for $t \to \infty$ the density of the eigenvalues of Γ tends to a limit, which is the spectrum of the stochastic process. We thus obtain the following formula for the gaussian entropy rate:

$$\eta(X) = \frac{1}{2} \log 2\pi e + \frac{1}{4\pi} \int_{-\pi}^{\pi} log(f(\lambda)) d\lambda \tag{4.35}$$

which is nothing but a restatement of Kolmogorov formula (4.13). In fact we have the relationship

$$\sigma^2 = \frac{1}{2\pi e} e^{2\eta(X)} \tag{4.36}$$

σ^2 being the *stochastic variance* used in the Wold decomposition. Stated another way, the entropy rate measures the stochasticity of the process, i.e. it is related to the variance of the best estimator given an infinite past.

For fMRI data, the process is of course of finite length, but, as the stochastic variance, the entropy rate can be computed through the estimated spectrum of the time series. As indicated in section 4.2.3, it is nevertheless preferable to use explicit prediction schemes; in any case, adding a correction term for the parameterization of the process, the entropy rate can be generalized into a complexity measure. Minimizing the latter is the goal of the *Minimum Description Length* approach.

4.3.2 Minimum Description Length

Let $\hat{x}(t)$ be the optimal predictor of $x(t)$ as in equation (4.24)

$$\hat{x}(t) = u(t) + a(t) = \sum_{l=1}^{L} \alpha_l x(t-l) + \sum_{m=0}^{M} \sum_{c=1}^{C} \beta_m \gamma_c P_c(t-m) \tag{4.37}$$

Let $q = L + M + C - 1$ be the number of parameters used in equation (4.37) and θ be the set of coefficients $((\alpha_l), (\beta_m), (\gamma_c))$.

Defining a prediction model for x is in fact equivalent to assuming a form for the joint probability of $x(1), .., x(T)$, the joint law being parameterized by θ; the complexity of the model together with the uncertainty of the data given the parameterization yield a complexity form, which is, as in Kolmogorov's initial work on complexity, stated in terms of description length; more precisely, we have the following theorem, stated and proved by J. Rissanen in [189] (see also [186], [187], [188]). Under the (fairly general) assumptions:

- θ ranges over a compact subset of $\mathbb{R}^q$ with a nonempty interior,

- the joint density depends smoothly on θ,
- the central limit theorem holds for Maximum Likelihood (ML) estimators of θ everywhere,

Letting L be any length function satisfying Kraft's inequality for all T, then for all q and $\epsilon > 0$,

$$\frac{1}{T}\mathbb{E}(L(X)) \geq H(X(1),..,X(T)) + (\frac{1}{2} - \epsilon)\frac{q}{T}\log T \tag{4.38}$$

for all θ except within a set whose volume goes to 0 as $T \to \infty$. Moreover, there exist integers n_ϵ and length functions for which the opposite inequality ($<$) holds for all negative ϵ when $n > n_\epsilon$ and for all θ of all dimension.

Kraft's inequality is simply a technical assumption about coding functions: $\sum_{\{X(T)\}} 2^{L(X)} \leq 1$, where the sum is taken over all strings of length T.

This theorem states that, no matter which universal code one uses, the mean code length is bounded below by the expression $\mathcal{C}(X, \theta, q)$, for any particular processes defined by θ. This can thus be taken as the intrinsic complexity of the process:

$$\mathcal{C}(X, \theta, q) = H(X) + \frac{q}{2T}\log(T) \tag{4.39}$$

The computation of the optimal predictor for the process X becomes then the solution of the problem $\min_{\theta,q} \mathcal{C}(X, \theta, q)$. The final predictor z can be viewed as the *Minimum Description Length* (MDL) predictor of the process $X(t)$ within the family of the predictors defined by equation (4.37). Note that the definition of length that we use here keeps only the first two terms of a more complicated expression, which is valid only for T big enough.

An analogous criterion for model selection, known as Bayesian Information Criterion (BIC) has been derived, using a non absolutely continuous prior in the parameter space [195]:

$$BIC(x, \theta, q) = -\mathcal{L}(X|\theta, q) + \frac{q}{2T}\log(T) \tag{4.40}$$

where $\mathcal{L}(x|\theta, q)$ stands for the log-likelihood of X; this is equivalent to (4.39) since $\mathbb{E}(\mathcal{L}(x|\theta, q)) = -H(x)$

This criterion is maybe less famous than Akaike's Information Criterion (AIC) [3]:

$$AIC(x, \theta, q) = -\mathcal{L}(X|\theta, q) + \frac{q}{T} \tag{4.41}$$

which is the asymptotic, expected value of the log-likelihood. But AIC does not take into account the uncertainty of the parameter estimation when estimating the uncertainty about predictions; BIC-MDL is in fact more conservative, especially for long time series. The BIC-MDL criterion is an asymptotic approximation of the complexity of the generative process. In [20], the penalty term $\frac{q}{2}\log(T)$ has been identified with the *subextensive entropy* of the time series $(x(1),..,x(T))$, which is a universal measure of the complexity of the underlying process; we develop this point of view in Appendix D.

4.3.3 Making it work in practice

The problem of process identification is now well-posed, we can now introduce further simplifications to obtain a quick algorithm. First, let us notice that the estimation problem (4.37) divides into two steps:

- Estimation of the task-related response defined in equation (4.20). This involves M, (γ_c), (β_m). This step is more difficult since the model is bilinear in the variables $(\gamma_c) \times (\beta_m)$.
- Estimation of the autoregression parameters $L, (\alpha_l)$, and derivation of the final variance $var(Z)$, following equation(4.20).

We propose that each step can be made efficiently through using orthogonal regressors. In the first step, we consider a space of regressors $P_c(t-m), c = 1..C, m = 1..M$ where M is sufficient to represent adequately the impulse response function. In order to obtain the orthonormal basis $\Omega = (\omega_{m,c}(t)), c = 1..C, m = 1..M$, we perform a PCA of the set of regressors $P_c(t-m), c = 1..C, m = 1..M$. This family is *overcomplete*, since it is of dimension MC instead of $M + C$; but the complexity criterion prevents from overfitting, and only a slight correction is made post hoc (see equation (4.49)) to constrain the factorized solution. Then the minimization of the complexity criterion

$$\mathcal{C}(X, M, (\gamma_c), (\beta_m)) = \frac{1}{2} \log \left(var(x - \sum_{m=0}^{M} \sum_{c=1}^{C} \beta_m \gamma_c P_c(t-m)) + \frac{C+M}{2} \frac{\log T}{T} \right. \tag{4.42}$$

boils down to the sum

$$\begin{aligned} \mathcal{C}(X, M, (\gamma_c), (\beta_m)) = \mathcal{C}(X, I, (\delta)) &= \frac{1}{2} \log \left(var(x - \sum_{i=1}^{I} \delta_i \omega_i) \right) + \frac{I}{2} \frac{\log T}{T} && (4.43) \\ &= \frac{1}{2} \log \left(var(x) - \sum_{i=1}^{I} \delta_i^2 \right) + \frac{I}{2} \frac{\log T}{T} && (4.44) \end{aligned}$$

where δ_i is the fit value associated with ω_i ($\hat{\delta} = X.\Omega^T$). $\hat{\delta}$ is thus obtained readily, and ordered $\delta_1 > .. > \delta_{MC}$. Then the criterion (4.44) depends only on the number I of selected regressors. Its minimization represents the tradeoff between goodness of fit -minimization of the variance of the residual $x - \sum_{i=1}^{I} \delta_i \omega_i$- and simplicity of the structure measured by $\frac{I}{2} \frac{\log T}{T}$. One can also notice that criterion (4.44) is convex with respect to I if the following property holds for $I \geq 1$:

$$(var(x) - \sum_{i=1}^{I-1} \delta_i^2)(\delta_I^2 - \delta_{I+1}^2) > \delta_I^4. \tag{4.45}$$

If this is true, the minimization can be performed by increasing I until

$$\log \left(1 - \frac{\delta_{I+1}}{var(x) - \sum_{i=1}^{I} \delta_i^2} \right) > -\frac{\log(T)}{T}, \tag{4.46}$$

which can usually be approximated by

$$\frac{\delta_{I+1}}{var(x) - \sum_{i=1}^{I} \delta_i^2} < \frac{\log(T)}{T}. \tag{4.47}$$

In more general cases, the criterion has to be computed for all values of I.

Once I is obtained, one comes back to the original basis

$$\sum_{i=1}^{I} \delta_i \omega_i = \sum_{m=0}^{M} \sum_{c=1}^{C} \zeta_{m,c} P_c(t-m) \tag{4.48}$$

The solution in terms of $(\gamma_c), (\beta_m)$ is then the best rank one approximation of the coefficients $(\zeta_{m,c})$:

$$((\hat{\gamma}_c), (\hat{\beta}_m)) = max_{(\beta),(\gamma)} (\zeta_{m,c} - \beta_m \gamma_c)^2 \tag{4.49}$$

which is simply provided by the SVD of $(\zeta_{m,c})$ once written as a $M \times C$ matrix (from our experiment, the approximation is always very accurate). All these tricks convert a non-convex non linear problem into a convex linear one. Moreover, if many time courses have to be analyzed, the regressors (ω_i) are computed once and for all.

Step 2 is solved in the same way, though the setting is more simple:

$$\min_{L,(\alpha)} \frac{1}{2} \log \left(var(z - \sum_{l=1}^{L} \alpha_l z(t-l)) \right) + \frac{L}{2} \frac{\log T}{T} \tag{4.50}$$

where $z(t) = x(t) - \sum_{m=0}^{M} \sum_{c=1}^{C} \beta_m \gamma_c P_c(t-m)$ the regressors $x(t-m)$ are orthogonalized using the usual Gram-Schmidt procedure. This is because the order of the regressors is important: $x(t-1)$ conveys a priori more information on $x(t)$, so that it is simply normalized; then $x(t-2)$ is orthogonalized with respect to $x(t-1)$ and normalized, $x(t-3)$ is orthogonalized with respect to $x(t-1)$ and $x(t-2)$ and normalized and so on. The minimization of the criterion (4.50) involves z instead of the input data x; this means that the fraction of variance that could be attributed to paradigm-related activity has been subtracted away. In turn $var(z - \sum_{l=1}^{L} \alpha_l z(t-l))$ is an improved estimation of the true stochastic variance of the data.

Then step 1 and step 2 are iterated until convergence, x being replaced with $x - \sum_{l=1}^{L} \alpha_l z(t-l)$ in step 1 to have an improved stochastic variance estimation. Convergence is not proved, but is quick in practice (there are only two discrete parameters left, I and L). One example of the resulting procedure is given in figure 4.3.

The philosophy of the minimum complexity approach is illustrated and discussed in figure 4.4, where it is shown on a simulation that the accuracy of the hemodynamic repose model, defined by the number of regressors that survive the minimum complexity test (equation(4.44)), is related to the signal to noise ratio (SNR): stated differently, this means that higher SNR allow for more sophisticated, hence more accurate hemodynamic response estimation.

We discuss some limits of this model and generalize it when multiple realizations of the data are given in input in section 9.1.

4.4 Some estimation issues

A very important problem is the set of admissible hypotheses that can be made for the mathematical treatment of the functional MR signal. This question has two counterparts:

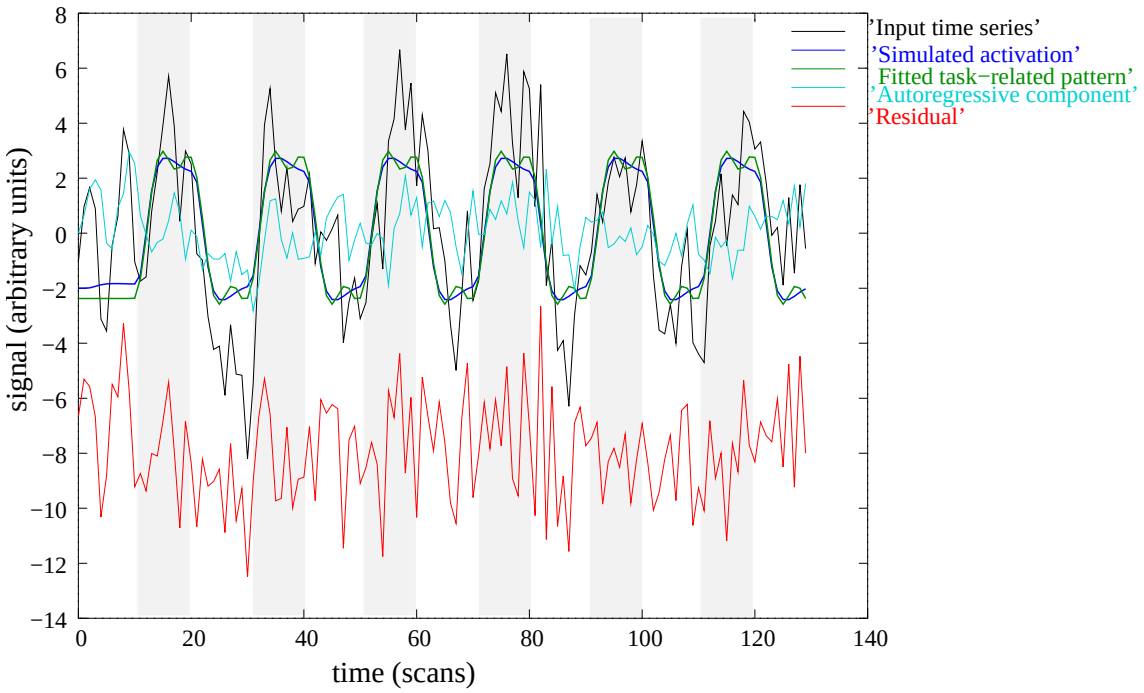


Figure 4.3: Illustration of the temporal model.
Given the fictive experimental paradigm (represented in grey, block design) a time course of length $T = 130$ has been generated by adding a task-related pattern (blue time curve), a Brownian signal and white noise, so that the signal to noise ratio (SNR) is approximately 0.5. This signal is detrended, providing the input signal for the algorithm (in black). The algorithm provides two components: the estimated task-related component (in green), that is related to experimental design, and the autoregressive one (in cyan). The residual of the model (in red) has been lowered by 8 units in order to facilitate visualization. Note that the task-related component is correctly approximated by the model, in spite of the noise level.

- These hypotheses should be relevant given the knowledge that we have from the data.

- They should yield computationally tractable models given the nature of the data (noise, length of the available series).

With these two points of view in mind, we address some classical questions about temporal modeling of fMRI data: the normality of the noise, and the linearity of the response with respect to the stimulation. Last, we discuss the use of more technically advanced techniques issued from the theory of dynamical systems.

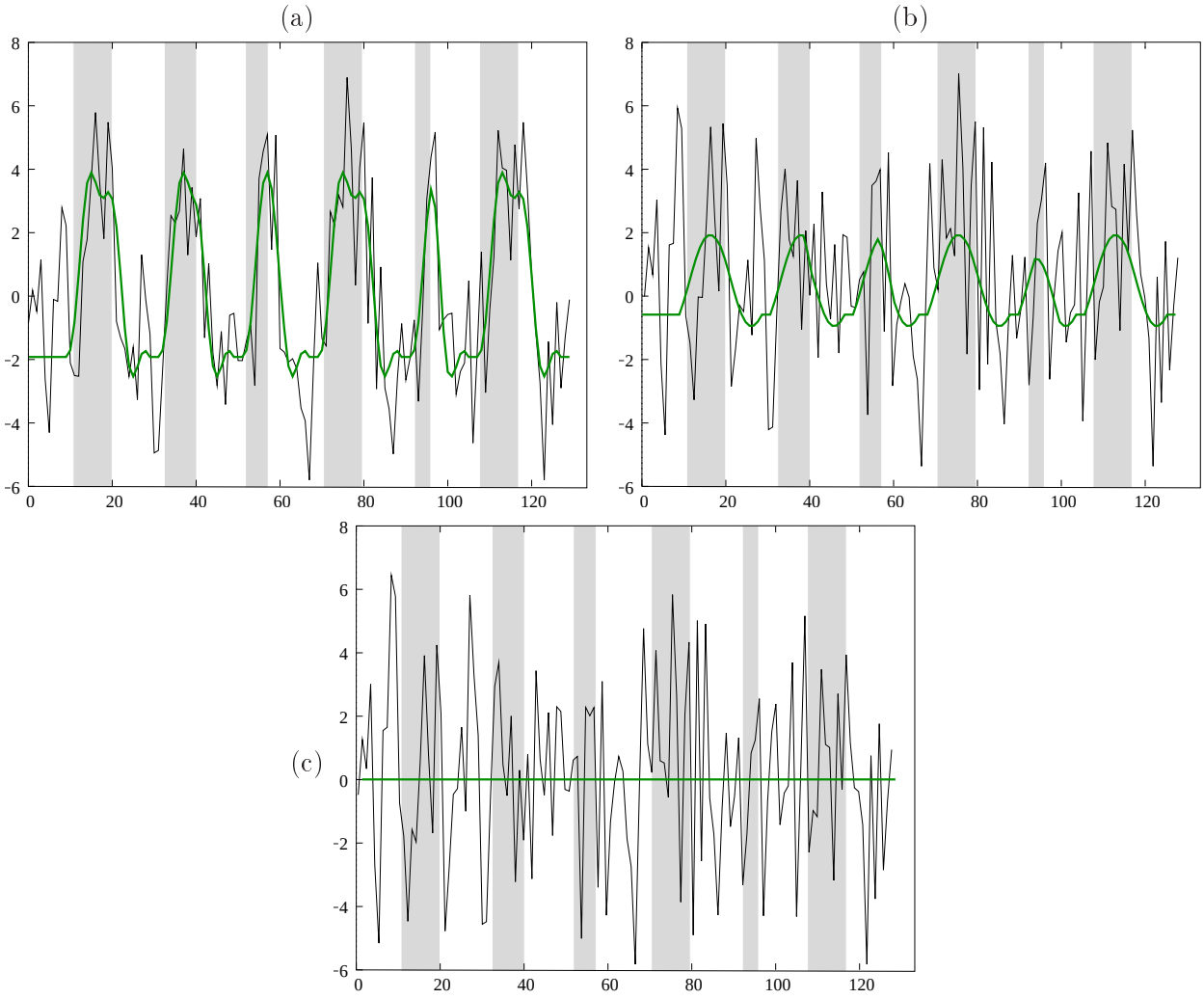


Figure 4.4: Minimum Complexity and Signal to Noise Ratio (SNR)
Given a fictive experimental paradigm (gray strips), we simulate task-related activity with different Signal to Noise Ratios: 1 (a), 0.33 (b) and 0.1 (c), and analyze the resulting time courses (in black) with the method presented in 4.3.3. The estimated task-related signals are represented in green. As expected, the amplitude of the resulting pattern is related to the SNR, but the accuracy of the model is also highly dependent on the SNR: The model in (a) has been estimated with 3 coefficients, the model in (b) with 2 coefficients, the model in (c) with 0 coefficient. The interpretation is that in the minimum complexity perspective, higher SNR allow for more sophisticated, hence more accurate models.

4.4.1 The gaussian noise hypothesis

Is it reasonable?

Actually, it is extremely important to define what is assumed normal or not: for example, the residual from a regression analysis applied to raw fMRI data is, in general, far from

gaussian. However, if a more complete data decomposition is applied, as in the Wold model (4.12), there will likely be little deviation from normality in the residual. But, does the difference between these two cases lie in the temporal structure (autocorrelation), or in the statistical distribution ? Using the vocabulary of information theory, we notice that *both deviation from normality and whiteness can be interpreted as the presence of information in the signal.* Indeed:

- From the statistical perspective, the gaussian distribution is the one that has maximal entropy (i.e. dispersion) for a given variance [48], which we interpret the following way: among all signals of identical energy, the gaussian is the one that carries minimal information. The notion of information becomes clear if we compare the gaussian distribution with a bimodal distribution, that can code e.g. for the state of a binary system.

- From a dynamical perspective, the entropy rate, i.e. the intrinsic randomness of a process, decreases when the autocorrelation increases (for an AR(1) gaussian process of coefficient ρ, the entropy rate is equal to $-\frac{1}{2}\log\sqrt{1-\rho^2}+cste$).

Moreover, these two aspects are not independent in practice; for example, the density of a sinusoidal signal is $d(u)=\frac{2}{\pi}\frac{1}{1-u^2}$, thus certainly not gaussian.

We restate thus the question in the following manner: does deviation from normality in fMRI data have a meaning *per se*, or is it rather related to the temporal structure ? We plead for the second explanation, due to the known predominance of low frequency -probably aliased- signals in fMRI data. On the other hand, there is to our knowledge no particular source of non-gaussian randomness in fMRI.

In practice, this means that temporal models that account well for trends and data correlation can be assumed to have gaussian residuals. This clearly eases the estimation procedures.

How can one manage data analysis without this hypothesis?

Assuming that the gaussian hypothesis should be rejected, we would have to face the problem of estimating the signal distribution from empirical data. It is then advisable to use non parametric methods, such as kernel estimation as in [63]: for example, let us assume that we would like to estimate the statistical density $\mathcal{D}$ of a random variable X given T samples $X(1),..,X(T)$: we can then use the estimation

$$\mathcal{D}_\sigma(u)=\frac{1}{T}\sum_{t=1}^{T}K_\sigma(u-X(t)) \tag{4.51}$$

where K is a kernel, e.g. a gaussian kernel, of width parameter σ. The entropy of the distribution is then easily derived

$$H(X)=H(\mathcal{D})=\int log(\mathcal{D}(u))\mathcal{D}(u)du \tag{4.52}$$

Thus there remains only to derive the parameter σ. We have the following theorem, stated and demonstrated in [26] for the general case of a multivariate process X of dimension d:

If X is stationary and if $\mathcal{D}$ decays rapidly at ∞, then, setting $\sigma = c\left(\frac{\log(T)}{T}\right)^{\frac{1}{d+4}}$ $(c > 0)$,

$$\frac{1}{\log T}\left(\frac{T}{\log T}\right)^{\frac{2}{d+4}} sup_{x\in\mathbb{R}^d}|\mathcal{D}_\sigma(x) - \mathcal{D}(x)|_{T\to\infty} \to 0 \ \ a.s. \tag{4.53}$$

In practice, for $d = 1$, setting $\sigma = \sqrt{var(X)}T^{-\frac{1}{5}}$ gives a good approximation of the density. But from our experiments, this is not necessary for fMRI data.

4.4.2 The linearity hypothesis

Non-linearity in the BOLD response

We have insisted a lot on this question in the beginning of this chapter, in sections 4.1.1 and 4.1.2. We simply recall that for an inter-stimulus interval of more than 4 seconds, the deviation from linearity is expected to be small.

What if we abandon the linearity hypothesis?

Assuming that temporal resolution of fMRI experiments improves, this question may become recurrent in future analysis. Different methods are then possible.

- Find a parametric model of the non-linear activation, and estimate the parameters given the data [76]. This approach has the advantage of being well-founded, yielding then interpretable results, but relies on a possibly wrong or incomplete model, and is often hardly tractable.
- Find a general model for non-linearities that allows for quick estimations, e.g. Volterra series expansions [83]. This method is more efficient but less interpretable than the first one, except if they can be linked together [85].
- Find a more general -non parametric- estimator of the non-linearity. This more exploratory approach can be made efficient with adapted techniques. It has probably little power for inference, but it may be interesting for exploratory analyses. A typical way to do it is to introduce a non-linear function between the predictor variable $\hat{x}$ and the output data x

  $$x(t) = \phi(\hat{x}(t)) + \epsilon(t) \tag{4.54}$$

  ϕ can be estimated by a kernel procedure

  $$\phi(u) = \mathbb{E}(X|\hat{X} = u) = \frac{\sum_{t=1}^{T} x(t)K_\sigma(u - \hat{x}(t))}{\sum_{t=1}^{T} K_\sigma(u - \hat{x}(t))} \tag{4.55}$$

4.4.3 Non-parametric dynamical system

It is tempting to consider fMRI time series as being generated by a complex, unknown and non-linear dynamical system. The question is then to characterize this system.

Dynamical systems known through their outputs can be characterized in several ways:

- The Lyapunov exponents that describe the time series [180].

- The correlation integral defined in [99], and used in [50].
- Invariant measures or all other quantities that describe physical systems [55].

However, we have to stress that these methods are of no use here: for fMRI data, one should keep in mind three basic characteristics: *i)* the time series are short, *ii)* the data is noisy, and the noise level is comparable in magnitude with respect to the signal of interest, *iii)* the data is quite redundant (multiplicity of the time series). Both *i)* and *ii)* readily imply that the use of nonlinear dynamical systems characterization is hopeless; in other words, stochastic effects are stronger than dynamic ones. The only hope is that the redundancy of the information can help for its characterization, but this is matter for the next chapter.

Conclusion on temporal modeling

Finally, our proposition for the temporal modeling of fMRI data results in a combination of FIR fitting methods with Wold decomposition and Minimum Description Length. This stands as a compromise between flexibility -motivated by the a priori unknown structure of the hemodynamic response- risk of over-fitting the data and interpretability of the temporal components. One of the interesting results of this approach is the definition of the *stochastic variance* of the time series, which quantifies its randomness, even if the generative process is not known. Moreover, via the entropy rate, this quantity is related to the complexity of the process, a quantity that emerge from different approaches ([189], section D), and seems to yield a universal characterization of temporal processes.
Note that the modeling proposed here is not concerned with the elimination of *false positives*; it rather tries to give an explicit form to the structured components that emerge from the study of the time series. The problem of false positives control, as well as the extension to multi-session data, will be studied in more detail in section 9.1.

Chapter 5

Dealing with multivariate data

In this chapter, we discuss the second question which is essential in fMRI analysis, i.e. the way of taking into account the redundant information contained in the sequence of images. In other words, considering the image sequence as a statistical process, we should consider this process as multivariate ($N-$variate if we consider N voxels of the dataset). In fact, since current multivariate methods do not really consider temporal modeling, we will temporarily consider the dataset as N dimensional random variable, thus losing temporal information.

This basic approach does not consider the spatial structure of the data. This is of course a weakness, but we have preferred not to address this question, since the spatial organization of the data is related to the anatomy, which requires specific approaches.

Leaving spatial structure apart, we simply have statistical tools to model the data. We have briefly described them in chapter 2, we will develop here some more technical questions; however, we will insist on the theoretical basis of these methods by relating them to information theory, in three instances: *i)* in the Canonical Correlation Analysis (CCA) framework, which makes a gaussian signal hypothesis, *ii)* in the ICA framework (we will discuss some of the current algorithms), which takes exactly the opposite point of view of non-normally distributed data and *iii)* in compression theory, since we propose an information bottleneck algorithm for data clustering.

5.1 The second order approach: SVD

We start this chapter with a few words on the Singular Values Decomposition (SVD), since it can be seen as an elementary independent component analysis method. We recall its main features, then some difficulties that arise in practice, and possible solutions. Last we make the connection with information theory concepts by considering Canonical Correlation Analysis.

5.1.1 SVD : the simplest data decomposition technique

Interpreting a SVD of fMRI data

Let X be a dataset, i.e. a $N \times T$ matrix (N = number of voxels considered, T = length of the time series). The SVD of X is

$$X = U\Sigma V^T; \tag{5.1}$$

where U and V are $N \times N$ and $T \times T$ orthogonal matrices, and Σ a $N \times T$ matrix with non-zero elements only on its diagonal. The interpretation of (5.1) is that the columns of U and V are orthogonal spatial and temporal components respectively. The diagonal of Σ, i.e. the set of singular values ($\sigma_k, k = 1..T$), gives the amount of data described by the associated spatio-temporal components (i.e. the columns u_k and v_k of U and V associated with the singular values): in other words, the SVD basically solves the following optimization problem:

$$\sigma_k = max_{(u,v)\in\mathbb{R}^N\times\mathbb{R}^T} \frac{u^T.X.v}{\sqrt{u^T.u}\sqrt{v^T.v}} \tag{5.2}$$

under the constraint $u \perp span(u_1, .., u_{k-1})$ and $v \perp span(v_1, .., v_{k-1})$.

Practical advantages

The method has the advantage of its efficiency, since it can be achieved in $O(NT^2)$ operations, which makes it slightly heavier than the General Linear Model, for instance. Let us add that the method is non-iterative, so that there is no particular difficulty in applying it. Additionally, the outcome of the method is not very sensitive with respect to small variations in the data (e.g. eliminating some voxels from the analysis), at least if the singular values $\sigma_1, .., \sigma_k$ are sufficiently distant.

Natural "information bottleneck"

The consequence of equation (5.2) is that the SVD provides an optimized representation of the data in the sense that the first K components are those -among all possible rank K representations- that retain most of the data variance.

5.1.2 Practical difficulties with SVD of fMRI data

Orthogonality constraint

Clearly, in equation (5.2), all spatio temporal components are constrained to be mutually orthogonal (or decorrelated, in the statistical language). In fact, this provides an increasing constraint space when the rank of the component increases.
For fMRI data, the space of interest (task-related activity) is thus not necessarily well described through SVD: if the first component is explicitly the *main effect* present in the data, no such clear interpretation is possible for the second or third task-related component. In fact, the SVD does not unmix the data into mutually independent components, so that the separation of the different effects of the dataset is generally poor.

Order selection

Among the components of the decomposition, some of them are of interest and the others will be considered as noise. This induces two problems: the problem of subspace of interest selection (which can be simply solved by selecting the temporal patterns that well fit the data matrix) and the problem of order selection. If we denote the order of the final model by K, we need some additional way to do it.

A frequent way to deal with this is to use a rank test as Wilk's lambda [79], or Bartlett or, more generally, sphericity test [224] [126]. The underlying hypothesis is that the noise space or null space is *isotropic*, i.e. its singular values are approximately equal, whereas the signal space deviates from isotropy. But this heuristics is satisfactory only after projection of the data onto a space of interest (see the MLM method in 5.1.3.).

Another approach is the introduction of a generalization error criterion that includes a data fit and a rank penalty term, based on asymptotic approximations [103]. The authors moreover report that the analytical rank estimate obtained from the training dataset is too optimistic, and that data splittings are better suited for rank estimation than the analytical method based on only one training dataset.

The underlying hypothesis: gaussianity

Let us insist on the fact that SVD gives in general a bad description of the signal space. If the assumption that the noise space is gaussian is probably correct, there is no reason why a signal space should be gaussian. In fact, the presence of activations in the data is actually characterized by the *deviation from gaussianity* in the general linear model. But, precisely, the SVD description of that data is sufficient under the gaussian hypothesis (the SVD relies only on the covariance structure of the dataset, which is in turn equivalent to a gaussian hypothesis). For instance, the simple-minded synthetic example given in appendix A.1 violates this hypothesis.

Lack of priors

The basic SVD setting is simply based on the covariance of the data, and by no way enhances temporal effects of interest in the data. In other words, the SVD does not incorporate any temporal priors for data modeling. This issue has been solved by means of data projection in the CVA, PLS and MLM methods (see section 5.1.3).

Some pitfalls in SVD interpretation

Performing a SVD of the data has sometimes been interpreted as a study of *functional connectivity* of the subject [80]. By this, the authors simply meant the correlation structure of the data, which is neither the *anatomical connectivity* nor the *effective connectivity*, which has been elaborated in the analysis of multiunit recordings of separable neural spike trains. In fact, nothing supports an interpretation in terms of connectivity (at most could one speak of *correlation* between fMRI time courses), and the spatial maps evidenced by multivariate analysis are a complex compound of effects present in the data. The interpretation of such maps is simply indicative of the co-occurrence of some effect during the experiment. Moreover, spontaneous correlations seem to be related to low frequency

fluctuations in fMRI data [22]; but this observation blurs rather than it confirms the concept of functional connectivity as applied to fMRI or PET data (see [115] a discussion on functional connectivity).

A more subtle abuse is to interpret the spatio-temporal components as *different* modes of hemodynamic activity. In reality, nothing more can be said than that *the space spanned by the first components contains most of the signal variance*. This is a strong difference with ICA, where each component can be interpreted on its own. Statistically speaking, one can observe that decorrelated components can be treated as independent only if the data is gaussian, which is probably not the case if the dataset contains activation patterns.

5.1.3 Overcoming the lack of prior: the MLM

Some efforts have been made to overcome the lack of temporal information inherent to the SVD method. In particular, let us mention the Multivariate Linear Model (MLM) [126] (see also [224]), which additionally uses an over-specified design matrix G of the experiment. The MLM performs in fact the SVD of $X_w G_w^T$, with X_w and G_w being respectively the whitened versions of X and G. This procedure is in fact equivalent to performing the SVD of the whitened data X_w projected onto the rows of the design matrix, i.e. $X_w G^T (GG^T)^{-1} G$. The advantage is that the resulting spatio-temporal components can be interpreted in terms of linear combination of the rows of G, so that the SVD is used for selecting a sub-model of the design matrix which is tailored to the data (this is why it is advantageous to choose initially an *over-specified* design matrix).

This method optimally uses the compression properties of the SVD, but does not address some weaknesses intrinsic to the method (decorrelation constraint, order selection, assumed normality).

5.1.4 Mutual Information and Canonical Correlation

In order to interpret covariance-based methods from an information theoretic perspective, it is necessary to assume that the observed structure is gaussian. Indeed, if we assume that two multivariate variables (e.g. two datasets) X_1 and X_2 of size $n_1 \times m$ and $n_2 \times m$ are gaussian and centered, their mutual information is then

$$I(X_1, X_2) = -\frac{1}{2} \log \frac{\det \begin{pmatrix} X_1 X_1^T & X_1 X_2^T \\ X_2 X_1^T & X_2 X_2^T \end{pmatrix}}{\det (X_1 X_1^T) \det (X_2 X_2^T)} \tag{5.3}$$

i.e. the logarithm of the determinant of the covariance matrix of the joint variable $[X_1, X_2]$ divided by the products of the determinants of the covariance of X_1 and X_2. Letting Δ be the block diagonal matrix of the marginal covariances

$$\Delta = \begin{pmatrix} X_1 X_1^T & 0 \\ 0 & X_2 X_2^T \end{pmatrix}, \tag{5.4}$$

and assuming (without loss of generality, since X_1 and X_2 can be preprocessed) that Δ is full rank, one has

$$I(X_1, X_2) = -\frac{1}{2} \log \det \Delta^{-\frac{1}{2}} \begin{pmatrix} X_1 X_1^T & X_1 X_2^T \\ X_2 X_1^T & X_2 X_2^T \end{pmatrix} \Delta^{-\frac{1}{2}} \quad (5.5)$$

$$= -\frac{1}{2} \log \det \begin{pmatrix} I_m & (X_1 X_1^T)^{-\frac{1}{2}} X_1 X_2^T (X_2 X_2^T)^{-\frac{1}{2}} \\ (X_2 X_2^T)^{-\frac{1}{2}} X_2 X_1^T (X_1 X_1^T)^{-\frac{1}{2}} & I_m \end{pmatrix} \quad (5.6)$$

where I_m is the identity matrix of size m. The interesting thing is that $C_{1,2} = (X_1 X_1^T)^{-\frac{1}{2}} X_1 X_2^T (X_2 X_2^T)^{-\frac{1}{2}}$ is nothing but the correlation matrix of X_1 and X_2, i.e. the product of the matrices after whitening. Then noting the singular values of this matrix $\sigma_1, .., \sigma_m$, we have the following

$$I(X_1, X_2) = -\frac{1}{2} \log \det \begin{pmatrix} I_m & C_{1,2} \\ C_{1,2}^T & I_m \end{pmatrix} = -\frac{1}{2} \sum_{i=1}^{m} \log(1 - \sigma_i^2) \quad (5.7)$$

Indeed, it results from the definition of $C_{1,2}$ that $\forall i \in 1, .., m, 0 \leq \sigma_t \leq 1$. This observation can be used for data analysis. Let (U, Σ, V) be the singular value decomposition of $C_{1,2}$. Then the singular values σ_i are the correlation between the vectors $U_i X_1$ and $V_i X_2$, and $I_i = -\frac{1}{2} \log(1 - \sigma_i^2)$ is the associated mutual information. This method, known as Canonical Correlation Analysis (CCA), aims at recovering the components that are common to X_1 and X_2. In fMRI data analysis, it can be used in at least 3 instances:

- To compare two different datasets (e.g. different sessions) with similar time length and/or spatial support (this has not been done to our knowledge).
- To obtain maximally autocorrelated (at lag one) components of a dataset by letting $X_1 = X(1, .., T-1)$ and $X_2 = X(2, .., T)$ [72].
- To derive maximally autocorrelated spatial maps by letting $X_1 = X$ and X_2 a spatially smoothed version of X [72].

For any CCA method, the mutual information can be interpreted as the volume of the intersection of the space spanned by the columns of X_1 and X_2 after some normalization.

Interestingly, a generalization to more than 2 datasets has been proposed in [9]. To our knowledge, this has not be used in fMRI data analysis. This stems from the following generalization of (5.3). Let $X_1, .., X_K$ be K gaussian datasets

$$I(X_1, .., X_K) = -\frac{1}{2} \log \frac{\det \begin{pmatrix} X_1 X_1^T & .. & X_1 X_K^T \\ \vdots & & \vdots \\ X_K X_1^T & .. & X_K X_K^T \end{pmatrix}}{\det(X_1 X_1^T) .. \det(X_K X_K^T)} \quad (5.8)$$

The derivation on canonical components is performed through the generalized eigenvalue problem

$$Cu = \lambda \Delta u \quad (5.9)$$

where C is the empirical covariance (or Gram) matrix

$$C = \begin{pmatrix} X_1 X_1^T & .. & X_1 X_K^T \\ \vdots & & \vdots \\ X_K X_1^T & .. & X_K X_K^T \end{pmatrix} \tag{5.10}$$

and Δ the block diagonal matrix

$$\Delta = \begin{pmatrix} X_1 X_1^T & 0 & 0 \\ 0 & X_k X_k^T & 0 \\ 0 & 0 & X_K X_K^T \end{pmatrix} \tag{5.11}$$

A possible application of this technique is the derivation of some patterns in the data that are present across different scales by using a dataset $X_1 = X$ and increasingly smooth versions of the latter for $X_2, .., X_K$ (see next section). A second application is the derivation of repeatable patterns across many sessions of an experiment for a given subject (the experimental paradigm does not need to be the replicated).
Last, as PCA, CCA suffers from the underlying gaussian hypothesis. For non-gaussian data, it may yield non obvious results.

5.1.5 Illustration on a synthetic example

We have used the spatial CCA method on the synthetic dataset described in Appendix, section A.1. The method finds the most correlated components between the original dataset and the same dataset smoothed in 3D. In fact, this yields the details which are present throughout the image sequence, at the scale specified by the smoothing filter. We have made experiments with different filter width ranging from 0.5 to 10 voxels, or by combining different scales (equation (5.8)). An example is displayed in figure 5.1. The first two components are invariably the same, forming a basis of the two dimensional subspace of activated patterns. However, the method does not explicitly predict the dimension of the final space of interest, which has to be chosen by *ad hoc* means. Last, it embeds the three modes activation pattern into a two dimensional space. This is correct, but suboptimal from the perspective of activation space description.

The limitation of that method is obviously the choice of a particular spatial kernel, which amounts to assuming a priori a certain structure for the components of interest.

5.2 Information and non-normality: Independent Components Analysis

In this section, we study Independent Components Analysis (ICA), which is a linear decomposition method that is connected to information theory. It is more sophisticated than Principal Components Analysis (i.e. SVD), and corresponds more to the intrinsic concept of a generative model of the data. As mentioned in chapter 2, it is quite successful for fMRI data analysis. We start by stating the main features of this model; then we mention some theoretical difficulties that arise with the use of ICA on fMRI data, and then some more practical concerns.

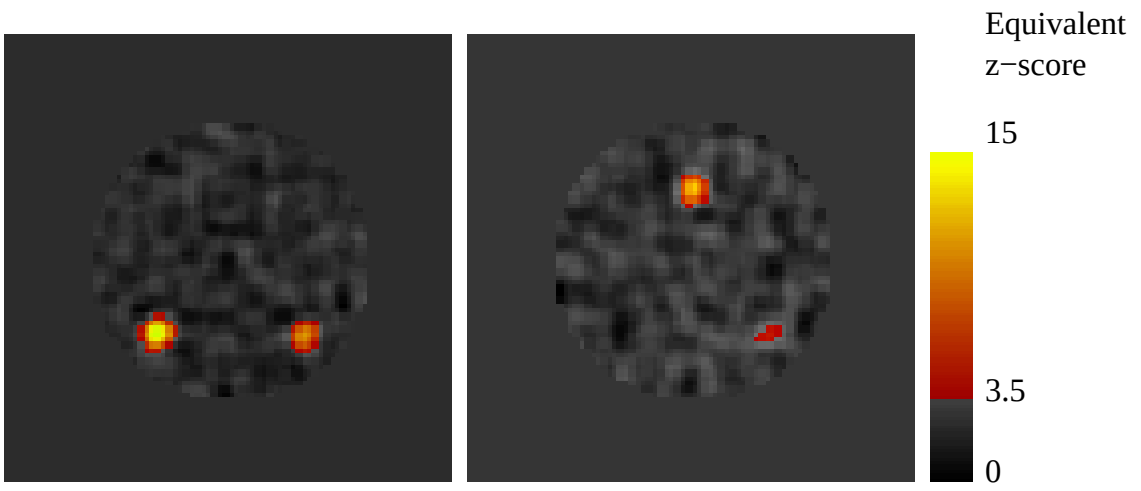


Figure 5.1: Two main resulting images of the spatial CCA method applied to the synthetic dataset.
These images, converted into z-scores and thresholded as indicated in appendix B, are a basis of the spatio-temporal activation patterns present within the data. The 3 modes structure of the activation appears through the different intensity of the 3 clusters within the two images, but the decomposition itself is not able to disentangle the 3 modes. This can be viewed as a fundamental limitation of the covariance-based methods, which are optimal only for gaussian data.

5.2.1 ICA foundations

We derive here the equations for spatial ICA of fMRI data. The basic setting is the following: the dataset X is a set of images $X(t)$ which are viewed as the *superposition* of independent images, which are called *sources* and noted S in the ICA language:

$$X = MS + E \tag{5.12}$$

The superposition is modeled by a mixing matrix M, and an additional noise term E. In the most general setting, X, M, S and E are $T \times N$, $T \times K$, $K \times N$ and $T \times N$ matrices of rank T, $K < T$, K and $T - K$ respectively. K is the number of independent sources that have generated the observed data (it is an unknown parameter), and N is the residual noise. The solution of the problem consists in estimating the matrices W and S so that

$$S = W(X - E) \tag{5.13}$$

The $K \times T$ matrix W is called the *unmixing* matrix and can be viewed as a generalized inverse of the mixing matrix M. Especially, the resulting matrix S can be interpreted as a set of *activation maps* that correspond to different independent components present within the dataset.

Let us focus on the criteria that may be used to enforce the derivation of statistically independent components. We first deal with the case of noise free mixing.

Noise free mixing We assume here that $K = T$; in practice this means that the data has been previously reduced to a $N \times K$ matrix, for example by PCA data reduction. Consequently, the noise matrix E in (5.12) vanishes; moreover $W = M^{-1}$. The following deviation is inspired by the maximum likelihood approach of [39]: the likelihood of the data, $P(X)$, derives from (5.12) and the -unknown- probability density q of the source S by the following model:

$$P(X|q, M) = |\det(M^{-1})| q(S) \tag{5.14}$$

taking the log and denoting the log-likelihood of the data $\mathcal{L}(X)$, one obtains

$$\mathcal{L}(X|q, M) = \log|\det M^{-1}| + \log(q(M^{-1}X)) \tag{5.15}$$

which has to be optimized with respect to M. Let us note that q is an unknown K-dimensional distribution with independent entries. Now, a key hypothesis is that the voxels are independent realizations of the random variable X. Then one can normalize equation (5.15):

$$\frac{1}{N}\mathcal{L}(X|q, M) = \log(|\det M^{-1}|) + \frac{\sum_{n=1}^{N} \log(q(M^{-1}X(n)))}{N} \tag{5.16}$$

Due to the law of large numbers -and given the hypothesis that the N realizations $X(1), .., X(N)$ of the process are independent- the above expression converges when $N \to \infty$ towards its expectancy, i.e.

$$\begin{aligned} \lim_{N\to\infty} \frac{1}{N}\mathcal{L}(X(1..N)|q, M) &= \log(|\det M^{-1}|) - \mathbb{E}(\log(q(M^{-1}X))) & (5.17)\\ &= \log(|\det M^{-1}|) - K[\mathcal{D}(M^{-1}X)|q] & (5.18) \end{aligned}$$

where $K[d|q]$ stands for the Kullback-Leibler divergence between the densities d and q (see appendix C.1.2), and $\mathcal{D}(M^{-1}X)$ for the probability density of the vector $M^{-1}X$. No prior is available for q. However, letting $\tilde{D}$ be the density that has the same marginals as $\mathcal{D}(M^{-1}X)$, but with an additional independence property (i.e. $\tilde{D} = \prod_{k=1}^{K} \mathcal{D}(M^{-1}X)_k$), one has the classical property of the Kullback-Leibler divergence, known as the Pythagorean theorem [48, chapter 12]:

$$K[\mathcal{D}(M^{-1}X)|q] = K[\mathcal{D}(M^{-1}X)|\tilde{D}] + K[\tilde{D}|q] \tag{5.19}$$

$K[\tilde{D}|q] > 0$ is not accessible to us, so that there remains only to minimize $K[\mathcal{D}(M^{-1}X)|\tilde{D}]$, which depends only on M and X. The quantity $K[\mathcal{D}(M^{-1}X)|\tilde{D}]$ is also known as the mutual information between the reconstructed sources $S_k = (M^{-1}X)_k, k = 1..K$, which we denote by $\mathcal{I}(M^{-1}X)$. $\mathcal{I}(M^{-1}X)$ emerges thus as the natural measure for the statistical independence of the components of $M^{-1}X$. Therefore, the computation of the independent components consists in minimizing the functional

$$F_{ICA}(M) = \mathcal{I}(M^{-1}X) + \log(|\det M^{-1}|) \tag{5.20}$$

with respect to the mixing matrix M, which can now be addressed.

The more frequent way to deal with the minimization of (5.20), is to additionally impose that the reconstructed sources $(M^{-1}X)_k$ are decorrelated, which is implied by their mutual independence. In practice, this choice has 3 positive consequences:

- If the input vectors X are whitened -as is often done in practice- and if the decorrelated sources are chosen to have unit variance, the mixing matrix M becomes orthogonal, so that $\log(|\det M^{-1}|) = 0$, which solves half of the problem.

- Though this feature is rarely used in traditional methods, the orthogonal matrix M can be estimated within the manifold of orthogonal matrices, which is a smaller space than the linear group (see below Chef d'Hotel's method).

- The term $\mathcal{I}(M^{-1}X)$ simplifies to $\sum_{k=1}^{K} H((M^{-1}X)_k) - H(M^{-1}X)$, where the notation $H(Y)$ is used to denote the entropy of the probability density of the vector Y. Now, the entropy of the joint sources is $H(M^{-1}X) = H(X)$ since M is orthogonal.

The criterion (5.20) simplifies to

$$F_{ICA}^{\perp}(M) = \sum_{k=1}^{K} H((M^{-1}X)_k) \tag{5.21}$$

This particularly simple form is easy to interpret: since the variance of the reconstructed sources S_k is equal to 1, the minimization of their entropy for a constant variance makes their probability density more distant from the gaussian distribution, i.e. it emphasizes the non-normality of the data. Indeed, the gaussian distribution is the one that has maximal entropy for a given variance. In fact, there is an underlying idea which is that *the mixture (linear superposition) of independent statistical variables is closer to a gaussian distribution than the independent variables themselves.* This is nothing but a restatement of the central limit theorem. Technically, one can also consider that the criterion (5.21) is intended to maximize the negentropy of the marginals of $M^{-1}X$, i.e. the sum of the differences between the entropy of the unit variance univariate gaussian and the entropy of the actual source

$$F_{neg} = \sum_{k=1}^{K} (H^g((M^{-1}X)_k) - H((M^{-1}X)_k) = \sum_{k=1}^{K} \left(\frac{\log(2e\pi)}{2} - H((M^{-1}X)_k) \right) \tag{5.22}$$

This equivalent criterion makes explicit the deviation from normality of the sources. But a consequence of this analysis is that such ICA algorithms cannot unmix gaussian sources. Before going to the algorithms used for the solution of this problem, let us introduce the notation:

$$\psi(X) = \nabla H(X) = (\log(\mathcal{D}))'(X) = \frac{\mathcal{D}'}{\mathcal{D}}(X) \tag{5.23}$$

where $\mathcal{D}$ is the probability density of X. $\psi(X)$, known as the *score vector* of X, is the natural gradient of the entropy criterion (see appendix C.1.3). For a gaussian variable X^g, one has in particular $\psi(X^g) = \frac{1}{var(X^g)} X^g = X^g$ if the variance of X^g is fixed to 1; therefore non-gaussianity is equivalent to non-linearity of the gradient with respect to X. Letting $\mathcal{D}$ evolve in the opposite direction of its (supposedly non-linear) score vector minimizes locally its entropy. Thus $-\psi$ plays the role of an empirical contrast function that stirs away the density from its gaussian approximation. The estimation of $\mathcal{D}$, and thus of ψ is carried as described in appendix C.2.

Some algorithms for ICA

The minimization of criterion (5.21) can be achieved in many ways. Several algorithms have been proposed:

a) FastICA [117]: This algorithm does not use the natural gradient of the criterion (5.21), but an arbitrary gradient -or contrast function- that does not require the computationally heavy estimation of the density $\mathcal{D}$. For instance

- $\psi_1(X) = 4X^3$ is the derivative of the kurtosis of the distribution of X with respect to X. Since the kurtosis of a gaussian variable is 0, this simple criterion is a consistent way of maximizing negentropy.
- $\psi_2(X) = tanh(\alpha X)$ which biases the resulting density of S towards a $\frac{1}{cosh(\alpha X)}$ distribution, but requires a choice for α. The resulting density has heavier tails than a gaussian density.
- $\psi_3(X) = X \exp\left(-\frac{X^2}{2}\right)$ which introduces another explicit non-linearity in the score function, which puts more weight on the tails of density.

The update rule for the unmixing matrix $W = (W_1, .., W_K)$ is then

$$W_k \leftrightarrow \mathbb{E}X\psi(W_k^T X) - \mathbb{E}\psi'(W_k^T X)W \tag{5.24}$$

with the desired form for ψ, followed by an orthogonalization procedure, until convergence of the algorithm at the fixed point. Note that this method is derived from a Newton optimization approach [117]. This procedure used is for fMRI data in [37], [68], and [15] for instance.

b) The infomax algorithm: A second approach is the computation of the unmixing matrix directly from (5.20) (i.e. without imposing M and W orthogonal). Formally, this algorithm intends to *learn* the true sources and mixing matrix by iteratively updating their current estimates.

$$\delta W = \epsilon(I_K + \psi'(WX).(WX)^T)W \tag{5.25}$$

where $\psi(x) = \frac{1}{1+\exp(-x)}$ is an arbitrary nonlinear function and ϵ is a learning rate - i.e. a rate of change in the estimation procedure. Note that this form of the learning maximizes the entropy $H(WX)$, but does not necessarily minimize the mutual information of the reconstructed sources $\mathcal{I}((WX)_k) = \sum_{k=1}^{K} H((WX)_k) - H(WX)$, since the marginal entropies are not controlled. As a consequence, this algorithm is theoretically suboptimal from the above perspective. However, it has been used extensively in the fMRI literature [153] [154] [101] [35] [38] [212].

A study has shown that the outcome of the ICA procedure depends very weakly on the algorithm used [57].

c) Chef d'Hotel ICA algorithm. This more recent procedure has still not been used in the fMRI literature, but we mention it, since it brings several improvements in the minimization of the problem (5.21). After whitening of the sources, the noiseless ICA

problem amounts to finding the orthogonal matrix W in order to minimize a criterion of the output marginals (5.21). In fact the problem is formally the learning of an optimal orthogonal unmixing matrix given an objective function:

$$min_{\delta_W} \sum_{k=1}^{K} H\left((W+\varepsilon\delta_W)X\right)_k \mid (W+\delta_W) \in O(K) \tag{5.26}$$

Where ε is some positive constant and $O(K)$ the orthogonal group of dimension K. To simplify matters, let us denote $S = WX$ and $\delta_W = \delta_A W$ If we denote ψ the natural gradient of the problem in S (as in equation (5.23)), the optimal δ_A is defined by

$$\widehat{\delta_A} = argmin_{\delta_A}\|\delta_A WX - \psi\| = argmin_{\delta_A}\|\delta_A S - \psi\| \tag{5.27}$$

for some norm, under the constraint that $\delta_W \in T_W O(K)$, i.e., that δ_W belongs to the tangent space of the orthogonal group in W; this is equivalent to requiring that δ_A belong to the tangent space of the orthogonal group at the identity, which is the space of antisymmetric matrices. A straightforward solution is to take $\delta_A = \psi S^T - S\psi^T$. Then $W + \varepsilon\delta_W = (I + \varepsilon\delta_A)W$ remains approximately orthogonal. The new idea of Chef d'Hotel is that $I + \varepsilon\delta_A$ can be advantageously replaced $\exp(\varepsilon\delta_A)$. The main advantage is that W remains strictly orthogonal (and not up to the first order), since $\exp(\varepsilon\delta_A)$ is orthogonal $\forall\varepsilon$. Straightforwardly, this allows to increase the learning rate ε without deviating from the manifold of orthogonal matrices. This is not a *trick*, but stems from the observation that the curve

$$\varepsilon \rightarrow \exp(\varepsilon\delta_A) \tag{5.28}$$

is in $O(K)$ the geodesic -for the usual norm- with tangent vector δ_A at the origin ($\varepsilon = 0$). Last, this exponential is well approximated by

$$\exp(\varepsilon\delta_A) \simeq (I - \frac{\varepsilon}{2}\delta_A)^{-1}(I + \frac{\varepsilon}{2}\delta_A) \tag{5.29}$$

This approximation is fortunately an orthogonal matrix. In our implementation of spatial ICA, we use this method, which takes the best advantage of the mathematical framework of the problem.

Temporal ICA

It is probably unsafe to use the simple transposition of the spatial ICA into temporal ICA for fMRI data. The reason is that the statistical distribution of spatial maps relies on N samples, N being the number of voxels, and is thus quite reasonable, while the number of time samples is only $T << N$. This procedure is nevertheless achieved in [23] [37], though with little details. More realistically, some authors have used specifically temporal methods based on signal autocorrelation as in [171] [145]. The temporal Canonical Correlation Analysis (CCA) (see 5.1.4), is essentially another implementation of the same idea: the temporal signals of interest are strongly autocorrelated, so that temporal ICA can be performed by autocorrelation maximization. Last, the combination of both spatial and temporal ICA has been proposed in [201], but the choice of ad hoc weighting and contrast functions as well as the lack of intrepretability (no clear definition of the solution) weaken the credibility of this proposition.

5.2.2 ICA applied to fMRI: some pitfalls

An overview of the questions related to ICA of fMRI data can be found in [34]. However, we discuss here some points drawn from our own experience.

Theoretical point of view

Non-compatibility with spatial smoothing In the derivation of the ICA criterion, we have noticed that the statistical independence of the voxels was necessary. This is not completely true in practice, since there are spatial correlations embedded in the data. More importantly, this hypothesis is violated when prior spatial smoothing is applied to the data. However, the smoothing procedure has been applied in several studies [37] [35] [36] [38] [15], [201] which is a methodological inconsistency.

Arbitrariness in the contrast function The learning process associated with ICA should be based on the natural gradient of the problem (5.23). However, the -relative- difficulty of the estimation of the latter has induced the use of contrast functions that introduce some deviation from normality of the statistical distribution of the reconstructed components. This can be interpreted as the introduction of priors in the statistical distribution of the true sources (see for example [153]), but, to our knowledge, the validity of this argument has not been investigated for fMRI data.

Independence: a good criterion for fMRI data ? The ultimate justification for the use of ICA in fMRI data analysis is that the data is the superposition of statistically independent spatial patterns, which correspond to some effect. This is probably false in the sense that the spatio-temporal structure of fMRI datasets cannot be completely factorized into spatial and temporal components; rather there exists a kind of functional connectivity that governs simultaneous and successive neural -and thus hemodynamic- activations. What is criticized here is that ICA tries to uncover a *generative process* of the data, and that a description in terms of spatially independent components probably fails to reach a realistic generative process [158].

Practical point of view

Dimension reduction The main ICA algorithms have been denoised to use invertible, and thus square mixing matrices, so that the number of sources K is the dimension of the ambient space. This yields quite immediately two related difficulties: how to choose K, and how to choose the ambient space ? The second question is often solved by choosing the number of components of PCA, which is used for data reduction. The problem is then exactly the problem described in 5.1.2, for which only heuristics are currently available. It is worse however for ICA, since the use of an ICA algorithm with multiple gaussian distributions is likely to create spurious components. On the other hand, ignoring too many components may yield insufficient reconstruction of the true signal. To date, no systematic approach is available to answer this question.

Interpretation of ICA outcome A difficult point is to deal with the outcome of an ICA decomposition, since nothing tells us which components are of interest. This is due

to the fact that ICA never considers the informative content of the time courses, but only the statistical structure of the dataset. We are in a situation where we lack priors. Some solutions have been proposed:

- To study the information structure of the results (connectivity of spatial components, autocorrelation of the associated temporal components, measure of the deviation from normality of the spatial components (e.g. kurtosis)) [68]. For example, one can conclude from figure 5.2 that the marginal entropy of the sources is a possible indicator for the presence of a particular effect.

- More simply, to study the temporal components by linear regression, or by correlation with the hypothesized response - which is of course done more or less implicitly by all authors. But then, is it really worth to perform ICA ?

- Building an hybrid model in between the general linear model and the ICA decomposition [152]. The result is still difficult to interpret, so that the method has not been considered in practical applications.

On the other hand, one can also notice that the absence of priors is an advantage in the absence of an experimental paradigm. In particular, ICA can be used in the study of *spontaneous* physiological fluctuations [130].

5.2.3 Some improvements of the ICA methodology

Probabilistic ICA

Probabilistic ICA (PICA) embeds the usual spatial ICA algorithm in a well-defined probabilistic framework ([15], [14]). In fact, it is based on the initial noisy mixing model (5.12). It gives a principled way to achieve model order selection through the initial PCA, the incorporation of prior information about data structure and the thresholding of the resulting spatial maps.

- Model order selection relies once again on a sphericity test, but the natural spread of the spectrum is now taken into account. The reproducibility of the results across different validation techniques is promising. Still, one may wonder whether any information of interest is contained outside the first principal components of the data. Moreover, systematically removing the first components of the data from the noise can potentially yield an underestimate of the latter.

- Prior information: the confidence about the voxel information can be tuned through anatomical considerations, and, less convincingly when dealing with ICA, with neighboring terms.

- The thresholding of the spatial maps is improved by the normalization of voxel values by the estimated noise level. The assessment is also improved by the mixture modeling on the spatial maps and the formulation of alternative to the absence of activation.

Another implementation of Probabilistic ICA can be found in [112].

Mean field ICA

Mean field ICA [114] is a generalization of probabilistic ICA that explicitly deals with the estimation of noise covariance, and the case of correlated sources - besides the usual unmixing problem. This approach requires the introduction of a prior model for the sources; then the mean field theory is used for the estimation of the sources. Next, the mixing matrix and noise covariance are maximum a posteriori estimates. To our knowledge, this promising technique has not been used for fMRI data analysis.

Convolutive ICA

The main reproach that one may want to address to spatial ICA on fMRI data is the fact that it neglects the temporal structure of the data. On the other hand, temporal ICA completely overlooks the coupling and interactions among components. In particular, the temporal ICA model implies a trivial structure for the covariance matrix. While spatio-temporal ICA [201] has a blurring effect on ICA interpretation, convolutive ICA [102] seems to be much more adapted to fMRI data; it allows for both the characterization of joint activations among different regions, and the study of different temporal behavior. If this technique raises new difficult methodological issues (selection of components, complex estimation of the spatio-temporal filter) it seems very promising given the nature of fMRI data.

5.2.4 Experiment with a synthetic example

We have used Chef d'Hotel's spatial ICA algorithm for the same dataset (A.1), and in the same conditions as the spatial CCA method. The spatial ICA algorithm typically had a better convergence speed and yielded more denoised components than CCA (see figure 5.2). It has excellent reproducibility under different random initializations. Last, the non-gaussian components can be very easily determined from the observation of the final sources entropies. However, since the space of activated patterns has dimension 2, the algorithm is not able to separate the 3 modes of activation. To obtain this, an overcomplete representation would be necessary.

One can conclude from this that spatial ICA is good for separating different effects, but it is not well adapted to a precise study of the activation space -indeed, different activation patterns, as those simulated in the dataset, are distinguishable, but not independent.

5.3 Clustering vs Vector quantization: what information theory tells us.

Let us recall that clustering is another exploratory method based on the following statistical viewpoint: the dataset X is a set of N features (the temporal time courses) that belong to a given signal manifold or feature space $\mathcal{F}$. The distribution of the data in $\mathcal{F}$ can be modeled as a multi-modal distribution; each mode will be characterized as a data cluster. One can notice that clustering cannot provide us with a generative model of the data, since the different processes that generate the data are not identified. Rather, clustering can be thought of as a way to give a simplified overview of the data.

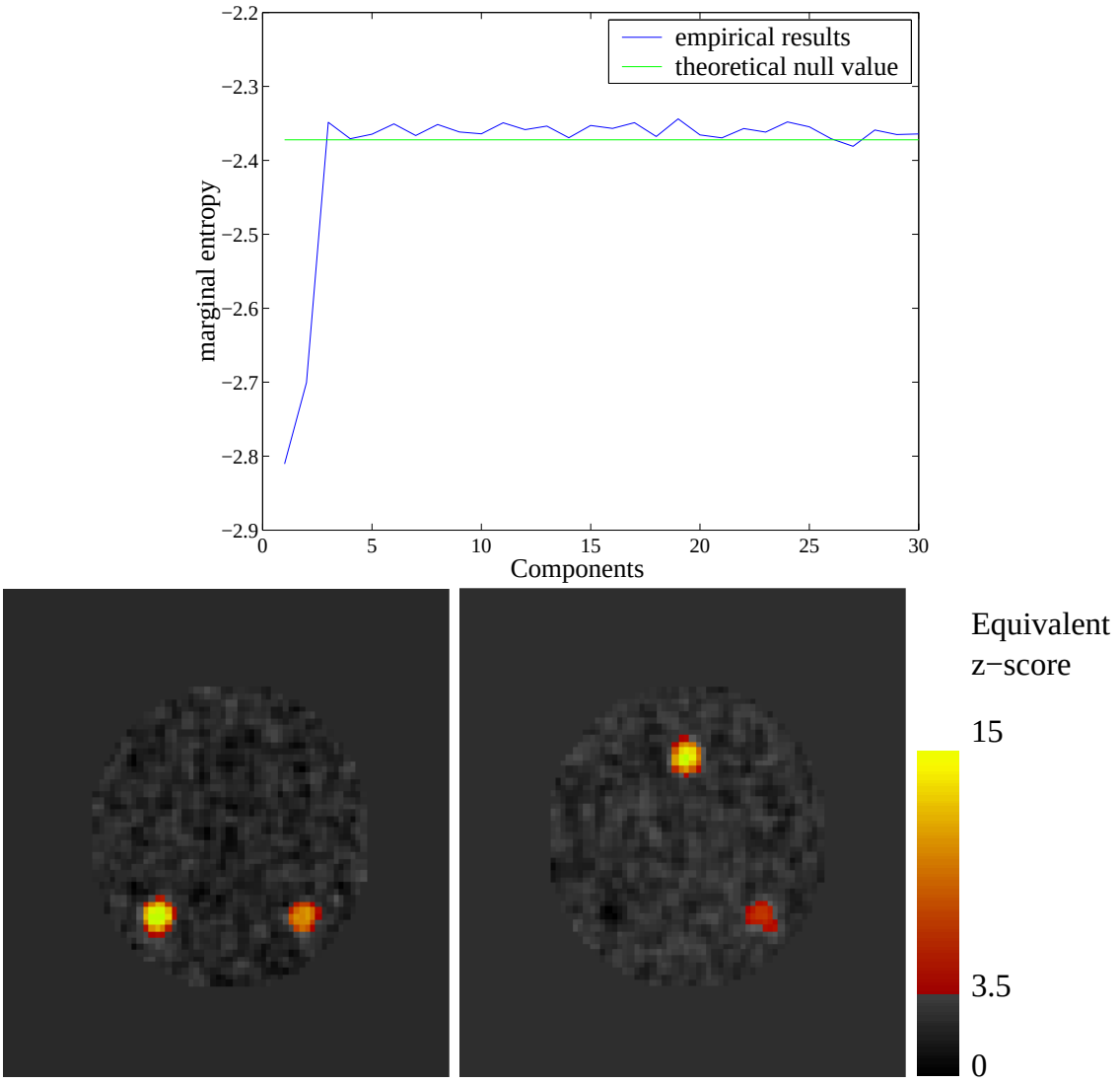


Figure 5.2: Outcome of the ICA algorithm applied to the synthetic dataset. (Top) Marginal entropies of the output spatial components versus the null (i.e. gaussian with the same variance) entropy. There are clearly only two non-gaussian components. Two main resulting images of the spatial ICA method applied to the synthetic images after PCA reduction to 30 components. These images, converted into z-scores and thresholded as indicated in appendix B, are a basis of the spatio-temporal activation patterns present within the data. The 3 modes structure of the activation appears through the different intensity of the 3 clusters within the two images, but the decomposition itself is not able to disentangle the 3 modes. This would require an overcomplete representation of the data by the decomposition, which is not possible with usual ICA algorithms.

5.3.1 Clustering: a fruitful approach for fMRI data?

As noticed earlier, clustering is a very flexible approach. It does not require the use of a linear decomposition of the data, as PCA or ICA. But the price to pay for this flexibility is quite heavy:

- The first difficulty is the definition of the feature space $\mathcal{F}$, that is, the metric that is used to quantify the similarity between time courses. If non Euclidean metrics are preferable, the convergence of the algorithms can be problematic in that case (for example, the update rules of C-means/ fuzzy C-means algorithms yield convergence in the case of an Euclidian metric, but not necessarily for arbitrary metrics).
- The quality of clustering results is difficult to assess (see the discussion in section 3.3.3).
- Determining the number of clusters (main modes of the data density in Γ) is not obvious. It results in fact in a bias/ variance trade-off: the more clusters you use to describe the dataset, the less bias you have in the resulting clusters -i.e. each sample resembles more in average to its cluster- but the more variance you have in the precise determination of the clusters: doing the clustering procedure with different realizations of the data more likely results in different configurations. However, this tradeoff is usually ruled implicitly by the blind choice of the number of clusters.
- If some undesirable effect (trend, spike) is present in the dataset, then potentially all clusters can be affected, while ICA or PCA can -at least in theory- isolate such a pattern.
- More generally, the influence of noise on clustering procedures is not easily identified.
- It is difficult to make inference on clustered data (most authors simply claim that a cluster of voxels is activated if its centroid is correlated with the reference time course, but what about *all* the voxels of the cluster ?).

In the next section, we propose thus a clustering method which is based on the explicit definition of a low dimensional space of interest in which each voxel is not represented by a point (i.e. its projection onto the feature space), but by a probability density, which reflects the uncertainty about the exact position of the point within the space. A key feature of this method is that it makes explicit the bias/variance inherent to any clustering method. This correspond to the work published in [209].

5.3.2 Making the compactness/precision trade-off explicit: the Information Bottleneck approach

The feature space:

We assume here that some preprocessing has been done on the data, so each observation is reduced to a low dimensional feature γ, i.e. $\mathcal{F} = \Gamma$. A typical case is the choice of $\gamma = b$ as the outcome of linear regression described in equations (3.3)-(3.4). The dataset is thus represented by a set of voxels X isomorphic to $[1, .., N]$, and the conditional distribution

$$p(\gamma|n) = \mathcal{N}(\hat{\gamma}(n), \Lambda_\gamma(n)) \tag{5.30}$$

where $\mathcal{N}(\hat{\gamma}(n), \Lambda_\gamma(n))$ is the normal distribution of mean $\hat{\gamma}(n)$ and variance $\Lambda_\gamma(n)$, these quantities being typically the least square estimates and dispersion of a parameter of interest.

The Information Bottleneck method

The Information Bottleneck (IB) method, described in [214], is a solution to the problem of optimal lossy data compression. It addresses the following problem: given a discrete dataset $X = [1, .., N]$, a discrete space of interest Γ, and the conditional densities $p(\gamma|n)$, find the *codewords* $\tilde{X}$ that maximize compression while retaining most of the information on $p(\Gamma|X)$. In mathematical terms this leads to the minimization of the quantity

$$I(X, \tilde{X}) - \beta I(\tilde{X}, \Gamma) \tag{5.31}$$

with respect to $\tilde{X}$, where $I(X, \tilde{X})$ is the mutual information between the dataset and its compressed representation, $I(\tilde{X}, \Gamma)$ is the mutual information between the compressed representation and the variable of interest, and β a positive scalar. We will see that β controls the bias/variance tradeoff. The minimization of $I(X, \tilde{X})$ yields compression of the original data X into $\tilde{X}$, while the maximization of $I(\tilde{X}, \Gamma)$ implies that the compressed data must preserve as much information as possible on Γ.

This problem has been shown to have a formal solution, which is obtained by differentiating equation (5.31) with respect to $p(\tilde{x}|x)$. In terms of $p(\tilde{x}|x)$ it satisfies the equation

$$p(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x, \beta)} \exp\left(-\beta \sum_\gamma p(\gamma|x) \log \frac{p(\gamma|x)}{p(\gamma|\tilde{x})}\right) \tag{5.32}$$

where

$$p(\gamma|\tilde{x}) = \sum_x p(\gamma|x)p(x|\tilde{x}) = \frac{1}{p(\tilde{x})} \sum_x p(\gamma|x)p(\tilde{x}|x)p(x) \tag{5.33}$$

(the first equality is postulated, while the second is the application of the classical Bayes theorem) and

$$Z(x, \beta) = \sum_{\tilde{x}} p(\tilde{x}) \exp\left(-\beta \sum_\gamma p(\gamma|x) \log \frac{p(\gamma|x)}{p(\gamma|\tilde{x})}\right) \tag{5.34}$$

is the normalization function.

Let us note that $\sum_\gamma p(\gamma|x) \log \frac{p(\gamma|x)}{p(\gamma|\tilde{x})}$ is nothing but the Kullback-Leibler divergence between the two probability distributions $p(\gamma|x)$ and $p(\gamma|\tilde{x})$, also noted $K[p(\gamma|x)|p(\gamma|\tilde{x})]$ in section 5.2.1. We write it henceforth $d(x, \tilde{x})$. Then, equation (5.34) rewrites

$$Z(x, \beta) = \sum_{\tilde{x}} p(\tilde{x}) \exp(-\beta d(x, \tilde{x})) \tag{5.35}$$

The practical solution of this problem does not have a closed form. Nevertheless, the following result holds [214]:

Equation (5.32) is satisfied at the minima of the functional

$$F(p(\tilde{x}|x), p(\tilde{x}), p(\gamma|\tilde{x})) = -\left\langle \log Z(x, \beta) \right\rangle_{p(x)} = I(X, \tilde{X}) + \beta \left\langle d(x, \tilde{x}) \right\rangle_{p(x, \tilde{x})} \tag{5.36}$$

where $\langle S(a) \rangle_{p(a)}$ stands for the expectation of the quantity S under the law p. The minimization can be done independently over the convex sets of the normalized distributions $p(\tilde{x})$, $p(\tilde{x}|x)$ and $p(\gamma|\tilde{x})$ by the converging alternating iterations (t being here the iteration step):

$$p_t(\tilde{x}|x) = \frac{p_t(\tilde{x})}{Z_t(x,\beta)} exp(-\beta.d(x,\tilde{x})) \tag{5.37}$$

$$p_{t+1}(\tilde{x}) = \sum_x p(x) p_t(\tilde{x}|x) \tag{5.38}$$

$$p_{t+1}(\gamma|\tilde{x}) = \sum_x p(\gamma|x) p_t(x|\tilde{x}) \tag{5.39}$$

An excessive number of clusters are generated randomly at the beginning. The IB algorithm (equations (5.37), (5.38), (5.39)) is applied to the data until convergence (typically a few hundred iterations). We use then the final probability laws $p(\tilde{x}|x)$ for a hard clustering of the data ($cl(x) = argmax_{\tilde{x}}\, p(\tilde{x}|x)$). The final number of clusters is given by the ones whose probability has not canceled during the iterations (i.e. $\{\tilde{x}/\exists x/\tilde{x} = cl(x)\}$). The number of remaining clusters is thus provided by the algorithm and depends highly on the choice of β, whose interpretation as a scale parameter is obvious. Consequently, the bias/variance trade-off in this quantization model is completely governed by the parameter β.

In practice, the use of a finite grid for the sampling of the probability density functions (pdfs) is important. From our experiments, it seems that the grid precision does not have much importance on the final result, as long as it is not coarser than the intrinsic data dispersion.

5.3.3 Making inference

Once data quantization is achieved, each surviving cluster $\tilde{x}$ is naturally represented by the density in the feature space

$$p(\gamma|\tilde{x}) = \sum_{cl(x)=\tilde{x}} p(\gamma|x) p(x) \tag{5.40}$$

This can be used to decide which clusters are significantly far from the null hypothesis; this means that one can check that $p(\gamma = 0|\tilde{x})$ is significantly small. This can be done by using the Highest Posterior Density (HPD) regions method (see [124], [98]). One considers the region $H_0 = \{\gamma/p(\gamma|\tilde{x}) > p(0|\tilde{x})\}$ and the associated probability $1 - \alpha = \int_{H_0} p(\gamma|\tilde{x}) d\gamma$. Then the null hypothesis can be rejected with probability α for cluster $\tilde{x}$. Of course, this method is nothing but a heuristic, but it is useful in practice to characterize activation at the cluster level.

If a hypothesis can be represented as a hyperplane splitting the feature space, then the inference is more simple. Let H_1 and $\bar{H_1}$ be the two alternatives and F_1 and F_2 be their image in the feature space. One obtains easily

$$p(H_1|\tilde{x}) = \int_{F_1} p(\gamma|\tilde{x}) d\gamma \tag{5.41}$$

and $p(\bar{H_1}) = 1 - p(H_1)$. This makes hypothesis testing easy.

5.3.4 Some experimental results

Once again, we use the synthetic dataset presented in A.1. Through equations (3.3) (3.4), we obtain a $S = 2$ dimensional feature-space that corresponds to the amount of signal associated with each experimental condition. We have displayed the estimated feature at each voxel in figure 5.3(a). Then, we have discretized the feature space on a (20×20) grid and analyzed it with the IB method. To study the dependence of the number k of final clusters on β, we present the cluster hierarchy, indexed by β, in figure 5.3 (b).

Figure 5.3 (b) shows that the 4 clusters configuration is the main non-trivial one. The associated pdf $p(\gamma|\tilde{x})$ (figure 5.3(c)) confirms the pertinence of this model. For comparison, we have applied a fuzzy-C-Means algorithm with 4 clusters on the same feature space, with 10^4 independent random initializations. In no case did we obtain the results described in figure 5.3 (d). This may be attributable to the small number of activated voxels, and to the inadequate choice of the Euclidean metric.

Conclusion

Let us emphasize the benefits of information theory in multivariate analysis of fMRI data: it gives a principled way to handle some bias/variance tradeoffs that inevitably emerge in multivariate modeling: For example, the Bayesian Information Criterion (BIC) D is a practical solution for the selection of the meaningful components of linear decompositions (see e.g.[103]). Figure 5.2 shows that the negentropy (5.22) of spatial maps is a possible measure for their deviation from null (i.e. gaussian) maps; last, the Information Bottleneck method shows that mutual information provides usable criteria for handling the bias/variance tradeoff inherent to any clustering method.
In the sequel of this thesis, we try to integrate multivariate and temporal modeling approaches to constrain the results of data decompositions towards interpretable phenomena.

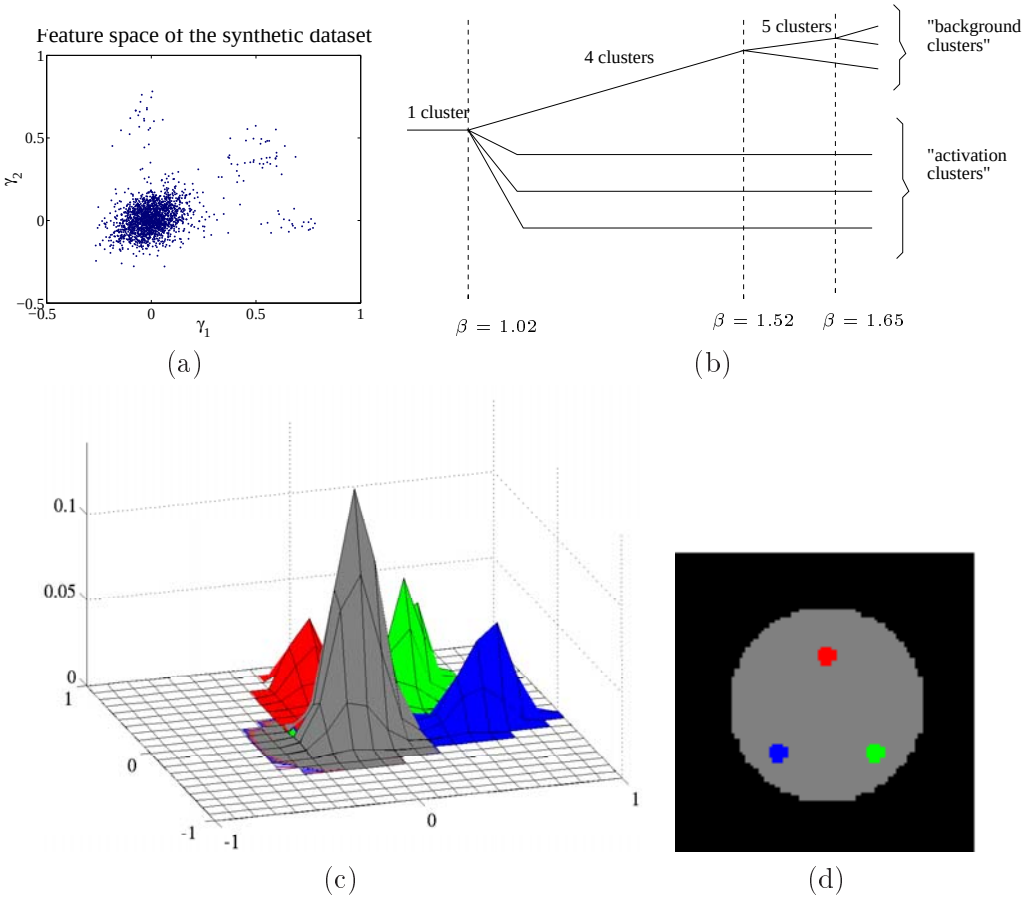


Figure 5.3: Cluster analysis of FMRI data with the Information Bottleneck method (a) Estimated features at each voxel $\hat{\gamma}(n) = (\hat{\gamma}_1(n), \hat{\gamma}_2(n))$ (each point represents the center of a gaussian density, and is thus associated with a dispersion, which is not represented here). (b) Cluster hierarchy obtained by letting the scale parameter β vary. Clusters appear by successive bifurcations or splittings. The terms *activation* clusters and *background* clusters refer to post hoc inference. The configuration with 4 clusters is stable over a large scale interval; we refer to this configuration in figures (c) and (d). The associated spatial map $cl(n)$ (d) is in fact identical to the original activation map(c). Probability density functions are associated with the four clusters $p(\gamma|\tilde{x})$. Note that they correspond readily to the main mode and the three "arms" of the feature distribution clearly visible in figure (a). Colors of figure (c) and (d) match.

Part II

Analysis

Chapter 6

Towards a spatio-temporal understanding of fMRI data: A state-space approach

In the previous two chapters, we have successively developed temporal models adapted to univariate modeling of fMRI data, and multivariate methods that give global descriptions of the data. A question of interest is naturally to build an integrated framework where both approaches are used. We reintroduce here the concept of generative model, and adopt a state-space formulation. We then describe some solutions to the problem, and three practical applications of this approach. We first state the problem in terms of dynamical components analysis.

6.1 Reformulation of the problem

Here we clearly state the general question of fMRI data analysis as a joint mixing/evolution estimation problem, or equivalently as the estimation of a generative model.

6.1.1 Dynamical Components Analysis

Let X be a fMRI dataset, where each $X(t)$ is an image which is simply written as an N-dimensional vector. N is thus the number of voxels considered in the analysis, and the length of the series is T. As stated in [207], the dynamical components analysis of X is a decomposition:

$$X(t) = \sum_{k=1}^{K} M_k z_k(t) + W(t) \tag{6.1}$$

where $0 < K < min(N, T)$, $\{M_k\} k = 1..K$ and each $W(t)$ are N-dimensional vectors and $\{z_k(t)\}, k = 1..K$ are temporal signals.

Equation (6.1) means that the dataset is decomposed into meaningful signals plus a noise term. The problem is to estimate K, $M = \{M_k\}$ and $Z(t) = \{z_k(t)\}$ given the data X and some prior information on the experiment (e.g. the experimental paradigm). The experimental paradigm will be represented as a $C \times T$ matrix P, where C is the number of experimental conditions. Each row of P is the time course of an experimental condition.

First, let us notice that this model is typically a generative model of the data: one can indeed assume that the empirical data is the compound of different processes; the general problem consists in finding those processes, together with the mixing model that relates them with the observations.

Second, the question is whether the processes of interest should be considered as independent or not. If yes, the problem essentially reduces to temporal ICA - with some possible constraints on the spatial counterparts of the decomposition. In particular, in [207], each component is treated as a univariate model analogous to those presented in chapter 4. Typical solutions to this problem are described in [175]. But it is probably more realistic to consider that the dynamical components can interact, so that their joint evolution does not simply factorize into spatio-temporal components. The different voxels of the dataset present different mixtures of this multi-dimensional system. In section 6.1.2, we develop a model based on this second idea; this model has been presented in [210].

Third, this problem is difficult, since all the quantities involved in the right-hand side of equation (6.1) have to be estimated. Moreover, the short time series derived from fMRI data are not well suited for the unambiguous definition of temporal effects. This is a common pitfall to all temporal ICA methods applied to fMRI data. We develop these questions in section 6.1.3.

6.1.2 The state-space formulation

The next step is to add to equation (6.1) another equation that defines the evolution of the system, which is the "dynamical" part of the problem. Let Φ be a function from $\mathbb{R}^{K+C}$ to $\mathbb{R}^K$. We propose the following model for the data:

$$Z(t+1) = \Phi(Z(t), P(t)) + V(t) \tag{6.2}$$

$$X(t) = MZ(t) + W(t) \tag{6.3}$$

where each $K-$dimensional vector $V(t)$ is a second noise term, known as innovation process. Note that equation (6.3) is exactly equation (6.1), which we will call the mixing equation, since it interprets the observation X as a noisy mixing of the $K-$dimensional state variable Z. Equation (6.2) is the evolution equation, which models the temporal evolution of the data. Z is called the state of the system, since it indeed gives all the relevant temporal information about the system. The covariance matrices of V and W will be denoted Λ_V and Λ_W, respectively.

We now have to deal with the estimation of Φ, which requires the use of more assumptions. Let us look for a linear approximation of Φ which will be written as a $K \times (K+C)$ matrix A. Equation (6.2) becomes :

$$Z(t+1) = A[Z(t), P(t)] + V(t). \tag{6.4}$$

Letting $A = [A_1^T, A_2^T]^T$, equation (6.4) rewrites

$$Z(t+1) = A_1 Z(t) + A_2 P(t) + V(t). \tag{6.5}$$

One can recognize that our evolution equation is a well-known multivariate AR(1) model, excited by the input $P(t)$. More in the spirit of the model presented in chapter 4, we can also make separate models for autoregressive components and task-related ones.

Importantly, we can notice that the multivariate state Z is in fact a signal subspace, and that the marginals do not have necessarily any meaningful interpretation.

An obvious limitation of the system (6.2-6.3) is the restriction to a first order model. One formal answer is to translate higher order evolution models into a first order one by introducing an auxiliary variable $Y(t) = (Z(t), Z(t+1))$ and solving a similar system in Y. However, this is not always practical, since it introduces more parameters to estimate in the model.

Last, it can be observed that the solution space of equations (6.2-6.3) is invariant under the action of the linear group $Gl(K)$: if $N \in Gl(K)$, and $(Z, A_1, A_2, M, \Lambda_V, \Lambda_W)$ is a solution, then
$(NZ, NAN^{-1}, NA_2, MN^{-1}, N\Lambda_V N^T, \Lambda_W)$ is also a solution.

This invariance is not incidental; it is the mathematical counterpart of the fact that the state is not a variable, but a space; the invariance under the action of $Gl(K)$ being the fact that any basis of that space gives an equivalently correct representation. Note that this invariance is somehow opposite -or complementary- to the concept of independent components. However, the practical solution of the problem will yield a particular choice for Z.

To our knowledge, state-space models have been proposed for fMRI data only in [113] and [95]. In [113] the authors propose a univariate modeling of the data, without prior knowledge of the experimental paradigm. In fact, the paradigm is formally replaced by a discrete state variable which is estimated from the data. In contrast, we use a multivariate model, which is probably better suited for fMRI data, and use explicitly the information about the experimental paradigm. In [95], the authors introduce a time-varying modulation of the response amplitude. This certainly improves data fitting, but prevents statistical testing and even simple interpretation of the data.

6.1.3 Strategy in the solution of the problem

Before turning to a description of linear estimation procedures, let us notice that this problem has received recently much attention [8] [7] [216] , and that non-linear estimation procedures have been proposed, in the spirit of artificial neural networks (i.e. models with non linear evolution and mixing models). Such models can be very successful for the study of non-linear dynamical systems (e.g. Lorenz attractors), but they are not usable for fMRI data at least for 3 reasons:

- These methods are built to deal with low dimensional systems with long observation intervals (i.e. $N \ll T$), which is never the case for fMRI data.
- These methods deal with high SNR data, once again in contrast with fMRI.
- fMRI signals are not exactly stationary, so that any investigation method should be sufficiently robust to weak non-stationarities. This is very challenging for non-linear methods.

Instead, we follow the classical point of view developed in [61]: the identification of the linear dynamical system enjoys a unique -modulo the above indeterminacy under the action of $Gl(K)$- optimal solution, namely the Kalman solution. Moreover, the systematic study of this model, namely the time series with a Markovian representation, shows that

all the information about the underlying dynamics and optimal estimators are in fact given by the study of the covariance

$$\Lambda_X(\tau) = \mathbb{E}(X(t)X(t+\tau)^T) \tag{6.6}$$

Though we will present in section 6.2.2 a simplified model, it is interesting to notice that the methodology proposed in [61, section 8.7]:

- Derivation of the autocovariance function $\Lambda_X(\tau)$
- Identification of the model given the autocovariance function
- Derivation of the state Z,

is the one that we follow quite closely in our work.

6.2 Solving the problem in practice

There are at least two ways to estimate the quantities involved in equations (6.2-6.3). The first one is an extension of the classical Kalman procedure, which iteratively estimates the state variable Z, and the related quantities $M, A, \Lambda_V, \Lambda_W$. It requires a value for the state dimension K. The other model is based on simplifying assumptions, and is thus less complicated, and moreover non-iterative. It is thus easier to apply, and gives good approximations of the global solution. Moreover, it can be used to estimate the state dimension K.

6.2.1 E.M. Kalman method

The first solution of the problem consists in implementing an Expectation Maximization version of the Kalman method [92]. This approach requires the prior knowledge of K. We use the following formulation of the problem

$$\begin{aligned} Z(t+1) &= A_1 Z(t) + A_2 P(t) + V(t) && (6.7)\\ X(t) &= M Z(t) + W(t) && (6.8) \end{aligned}$$

for $t = 1..T-1$ where the innovation and observation noise processes are not auto-correlated but have covariance matrices Λ_V and Λ_W, respectively. Moreover, we assume that the initial state is gaussian with mean μ_1 and covariance Λ_1. Then, given $\Lambda_V, \Lambda_W, \mu_1, \Lambda_1, A_1, A_2, M$ the log-likelihood of the joint state and observation writes

$$\begin{aligned} \mathcal{L}(Z,X) = & -\frac{1}{2}\sum_{t=1}^{T}([X(t)-MZ(t)]'\Lambda_W[X(t)-MZ(t)]) - \frac{T}{2}\log|\Lambda_W| \\ & -\frac{1}{2}\sum_{t=1}^{T-1}([Z(t+1)-A_1Z(t)-A_2P(t)]'\Lambda_V[Z(t+1)-A_1Z(t)-A_2P(t)]) \\ & -\frac{T-1}{2}\log|\Lambda_V| \\ & -\frac{1}{2}[Z(1)-\mu_1]'\Lambda_1[Z(1)-\mu_1] - \frac{1}{2}\log|\Lambda_1| - \frac{T(p+K)}{2}\log 2\pi \end{aligned} \tag{6.9}$$

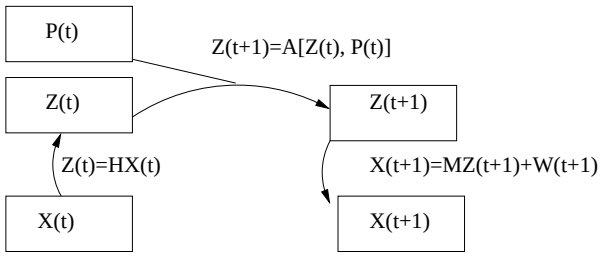


Figure 6.1: Basic mixing/evolution model in the case of a deterministic ($V = 0$) evolution.

The EM algorithm maximizes $\mathcal{L}(Z, X)$ alternatively by estimating Z given $(X, \Lambda_V, \Lambda_W, \mu_1, \Lambda_1, M)$ by the Kalman method (E step), then by estimating $(\Lambda_V, \Lambda_W, \mu_1, \Lambda_1, A_1, A_2, M)$ given (Z, X) (M step). The M step is provided by standard derivation of the expectation of criterion (6.9). We do not restate the formulas, which are given in [92].

A few comments on this method This method is somehow cumbersome in practice, since it artificially increases the number of variables. Moreover, it requires an initial guess for the dimension K of the state. A possible solution is to use information criteria together with equation (6.9) to validate K among many possibilities. What we propose is rather to use the procedure described next, which moreover yields a good approximation of the state vector in the sense of equation (6.9), and to keep the Kalman filter for the optimization of the solution, when necessary.

6.2.2 The linear method

We start by recalling the linear estimation procedure introduced by Soatto and Chiuso [198] in the context of dynamical textures analysis. This method is in fact analogous to the temporal CCA method (see section 5.1.4), or to some versions of temporal ICA [160] [200]; however the state space formalism emphasizes the concept of generative model, which is important in our context; in particular, we address explicitly the problem of state dimension estimation. The method proposed by Soatto and Chiuso is to first derive estimates for M and Z and only then for A; W and V are residuals of the estimation procedure. Assuming a deterministic evolution $V = 0$, we are in the situation described in figure 6.1. Note that the following equations assume that the dataset is reduced (e.g. by PCA) to $N < T$ entries. We come back to that point at the end of the section.

We start by neglecting the experimental paradigm $A_2 = 0$. Given the observation noise covariance Λ_W, the model represented in figure 6.1 yields the following problem:

$$min_{M,H} \|X(t+1) - MA_1HX(t)\|_{\Lambda_W}, \tag{6.10}$$

where H is the unmixing matrix associated with the mixing matrix M ($Z(t) = HX(t)$), and $\|x\|_{\Lambda_W}$ stands for $\sqrt{x^T \Lambda_W^{-1} x}$.

Since Λ_W is not known, one has to approximate this solution; one possibility is to make sure that the noise is white. Therefore, we first whiten the data X in order to approximate the situation where W is white ($\Lambda_W \sim I_N$).

Let $X_1 = [X(1), .., X(T-1)]$, $Z_1 = [Z(1), .., Z(T-1)]$, $X_2 = [X(2), .., X(T)]$ and $Z_2 = [Z(2), .., Z(T)]$. Let L_1 and L_2 be the Cholesky decomposition of X_1 and X_2 ($L_i L_i^T = X_i X_i^T$, $i = 1, 2$, and $L_i, i = 1, 2$ is trigonal, L_i=Cholesky(X_i)). Equation (6.10) is thus reformulated as

$$\widehat{M}, \widehat{H} = argmin \|L_2^{-1}(X_2 - MA_1HL_1L_1^{-1}X_1)\|_2 \tag{6.11}$$

One can introduce the whitened data $Y_1 = L_1^{-1}X_1$ and $Y_2 = L_2^{-1}X_2$ to rewrite the equation

$$\widehat{M}, \widehat{H} = argmin \|Y_2 - L_2^{-1}MA_1HL_1Y_1\|_2 \tag{6.12}$$

Since we solve the problem first in terms of M and $Z_1 = HX_1$, without prior knowledge on A_1, we make further assumptions on A_1: its singular values should be less than or equal to 1, otherwise, the system would be unstable. On the other hand, what makes the difference between the state and the noise is that the state of the system is temporally structured, so that the singular values of A_1 are close to 1. Hence, we make the hypothesis that A_1 is an orthogonal matrix. It can then be incorporated into H. This yields the following problem:

$$\widehat{M}, \widehat{H} = argmin \|Y_2 - L_2^{-1}MHL_1Y_1\|_2 \tag{6.13}$$

$$= argmin \|Y_2Y_1^T - L_2^{-1}MHL_1\|_2 \tag{6.14}$$

The latter equation resulting from the orthogonality of Y_1. It is a classical result [94] that the singular value decomposition (SVD) of $Y_2Y_1^T$ provides us with the best estimate of M and $Z_1 = HX_1$ in the sense of Frobenius, i.e. in the least squares sense:

$$Y_2Y_1^T = U\Sigma\Omega^T \tag{6.15}$$

Where U and Ω are orthogonal matrices, and Σ is diagonal. Note that by construction the singular values $\sigma_1 > .. > \sigma_N$ are between 0 and 1 and represent the correlation between the data components at time t and $t+1$. They can equivalently be interpreted in terms of mutual information between these components through the formula $I_i = \frac{1}{2}\log(\frac{1}{1-\sigma_i^2})$. The whitening procedure amounts to analyzing the empirical cross-correlation matrix of the data ($Y_2Y_1^T = L_2^{-1}X_2X_1^TL_1^{-T}$) instead of its empirical cross-covariance ($X_2X_1^T$).

In the ideal case of noise-free mixing, one has $\sigma_1 > .. > \sigma_K > 0$ and $\sigma_{K+1} = .. = \sigma_N = 0$ (see the above hypothesis on A_1). In practice, one has to set a threshold. This question is addressed later. This being done, one reduces the matrices U, Σ and Ω to their first K rows, which yields U_K, Σ_K and Ω_K.

$$\widehat{M} = L_2U_K\Sigma_K^{1/2} \tag{6.16}$$

and

$$\widehat{Z}_1 = \Sigma_K^{1/2}\Omega_K^TL_1^{-1}X_1 = \Sigma_K^{-1/2}U_K^TL_2^{-1}X_2 \tag{6.17}$$

The estimation (6.16) of the mixing matrix M allows for a least squares solution for Z and W, given equation (6.3) (note that in [198], a reprojection is proposed for the estimation of Z_2, but this has little impact on the final results). Given an estimate of Z_1 and Z_2, an estimate of A_1 follows:

$$\hat{A}_1 = \widehat{Z}_2.\widehat{Z}_1^T\Sigma_K^{-1} \tag{6.18}$$

This comes from the fact that $\widehat{Z}_1\widehat{Z}_1^T = \Sigma_K$ by equation (6.17). An estimation of V follows immediately. Note that Z has indeed a diagonal covariance matrix. In contrast, A_1 is not necessarily diagonal, which means that some coupling between the state components is allowed (the state variable Z is a true multivariate process).

Introducing priors in the analysis The next idea is to introduce prior knowledge in the estimation procedure: we know that the experimental paradigm P is a potential factor of the data dynamics, which has to be taken into account in the definition of the state variable Z and the mixing matrix M. To do this we replace equation (6.11) by:

$$\widehat{M}, \widehat{H} = argmin\|L_2^{-1}(X_2 - M(A_1HX_1 + A_2P))\|_2 \tag{6.19}$$

Writing $P_1 = [P(1), .., P(T-1)]$ and $\bar{X_1} = [X_1^T, P_1^T]^T$, one computes $\bar{L_1} = cholesky(\bar{X_1})$ and

$$\bar{Y_1} = \bar{L_1}^{-1}\bar{X_1}, \tag{6.20}$$

which allows us to rewrite equation (6.19):

$$\widehat{M}, \widehat{\bar{H}} = argmin\|L_2^{-1}(X_2 - MA\bar{H}\bar{L_1}\bar{Y_1})\|_2 \tag{6.21}$$

where $\bar{H}$ is the generalized -i.e. completed by the identity on its C last rows- unmixing matrix. This problem is solved exactly the same way by computing the SVD of $Y_2\bar{Y_1}^T$, and identifying $\widehat{M}$ and $\widehat{Z}_1$. Once Z is estimated, A_1 and A_2 are the least squares estimates obtained from equation (6.5).

A perhaps more intuitive way to understand this method is to interpret it in terms of projection: from equation (6.14), one has

$$MZ_1 = MHX_1 = L_2Y_2Y_1^TL_1^{-1}X_1 = X_2Y_1^TY_1 \tag{6.22}$$

Since $Y_1^TY_1 = X_1^T(X_1X_1^T)^{-1}X_1$, $Y_1^TY_1$ is nothing but the projector onto the rows of X_1; thus equation (6.22) simply means that the product of the mixing matrix with the state matrix is actually the projection of the data at time $t+1$ onto the data at time t. Introducing priors in the model consists simply in *augmenting* the projection operator: one projects X_2 onto the rows of X_1 and P_1 instead of X_1 only.

Estimation of K The most important part of the procedure is to determine the correct value for K, given the singular values $1 \geq \sigma_1, .., \sigma_N \geq 0$. However , the problem is ill-posed: as soon as $N > \frac{T}{2}$, one has necessarily $\sigma_1 = 1$; in fact, the rows of Y_1 and Y_2 generate two subspaces of dimension N in $\mathbb{R}^T$, thus they share at least one vector. It becomes actually impossible to test for the presence of coherent sources in the process that generated the data. We solve this issue with the recursive model (see below). Next, assuming -as all authors do, whether explicitly [72] or not- that $N \ll T$, the general method is to estimate the distribution of σ_1: Keeping the projection interpretation given in equation (6.22), σ_1 is simply the highest correlation between any two vectors belonging to two N-dimensional subspaces in $\mathbb{R}^{T-1}$ (spanned by Z_1 and Z_2 respectively). Hence its value under the null hypothesis -no underlying dynamical structure in the data- depends only on N and T. This problem is tentatively illustrated in figure 6.2.

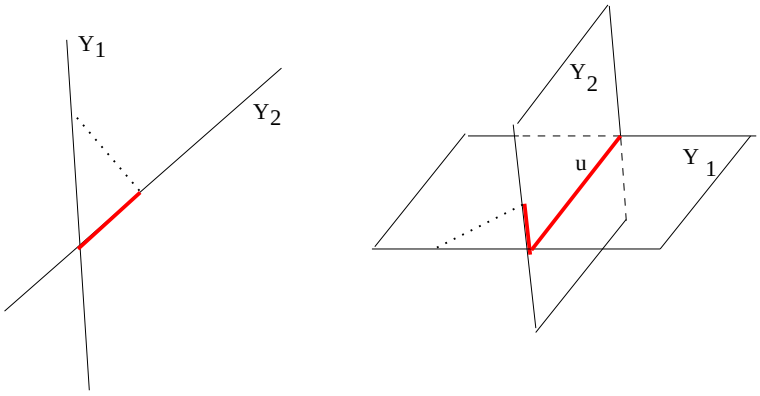


Figure 6.2: Canonical correlation and projection.

Given equation (6.22), one may interpret MZ_1 as the orthogonal projection of the space spanned by X_2 onto the space spanned by X_1 (Y_2 and Y_1 are nothing but orthonormal bases of these spaces). As an illustration, in $\mathbb{R}^3$, if these two spaces are of dimension 1 (left), the operation reduces to the projection of one vector onto the other, and the eigenvalue σ_1 is distributed as the correlation between any two random unitary vectors of $\mathbb{R}^3$. If the two subspaces have dimension 2 (right), they have necessarily one vector u in common, so that $\sigma_1 = 1$, and σ_2 is the correlation between the remaining vectors that belong to $u^\perp \equiv \mathbb{R}^2$.

A simple way to obtain the distribution of σ_1 is to generate random matrices X and derive the values of σ_1. The repetition of this procedure gives an empirical distribution for σ_1, and in particular, a confidence interval I_1 , so that $\sigma_1 \in I_1$ with sufficient probability. A simplification of this procedure is to derive only the empirical mean m_1 and standard deviation s_1 of the law of σ_1; then, approximating this distribution by a gaussian, $I_1 = [m_1 - 3.29s_1, m_1 + 3.29s_1]$ is an interval with P-value 10^{-3}. Since only the empirical mean and variance have to be estimated, this allows for quick simulation procedures.

However, let us point out the fact that the distribution of σ_1 is not gaussian (in particular, it is bounded, since $0 \leq \sigma_1 \leq 1$). It is known that the singular values of random gaussian matrices have a Wishart distribution [56], so that the distribution of σ_1 is probably also close to a Wishart distribution; however a gaussian approximation of the latter distribution is acceptable in practice.

One tests the null hypothesis on the first eigenvalue σ_1; if the test is negative, one recursively tests the null hypothesis for σ_K, in $\mathbb{R}^{N-K}$ and $\mathbb{R}^{T-K}$ until the null hypothesis is no longer rejected. This simple procedure gives excellent results on test samples.

An analytical approach Last, one can propose an approximate test for the significance of the canonical correlations. First, let us notice that the singular values defined by (6.15) are also the singular values of the matrix $Y_2 Y_1^T Y_1 = L_2^{-1} X_2 Y_1^T Y_1$. Indeed we have noticed that the multiplication by $Y_1^T Y_1$ is nothing but the projection of the rows of Y_2 onto a subspace of dimension N of $\mathbb{R}^T$. Let us write $y_2 = Y_2 Y_1^T Y_1$. The first singular value of

(6.15) is thus exactly

$$\sigma_1 = \sqrt{max_{\|w\|=1} w^T y_2 y_2^T w} = \sqrt{max_{w\in\mathbb{R}^N} \frac{w^T y_2 y_2^T w}{w^T w}} \tag{6.23}$$

Now, let $x_2 = L_2 y_2$. Applying the change of variable $u = L_2^{-T} w$ ($u \in \mathbb{R}^N$) in (6.23), we obtain

$$\sigma_1 = \sqrt{max_u \frac{u^T x_2 x_2^T u}{u^T L_2 L_2^T u}} \tag{6.24}$$

but $L_2 L_2^T = X_2 X_2^T$, so that

$$\sigma_1 = \sqrt{max_u \frac{u^T x_2 x_2^T u}{u^T X_2 X_2^T u}} \tag{6.25}$$

Now, under the null hypothesis, the entries of X_2 are i.i.d. normal and centered with variance v. Note that (6.25) does not depend on v so that we can assume that $v = 1$. We propose to approximate (6.25) by

$$\sigma_1' = \sqrt{\frac{max_u (u^T x_2 x_2^T u)}{\mathbb{E}(u^T X_2 X_2^T u)}} \tag{6.26}$$

This approximation is correct for $N \ll T$ (and thus $\sigma_1 \ll 1$). Without loss of generality one can impose that $\|u\| = 1$. Since $u^T X_2 X_2^T u$ is a χ_2 variable with T degrees of freedom, $\mathbb{E}(u^T X_2 X_2^T u) = T$.

Moreover, since $x_2 = L_2 y_2 = X_2 Y_1^T Y_1$ results from the orthogonal projection of the rows of X_2 on an N-dimensional space, there exists a matrix A of size $N \times N$ so that $u^T x_2 x_2^T u = u^T A A^T u$ and the entries of A can be assumed i.i.d. $A_{ij} \sim \mathcal{N}(0,1), 1 < i, j < N$. Let $q_1 = max(u^T A^T A u)$.

From [122], we have the following result: the distribution of q_1, has a limit when N goes to ∞. Indeed, letting $\mu = (\sqrt{N} + \sqrt{N-1})^2$ and $s = (\sqrt{N-1} + \sqrt{N})(\frac{1}{\sqrt{N-1}} + \frac{1}{N})^{1/3}$, then $\frac{q_1 - \mu}{s}$ approaches a Tracy-Widom law of order 1, whose values are tabulated, so that confidence intervals can be found for q_1. Though the result holds for $N \to \infty$, it is already correct for $N \sim 10$.

This upper bound can then be used to derive an upper bound for $\sigma_1' = \sqrt{\frac{q_1}{T}}$, which is an approximation for the upper bound of σ_1.

A comparison of the gaussian model with the empirical estimate is given in figure 6.3. The analytical is quite realistic for $N \ll T$.

6.2.3 A refinement of the linear method: the recursive model

As noted in the previous section, the procedure to estimate the state rank K breaks down when $N > \frac{T}{2}$ (and before in practice). This may depend on the way the state-space formalism will be used, but in general, for fMRI, the number of simultaneous observations exceeds by far the number of time samples, i.e. $N \gg T$. Since the rank of the data is less than or equal to $min(N, T) = T$, one can consider that $N \leq T$. In [72], the authors simply reduce their data by PCA in order to achieve $N \ll T$. However, this is a bit crude.

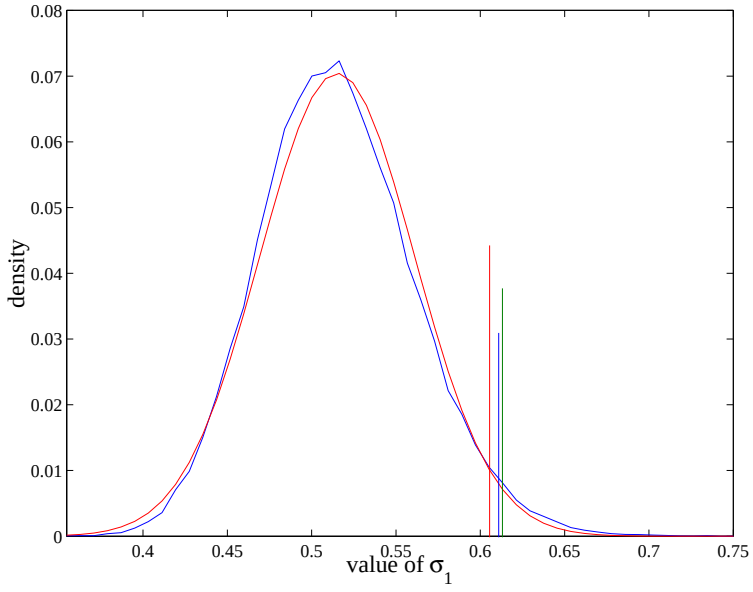


Figure 6.3: Empirical vs analytical estimation of the singular values distribution (blue) Empirical distribution of σ_1 for surrogate white gaussian data. (red) gaussian distribution with the same mean and variance. The parameters here are $T = 120$ and $N = 10$; 50000 independent surrogate datasets have been simulated. In practice, the gaussian approximation for the density yields an acceptable approximation. Vertical lines : upper bound of the 99% confidence interval: surrogate data (blue), surrogate data+gaussian hypothesis (red), analytical estimate (green).

A more cautious alternative is to keep N high enough, but to add some constraints in the estimation procedure to bypass this limitation. For example applying recursively the evolution model (6.5) provides a system of equations:

$$Z(t+p) = A_1^p Z(t) + \sum_{\tau=0}^{p-1} A_1^{p-1-\tau} A_2[P(t+\tau)] + \sum_{\tau=0}^{p-1} A_1^{p-1-\tau} V(t), \quad p \geq 1. \tag{6.27}$$

this model simply means that the prediction procedure that is used to predict $X(t+1)$ from $X(t)$ and $P(t)$ could also be extended to predict $X(t+p)$ from $X(t)$ and $P(t), .., P(t+p-1)$ -without making the model more complex.

In practice, the method is thus to generate the successive observation matrices X_1,..,X_p (instead of X_1 and X_2 only), their whitened counterparts $Y_1, Y_2..Y_p$ -which may also incorporate the experimental paradigm as in (6.20)-, and then to apply recursively the projection equation (6.22):

$$\widetilde{Y}_2 = Y_2 Y_1^T Y_1 \tag{6.28}$$

..

$$\widetilde{Y}_p = Y_p \widetilde{Y}_{p-1}^T \widetilde{Y}_{p-1}. \tag{6.29}$$

then we simply have:

$$L_p^{-1} M Z_1 = \widetilde{Y}_p \tag{6.30}$$

The mixing matrix M is then estimated by computing the SVD of $\tilde{Y_p}$, and the estimation of Z and other quantities simply follows by least squares methods.

The reason why this works is that one recursively projects a N-dimensional subspace onto a second one, then a third one ... so that the probability of having by chance a vector common to all subspaces becomes negligible. The empirical rank test described previously adapts to this generalization, by using the same iteration procedure for the surrogate data. As could be expected, the variance of the estimators increases with p.

Given N and T, we first estimate p, i.e. how many projections will be necessary in order to enable a test for the value of K, which amounts to requiring $\sigma_1 < 1$ with a given P-value (see figure 6.4). Then, one applies the recursive projection procedure to the data, possibly with the experimental paradigm. The rank of the state process is then determined by comparison with the empirical null distribution of the eigenvalues.

6.2.4 Validation on a synthetic example

We validate here our model, based on the synthetic example presented in appendix A. The main question of interest is of course the estimation of the dimension of the signal space, and then the estimation of a basis of this space. The dataset is reduced by PCA to N dimensions; we derive an estimate of K given this subspace. This involves the choice of adapted values for p (see section 6.2.3). Our result is given in table 6.1. Let us recall that the true dimension of the generative process is 2.

The dimension of the state has been estimated to 2 in all but three cases, in which it has been estimated to 1 only. This means that our rank test is correct in general, though a bit too conservative for high values of N. But this should be contrasted with sphericity tests, that do not succeed at all on such datasets, due to the flatness of the spectrum.

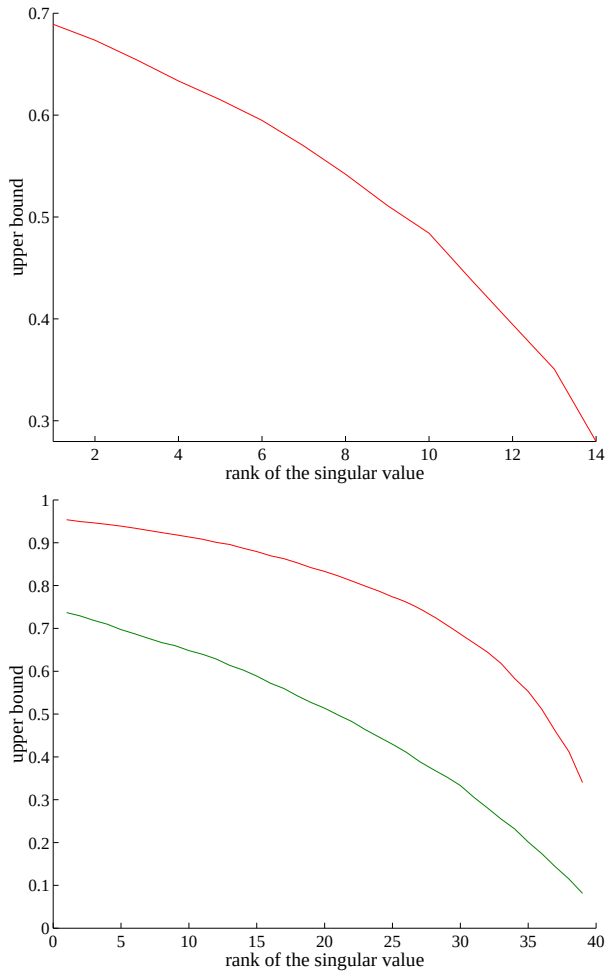


Figure 6.4: Order selection for a dataset.
Left: $T = 120$, $N = 15$; the red line represents the gaussian upper bound at $P = 10^{-2}$ for each singular value, under the null hypothesis: we are in the situation where $N \ll T$ and the basic projection step works. Right: here $T = 120$ and $N = 40$; the red line represents the gaussian upper bound at $P = 10^{-2}$ for each singular value, under the null hypothesis; (green) the same thing, but after two projections ($p = 2$ instead of 1). In this case, a second order model ($p = 2$) is recommended for the estimation of the rank K of the system.

N	p	R
10	1	2
20	1	2
30	1	2
40	1	2
50	1	1
60	2	2
70	2	2
80	2	2
90	2	2
100	3	2
110	3	1
120	3	1

Table 6.1: Estimation of the rank of the state of the synthetic dataset

The associated signal basis is given in figure 6.5, and is a correct approximation of the true activation basis: the correlation of each component with the input activation signals is respectively 0.835 and 0.739. This illustrates the good noise reduction ability of the state-space approach.

6.3 Three applications of the state-space approach

6.3.1 Analyzing multi-session data

This first application of the state-space framework is the following: we concentrate on a voxel, which is observed during several sessions, during which the subject undergoes the same experimental paradigm. It is known that the use of multi-session data increases the sensitivity of the analysis [193]. Here, the problem is to characterize the activation pattern given the multisession observation. A trivial way to condense multi-session observation is averaging, but this removes the -essential- information about session-wise variability; moreover, this is sensitive to gross artifacts. Alternatively, state-space modeling can be applied:

$$Z(t+1) = A_1 Z(t) + A_2 P(t) + V(t) \quad (6.31)$$

$$X(t) = M Z(t) + W(t) \quad (6.32)$$

with $N = S$, the number of sessions. The motivation for this model is that the reproducible pattern $Z(t)$ among the time series $X_1(t), .., X_S(t)$ taken from different sessions is likely to be also the task-related pattern -whereas most autoregressive patterns are likely not to be reproducible.

The fact that M is not constant allows for different signal intensities across different sessions (in other words, response amplitude is treated as a random effect). In that case, the estimation method can be viewed as a kind of averaging of the data, but that uses

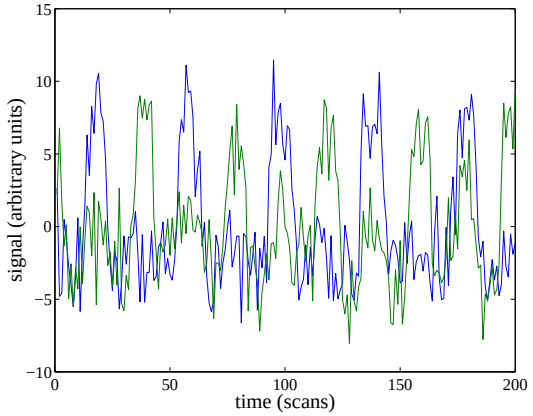


Figure 6.5: Basis of the state-space of the synthetic dataset.
This basis is a correct approximation of the true basis (compare with figure A.1): the resulting signals have high correlation levels with the activation signals used in the generation of the data.

prior information (the experimental paradigm) to give more or less confidence to each observation. By contrast, a SVD of the data gives more weight to the time series that have greater variance, while simple averaging gives equal weight to all time series.

We have applied this analysis on a voxel of the dataset taken from appendix A.3. After spatial registration of the sessions, we obtain for the given voxel 12 time series -one for each session. These time courses, acquired with the same experimental paradigm, are represented in figure 6.6.

The algorithm is applied, with the inclusion of the experimental paradigm and $p = 1$ since $12 = N \ll T = 160$. The singular values obtained from (6.15) are displayed in figure 6.7, together with their analytically derived expected and maximal value. Unambiguously, one notices that $K = 1$; interestingly, the analytical method slightly overestimates the next singular values. This is attributable to the fact that the first component makes up an important proportion of the data variance, so that the effective variance of the residuals is smaller than what would be true theoretically.

The ensuing state time series is given in figure 6.8, together with the experimental paradigm. there is clearly an activation pattern, which is confirmed by examining the A matrix: $A_1 = 0.61$, and $A_2 = [-11.95 \; -7.21]$, after normalization with respect to the innovation process ($\Lambda_V = 1$). Note that the values in A_2 are not the t values associated to the two experimental conditions. However, one can still interpret equations (6.31-6.32) as a linear model, and test whether the values in A_2 are significantly different or significantly far from 0. The resulting statistic should take into account the variance of the state estimate Λ_V and the variance of the unfitted residual Λ_W. For example, the significance of the difference between $A_2(1)$ and $A_2(2)$ can be tested with

$$t = (A_2(1) - A_2(2)) / \sqrt{\Lambda_V + \frac{trace(\Lambda_W)}{(S-1)}}, \tag{6.33}$$

Which is student distributed with $ST - 4$ degrees of freedom. However, it is not very

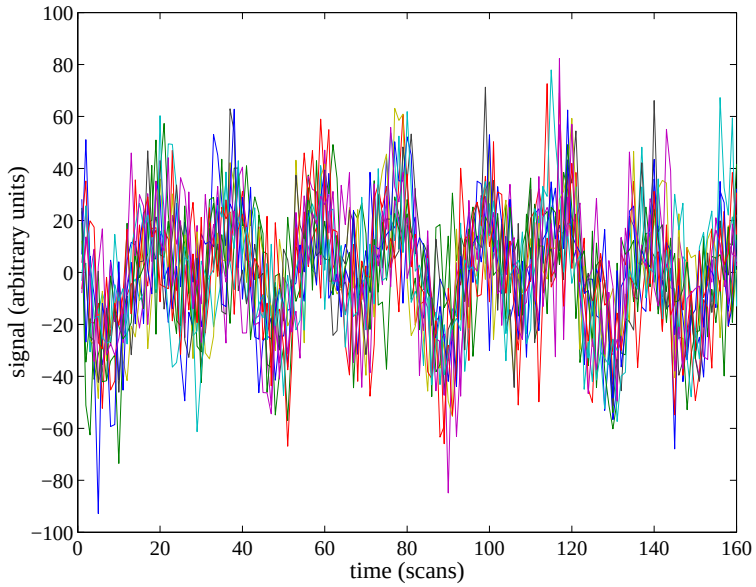


Figure 6.6: Typical input for the multi-session state-space model
The time course at a given voxel of the data recorded during 12 sessions , but with the same experimental paradigm. State-space procedures disentangle the dynamics that generated this data.

powerful, since it relies on a very rough approximation of the true signal, and the state estimation does not completely make a difference between task-related signals and potential confounds. More accurate models will be presented in chapter 9 for the derivation of accurate voxel-based multi-session hemodynamic models. This model has the advantage of using very weak priors on the signal.

Last, let us notice that using the E.M. Kalman procedure on this result did not modify it. A possible reason is that the eventual suboptimality in the estimation of the state variable is less important than the uncertainty about the other quantities of the methods, so that the Kalman procedure does not improve the estimation.

Robustness with respect to noise Here we check the robustness of our method with respect to noise in the data. To do this, we keep the same dataset, but replace the last six session data by gaussian white noise with the same variance: this means that exactly half of the data no longer carries any information. Our procedure yields once again one component ($\sigma_1 = 0.87$, which is high above the threshold in figure 6.7). Moreover, the corresponding time course is very close to our ground truth -the time course showed in 6.8- while the average signal over the 12 sessions is now quite noisy, as can be seen in figure 6.9.

We obtained the same result by replacing 9 of the input time courses by noise (in that case $\sigma_1 = 0.77$). This shows that the method is robust with respect to the massive presence of confounds in some of the sessions. Once again, the EM-Kalman method did not further improve the outcome.

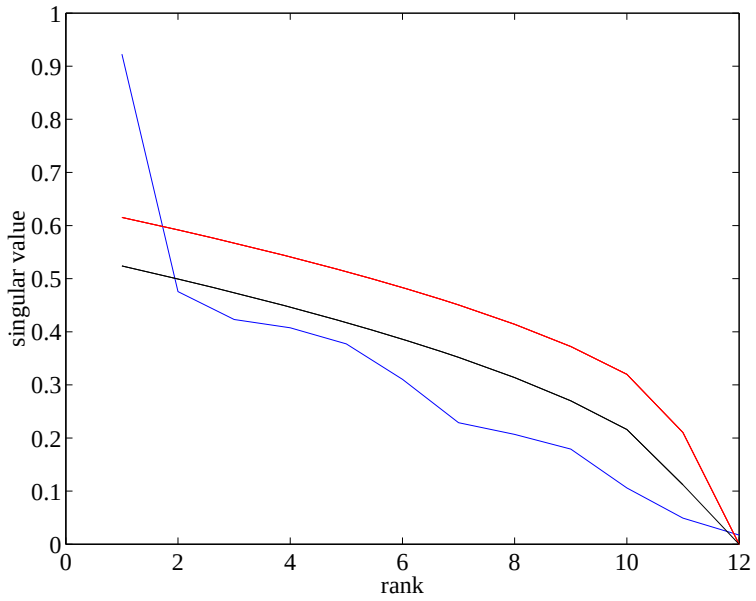


Figure 6.7: Empirical singular values and estimation of the rank of the state-space Distribution of the singular values that result from the analysis of the dataset of 6.6 (blue), together with the analytical expectation for each value (black), and the maximum of the value under the null hypothesis, with P-value of 10^{-4} (red). Only the first singular value is above the theoretical threshold, so that $K = 1$. Simulated distributions (not shown) give very similar estimations to those displayed in the figure.

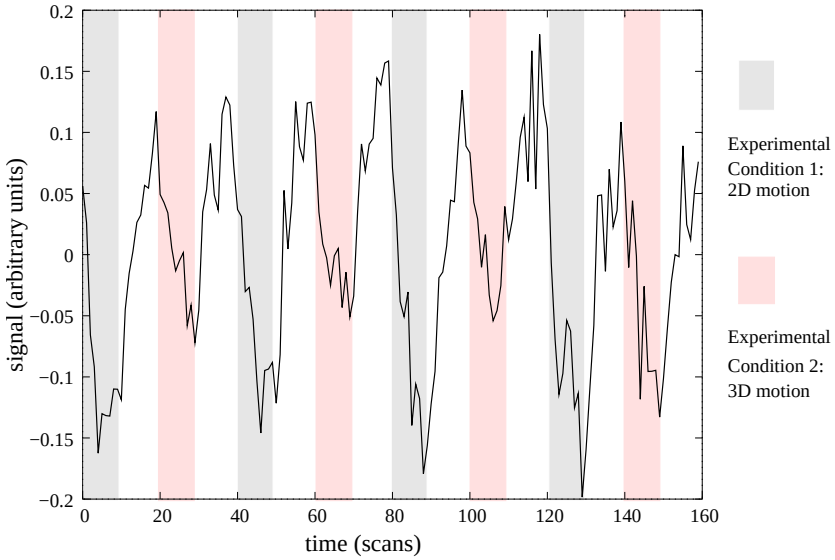


Figure 6.8: Resulting estimation of the state vector from the multi-session data. This time course is a kind of summary of the input data. It remains somewhat noisy, due to the innovation process, but there is undoubtedly a relationship between the experimental paradigm and the dynamics of the state time course.

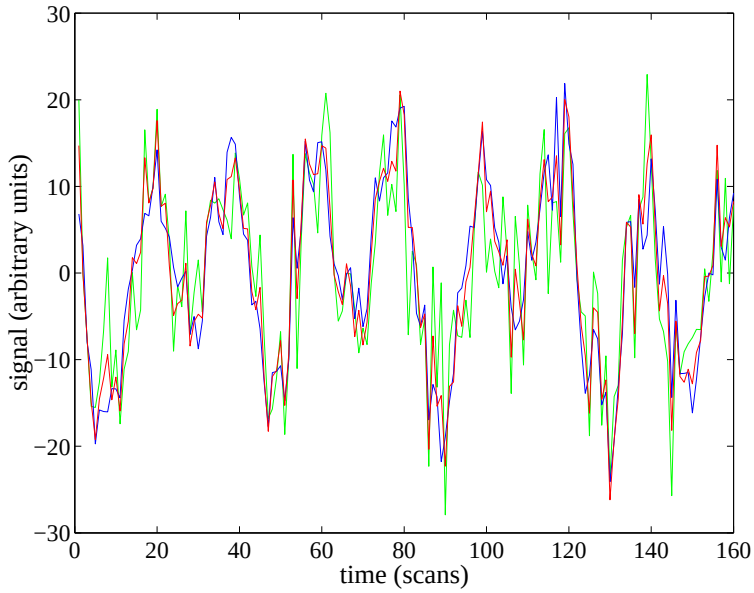


Figure 6.9: Robustness of the estimation of the state in presence of noise. The state of the system has been re-estimated on data, where half of the sessions have been replaced by gaussian white noise. The estimate of the state (red) remains close to the original one, given in figure 6.8 (displayed in blue) -which serves as a ground truth. In contrast, the average time course (in green) is quite contaminated by noise.

6.3.2 Analyzing data locally

This application is based on the same model (6.31-6.32), but uses the time courses in the neighborhood of a given voxel instead of different sessions. The underlying hypothesis is that the hemodynamic response to the stimulus varies slowly within this neighborhood. This method can be seen as a regularization procedure for the estimation of the hemodynamic response at the voxel. The potential advantage is that the method is not isotropic (each time course has a natural weight given by the mixing model). Note that a very similar setting, termed *maximum correlation modeling* has been presented in [71] and improved in [73].

The neighborhood model can be based on the 3D structure of the data, or anatomically informed distances, or parcels.

One application in retinotopy Retinotopy by fMRI [196] is potentially a good field for the application of local estimation techniques:

- It relies heavily on the spatial structure of the dataset.
- It involves a parameter estimation, which can be improved by introduction of spatial constraints.
- It yields relatively short time series, making signal extraction challenging. This is related to the fact that retinotopy is now considered as a calibration experiment for studies of cortical vision, hence should be performed quickly. $T = 72$ in the dataset studied here.

We use here the dataset described in appendix A.4. Let us recall that retinotopy through fMRI essentially amounts to the computation of the phase at each voxel time course at the stimulus frequency ω. The state dynamics are thus constrained to belong to the space defined by $(sin(\omega t), cos(\omega t))$. We apply here the state-space model with $A_1 = 0$ (no autoregressive term). An example of the time courses of 7 neighboring voxels together with the fitted sinusoid is given in figure 6.10.

From that study, one can obtain retinotopic maps of the subject for both eccentricity and polar representations of the visual field. In figure 6.11, we give the eccentricity maps obtained from the standard univariate procedure, the standard procedure applied to smoothed data, and our local estimation procedure, respectively. Next, in figure 6.12, we give the polar maps obtained from the standard univariate procedure, the standard procedure applied to smoothed data, and our local estimation procedure, respectively. In both cases, functional results have been projected on the inflated left hemisphere of the subject. Results are symmetrical on the right hemisphere.

The results of these experiments suggest that local estimations improve the quality of the resulting maps, without introducing too much bias. This can be checked by looking at reference maps for both experiments (see figure 6.13). The state-space model is a satisfactory framework for such purpose.

6.3.3 Analyzing data globally

Last, the most general use is to deal with the whole dataset. As explained previously, it can be done only after severe dimension reduction. The application starts thus with a Singular

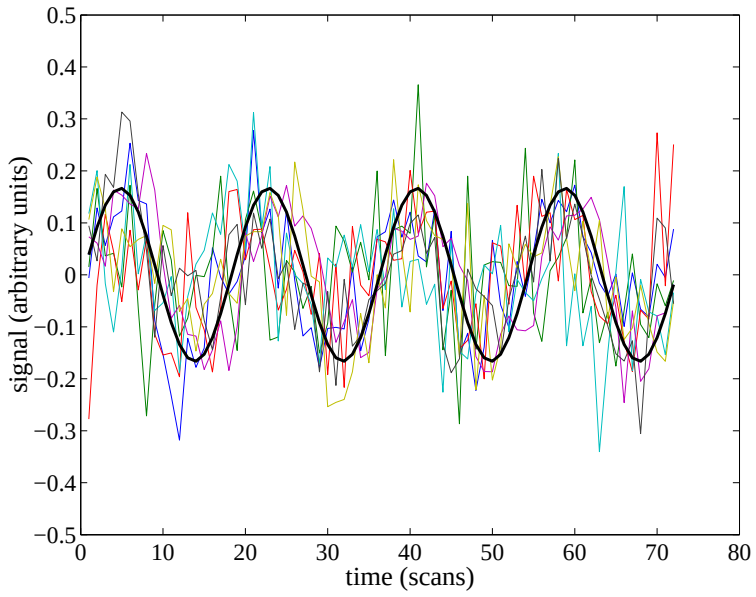
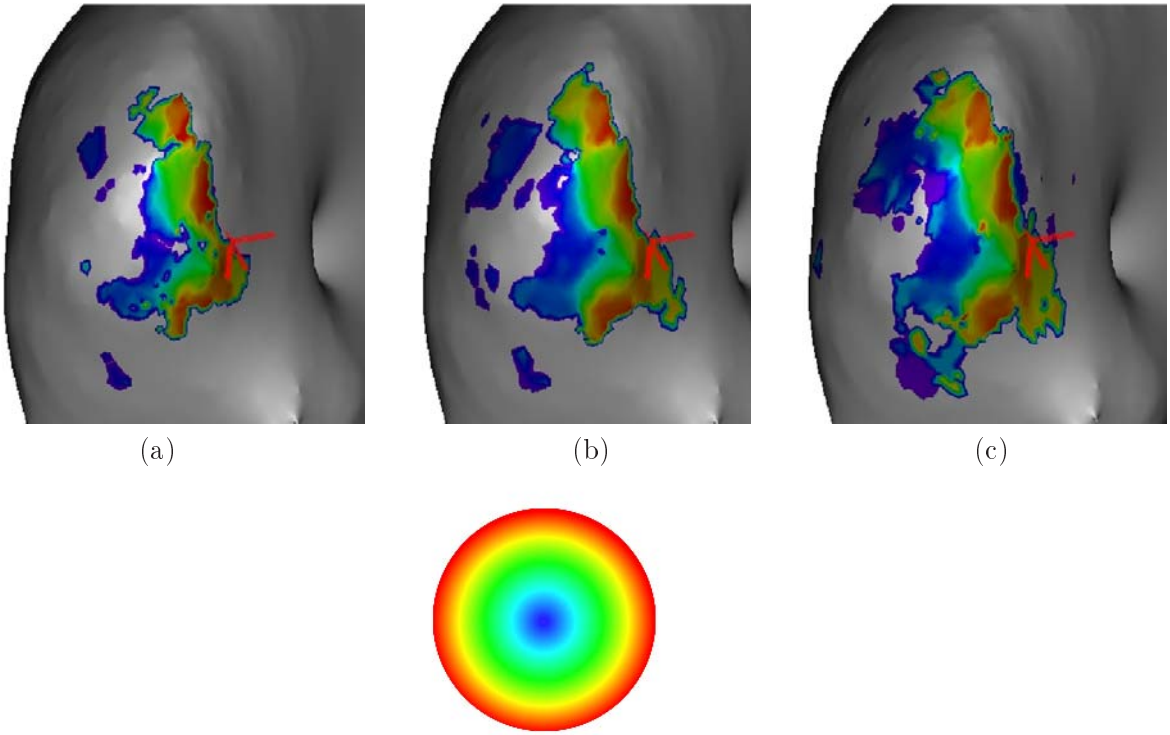


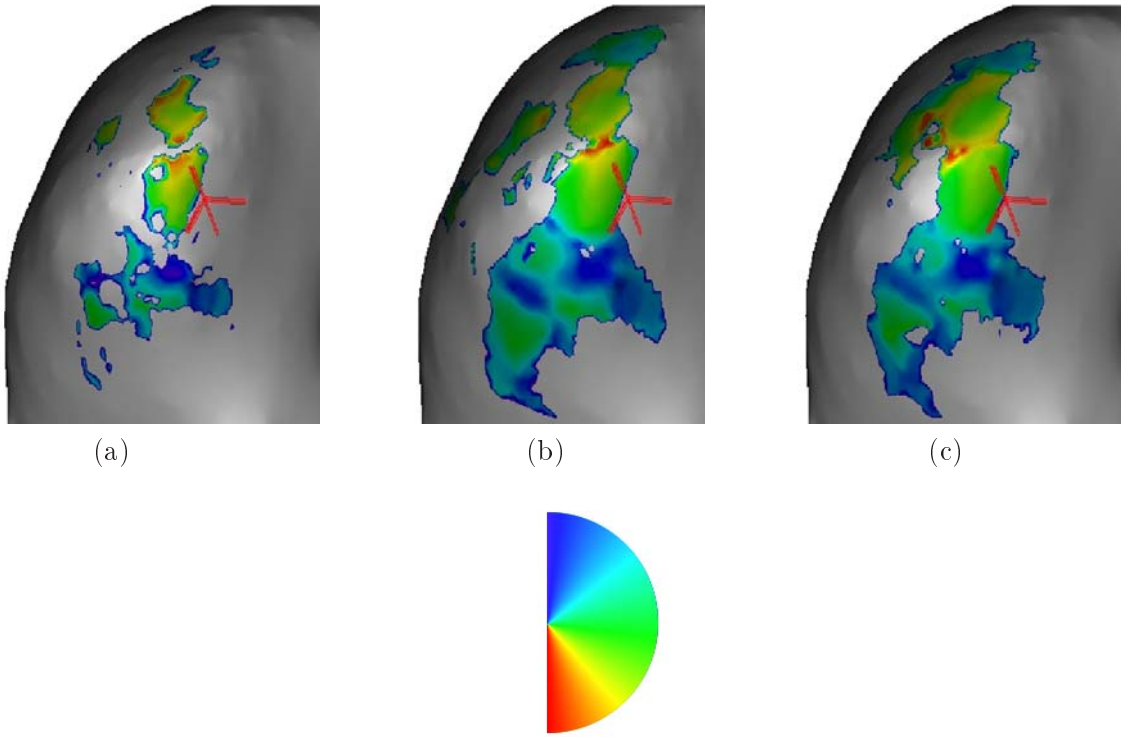
Figure 6.10: Example of state-space model for local data
The time courses of 7 neighboring voxels are represented, together with the best-fitting sinusoid obtained through the state-space procedure. The phase estimate of the sinusoid provided by the method has less dispersion than the traditional univariate estimate, provided that there is not too much discrepancy between neighboring voxels (this is a key hypothesis for retinotopy).



(a) (b) (c)

(d)

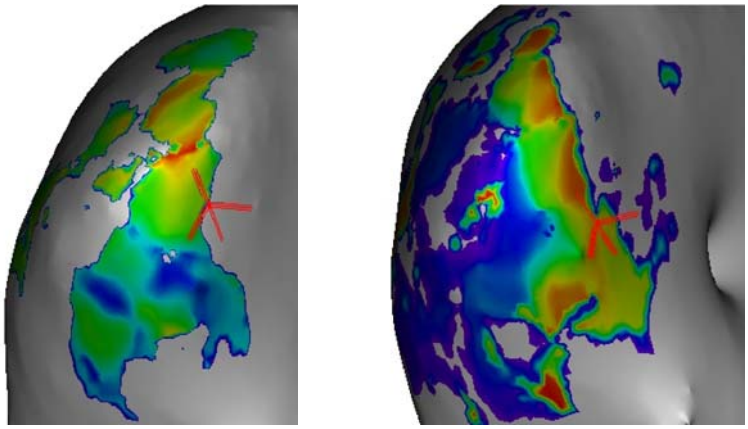
Figure 6.11: Eccentricity maps obtained from dataset A.4 with three spatial models (a) standard univariate method, (b) standard univariate method after spatial smoothing, (c) local state estimation procedure. The color code of the maps is given in (d). The local procedure yields more activated areas without blurring too much the initial map. The resulting map conforms itself to current knowledge about retinotopy and to our reference result (see figure 6.13). In theory, one expects a monotonic variation of the phase from left to right.



(a) (b) (c)

(d)

Figure 6.12: Polar maps obtained from dataset A.4 through 3 estimation procedures (a) standard univariate method, (b) standard univariate method after spatial smoothing, (c) local state estimation procedure. The color code of the maps is given in (d). The local procedure yields more activated areas without blurring too much the initial map. The resulting map conforms itself to current knowledge about retinotopy and to our reference result (see figure 6.13). Changes in the monotonicity of the variation of the polar angle are used to indicate different visual areas.



(a) (b)

Figure 6.13: Reference maps for the *wedge* (a) and *ring* (b) experiments.
These maps have been obtained on the same subject, but with a twice longer experiment. The functional data have been smoothed.

N	10	15	20	25	30	35	40	45	50	60	70
p	1	1	2	2	2	2	2	3	3	3	4
K	3	3	4	4	4	5	5	4	4	4	4

Table 6.2: Estimation of the rank of state of the real dataset.

Value Decomposition (SVD) of the data. But up to $T/2$ components can now be retained for further analysis, in contrast with the basic procedure [72]. The state-space model extracts a low dimension of the data that retains most of the dynamical (task-related, or autocorrelated) information of the data.

There remains an intrinsic difficulty in the method: the exploitation of the state model for interpretation. Indeed, there is no best choice of a particular state value and corresponding spatial maps (provided by the mixing matrix). Finding the best representation can be made by exogenous methods: temporal ICA, spatial ICA or clustering. We develop some ideas about such methods in the next chapters. However, the state-space model gives a good starting point for such advanced procedures, since it reliably estimates the state dimension.

For example, we have applied the procedure to one session of the dataset presented in A.2. The session data is first corrected for the presence of trends and then reduced to dimension N by SVD, and the algorithm is applied.

Rank estimation. The challenge in the estimation of the rank of the system is the presence of temporal correlation in the data. Our hypothesis is that a low-dimensional signal space can account for it. K has been estimated for different values $N = 10$ to 70 - given that $T = 120$. The number p of recursions in the estimation of the state, as well as the results for K are given in table 6.2. One can expect that the estimated value of K should increase with N. In fact, we rather observe that this number is stationary, indicating that

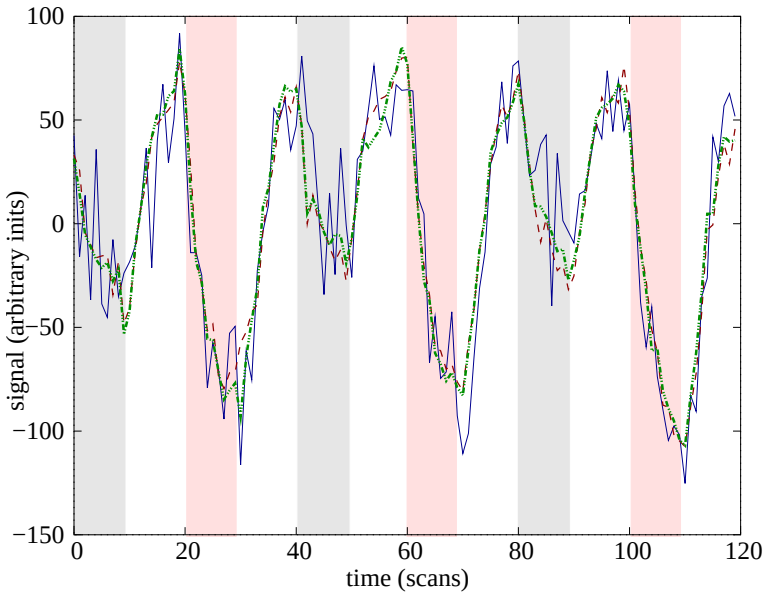


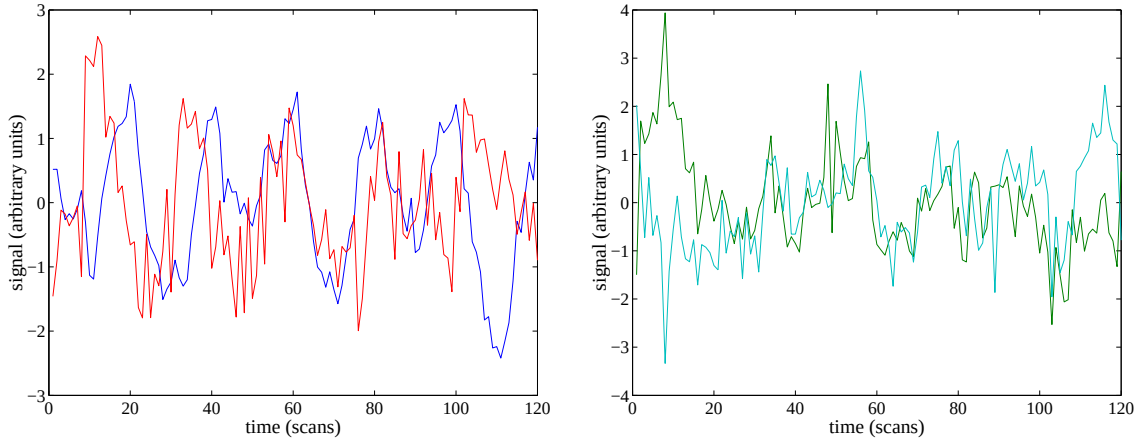
Figure 6.14: Approximation of a time course of interest (blue) by the state model (green, dashed) Reconstructed time course, after reduction to $N = 20$ components by PCA; (red, dashed-dotted) the same, after reduction to $N = 40$ components by PCA. Clearly, the main dynamical feature of this time course have been well preserved by the State-space model; the differences between both reconstructions are small.

probably most of the autocorrelated components lie in the subspace generated by the first principal components of the data. It is not clear whether $K = 4$ or 5. In the following, we choose $K = 4$.

Data fit. It is interesting to compare the state estimations for different values of N. Since there is no obvious criterion for such a comparison, we can qualitatively study how both models approximate the signal of a voxel of interest. The time courses of a voxel $X_n(t)$ of interest, as well as the reconstructed time courses (after reduction to $N = 20$ and 40) $M_n Z(t)$ are displayed in figure 6.14. We notice that both state-spaces provide a good reconstruction model for the data, in the sense that the activation pattern is well fitted.

The state-space. We present the state basis given by the algorithm in figure 6.15. There is one consistently task-related component, one which is at least transiently task-related. The two remaining ones seem to be dynamically coupled; they could be related to some biologically rhythm. All four components are predominantly low frequency.

Are there motion-related artifacts in the data? Another question of interest is the effect of body motion. To study this, we can simply make a correlation analysis between the state and motion spaces, i.e. the six rigid realignment parameters estimated by the software of Roche et al. (see [70] for related work). The projection of the motion estimation onto the state-space shows that two components can be viewed as common; they



(a) (b)

Figure 6.15: A basis of the state-space for one session of the dataset A.2, after detrending and dimension reduction.
(a) Two task-related components, the blue one being consistently task-related, and the red one only transiently task-related. (b) Two remaining components without obvious interpretation. Note that they appear to be tightly coupled -i.e. dynamically and statistically dependent. All four components are predominantly low-frequency signal.

are displayed in figure 6.16. Fortunately, they are not correlated with task-related activity.

Finally, one can notice that the space representation is a summary of the main features of the dataset, which is much easier to use thanks to its low dimension. Moreover, we have found practical solutions for the estimation of the state dimension, which solves a very difficult problem in fMRI data analysis. This also avoids blind data reduction by PCA. There remains the concern of optimal state representation, which has not been solved so far.

6.3.4 Conclusion: State-space models of fMRI data

First of all, we can notice that the state-space formalism is essentially a reformulation of existing techniques, e.g. temporal canonical correlation analysis, applied for fMRI data. The pleasant feature is that it is naturally interpreted in terms of generative process, the latter being any cause of the observed signal (task-related response, body motion, biological rhythms). The scheme that we propose for the estimation of the state-space and its dimension is moreover computationally efficient.
The problem of optimal state representation requires other approaches. Though ICA seems a natural solution, we propose in 7.3.2 an alternative based on kernel PCA.
Last, since the simple formulation that we have chosen does not allow for the precise characterization of the hemodynamic response, in chapter 9, we use state-space models for the detection and suppression of autocorrelated confounds, and use explicit (FIR) models for the hemodynamic response.

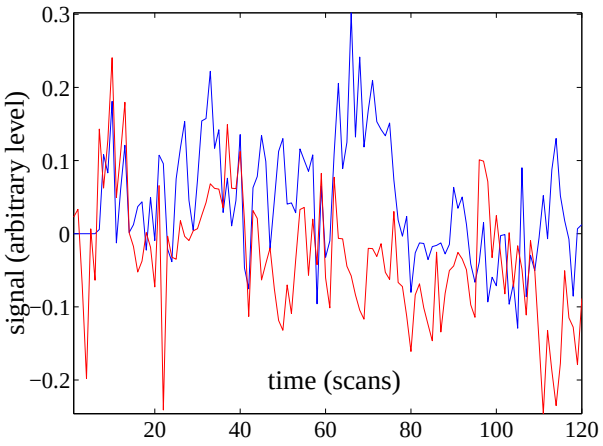


Figure 6.16: Two patterns are common to both the state-space and the realignment parameters.
They have been obtained through the CCA technique explained in 5.1.4.

Chapter 7

Kernel PCA and non-linear mixing

Our next proposition is the introduction of non-linearities in the mixing model through the method of kernels. Indeed, it has been shown in chapter 4 that decomposition methods based on linear mixing models (PCA, CCA, ICA) were not able to disentangle the structure of even simple signal spaces, such as the space represented by the example defined in A.1. Then chapter 5 has shown that we were able to isolate a low dimensional representation of the data that keeps all the -temporal- information of interest; but the problem of optimal state representation was not solved. Its is likely that a nonlinear mixing method could yield some solutions to that problem.

Several issues need to be addressed when dealing with nonlinear methods: the choice of a non-linearity model that should be meaningful; the possibility of estimating such a model from empirical data; the computational load. Our proposition [208] is the introduction of kernel PCA, which has the advantage of an explicit model for the non-linearity, which can be well controlled, and the clear definition of the solution. We first present the kernel PCA model and show on a synthetic example the unmixing ability of the method; then we develop some more technical points. We end this chapter with three examples that show how to use the methods on real datasets, in different frameworks: *1)* the study of preprocessed time courses, *2)* the derivation of an adequate overcomplete representation for the state-space model and *3)* a comparison between the Multivariate Linear Model (MLM) and kernel PCA, both being based on a standard analysis performed with the General Linear Model.

7.1 Kernel PCA

7.1.1 Nonlinear mixing, overcomplete representation and feature space

As we have mentioned, the difficulties encountered with PCA and ICA in chapter 5 can be overcome by the introduction of kernel PCA [194]. To understand this, we have to introduce the concepts of feature space and overcomplete representation, and link them with the concept of nonlinear mixing.

Feature space and nonlinear mixing Let X be our input data - an fMRI dataset. It can be seen as a collection of voxel time series $X_n(t)$, which are N items of a given space. One would like to analyze X through a generative model approach (as in 6.2-6.3)

$$X = MS + E \tag{7.1}$$

But this linear mixing model may be inadequate. For example, temporal signals $X_n(t)$ may stem from different effects in data generation (brain activity, other physiological events, data acquisition, artifacts, preprocessing, see chapter 3), so that it is awkward to consider all these effects at the same level of analysis. Rather, it may be worth considering that the true data generation involves a complex -multi-layer, high dimensional- structure, the data $X_n(t)$ being a kind of *embedding* of this high dimensional underlying process into the observation space. Of course, nothing is known about this complex generative process - but the data itself. Kernel PCA relies on the following conjecture: *Assuming that one can access the data in the feature space $\mathcal{F}$ through a certain mapping Φ, then the independent components of the process are provided by Principal Components Analysis of* $(\Phi(X_n))_{n=1,..,N} \in \mathcal{F}$. Consequently , the problem becomes the specification of

$$\kappa_{i,j} =< \Phi(X_i), \Phi(X_j) > \tag{7.2}$$

which represents the empirical covariance (or Gram) matrix of the data in $\mathcal{F}$. The interesting thing is that only κ, not Φ, is necessary for the solution of this problem. The *trick* consists in deriving κ directly from the input items (X_n), bypassing the specification of Φ:

$$\kappa_{i,j} = K(X_i, X_j) \tag{7.3}$$

Where K is the kernel that accounts for the nonlinearity associated to the non-linearity of Φ. This is fortunate, since the specification of Φ is in fact very cumbersome: $\mathcal{F}$ can be infinite dimensional, the only requirement being that it has an Hilbertian structure. The feature space is fully determined by the definition of a metric, or equivalently, by the definition of a generalized covariance $K(X_i, X_j)$ for any $(i, j) \in [1, N] \times [1, N]$. Then the decomposition consists in diagonalizing the covariance matrix $\kappa(i, j) = K(X_i, X_j)$ for $i, j \in [1, N] \times [1, N]$. The eigenvectors w of κ will represent the spatial modes of coherent dynamical activity in the input data X.

Overcomplete representations and nonlinear mixing More realistically, one can simply consider the following argument: the usual bilinear covariance structure in $\mathbb{R}^T$ yields a Gram matrix $\kappa = (< X_i, X_j >)_{(i,j)=1..N}$ of rank T. Instead, the matrix (7.3) has a higher rank, and possibly full rank N (in which case there are as many dimensions as data items, so that the dimension of $\mathcal{F}$ is potentially infinite). Letting $(w_i)_{i=1..N}$ be a basis of $\mathbb{R}^N$ that diagonalizes κ, it is clear that this generative family is overcomplete once considered in the original space $\mathbb{R}^T$.

Finally, Kernel PCA is then nothing but a way to produce an overcomplete representation of the signal. This may be disadvantageous for the sake of sparsity in data representation, but it has been shown in some works [138] that it has better representation properties when the data is not gaussian.

Choosing a kernel The important question now is: what choice for K will produce meaningful eigenvectors? We start with the idea that distinct time courses are the realization of different underlying dynamical phenomena, and thus should be distinguished by the kernel (this is not actually the case with PCA). Let $X_i(t)$ and $X_j(t)$ be two time courses from the dataset. First, let us notice that it will be useful to replace them by models (i.e. after noise removal), denoted by z_i and z_j respectively; these models can be derived with the techniques presented in chapters 4 and 6. The classical covariance, noted K_∞ -$K_\infty(z_i, z_j) =< z_i, z_j >$- can be written as $K_\infty(z_i, z_j) = cc(z_i, z_j)|z_i||z_j|$, where $cc(z_i, z_j) \in [-1, 1]$ is the correlation between the time courses. $cc(z_i, z_j) = 1$ means that the time courses are identical, up to a positive factor, $cc(z_i, z_j) = -1$ that they are opposite, up to a positive factor, and $cc(z_i, z_j) = 0$ that they are uncorrelated. In fact, correlation can be thought of in terms of shared information between the time courses. We propose to introduce the non-linearity by penalizing the values of cc that are far from 1. To do that, we multiply the usual covariance matrix $K_\infty(z_i, z_j)$ by a function $\phi(cc(z_i, z_j))$ so that $\phi(1) = 1$ and $\phi(cc)$ decreases to 0 rapidly when cc gets smaller. The fact that $K(z_i, z_j) = K_\infty(z_i, z_j).\phi(cc(z_i, z_j))$ decreases rapidly when the correlation decreases, means that the time courses will be viewed as orthogonal features as soon as their correlation will be under a given threshold, say $1 - 3\sigma$. This is justified if one thinks that these time courses are the realization of different underlying dynamical phenomena. Here we choose

$$\phi_\sigma(cc(z_i, z_j)) = e^{\frac{cc(z_i,z_j)-1}{\sigma}}, \tag{7.4}$$

yielding the kernel-covariance

$$\kappa_{i,j} = K_\sigma(X_i, X_j) = K_\infty(z_i, z_j)\phi_\sigma(cc(z_i, z_j)) =< z_i, z_j > e^{\frac{cc(z_i,z_j)-1}{\sigma}} \tag{7.5}$$

Let us notice that the parameter σ controls the amount of nonlinearity in the kernel: $\sigma \longrightarrow \infty$ corresponds to the classical PCA ($K = K_\infty$), while $\sigma \longrightarrow 0$ means that any two distinct -up to a positive factor- time courses are *orthogonal*. A particularity of our choice is that opposite time courses ($cc = -1$) are treated as orthogonal, hence distinct phenomena. This is related to the hypothesis that an effect present in the dataset should not appear as *positive* for some voxels and *negative* for some others (which makes sense if one avoids some corrections as global scaling, that induce the artefactual *opposite* signals).

Choice of σ, and sensitivity with respect to this parameter The choice for the optimal value of σ should be guided by the interpretation of this parameter. Given our model (7.4), the role of σ is to *cancel* the correlation values below $1 - 3\sigma$. Consequently, a smaller value for σ yields more scattering of the data variance into the different components (see figure 7.1), which does not mean that more components are actually of great interest.

Two strategies are possible: first, to compute and diagonalize the kernel with different values of σ (given that in practice the useful values lie in the range of $[0.02, 0.2]$); second to decide a pertinent value by plotting the histogram of the empirical correlation between all pairs of fitted signals across the dataset (see section 7.3.1).

Running the method Kernel PCA of the data is then performed by the following steps:

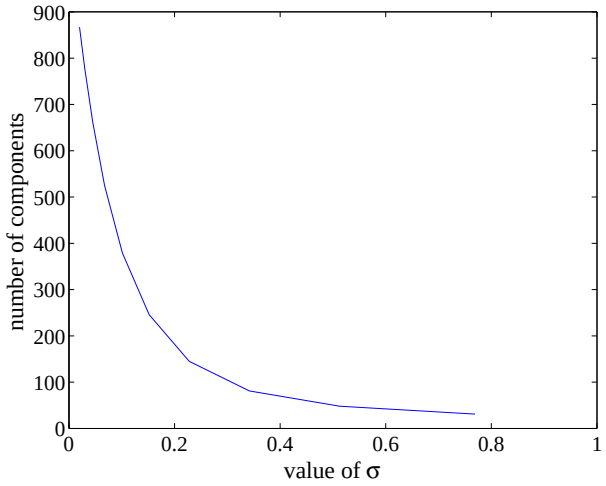


Figure 7.1: Number of components that are necessary to account for 95% of the total kernel-covariance as a function of σ.
This is taken from the data presented in section A.2. For comparison, the number for a PCA decomposition would be only 6 components. Note that this is not necessarily the number of components of interest in the dataset.

1. Derivation of the kernel-covariance matrix κ by (7.5).
2. Diagonalization of κ. this yields eigenvectors $w_1, .., w_N$ sorted by decreasing energy,
3. Selection of the first I components, and study of the spatial maps $(w_i)_{i=1,..,I}$.

7.1.2 On kernels

An important concern is to ensure that $\mathcal{F}$ is indeed a Hilbert space, that is that the kernel-Gram matrix κ is a covariance matrix, i.e. a positive definite matrix. This implies some conditions on the kernel K. Of course, the kernel should be symmetric $K(x, y) = K(y, x)$ and satisfy Cauchy-Schwartz inequality

$$K(x, y)^2 \leq K(x, x)K(y, y). \tag{7.6}$$

The additional conditions that defines admissible kernels is precisely the positivity condition: if $x^1, .., x^N$ are a dataset, then for any set $u_1, .., u_N$ of real numbers,

$$\sum_{i,j=1}^{N} u_i u_j K(x^i, x^j) \geq 0 \tag{7.7}$$

Consequently, the set of admissible kernels is stable by some operations (see [91] for example):

- Convexity: letting K_1 and K_2 be two admissible kernels, and α and β two positive numbers then
  $K = \alpha K_1 + \beta K_2$ is also admissible.

- Multiplication: with the same notations, $K = K_1 K_2$ is also an admissible kernel.
- Consequently, for any real polynomial with positive coefficients Π, $\Pi(K)$ is admissible if K is admissible.
- Consequently, the exponential of an admissible kernel is admissible.
- Moreover, for any real-valued function f, $K(x, y) = f(x)f(y)$ is a valid kernel, and so is $K_1(f(x), f(y))$, if K is an admissible kernel.
- Of course, any positive symmetric matrix A yields an admissible kernel by $K(x, y) = x^T Ay$.

The kernel given in (7.5) is thus valid, since it is the product of two valid kernels. Though we will not use it, let us mention the spectral representation theorem from [24] that is useful to decide for the admissibility of some kernels:

A stationary, i.e. translation invariant kernel $K(x, y) = K(x - y)$ is admissible if and only if

$$K(x - y) = \int_{\mathbb{R}^T} cos(\omega^T (x - y)) F(d\omega) \tag{7.8}$$

where F is a positive measure.

The nice thing is that (7.8) is nothing but the Fourier transform of F, so that a stationary kernel is admissible if and only if it has a positive Fourier transform. This is in particular the case for gaussian kernels

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right) \tag{7.9}$$

7.1.3 From theory to empirical data

Before going into more issues on kernel PCA, we propose now to give an illustration of the power of the method to disentangle some activation patterns. Once again, we use the synthetic example dataset described in A.1. To show the interest and the limits of the method, we compare the data decompositions obtained by processing with four different methods. The differences between the four methods lie in the definition of the corresponding kernels: in particular, for the data item $X_i(t)$, we introduce a fitted model $z_i(t)$ which is the time course after projection on a subspace of interest. z_i represents then a temporally pre-processed version of X_i.

The parameter σ is set to 0.1 for both K_3 and K_4.

Results To select the number of final components, we plot the amount of variance contained in each component, for all techniques, and select the number of components as the inflection points in the decreasing curve (see figure 7.2). Both kernels K_1 and K_2 yield two main components, while the kernel K_4 yields three main components. For K_3, it is not clear whether a particular subspace can be defined. In this case we display three components by analogy with K_4. We also checked that our selection procedure did not leave apart any structured spatio-temporal component for either method. The

method	temporal processing	procedure	corresponding kernel
1	none	PCA	$K_1(X_i, X_j) =< X_i, X_j >$
2	fitted time course	PCA	$K_2(X_i, X_j) =< z_i, z_j >$
3	none	KPCA	$K_3(X_i, X_j) =< X_i, X_j > e^{\frac{1}{\sigma}\left(\frac{<X_i, X_j>}{\|X_i\|\|X_j\|}-1\right)}$
4	fitted time course	KPCA	$K_4(X_i, X_j) =< z_i, z_j > e^{\frac{1}{\sigma}\left(\frac{<z_i, z_j>}{\|z_i\|\|z_j\|}-1\right)}$

Table 7.1: Four different kernels used for comparison on synthetic data.

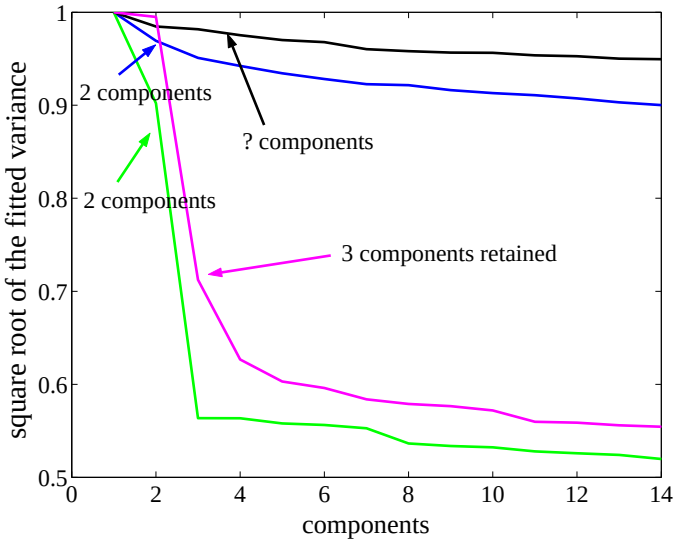


Figure 7.2: Relative amount of fitted variance of the first components.
The square root of the variance associated with each component is plotted after normalization with respect to the first component. The four curves represent the components obtained after the covariance estimate from K_1 (blue), K_2 (green), K_3 (black) and K_4 (magenta).

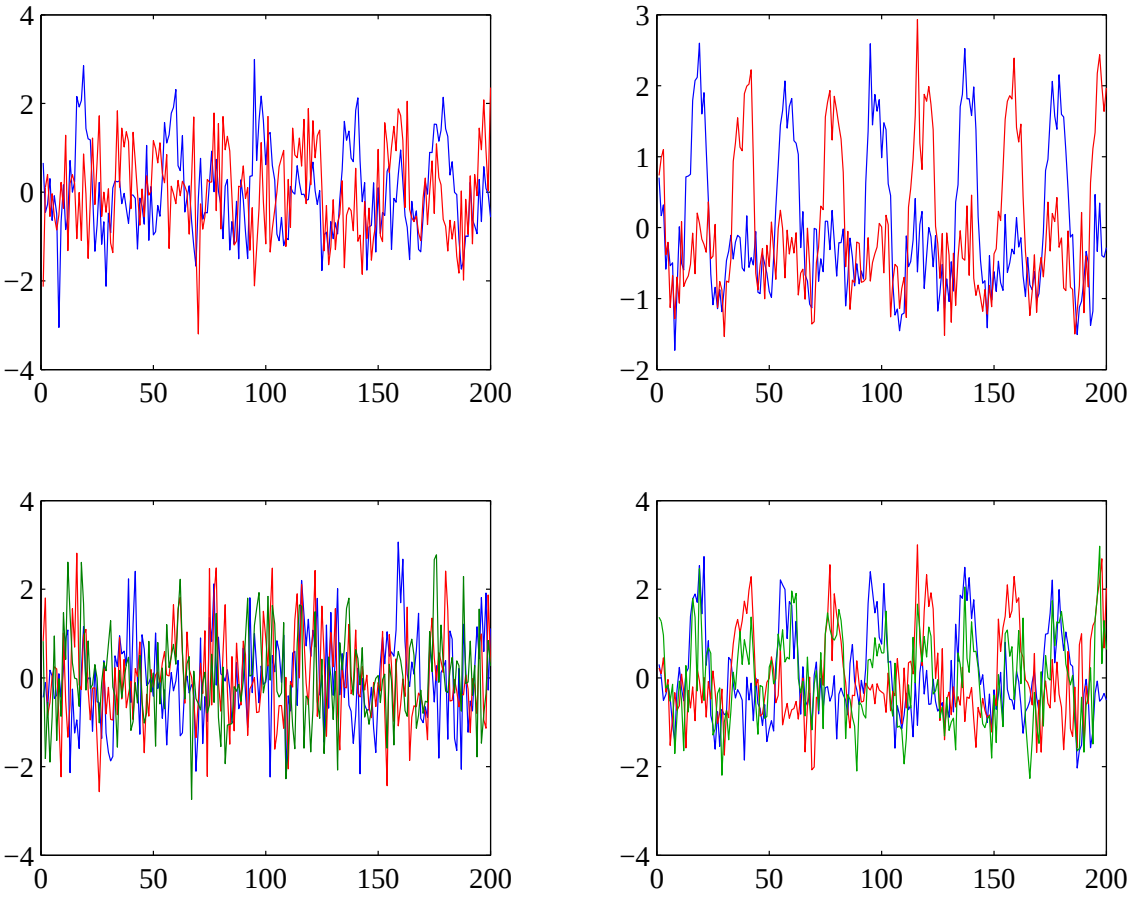


Figure 7.3: Main time courses obtained by the decomposition of the dataset using kernels K_1 (up left), K_2 (up right), K_3 (bottom left) and K_4 (bottom right).

resulting temporal patterns are displayed in figure 7.3, and have to be compared with those presented in figure A.1.

The corresponding maps, overlayed on the "ground truth" from figure A.1, are displayed in figure 7.4.

Comparison and interpretation. First, it is noticeable that kernel K_3, which performs the kernel PCA without prior analysis, gives the worst description of the dataset in terms of time courses. The reason is clear from figure 7.4: the spatial maps are reduced to one voxel. Clearly, the kernel K_3 overfits the data. This indicates that under low signal to noise ratio, the use of kernel PCA without a related modeling of the data will perform poorly.

The comparison of the time courses provided by kernels K_1 and K_2 shows that the introduction of a temporal model of the data prior to PCA greatly reduces the noise embedded within the components. But both methods perform poorly in unmixing the spatial components, as can be seen in figure 7.4.

Last, kernel K_4 is the only method that clearly indicates that the signal distribution has three main modes, and identifies the corresponding spatio-temporal components correctly. A reason is that the three "modes", represented by the three time curves of figure A.1, lie within a two dimensional space, which makes their identification problematic for a linear method, as PCA. In fact, PCA is optimal for gaussian distributed data, an hypothesis which is not met here.

7.2 Making it work in practice

In this section, we develop some technical details about the kernel PCA method: Computational problems related to the size of real datasets, the question of data centering and -last but not least- of dimension selection.

7.2.1 Size of the system

The main problem with kernel PCA is the amount of computation in the diagonalization of the kernel-covariance matrix (this is nothing but the counterpart of the additional ability of the method to detect more subtle spatial modes of activity): this comes from the fact that the $N \times N$ matrix κ has to be diagonalized, N being the number of voxels (ranging from 10^4 to 10^5 in practice !). This problem is the main bottleneck for the practical use of the method, since the Singular value Decomposition of a $N \times N$ matrix has complexity $O(N^3)$; on a standard PC, it currently takes around one hour for $N = 4000$. We successively explore two solutions to overcome the difficulty: *i)* The reduction of the dataset to reach a reasonable N value, and *ii)* The approximation of the first I components by an EM method.

Data reduction The first way to deal with the problem is to arbitrarily reduce the number N of voxels in the kernel estimation. The solution may seem crude, but it offers several advantages:

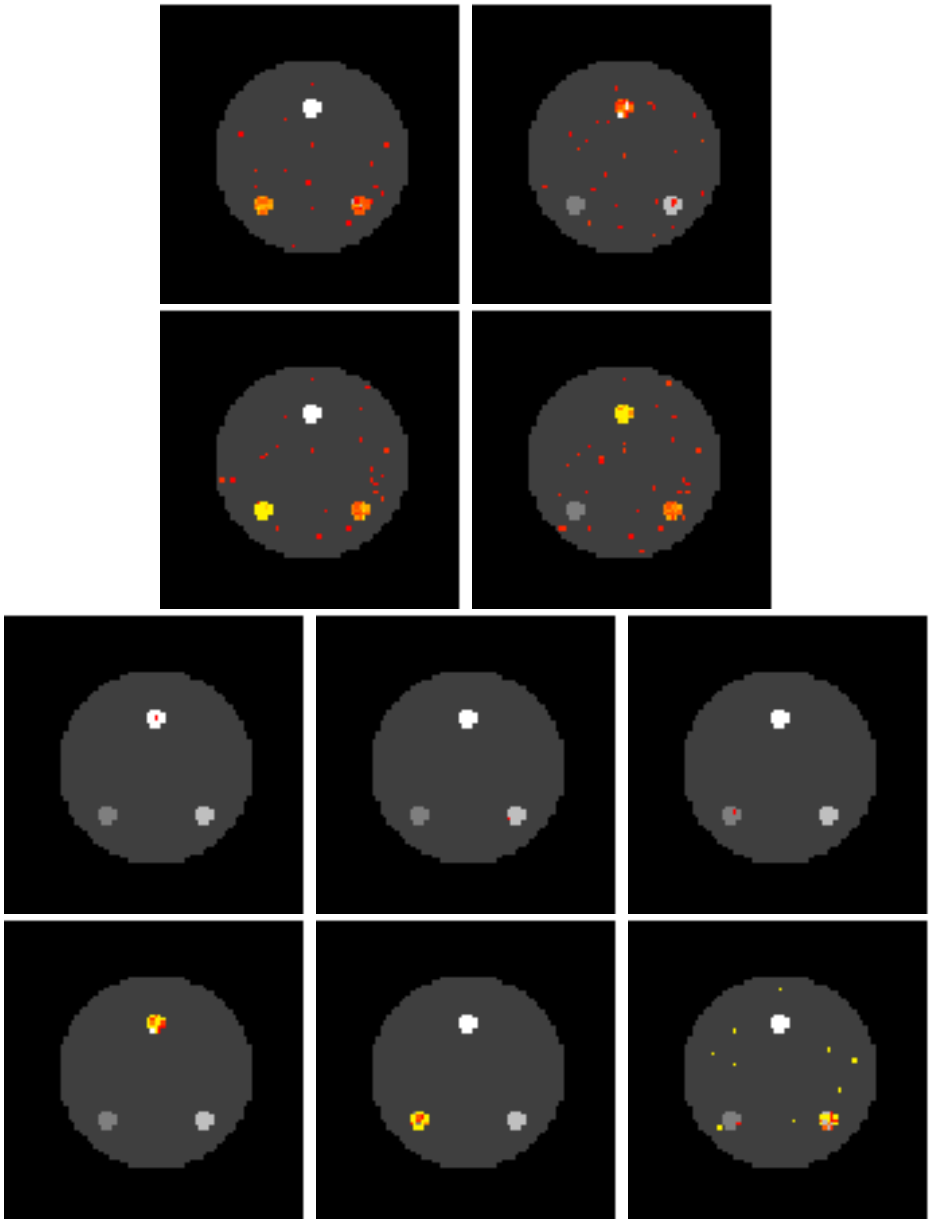


Figure 7.4: Spatial maps that define the main components of the decomposition for the different kernels.
These maps are thresholded at a significance value $P = 0.05$, using a gaussian random variable approximation. The first row contains the two maps obtained with K_1, the second row the two maps obtained with K_2, the third row, the first three maps of K_3, the fourth row, the three maps obtained with K_4.

- It does not only reduce the computation time, but also the storage cost for the matrix κ.

- The particular kernel that we have chosen in equation (7.5) is not very sensitive for weakly informative voxels: since $\kappa_{i,j} = |z_i||z_j|cc(z_i, z_j)e^{\frac{cc(z_i,z_j)-1}{\sigma}}$, $\kappa_{i,j}$ is small $\forall j$ if $|z_i|$ is. Thus eliminating the i^{th} row and column of κ does not change significantly the result of the PCA (this is the case for classical and kernel PCA). Consequently, eliminating voxels that carry little effect of interest does bias neither the eigenvalues of the PCA neither the associated spatio-temporal modes.

- The eliminated voxels can be *reintegrated* in the images w_k by the following formula:

$$w_k(i) = \frac{1}{\lambda(k)} \sum_{n=1}^{N} K(X_n, X_i) w_k(n) \tag{7.10}$$

  where $(X_n), n = 1, .., N$ is the reduced training set, w_k is the k^{th} spatial map, $\lambda(k)$ the associated eigenvalue of K and i a voxel that has not been included in $(1, .., N)$.

Let us mention that there could be some different ways to undersample more properly the data; for example, replacing voxels by regions of interest that have greater size reduces N without losing too much information ([66], [127]).

Using an approximation of the solution Recently, a method has been proposed in [159] for the approximation of the first components in a kernel PCA approach. The corresponding subspace $W_I = (w_1, .., w_I)$ is obtained in an Expectation-Maximization (EM) fashion by the following iterations:

$$\Psi = (W_I^T W_I)^{-1} W_I \kappa \tag{7.11}$$

$$W_I = \kappa \Psi (\Psi \Psi^T)^{-1} \tag{7.12}$$

whose complexity is no more than $O(IN^2)$ for each iteration. This simple procedure converges quickly towards the first eigencomponents of κ. Additionally, a whitening of W at each step greatly reduces numerical instabilities. This procedure can reduce the computation time by a factor 4 with respect to the SVD.

7.2.2 Centering the data

So far, it has been assumed implicitly that equation (7.5) represents the covariance of the data after applying the Φ function. In fact this is true only if $\mathbb{E}(\Phi(X)) = 0$. This is not necessarily satisfied, especially with kernels like (7.9). Consequently, a centering procedure has been introduced in [194] to correct for non-centering of the data; using the following approximation,

$$\sum_{n=1}^{N} \Phi(X_n) = 0, \tag{7.13}$$

such that the matrix κ is then slightly modified: Let $\mathbb{I}$ be the $N \times N$ matrix such that $\mathbb{I}_{i,j} = 1 \; \forall (i,j)$. We define the centered matrix $\tilde{\kappa} = (\tilde{\kappa}_{i,j})$ by

$$\begin{aligned} \tilde{\kappa}_{i,j} &= < \Phi(X_i) - \frac{1}{N}\sum_{n=1}^{N} \Phi(X_n), \Phi(X_j) - \frac{1}{N}\sum_{n=1}^{N} \Phi(X_n) > & (7.14) \\ &= \kappa_{i,j} - \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}_{i,n}\kappa_{n,j} - \frac{1}{N}\sum_{n=1}^{N} \kappa_{i,n}\mathbb{I}_{n,j} + \frac{1}{N^2}\sum_{n,m=1}^{N} \mathbb{I}_{i,m}\kappa_{m,n}\mathbb{I}_{n,j} & (7.15) \\ &= (\kappa - \frac{\mathbb{I}\kappa + \kappa\mathbb{I}}{N} + \frac{\mathbb{I}\kappa\mathbb{I}}{N^2})_{i,j} & (7.16) \end{aligned}$$

We have not applied this procedure with kernel (7.5).

7.2.3 Choosing the dimension

Choosing the optimal rank (the number of components kept in the final description of the dataset) for a data decomposition is a difficult problem. For PCA decompositions in neuroimaging, two methods have been proposed (see also section 5.1.2):

- The use of a sphericity criterion to control that the eigenvalues rejected into the null space can be considered as equivalent (the noise dispersion being approximately isotropic in the signal space) [224].

- The introduction of a *generalization error* criterion that includes a data fit and a rank penalty term, based on asymptotic approximations [103]. It is reported there that the analytical rank estimate obtained from the training dataset is too optimistic, and that data splittings are better suited for rank estimation than the analytical method.

However, these methods do not match the hypotheses introduced with our kernel PCA. For example, in [103], the authors assume that the *signal space* is gaussian, as well as the *noise space*. Here, we propose a different procedure: we compute the goodness of fit of the decomposition at rank r by the following formula:

$$G(r) = \sum_{n=1}^{N} \frac{\sum_{t=1}^{T}(X_n(t) - \sum_{k=1}^{r} I_k^n s_k(t))^2}{2T\|\varepsilon^n\|^2} + \frac{r\log(N)}{2} \qquad (7.17)$$

where $\|\varepsilon^n\|^2$ is an estimate of the variance of the residual noise found at voxel (n), e.g. the stochastic variance defined in section 4.2.2. If one assumes that the residuals are independent among voxels, and that the estimate of the noise variance is unbiased, this criterion is known as the Bayesian Information Criterion (BIC), and is in fact analogous to the MDL criterion used for the temporal model. The minimum of G with respect to r gives the optimal number of components.

Let us also mention the original procedure proposed in [107] which is based on the iteration of kernel PCA and ICA.

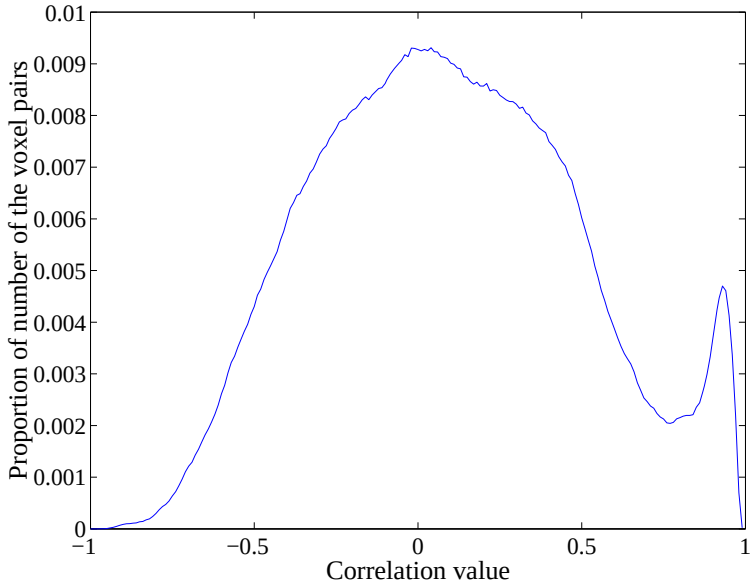


Figure 7.5: Histogram of the empirical correlations obtained from the dataset, after pre-processing of the time course
$cc(i,j) = \frac{<z_i,z_j>}{|z_i||z_j|}, 1 \leq i < j \leq N$. The width of the histogram bins is 0.01, and the population in each bin has been divided by the total number of pairs, i.e. $\frac{N(N-1)}{2}$.

7.3 Results with real data and discussion

7.3.1 Use in combination with a univariate model

We have applied the kernel PCA to a dataset which had been pre-processed. We have used the dataset described in A.2.

Materials and results

The dataset is first reduced by selection of the most relevant voxels. The criterion is the complexity criterion (4.39). This yields $N = 4079$ instead of 12320 [1]. Then a dynamic model z_i is derived for each voxel, by using the multi-session analysis method described in 6.3.1. The histogram of empirically obtained correlations is given in figure 7.5.

This histogram contains essentially a wide mode centered at 0, and a narrow peak around 0.95. This narrow peak contains the activated voxels, which are mutually positively correlated. Kernel PCA can then be used *i)* to separate the voxels contributing to this mode from the other voxels, and *ii)* to see whether there are separable models within this mode (one can notice that very few pairs have correlation equal to 1). A generalized covariance is then derived through the kernel (7.5), with $\sigma = 0.03$. After computation of the first components of κ, the rank is estimated with the function (7.17). This yields

[1] Interestingly, using $N = 2898$ yielded exactly the same results as $N = 4079$; this indicates that the results are not very sensitive to the addition of weakly informative voxel time courses.

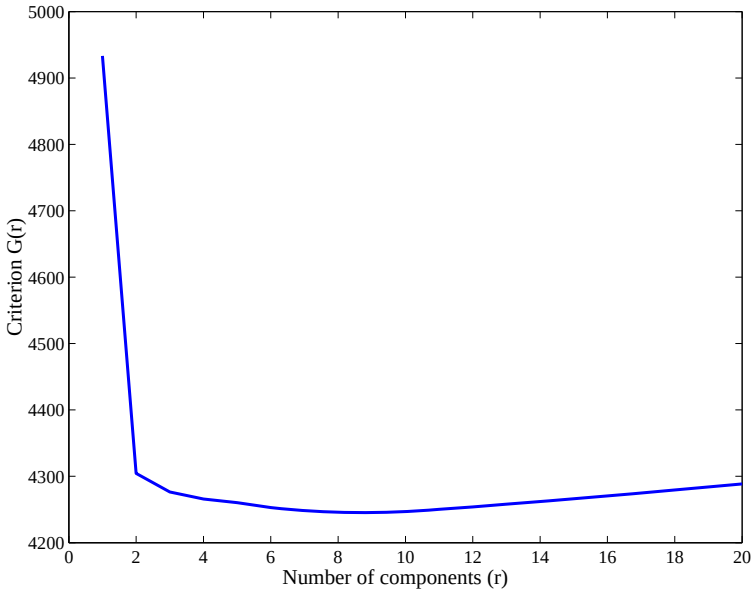


Figure 7.6: Selection of the number of components with the criterion $G(r)$ (equation (7.17)).
The minimum of the functional is reached for $r = 9$.

$r = 9$ components for the data, as can be seen in figure 7.6. We use this number for practical purposes, acknowledging the fact that there is no clear-cut indication of a true dimensionality.

The associated temporal patterns are then derived by least squares solution:

$$s_k(t) = \sum_{n=1}^{N} w_k(n) X_n(t) \tag{7.18}$$

The 9 patterns associated with the first components are presented in figure 7.7. For the sake of readability, they have been divided into three groups, organized according to their similarity.

To simplify the presentation of the associated spatial patterns, we present three correlation maps associated with the average of the three groups presented in figure 7.7. The maps, shown in figure 7.8, are derived by correlation of the voxel time course with the pattern time course, and thresholded at the level $P = 0.05$, Bonferroni-corrected for multiple comparisons.

Interpretation and discussion

On the spatial maps. In figure 7.8 we have plotted the color (blue, red, green) coded maps associated with the three main groups of patterns, together with a map of the *motion-static* contrast obtained from SPM analysis. There is a good correspondence between the positive activation areas and the green- and blue-labeled voxels; but also between the red-labeled and the areas that respond negatively to the t-test. It is noticeable that the labeled

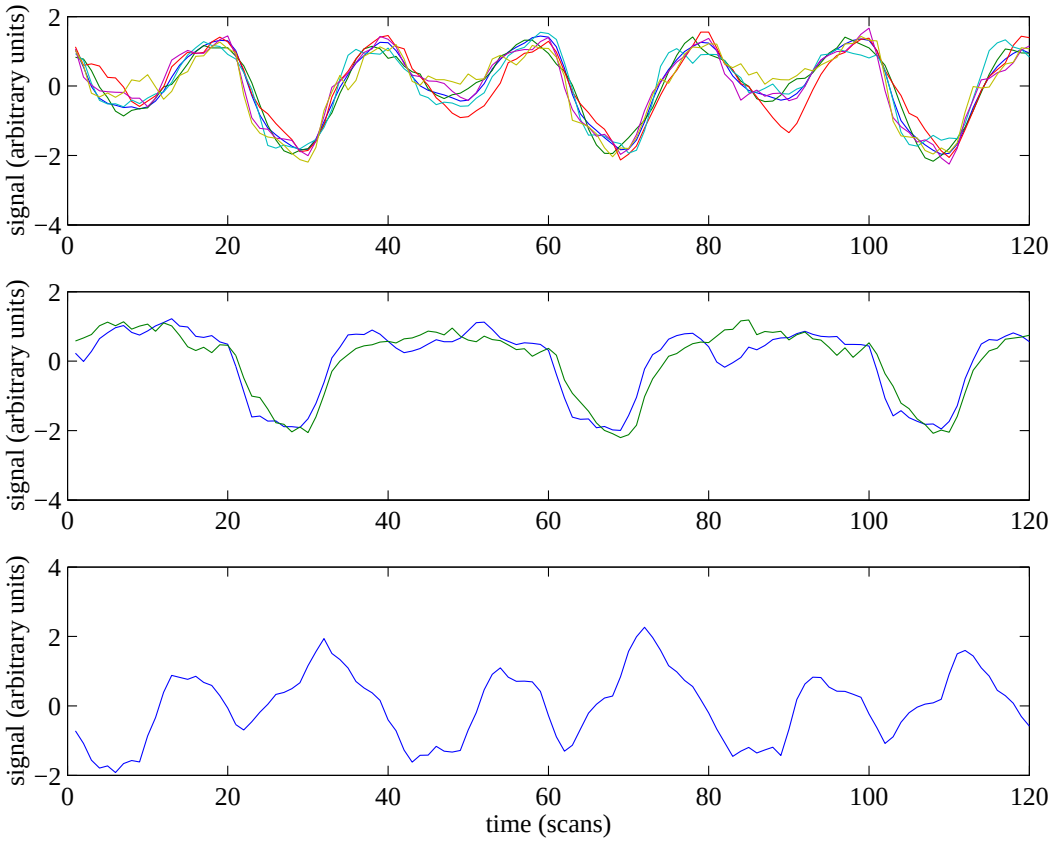


Figure 7.7: Main temporal patterns associated with the decomposition.
They have been sorted into three groups according to their similarities; the first group comprises six components, which present an activation for both experimental conditions, in a relatively homogeneous way. The differences among them concern only secondary features, e.g. the response shape or the amount of noise within the time course. The second group comprises two components, which are only driven by the *motion* condition (time onset: 20,60,100). The difference between them (timing or shape) is less interesting to interpret. The last group comprises only one component. It is noticeably anti-correlated with the first group, but the task-related pattern is less consistent than what can be seen within the first group.

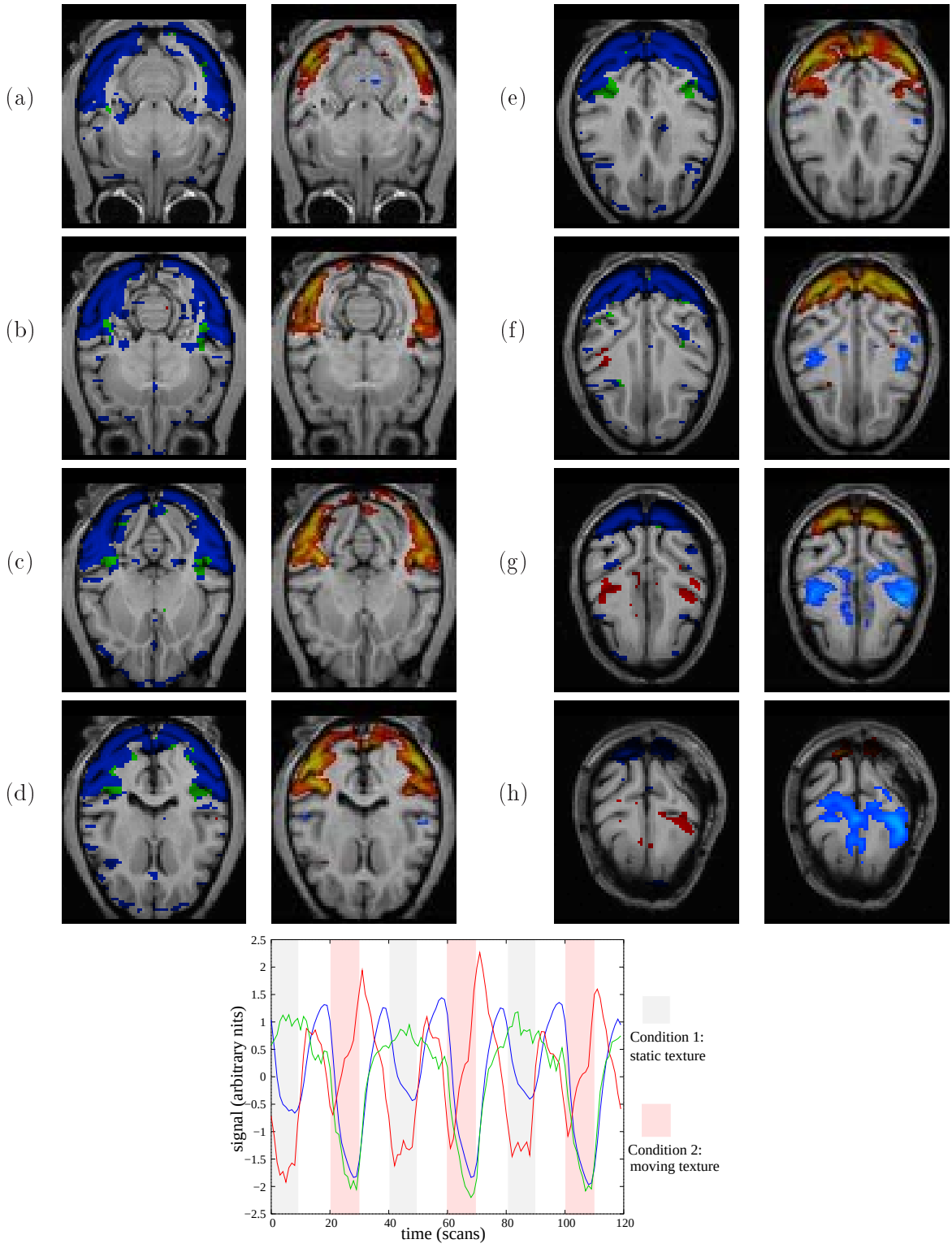


Figure 7.8: Results of kernel PCA as applied to dataset A.2
(top) Eight axial slices (a-h) of the label map associated with the first three components. The slices are presented in two main columns, each one showing our labels (left), and the corresponding SPM map of the (M-S) *moving-static* contrast (right). Both spatial maps are thresholded at a significance value of $P = 0.05$, corrected for multiple comparisons. Bottom: temporal patterns associated with the labels. The colors of the label map are the same as those of the temporal patterns.

areas are wider than the thresholded t-maps; this is understandable, given the fact that the conjunction of three models fitted to the data is likely to uncover more information than a single contrast. However, KPCA makes more precise some information provided in the SPM t-map:

- First, the distinction between the blue- and green-labeled areas: the time curve associated with the spatial maps recalls that the *green* activation pattern is uniquely driven by the *motion condition*, while the blue pattern was driven by both conditions, though much more by the motion condition. Our map shows a distinction between the blue-labeled primary visual areas (V1, parts of V2, V3, V4 and TE), and the green-labeled areas (MST, MT/V5). This confirms the selectivity of MT/V5 to motion stimulation, but also the observation that primary visual areas respond positively to the SPM contrast *motion-static* (in [218], it is reported that some layers of V1, V3 and some stripes of V2 contain a non-negligible proportion of direction-selective neurons). Though compatible with the SPM analysis, the label map gives thus a finer description of the responses of the visual system.

- Second, the areas that exhibit consistently negative responses to the SPM t-test yield difficulties in the interpretation of the experimental results. Clearly, these areas, that include the cerebellum and the parietal cortex, are not part of the visual system. If we identify them with the red-labeled areas, we see that the associated temporal pattern shows indeed a clear *static-motion* effect, which is nevertheless combined with a kind of trend. Consequently, this pattern is not consistently task-related, and thus could be the result of the combination of different effects (motion of the animal correlated to the task, biased motion estimate/correction, or even the competition between neural/hemodynamic events). We have noticed that the use of the standard SPM motion correction routine yielded systematically more consistent and wider negative areas (data not shown).

On the temporal patterns: Figure 7.7 illustrates the ability of kernel PCA to derive over-complete representations of the data. In particular, no linear decomposition method could yield such results, since all the components are (positively or negatively) correlated. One might ask whether such redundancy in the decomposition is really interesting; in particular, the first group of 6 time courses has been advantageously replaced by one model for the study of the spatial maps. However, this is because we mostly concentrate on the *static-motion* contrast. In a different context, or with different priors, finer description of the signal space may be interesting. Moreover, the distinction between group 1 and group 2 (or equivalently, between blue- and green-labeled areas in figure 7.8) is fundamental, though it cannot be achieved by usual methods (including spatial ICA): it can be obtained only by introducing temporal-preprocessing, and, in the case of KPCA, a narrow ($\sigma = 0.03$) kernel.

Finally, this study mainly outlines the potential of kernel PCA as an exploratory tool (even if it uses temporal pre-processing), rather than an inferential tool (it does neither produce a true generative model of the data nor statistically validated maps as SPM).

7.3.2 Use of kernel PCA in combination with the state-space model

Now we present how kernel PCA can be used in order to solve a problem encountered with state-space model (see chapter 6) : the problem of state representation.

A kernel for state-space models Let us assume that a state-space representation has been derived for a dataset Y, as given by equations (6.2-6.3). The problem is to find a meaningful representation of the dataset, which shows how the main dynamical features of the dataset are organized spatially. As noticed earlier, the state-space model alone does not solve the issue; we propose here to derive an overcomplete representation of the state X and the mixing matrix M of the model.

Within the state-space framework, each voxel n has a natural feature representation $f(n) = (MX)_n$. One can thus use an admissible kernel to derive $\kappa_{i,j} = K(f(i), f(j))$. But, in order to keep the advantage of the low dimensional representation of the dataset provided by the state-space, and also to show the generality of kernel method, we will use a different setting than previously. We choose the kernel

$$\kappa_{i,j} = K(f(i), f(j)) =< f_i, f_j >^3 \tag{7.19}$$

The advantage of the polynomial kernel is that the feature space remains finite dimensional, which allows for an explicit computation of the mapping Φ and avoids in fact the computation of κ. Besides, the cubic polynomial preserves the sign of the scalar product, avoiding a centering procedure. Let Q be the dimension of the state-space; the dimension of the feature space, i.e. the rank of κ, is here $d = \frac{Q(Q+1)(Q+2)}{6}$. Indeed,

$$\begin{aligned}
< f_i, f_j >^3 &= \left(\sum_{q=1}^{Q} f_i(q) f_j(q) \right)^3 \\
&= \sum_{q=1}^{Q} f_i(q)^3 f_j(q)^3 + 3 \sum_{q_1=1}^{Q} \sum_{q_2<q_1} f_i(q_1)^2 f_i(q_2) f_j(q_1)^2 f_j(q_2) + \\
&\quad +6 \sum_{q_1=1}^{Q} \sum_{q_2<q_1} \sum_{q_3<q_2} f_i(q_1) f_i(q_2) f_i(q_3) f_j(q_1) f_j(q_2) f_j(q_3) \qquad (7.20) \\
&= < \Phi(f_i), \Phi(f_j) > \qquad (7.21)
\end{aligned}$$

where

$$\Phi(f_i) = ((f_i(q)^3), \sqrt{3} \sum_{q_1=1}^{Q} \sum_{q_2<q_1} f_i(q_1)^2 f_i(q_2), \sqrt{6} \sum_{q_1=1}^{Q} \sum_{q_2<q_1} \sum_{q_3<q_2} f_i(q_1) f_i(q_2) f_i(q_3)) \tag{7.22}$$

is of dimension $d = Q + \frac{Q(Q-1)}{2} + \frac{Q(Q-1)(Q-2)}{6} = \frac{Q(Q+1)(Q+2)}{6}$ (it is in fact the dimension of the space of homogeneous polynomials of degree 3 in Q variables).

Here the KPCA method boils down to the SVD of the feature data $\Phi(f_i)$. The advantage is that no dimension reduction is necessary.

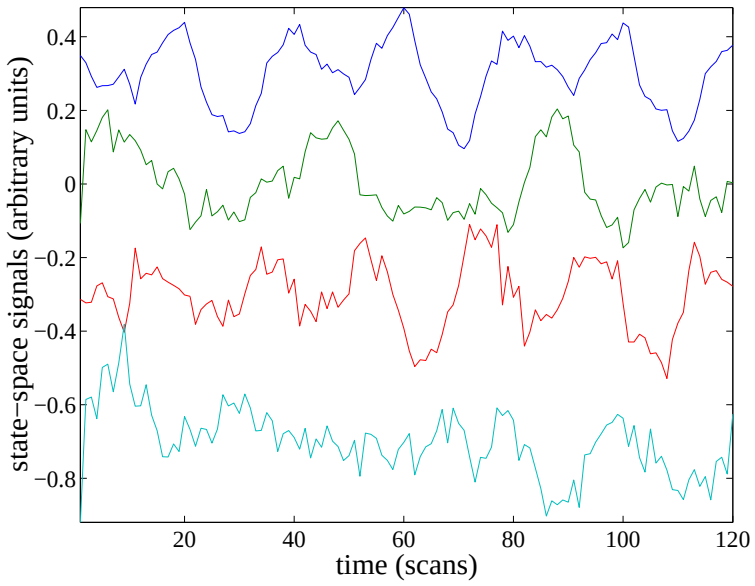


Figure 7.9: A particular basis of the state space obtained for the dataset.

If the interpretation of the first component is obvious in terms of task-related activity, it is not clear that the other temporal patterns are meaningful. The question is thus: how to relate these patterns to the empirical dynamics encountered in the dataset ?

Application to real data We use once again the dataset presented in A.2. The state-space model is applied to the data averaged over the different sessions, after reduction to $N = 20$ components by PCA. This results in a $Q = 4$ dimensional state-space. A basis of the time courses that span the state space is given in figure 7.9.

If the first time course is easily interpreted in terms of task-related activity, the interpretation of the other time courses is somehow more difficult; one of the reasons for that is that the time courses are not dynamically independent, as explained in chapter 6. This raises a natural question: what is/are the model(s) that best describe(s) the actual dynamical patterns within the dataset ? Let us study the answer given by kernel PCA. Since $Q = 4$, the dimension of the kernelled data is $d = 20$. The SVD of the matrix $\Phi(f_i), i = 1..N$ provides d spatial maps for the data. We now study the first three components, that make up 99% of the kernel variance. The three time courses $s_n(t), n = 1..3$ associated to the three maps are given in figure 7.10.

As expected, the first time course represents the main mode of activation (compare with figure 7.7). The second time course is largely more jittered; in fact it is essentially the derivative of the first pattern with respect to time: the correlation between $s_2(t)$ and $\frac{ds_1}{dt}(t)$ is 0.66. We face here a usual feature of the multivariate decomposition of time courses: a difference in timing results in a modulation through the time derivatives:

$$s_1(t + dt) = s_1(t) + dt\frac{ds_1}{dt}(t) + o(dt) \simeq s_1(t) + dts_2(t) + o(dt) \tag{7.23}$$

the sign of the spatial map associated with $s_2(t)$ will indicate this timing difference for each voxel. The third time course $s_3(t)$ is then the more interesting modulation according

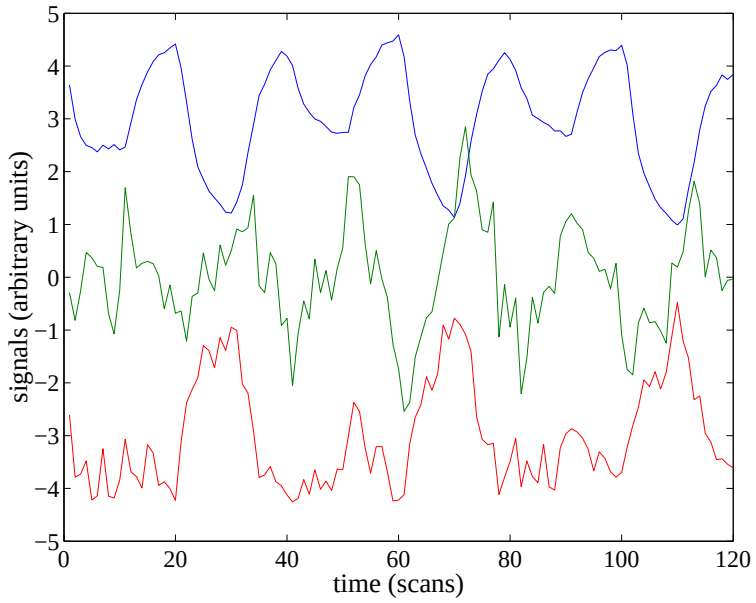


Figure 7.10: Time course associated with the first three components obtained by kernel PCA on the feature data.
The first model is the main activation pattern within the dataset; the second one is essentially its derivative with respect to time, and the third one presents a modulation with respect to the activation condition, much similar to what has been reported in 7.3.1. As usually with kernel PCA, these time courses are mutually correlated.

to the contrast stimulus1-stimulus 2 (see section 7.3.1). The spatial maps provided by kernel PCA are displayed in figure 7.11.

Discussion The interpretation of the maps is complicated by the difficulty of finding equivalent z-maps for the spatial components. We have used the method described in appendix B to derive an equivalent z-map based on robust statistics. In particular, we have applied the correction for non-normality of the null distribution for all maps, since there were in all cases strong deviations from normality. However, the resulting z values are probably conservative. Anyway, some important characteristics of the spatial maps displayed in figure 7.11 can be outlined:

- The first map is positive almost everywhere, by contrast with the others; as noticed from the time courses, the first component represents the main response within the dataset, so that it is not surprising that it is positive everywhere. This does not preclude the presence of negative patterns (see figure 7.8); but the latter are thus not exactly anti-correlated with the positive pattern, so that they are absent from the above map (this is a kernel effect).

- The second map represents essentially a left-right contrast. From the observation of the time course, one can deduce that a positive value indicates a delay in the response, while negative values indicate an anticipation in the response, with respect to pattern 1. Here, this may simply reflect slice acquisition order, which has not been corrected.

- The third map represents the effect of interest, namely the modulation by the contrast moving-static. Negative responses indicate a greater impact of motion condition on the time course, while positive values indicate a weaker impact. Spatially, this provides a nice interior/exterior contrast, with peripheral regions being less sensitive to motion than interior regions. This map is coherent with the spatial components displayed in figure 7.8, though the latter showed a more exclusive delineation of MT/V5 areas. A more thorough study of maps 1 and 3 of figure 7.11 would indeed confirm that these are the areas selectively activated by the motion condition.

Once can conclude from this that the supplementary degrees of freedom afforded by the kernel PCA technique allow for the derivation of more interpretable components from the state-space model presented in chapter 6. The use of a cubic kernel is advantageous, since it avoids the dimension explosion induced by non-polynomial kernel. Besides, it enhances strongly covarying structures, yielding almost *binary* maps that allow for simple interpretation.

7.3.3 Comparison with the MLM

The setting

We perform a comparison of linear and non-linear PCA methods -MLM and KPCA- on a real dataset (see also [211]). The latter is described in [42] and is taken from an experiment on 9 subjects analyzed in [126]. The information of interest is reduced to a contrast between two experimental conditions. In order to improve the estimation of the

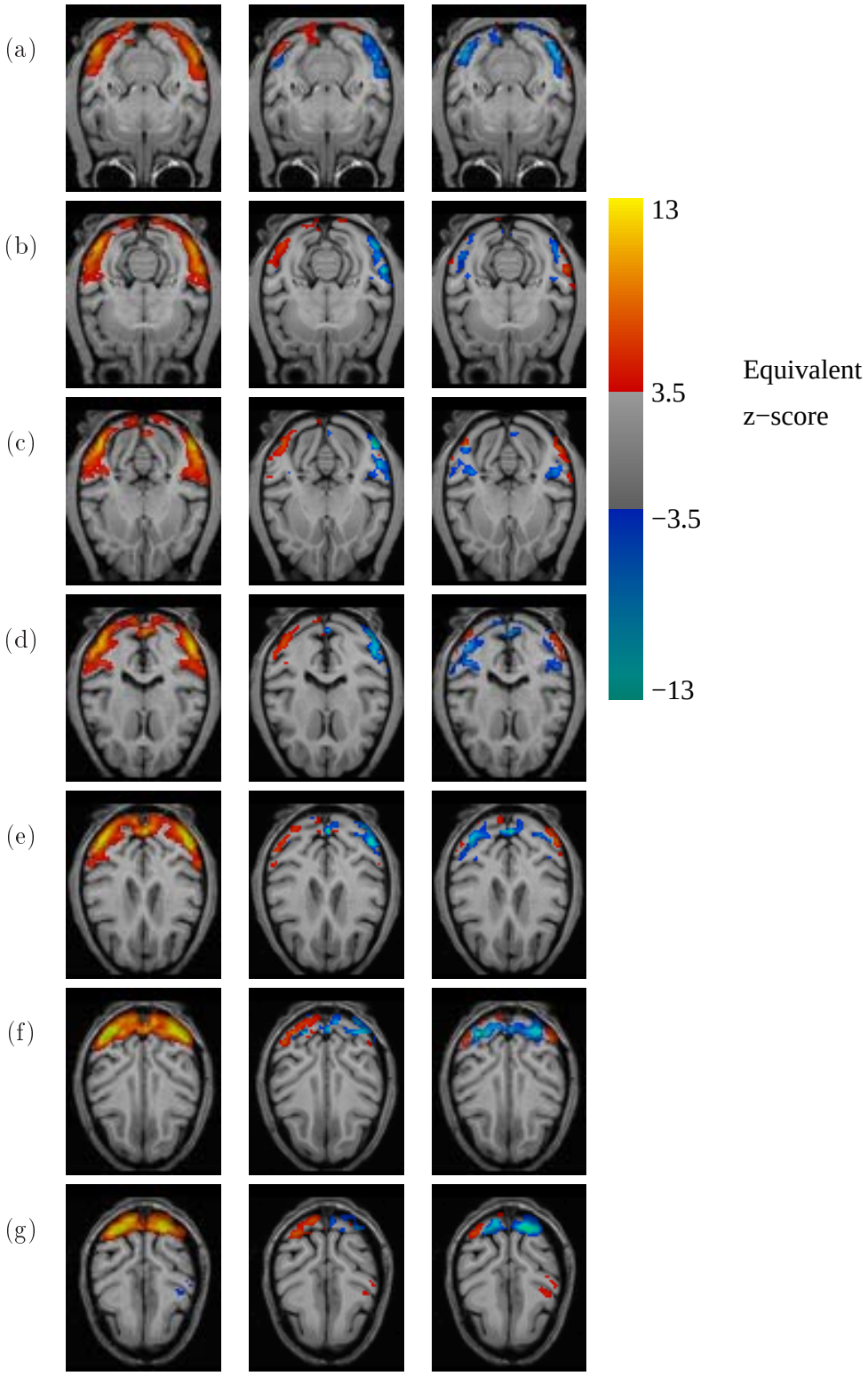


Figure 7.11: Seven slices of the first three spatial maps obtained with the state-space approach followed by KPCA technique.
As shown on the right hand side color bar, red-yellow clusters indicate positive values of the maps, blue colors indicate negative values of the maps.

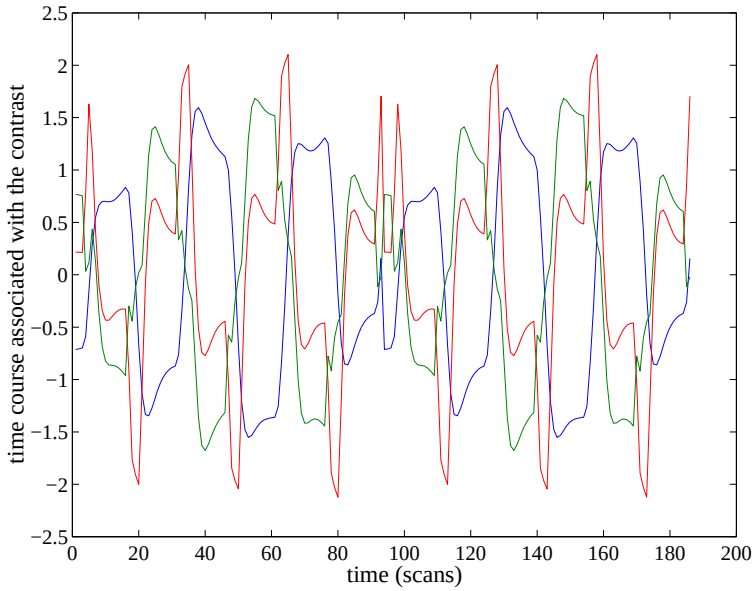


Figure 7.12: Time courses associated with the 3 contrasts studied here. Voxel-based responses to these contrasts make up the feature space $f_1(n), .., f_3(n)$.

latter, one proposes to introduce a 3-dimensional hemodynamic space that contains the SPM standard hrf and its first two time derivatives.

We work here on the analysis of a subject which has been shown to have low activation values. The number of voxels considered in the analysis for this subject is $N = 17721$.

The data is reduced to the regressors and residual noise by standard procedures (see section 3.3). Each voxel is then represented by a vector $X(n) = (\beta_1(n), .., \beta_R(n), \varepsilon(n))$, where (β) are the results of SPM regression, and $\varepsilon(n)$ the residual variance (see equations (3.3)-(3.4)). Then for both methods, a 3-dimensional contrast is derived. The contrast amplitudes estimated at each voxel are normalized by the standard deviation of the noise level $\varepsilon(n)$. The resulting 3-dimensional vector is now considered as the feature $f(n) = (f_1, f_2, f_3)(n)$. This feature has the structure of a signal to noise ratio, so that a feature with a high norm is a priori strongly informative. It is defined at each voxel; both KPCA and MLM are used to derive spatial maps $m(n), n = 1..N$ of the feature distribution.

For KPCA, we choose the kernel

$$K_\sigma(i,j) =< f_i, f_j > \exp\left(\frac{1}{\sigma}(\frac{< f_i, f_j >}{|f_i||f_j|} - 1)\right), \tag{7.24}$$

for $i, j = 1..N'$, with $N' = 3020$ results from the selection of the voxels that respond to the "effects of interest" with a P-value$< 10^{-3}$, uncorrected for multiple comparisons. Here $\sigma = 0.05$. K is then diagonalized

$$K_\sigma = M \Delta M', \tag{7.25}$$

where $\Delta = diag(\delta_1, .., \delta_N)$ is a diagonal variance matrix. Each resulting map can be

associated with a feature by

$$f(m) = \frac{\sum_{i=1}^{N} m(i) f(i)}{\delta(m)} \tag{7.26}$$

Next, the reduced representation can be extended back to the whole dataset by the generalization formula

$$m(n) = \frac{\sqrt{N'}}{\Delta(m)} \sum_{i=1}^{N'} K_\sigma(f_n, f_i) m(i),\ n = 1..N, \tag{7.27}$$

For KPCA, the maps are far from gaussian, and there is no well-defined way to make them gaussian (as in appendix B). For example, the maps are much more contrasted than what could be expected from a normal model. However, the thresholds derived in the normal case are still helpful to isolate the main clusters of coherent activity. This is due to the normalization of the map. We thus keep a similar setting and interpretation in terms of $z-$values.

For the MLM, the setting is much simpler, since no dimension reduction is necessary. Three maps are derived by diagonalization of the 3×3 feature covariance matrix; they can be analyzed as indicated in appendix B.

Results

MLM The three squared singular values, once properly normalized [126], yield map-wise feature squared norm averages of of 4.32, 2.02 and 0.98 respectively. This is significant for the first two ones. The three time courses (adjusted and fitted effect), reconstructed from the spatial maps in a standard way, are displayed in figure 7.13.

MLM map1 is represented in figure 7.17 with the first two KPCA maps; one slice of MLM map2 is represented in figure 7.18; in both cases, low thresholds have been used ($z > 3$ and $z > 2$ respectively). Spatial map 1 is associated with the strongest effect present within the dataset. Although there are clearly positive and negative clusters of points, only one cluster (on the negative side) is beyond the threshold $z = 4.7$, which corresponds to a corrected P-value $P = 0.05$; on the positive side $z < 3.4$ uniformly. For MLM map2, no voxel exceeds the absolute z-value $z = 3.8$. This means that the extrema of the maps are weak, i.e., that the SNR is very low in the dataset.

Note that the results are invariant if the same analysis is performed on a reduced dataset- as KPCA. This is logical, since the features close to 0 have low impact on the global covariance structure (data not shown).

KPCA An histogram of the empirical correlations $\frac{<f_i, f_j>}{|f_i||f_j|}, i \neq j$ is shown in figure 7.14. The statistical distribution of correlations is important if one considers equation (7.24), which indicates that correlations that fall below $1 - 3\sigma$ are canceled ($K \simeq 0$). The histogram shows the presence of two peaks, one around 1 and the other one around -1, indicating that the distribution of the data is essentially bimodal. Second, the fact that the histogram reaches nowhere 0 indicates the spread of the data around the two modes. $\sigma = 0.05$ seems to yield a correct characterization of the peaks. Note that considering the entire dataset would yield a different correlation histogram, probably with another mode around 0.

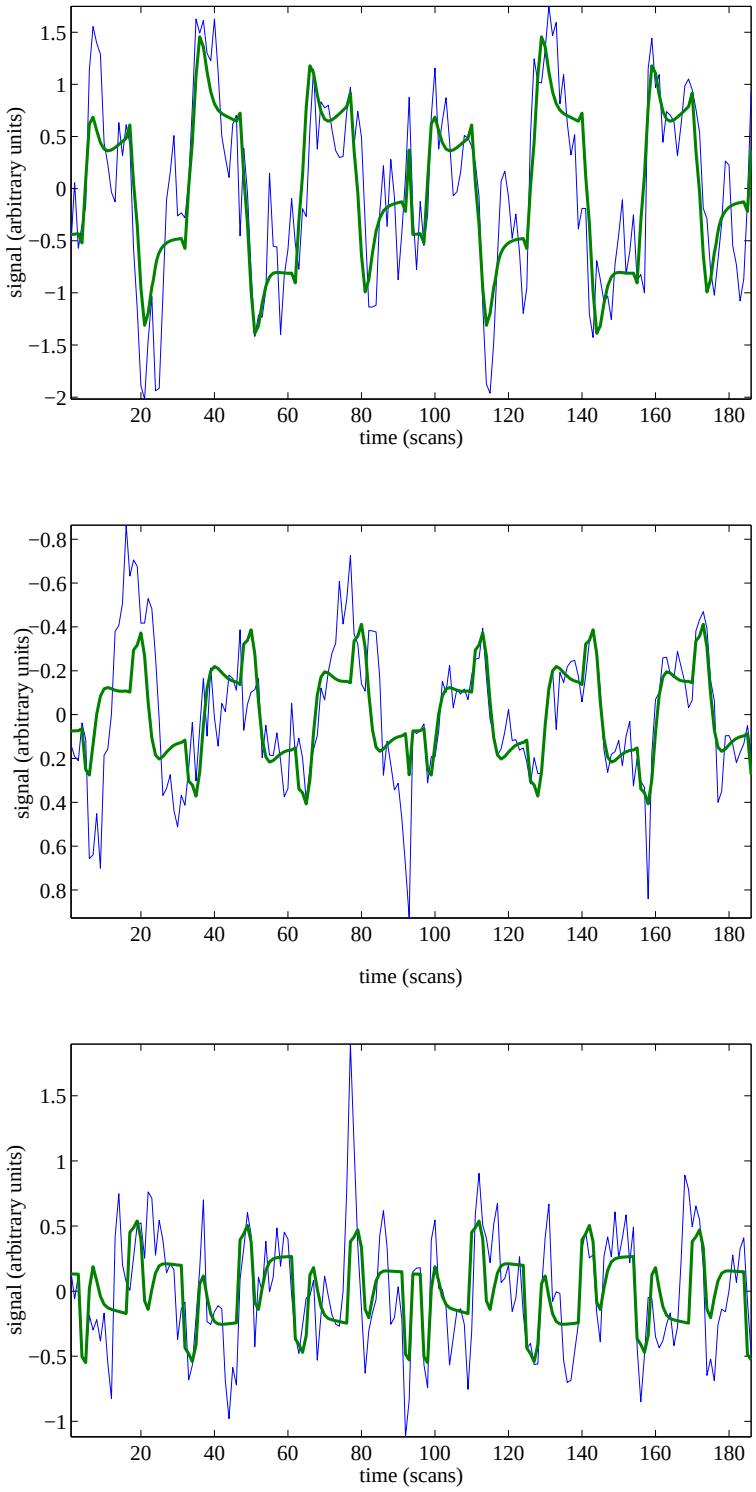


Figure 7.13: Set of time courses obtained with the MLM method.
Three components are represented, with for each a fitted (blue) and adjusted (green) time course for each component. They are associated with decreasing amounts of variance within the data.

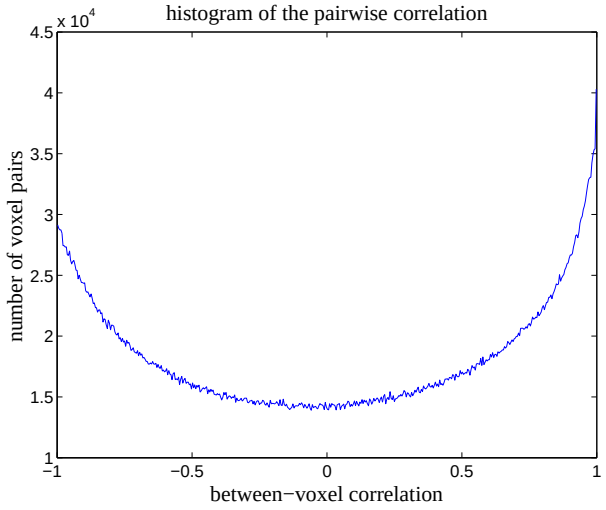


Figure 7.14: Histogram of the empirical cross-correlation of the dataset.
The maxima at $cc = 1$ and $cc = -1$ indicate that the data is dominated by two opposite components.

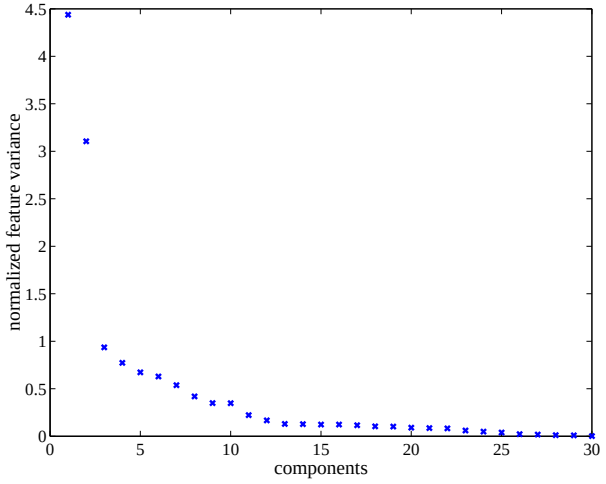


Figure 7.15: This curve characterizes the norm of the feature associated with each component.
The first two values are far above the others, indicating that there are two main modes in the signal distribution. They are also the two statistically significant ones. Further studies indicate that these two modes are approximately opposite in the feature space. In the sequel, we consider the first three components.

The amount of variance associated with each component is given in figure 7.15. It is clear that the first two components are far above the others and have significant values (note that these values are analogous to F statistics); nevertheless, we have decided to study the first three components. The associated time courses are given in figure 7.16. A particularity of kernel (7.24) is that opposite effects appear as different components, whereas they appear as one subspace with MLM. It is easy to deduce that the first component found by MLM is thus associated with the first two components obtained by KPCA, whose temporal time courses are almost opposite, as can be seen in figure 7.16. Therefore, we plot together $(-1\times)$MLM map1, KPCA map1 and KPCA map2 in figure 7.17. One slice of KPCA map3 is represented in figure 7.18 together with MLM map2; unlike the latter, it contains supra-threshold clusters of voxels.

Discussion

MLM and KPCA: similarities First of all, it is important to keep in mind that the methods perform comparable analyses of the data. Moreover, choosing $\sigma = \infty$ in kernel PCA yields exactly the MLM decomposition. Besides this definition, and setting MLM TC i for the i^{th} temporal component obtained with MLM, and the same notation for kernel PCA there is a striking similarity between :

- MLM TC1 and (-1×) KPCA TC1,
- MLM TC1 and KPCA TC2
- MLM TC2 and KPCA TC3,

Logically, these similarities are also true in the spatial domain. While this is evident in the first case ($(-1\times)$MLM map1 and KPCA map1), this is also true in the other cases, but hidden by the weakness of the MLM maps. The fact that component 1 of MLM is overwhelming is shown by both the distribution of the singular values in MLM and the structure of the kernel PCA solution (in terms of eigenvalue and fitted variance). This is of course consistent with the observations made on the histogram in figure 7.14.

MLM and KPCA: differences The introduction of a non-linearity in kernel PCA is related to the form of selectivity, which is clear from the study of the function $\frac{<f_i,f_j>}{|f_i||f_j|} = cc \in [-1,1] \longrightarrow \exp\left(\frac{1}{\sigma}(cc-1)\right)$. In other words, the components provided by kernel PCA are *narrowed* by the parameter σ. While this induces additional difficulties, e.g. in terms of computational load, selection of the number of components, this has advantages in terms of denoising and separation of effects. For example, KPCA TC1 and TC2 are not exactly opposite; TC1 has a stronger contribution of the first coordinate, while TC2 contains more information from the derivatives -and is thus probably less reliable than TC1 in terms of signal interpretation. The fact that the two components are gathered in the MLM model implies that their detection is less accurate.

Moreover, a consequence of the kernel higher selectivity is that the spatial maps are better defined in terms of activation clusters than with MLM. This is true both in the cases of MLM map1 and MLM map2 with respect to KPCA map2 and KPCA map3: these maps are similar, but the KPCA maps are more contrasted, and thus yield more

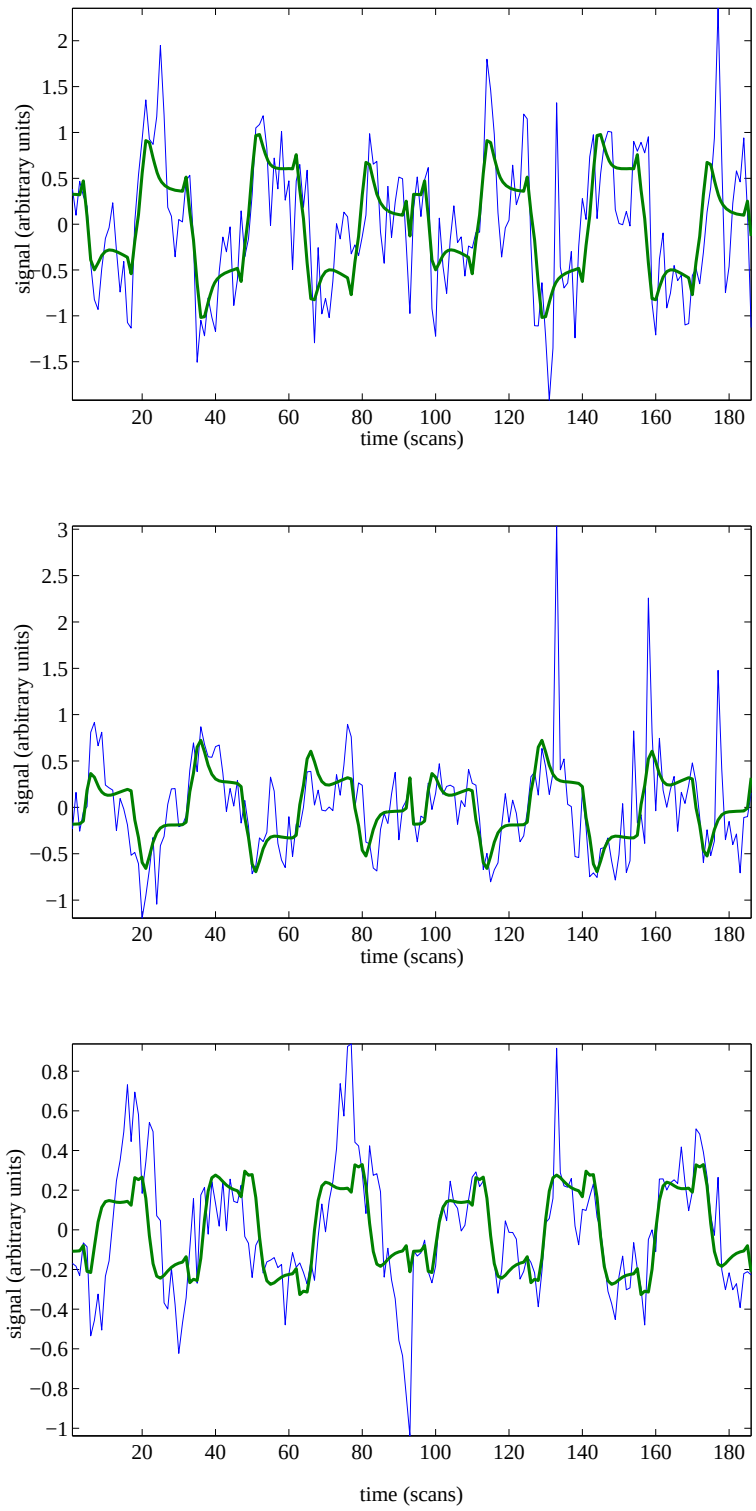


Figure 7.16: First three time courses -fitted (blue) and adjusted (green) effects- obtained by the kernel PCA decomposition.
Note that they are not constrained to be orthogonal. Moreover, the second model is approximately the opposite of the first one.

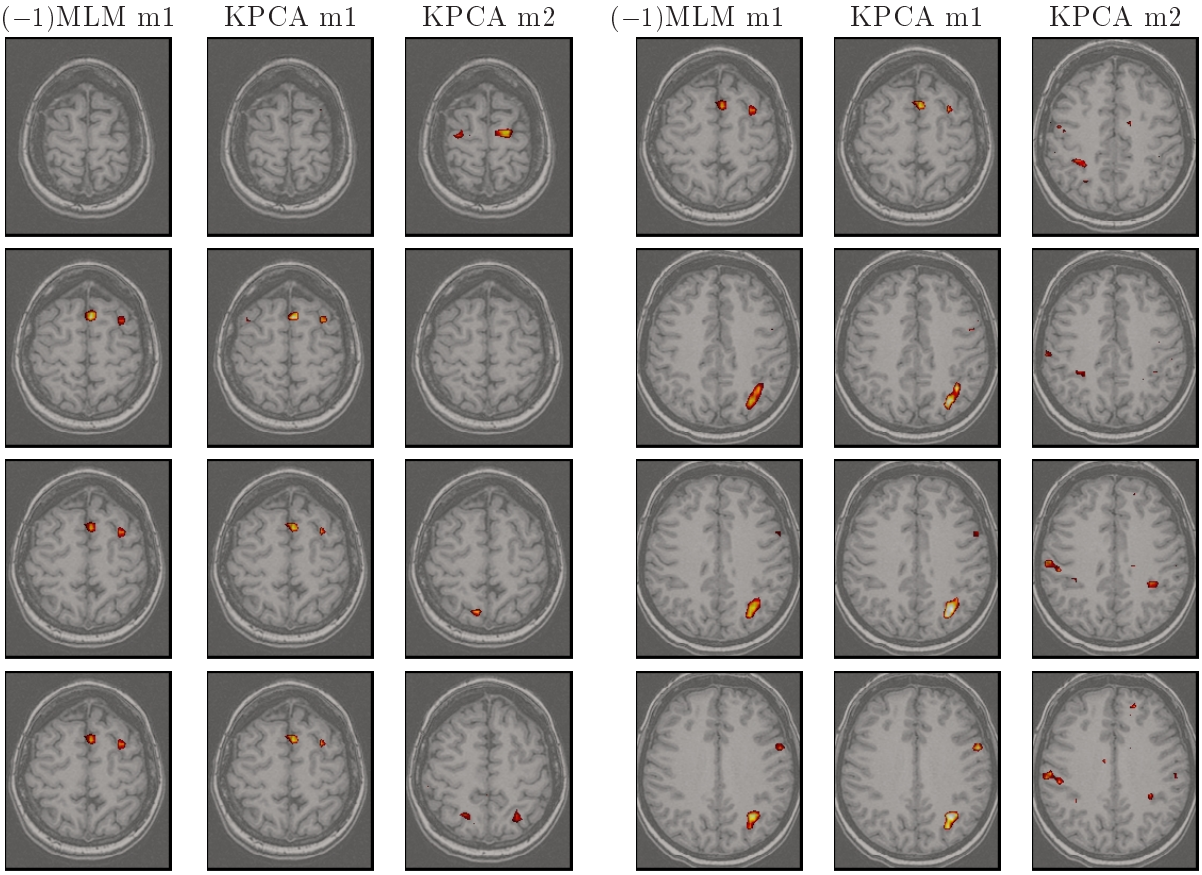


Figure 7.17: Main spatial maps given by the MLM, and by KPCA.
These maps have been reduced to 8 axial slices. The color map, identical in the 3 cases, is scaled from $z = 3$(red) to $z = 5$ (yellow-white). For MLM only the negative part of map 1 ($(-1\times)$MLM map1) is represented, since only 1 voxel reached the threshold in the positive part of MLM map1; for KPCA, only the positive part of the maps is significant. Clearly the maps ($(-1\times)$MLM map1) and (KPCA map1) coincide. Conversely, (KPCA map2) would coincide with(MLM map1), if the maxima of the latter were higher.

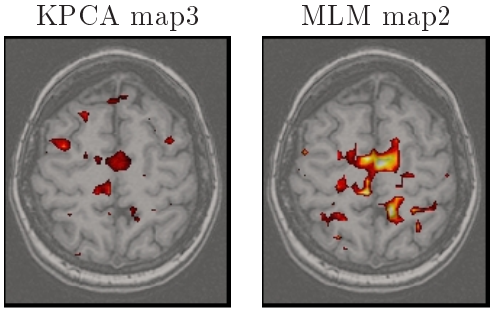


Figure 7.18: One slice of KPCA map3, and the same for MLM map2.
The color code for voxel intensity ranges from 2 to 5. So the structure is similar, the height of the central peak is 3.11 for MLM and 6.04 for KPCA.

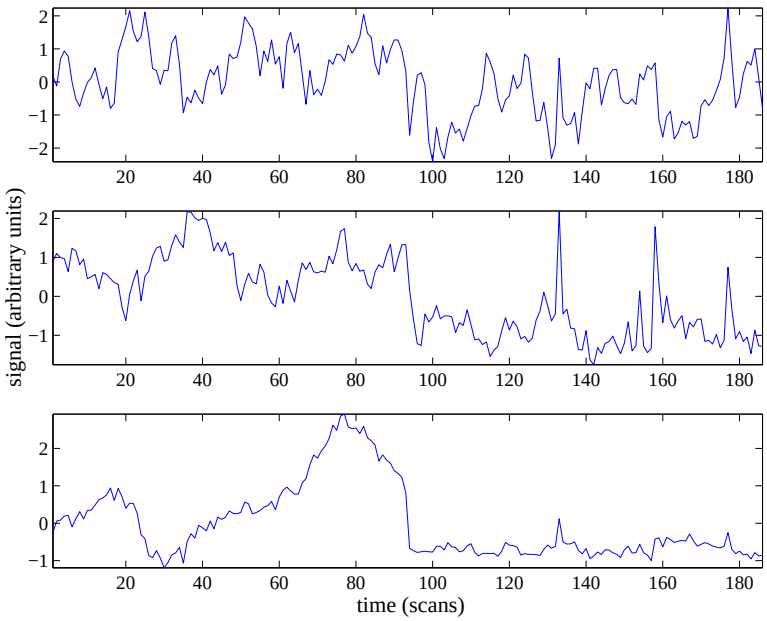


Figure 7.19: Time courses represented in figure 7.16 **before** filtering and fitting with the linear model
Obviously, components 2 and 3 are strongly mixed with non task-related effects: spikes on the second time course, and trend on the first half of the third time course.

supra-threshold clusters of voxels. This is not attributable to the selection of a subset of N' voxels in the case of KPCA. Indeed, performing the same selection with MLM does not change the outcome.

This difference is visible here because the SNR in the data is low: it has been noted from standard studies that activations are weak on this subject. To explain this, one could argue that

- The hemodynamic response to the experimental paradigm are weak, because the experiment did not work well (the subject did not respond,...).
- There are some confounds in the data that shadow the activation peaks.

Given our experience with kernel PCA and observation of the empirical time courses, i.e. before temporal processing (see figure 7.19), we plead for the second answer: activation patterns are indeed present, but they are shadowed by some confounds.

7.3.4 Conclusion

We have shown that the potential of kernel PCA to produce overcomplete data representations can be used to study datasets in greater detail than what is allowed by traditional linear methods (PCA, SVD). This comes at the price of an heavier computational cost.

In particular, section 7.3.3 shows that there is a balance between *i)* the simplicity of the MLM and *ii)* the additional power of kernel PCA to concentrate on some areas of the

signal space, which is also an immunity against confounds.

let us also recall the tradeoff between the ability of the MLM to reduce optimally (in terms of variance) high dimensional feature spaces and the capability of kernel PCA to unmix data components when the mixture is not gaussian. In parallel, the statistical interpretation of SVD-based decompositions is clear in terms of multivariate normal distribution, while kernel PCA is not simply associated to a particular statistical structure.

Kernel PCA can be recommended to analyze experiments where results are poor, or if more than one meaningful components can be expected from the dataset (e.g. in the case of many experimental conditions). It is thus essentially an exploratory tool. For example, we have shown in 7.3.2 that kernel PCA is an interesting post-processing after dynamical state-space analysis of the dataset: although all the information is indeed present within the state-space decomposition , kernel PCA usefully reveals some important features that are not obvious if one uses a naive signal basis.

Last, as in indicated in recent works [107], there is a great potential of the joint use of kernel PCA and ICA (see section 5.2): while kernel PCA produces overcomplete data bases, ICA finds more meaningful directions than those obtained by standard PCA techniques. The application of this idea to fMRI data has still to be done.

Chapter 8

The signal space as a manifold : exploration through the Laplacian graph technique

This chapter presents a new alternative in the study of nonlinear mixing models for fMRI data. This stems from the fact that kernel PCA lacks some guarantees of optimality in the representation of the data: the latter should be sparse and low dimensional. The final choice of the dimension is not necessarily robust with respect to the occurrence of non trivial data structures. The kernel parameters control the complexity of the representation, but there is little control on the tuning of the parameters. The Laplacian graph technique is a recent tool for nonlinear dimension reduction that can possibly optimize the representation of the signal space. Though it is closely related to some well-known techniques, as Multi Dimensional Scaling (MDS), Locally Linear Embedding (LLE) and Isomap algorithms (see section 8.2.1), it additionally has a meaningful geometrical interpretation, which makes it easier to interpret.

In the first section, we explain in detail the algorithm and its interpretation. Then we discuss briefly its particularities with respect to some closely related methods, and some technical details. We end this chapter with different applications of this technique to fMRI data. Another possible application of this technique for data visualization is presented in [28], for example.

8.1 Nonlinear dimension reduction with Laplacian graphs

We state here the main characteristics of the Laplacian graph technique. Most of this development comes from [16] [17]. As for kernel PCA, we consider a set of data items $(X_n, n = 1..N)$ that belong to a certain feature space $\mathcal{F}$. The question is to find a low-dimensional representation of the latter space that preserves the metric of the data within the feature space. More precisely, the key assumption is that the data is sampled from an unknown manifold $\mathcal{M} \subset \mathcal{F}$. For fMRI data, the feature space is naturally the signal space, endowed with an adequate metric. The input data is provided by voxel-based time courses.

8.1.1 The algorithm

Given the N points, we first construct a graph with N nodes, and derive the embedding map by computing the lowest eigenvectors of the graph Laplacian. Note that the graph and the graph Laplacian are nothing but the discrete counterparts of $\mathcal{M}$ and its associated Laplace-Beltrami operator $\mathcal{L}$.

Construction of the graph Two alternatives are possible: the first one, which we will refer to as the *metric* construction, consists in creating an edge between nodes i and j of the graph if X_i and X_j are close within $\mathcal{F}$. A possible construction is to connect i and j if

$$\|X_i - X_j\|^2 \leq \varepsilon \tag{8.1}$$

This procedure may yield a graph with several connected components, which is not necessarily a problem (each connected component can then be analyzed as a graph). The second alternative (which will be called henceforth the *topological* construction) is to create an edge between each node i and its k nearest neighbors (e.g. $k = 10$).

Choosing the weights The edges of the graph are given a weight W_{ij}. Typically

$$W_{ij} = \begin{cases} \exp -\frac{\|X_i - X_j\|^2}{2\sigma^2} & \text{if } i \text{ and } j \text{ are connected,} \\ 0 & \text{otherwise.} \end{cases} \tag{8.2}$$

Note the similarity with some kernel PCA methods, equation (7.9). However, the choice of σ is far from crucial here, so that $\sigma = \infty$ is possible (in which case all the neighbors of a given point have equal weight).

Eigenmaps For each connected component of the graph, we form the matrices D and L so that $D_{ii} = \sum_j W_{ji}$, and $L = D - W$ is the Laplacian matrix of the graph; L is diagonally dominant, hence it is positive semidefinite. The eigenmaps $(m_i), i = 1..d$ of the graph are derived through the generalized eigensystem

$$Lm = \lambda Dm \tag{8.3}$$

We sort the solutions of equation (8.3) by increasing λ. Note that $\lambda_0 = 0$, since by construction $Lm_0 = 0$ for the constant map $m_0 \equiv 1$. Then $i \to (m_1(i), .., m_d(i))$ is a low dimensional representation of the dataset. If d is chosen adequately, the geometrical interpretation is that $X_i \to M(i) = (m_1(i), .., m_d(i))$ is a map of the data manifold in $\mathcal{F}$. We now investigate the properties and the optimality of this mapping for data representation.

8.1.2 The variational approach

Given the graph that represents the signal, the question is to find a d-dimensional embedding of the latter that preserves most of its structure. In particular, $\|M(i) - M(j)\|$

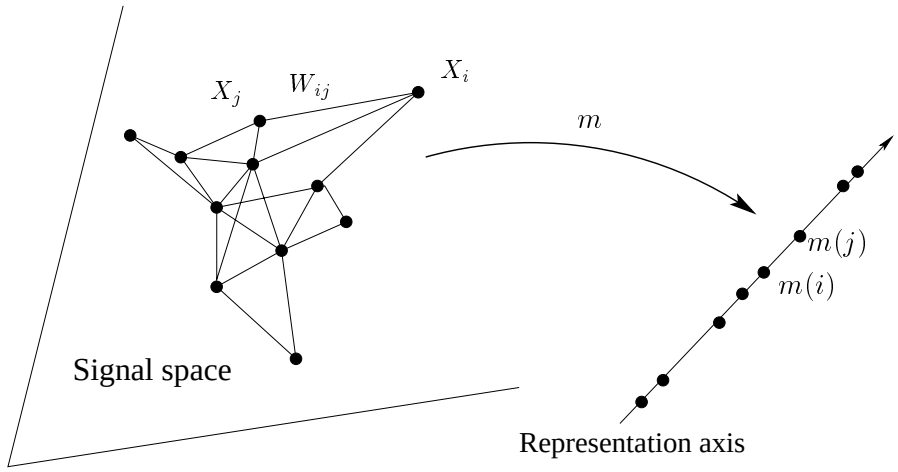


Figure 8.1: Illustration of the optimization problem solved by the Laplacian graph approach
Find the low dimensional embedding of the data that minimizes metric distortions.

should be small if $\|X_i - X_j\|$ is. Let us consider a one-dimensional embedding $M = m$. A natural objective function is then

$$\sum_{i=1}^{N} \sum_{j \in N(i)} (m(i) - m(j))^2 \exp\left(-\frac{\|X_i - X_j\|^2}{2\sigma^2}\right) = \sum_{i,j=1}^{N} (m(i) - m(j))^2 W_{ij}, \tag{8.4}$$

where $N(i)$ is a neighborhood of i, defined in the metric or topological way as previously. But, with the same notations as previously, we have $\frac{1}{2}\sum_{i,j=1}^{N}(m(i)-m(j))^2 W_{ij} = m^T L m$ (this simply results from the definition of L). Thus the optimal embedding -the one that minimizes 8.4- is provided by

$$argmin_m m^T L m \tag{8.5}$$

Of course there is an arbitrary scale factor in the embedding, that can be removed by the constraint $m^T m = 1$ or $m^T D m = 1$; the latter is preferable, since it gives more weight to highly connected voxels. With this normalization, the problem boils down to (8.3). The fact that the constant solution m_0 is trivial is clear in terms of embedding (the whole graph is mapped to a single point). An illustration of the general problem is given in figure 8.1.

The extension to a d-dimensional embedding M is straightforward: the optimal solution is spanned by the first d eigenvectors of equation (8.3).

8.1.3 Geometrical point of view: the signal space as a manifold

As mentioned before, the underlying model of this technique is that the signal lies within a low dimensional manifold $\mathcal{M}$ of $\mathcal{F}$, the *graph* being simply a discrete representation of the manifold. The natural metric on $\mathcal{M}$ is inherited from $\mathcal{F}$. Let us consider a twice differentiable one dimensional mapping $m : \mathcal{M} \to \mathbb{R}$. Let $(x, y) \in \mathcal{M} \times \mathcal{M}$. We have the

property

$$|m(y) - m(x)| \leq dist_{\mathcal{M}}(x,y)\|\nabla m(x)\| + o(dist_{\mathcal{M}}(x,y)), \tag{8.6}$$

where $dist_{\mathcal{M}}(x,y)$ is the *geodesic* distance between x and y on $\mathcal{M}$. But $dist_{\mathcal{M}}(x,y) = \|x-y\| + o(\|x-y\|)$ if the embedding of $\mathcal{M}$ in $\mathcal{F}$ is isometric (see [17] for more detailed explanations). Equation (8.6) implies that the locality preserving property of m is related to the minimality of the norm of its gradient: in other words, m yields an accurate representation of $\mathcal{M}$ if its gradient is small . Preservation of the locality on average yields the minimization of

$$\int_{\mathcal{M}} \|\nabla m(x)\|^2 dx, \ \|m\|^2 = 1, \tag{8.7}$$

which is in turn equivalent to the minimization of $m^T L m = \frac{1}{2}(m_i - m_j)^2 W_{ij}$ on the associated graph. We can also notice that the solution of (8.7) is equivalent to finding the eigenfunctions of the Laplace-Beltrami operator $\mathcal{L}$ on $\mathcal{M}$. Indeed, applying integration by parts, one has

$$\int_{\mathcal{M}} \|\nabla m(x)\|^2 dx = -\int_{\mathcal{M}} m(x) div(\nabla m(x)) dx = \int_{\mathcal{M}} \mathcal{L}(m) m dx \tag{8.8}$$

Thus the minimization of (8.7) is achieved for the first eigenfunctions of $\mathcal{L}$ on $\mathcal{M}$. As a classical result of Riemannian geometry, $\mathcal{M}$ being compact, the spectrum of the operator $\mathcal{L}$ is discrete. The fact that $\mathcal{M}$ *is* compact is an ad hoc hypothesis, which nevertheless makes sense for empirically derived data. The operator L defined on the functions of the graph is formally and in practice the discrete counterpart of $\mathcal{L}$. The first eigenfunction of $\mathcal{L}$ is the constant function that maps $\mathcal{M}$ to a single point. This is of course the counterpart of $Lm_0 = 0$ in section 8.1.2. If we denote the next eigenfunctions of $\mathcal{L}$ by $M = (m_1, .., m_d)$, we obtain an optimal d dimensional embedding for $\mathcal{M}$.

The last thing that remains to be explained is the particular choice of weight attribution (8.2) for the edges of the graph. In our geometrical framework, this choice receives a nice justification: using the connection of the Laplace-Beltrami operator with the heat equation ($\frac{dM}{dt} = \mathcal{L}M$), one can use a Taylor expansion of the solution

$$M(t) = M(0) + dt\frac{dM}{dt} + o(dt) = M(0) + dt\mathcal{L}M + o(dt), \tag{8.9}$$

and the explicit solution $M(t)$ of the heat equation by means of convolution with a gaussian kernel to estimate $\mathcal{L}M$. This yields:

$$\mathcal{L}M(x) \sim \frac{1}{t}\left[M(x) - (4\pi t)^{-\frac{d}{2}} \int_{\mathcal{M}} \exp\left(-\frac{\|x-y\|^2}{4t}\right) M(y) dy\right] \tag{8.10}$$

for small t. The discretization of the latter formula yields

$$\mathcal{L}M(X_i) \sim \frac{1}{t}\left[M(X_i) - \frac{1}{N}(4\pi t)^{-\frac{d}{2}} \sum_{j/\|X_i - X_j\| < \varepsilon} \exp\left(-\frac{\|x-y\|^2}{4t}\right) M(X_j)\right] \tag{8.11}$$

If one chooses $M = m_0 \equiv 1$, then $\mathcal{L}m_0 \equiv 0$, which implies that $\frac{1}{N}(4\pi t)^{-\frac{d}{2}} \sum_{j/\|X_i - X_j\| < \varepsilon} \exp -\frac{\|x-y\|^2}{4t} = 1$. Thus, given our choice (8.2) for W_{ij}, and replacing t by $2\sigma^2$, we obtain, up to a scale factor

$$\mathcal{L}M(X_i) \sim \sum_j W_{ij}(M(X_i) - M(X_j)) \tag{8.12}$$

This justifies a posteriori our choice (8.2) for W_{ij}.

8.2 Discussion and technical points

8.2.1 Other strategies: MDS, LLE, Isomap, KPCA

The Laplacian eigenmap method is a nonlinear dimension reduction technique, and can be compared to closely related methods: Multi Dimensional Scaling (MDS), which is a now well-established technique, especially for data visualization, but also Locally Linear Embedding (LLE) or Isomap methods. We also make a quick comparison with the Kernel PCA approach described in chapter 7. Note that there exist also some connections with the normalized cuts technique for images segmentation [129], but this is out of the scope of this discussion. Last, the discussion could also include Self-Organizing Maps (SOM) and Principal Manifolds, as in [109]. However, these methods are derived from different points of view, and do not enjoy as elegant and simple solutions as the embedding approaches.

Multi Dimensional Scaling Multi Dimensional Scaling (MDS) is a well-established technique for high-dimensional data visualization [25]. It has been used for the visualization of different fMRI datasets representing a population of subjects [125]. Technically, it is also based on the minimization of distortion criterion, which is called a *stress*. With the same notations as in the previous paragraph, it is formulated as:

$$\sum_{i,j=1}^{N} (|M(i) - M(j)| - f(\|X_i - X_j\|)) \tag{8.13}$$

where f is some monotonic nonlinear (often bounded) function. This is an alternative to functional (8.4). In some way, Laplacian eigenmaps solve the indeterminacy of MDS techniques about the structure of the penalty term. Moreover, with Laplacian eigenmaps, the problem is solved with a spectral technique, while classical MDS approaches require heavy non-convex functional minimizations [131]. Last, MDS does not put the same emphasis as Laplacian eigenmaps on the neighborhood in the definition of f: this means that the precise structure of the dataset in $\mathcal{F}$ is not really taken into account.

Locally Linear Embedding Locally Linear Embedding (LLE) techniques [190] rely on 3 steps:

- For each input point X_i, find out its neighbors (X_j), $j \in D(i)$ (e.g. its k closest neighbors).

- Compute empirical embedding weights W_{ij} by minimizing

$$\sum_{i=1}^{N} \|X(i) - \sum_{j \in D(i)} W_{ij} X(j)\|^2 \tag{8.14}$$

  under the constraint that $\sum_{j \in D(i)} W_{ij} = 1$. (W_{ij}) can thus be thought of as a system of local barycentric coordinates.

- Derive the embedding $i \longrightarrow M(i)$ by minimizing

$$\sum_{i=1}^{N} \|M(i) - \sum_{j \in D(i)} W_{ij} M(j)\|^2 \tag{8.15}$$

Though this technique is evidently close to MDS and Laplacian eigenmaps, it has some particular characteristics: the emphasis on the neighborhood relationships (which is related to the geometrical interpretation of the method), the choice of a locally linear representation. The practical solution also boils down to a sparse eigenvalue problem. But the interpretation of the resulting embedding is strictly local; moreover, the local linearity constraint may become a problem if the neighboring system is ill-specified.

Isomap The last algorithm of the family, Isomap puts more emphasis on the global geometrical structure [204]. Indeed, after performing the first step of LLE and deriving local distance, Isomap uses global distances using the theory of shortest paths within a graph. Then a spectral decomposition is performed on the matrix $H\Delta H$, where $\Delta_{ij} = \|X_i - X_j\|^2$ and $H_{i,j} = \delta_{i,j} - \frac{1}{N}$. The main disadvantage of this method is that it does not solve a sparse eigenvalue system, making the practical solution difficult when N is high.

Kernel PCA The kernel PCA approach, presented in the previous chapter, is quite different from the above mentioned methods. However, the comparison makes sense. In some instances, kernel PCA performs the decomposition of a matrix analogous to W, with W defined as in (8.2). The main difference is that kernel PCA searches for the direction in the input space where a maximum of data is present, while nonlinear dimension reduction techniques are more devoted to the derivation of sparse representation and on the preservation of the structure of the data.

8.2.2 Choosing the hidden parameters

In practice, there remains to deal with the *hidden* parameters in the method. There are two such parameters with the Laplacian embedding method: in the *metric* approach the neighborhood size parameter ε in (8.1) and the kernel parameter σ in (8.2) and in the *topological* approach, the number k of neighbors considered, and σ.

In the metric approach, the kernel parameter σ has almost no impact as soon as $\sigma > \varepsilon$ (it is even possible to choose $\sigma = \infty$ without creating inconsistencies in the solution); conversely, ε is fundamental, since it defines the neighborhood relationship, e.g. the graph or manifold structure of the data. It should thus be chosen taking into account the priors on the data generative process; for example, considering that the input vectors $X(i)$ are

time series, it could be reasonable to normalize them with respect to their stochastic variance, and then to choose $\varepsilon = 2$ or 3. The neighborhood size is thus related to the probability of a common underlying process for the two time series.

For the *topological* approach, choosing k between 5 and 20 seems satisfactory for the construction of the manifold, and $\sigma = \infty$ is an acceptable default solution. We study the impact of k in section 8.3.3.

Another implicit parameter of the method is related to the sampling of the manifold $\mathcal{M}$: the method performs correctly only if the data points are sampled at least approximately regularly on the yet unknown signal manifold; this is far from certain with fMRI data. A possibility is to cancel all the data points of non-interest that are grouped into a kind of null cluster. As could be expected, we have noticed from our experiments that the topological method is less sensitive to sampling variations in the feature space than the metric method.

8.2.3 Computational issues

Computationally, the main problem is the diagonalization of $D^{-1}L$ (or $D^{-1/2}LD^{-1/2}$ if one wants to keep the symmetry of the system). Though a solution by SVD is possible for reasonable numbers N of entries, the problem is more efficiently solved by iterative methods.

Solution by SVD

Let us recall that this solution remains acceptable for $N < 4000$. While this is not the case for fMRI datasets, it can be used for reduced datasets.
We have also mentioned the possibility that the data graph could be disconnected. In that case, the problem is split into 2 smaller problems, allowing for a computationally lighter solution. This situation is not unrealistic: in certain experiments, one can obtain positive and negative responses to a certain stimulation (those negative responses can be related to artifacts), yielding two components. We have noticed that this is more frequent in the *metric* approach than in the *topological* approach.
However, the solution by SVD is a bit awkward, given the sparsity of the graph, hence of the matrix L.

Iterative solution

By construction the normalized Laplacian matrix $D^{-1}L$ has the following property

$$\forall m,\ |D^{-1}Lm - m|_1 < |m|_1, \tag{8.16}$$

where $|m|_1 = \sum_{n=1}^{N} |m_1(n)|$. Letting I_N be the $N \times N$ identity matrix , this implies that the eigenvalues of $D^{-1}L - I_N$ lie within $[-1, 1]$. In fact, the smallest eigenvalues of $D^{-1}L$ are the greatest eigenvalues of $I_N - D^{-1}L$, all of which are smaller than 1 in absolute value; in other words, this application is contracting. This implies that the iteration of $m \leftarrow m - D^{-1}Lm$ yields the solution of the problem (moreover, one can show easily that $I_N - D^{-1/2}LD^{-1/2}$ is positive, so that $I_N - D^{-1}L$ is also positive. There is thus no problem with possible negative eigenvalues). The greatest advantage of this method is

that it can be completed with a sparse coding for L: only the *neighbors* of an entry have to be coded. Hence the system can be solved for real datasets $N \in [10^4, 10^5]$. Moreover, the convergence towards the solution is certain.

8.3 Application to fMRI data

8.3.1 Exploratory analysis

The first application we will consider is an unsupervised (hypothesis-free) description of datasets using the low-dimensional representation capabilities of the method. The idea is analogous to Self-Organizing Map or clustering methods presented in [144] [143]: build a representation of the raw data that preserves its topology. The difference is that our method converges deterministically towards its solution, and that no prior data reduction is necessary. We apply the Laplacian eigenmap method using the version that uses the $n = 10$ closest neighbors of each voxel, and $\sigma = \infty$ in (8.2), i.e. all closest neighbors have equal weight.

Synthetic data

We present a first experiment on the synthetic dataset described in A.1. We start by studying the distribution of the Laplacian eigenvalues in equation (8.3); these are given in figure 8.2. While the first one is 0, as expected, the next three ones are close to each other, and far below the following ones. This hints at a three dimensional representation of the data. Indeed, looking at the resulting three dimensional mapping of the data shows that the three-mode structure of the synthetic dataset has been correctly accounted for (see figure 8.2).

The study of the spatial maps (not shown) and associated temporal patterns (in figure 8.2 (bottom)) shows indeed that the Laplacian eigenmaps describe the dataset with much accuracy. Note that in contrast with the example given in 7.1.3, this is achieved without any prior modeling of the data.

This illustration shows that the Laplacian eigenmap is able to detect the main features of the dataset, even in the absence of pre-processing. Unlike kernel PCA, it is not subject to overfitting.

Real data

We now study the ability of the Laplacian eigenmaps techniques to reveal structures from a real dataset. We use the first session of the dataset A.2 (this implies relatively low SNR conditions, since other sessions are available for this dataset). In order to allow for a comparison, we compare it with PCA, which also performs an "optimal" dimension reduction, but in the sense of the data covariance. In both cases, we consider arbitrarily a 2D representation of the data. This arbitrariness is acceptable, since this analysis is purely exploratory; the Laplacian eigenvalues, not shown, simply indicate that the first dimension is much more informative than the other ones. Figure 8.3 gives the joint distribution of the N voxels on the main 2D maps obtained by PCA and Laplacian embedding ($N = 12320$).

It is striking that the Laplacian eigenmap representation is much more structured than the PCA representation. Moreover, the cluster shape in the case of PCA is curved, so

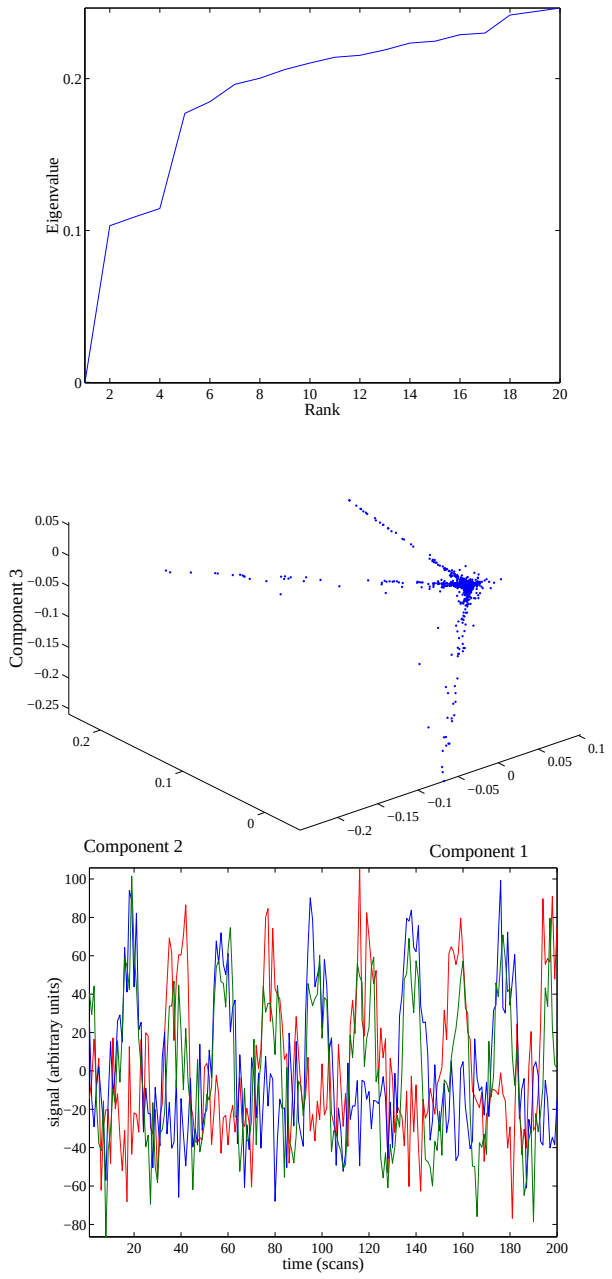


Figure 8.2: Analysis of a raw synthetic dataset with the Laplacian embedding method. (left) Sequence of eigenvalues obtained with the Laplacian eigenmaps method applied to unprocessed synthetic data; the first eigenvalue is 0 while the next three ones are close to each other; consequently, we opt for a 3D representation of the dataset. (right) 3D Laplacian map of the data; it results in a three-armed distribution, which describes the three modal data. (bottom) Consequently, the estimated components allow for a reconstruction of the time courses of the three activation patterns which is very close to the ground truth -see figure A.1.

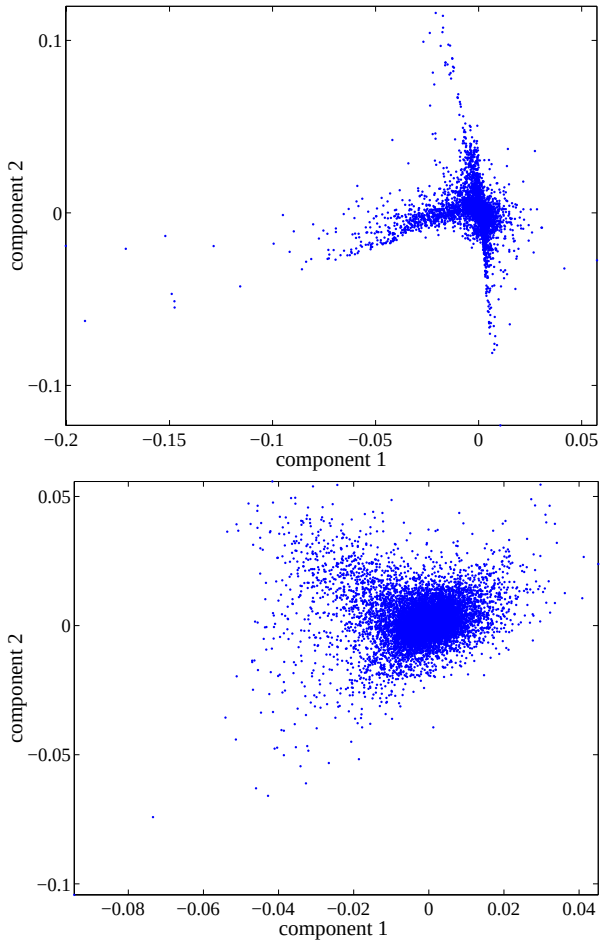


Figure 8.3: 2D representation of the real dataset, obtained through the Laplacian eigenmap embedding (left), and PCA (right).

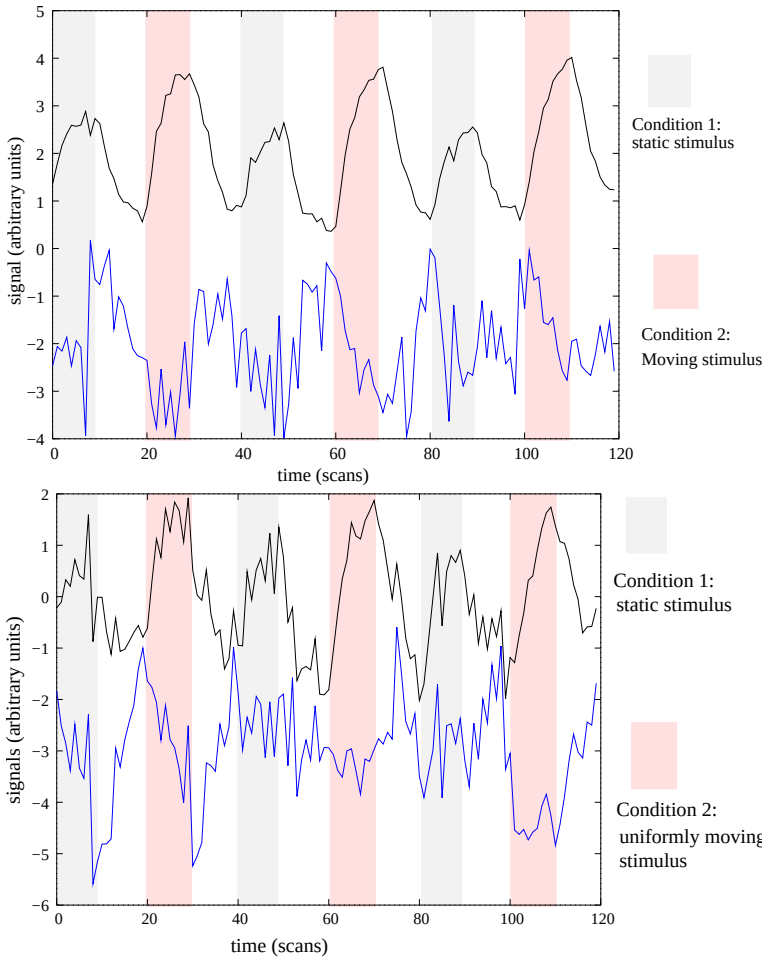


Figure 8.4: Main time courses obtained with both methods
(left) Two main time courses obtained from Laplacian embedding technique; they correspond to the first two Laplacian maps that are the axes of the feature space in figure 8.3. (right) Two main time course associated with the PCA decomposition.

that the representation in terms of main axes of the data does not fit well the data. The situation with the Laplacian eigenmap is much more simple, with a three arms star-shaped cluster.

Next we study the time courses obtained with both methods. In the case of PCA, these are the time courses of the two main axes. In the case of Laplacian eigenmaps, we similarly compute map-wise time courses by weighting the voxel time courses by the map value of each voxel. This results in figure 8.4.

There is a good correspondence between the first time course obtained with PCA and with the Laplacian embedding method. The latter describes -with an opposite sign- precisely the voxels of the *west arm* in figure 8.3. The difference is that there is much more noise reduction with the Laplacian model than with the PCA model; this is related to the stronger selectivity of the Laplacian map. Once again we meet the usual mean task-related

pattern present within this dataset. The second pattern also approximately matches the second PCA component. This gives -with a change of sign- a good description of the voxels from the *north* and *south* arms of figure 8.3. The interpretation of this temporal component is not obvious; it is more noisy than the first one, and the fundamental pattern has slightly lower frequency.

Next, we present the spatial maps obtained from the Laplacian embedding technique in figure 8.5. Those have been transformed into z-scores with the method described in appendix B. Both positive and negative significant z-scores are presented, after thresholding at an estimated *false discovery rate* $FDR = 0.05$, corresponding to $|z| \simeq 3.5$.

The first spatial map is mainly associated with negative areas (in blue), as could be expected from the joint map in figure 8.3. It encompasses the main visual areas (without making any visible difference about motion sensitivity). The second map outlines some regions at the rear part of the visual areas. Since the map is positive there, this region can be identified as the *north arm* in figure 8.3. Considering figure 8.4, it can indicate a different hemodynamic response within this region. The negative cluster of the second map reflects an inverse effect in the parietal cortex. This situation has already been described in chapter 7. The second map thus seems to correct slightly the first one, as can be deduced from the negative clusters within the primary visual cortex. This is also consistent with figure 8.3 where the arms are not exactly horizontal nor vertical.

These examples show that the Laplacian eigenmaps can reveal the main structures of a dataset, even without prior modeling of the data. However, priors are necessary in order to obtain finer descriptions of the signal space of interest.

8.3.2 Pre-processed data

Let us consider the following problem: given a space of signals of interest, one would like to model how the individual voxel responses are organized throughout this space for a given dataset. To do this, we can use a finite impulse response (FIR) filter model for the signal space as described in chapter 4, and then study the empirical signal manifold within the filter space.

Once again we consider the dataset A.2; a task-related signal space is defined through the filter approach described in 4.3.3. The dimension of the signal space is $L + C - 1$, where L is the length of the filter, and C the number of experimental conditions in the experiment. Moreover, the dataset is reduced according to the criterion (4.39), yielding $N = 1320$ filter models. To study the organization of these models, we use the optimal 2D representation of the dataset allowed by the Laplacian eigenmap technique (we use the topological construction with a $k = 10$ neighbor system) . The eigenvalues of equation (8.3) are given in figure 8.6. The embedding dimension is not clearly indicated by figure 8.6, but we found that the third dimension did not add much information on the manifold structure. Keeping the first two dimensions of the decomposition results in the cluster displayed in figure 8.7(a). The color code is arbitrarily defined as the polar angle of each point on the 2D map (we found that it gives a good representation of the data manifold). An arbitrary (but topology preserving) segmentation of the manifold in polar sectors yielded 8 clusters, each of which is represented in figure 8.7(b) by an average time course (reduced to $T = 40$ time samples; note that the signals are periodical).

The study of the individual models shows that the first eigenmap essentially introduces

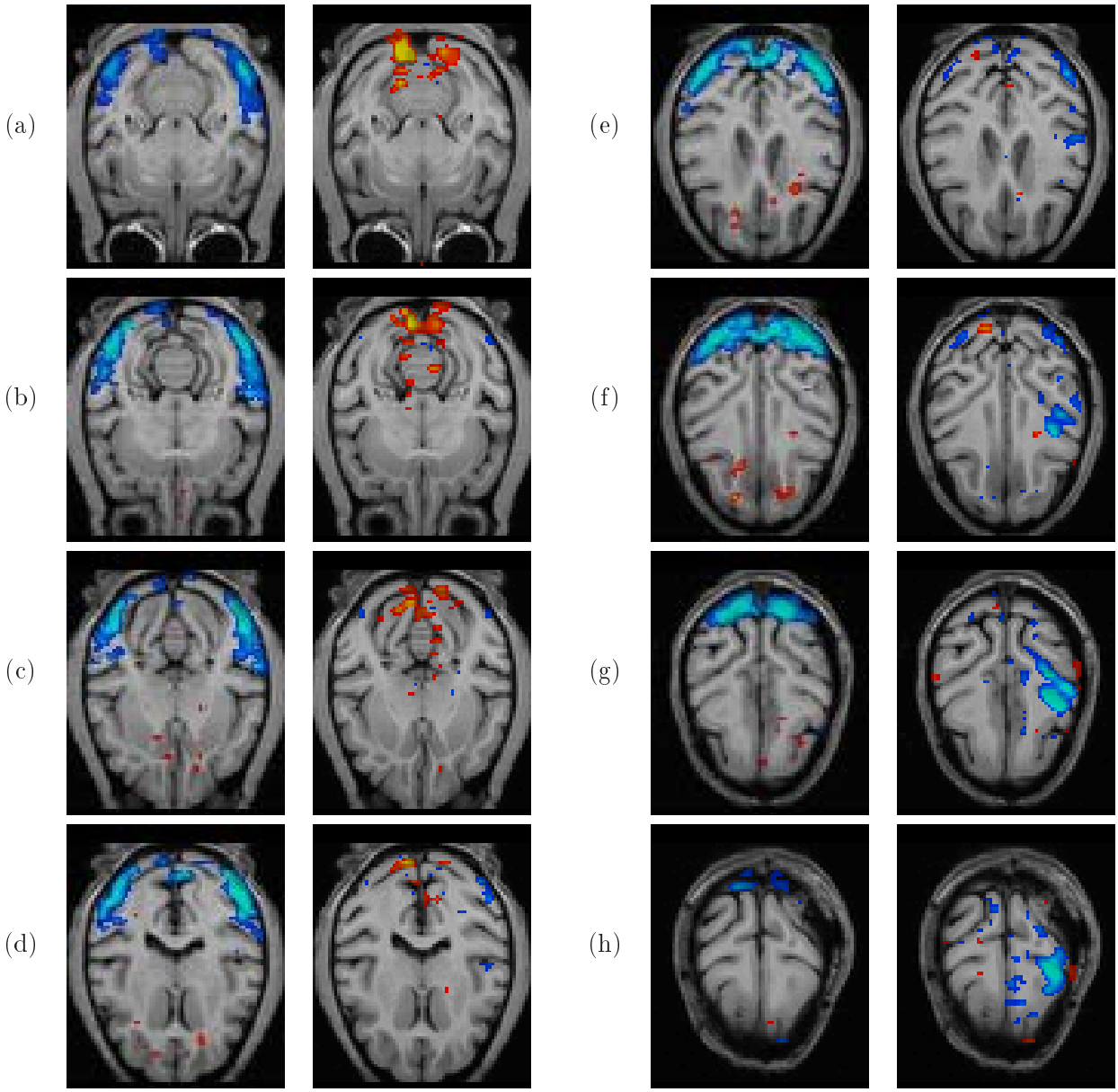


Figure 8.5: Eight axial slices (a-h) of the first two Laplacian maps of the dataset. Those have been transformed into z-scores with the method described in appendix B. Both positive and negative significant z-scores are presented in blue and red-yellow color respectively, after thresholding at $|z| \simeq 3.5$ ($FDR = 0.05$).

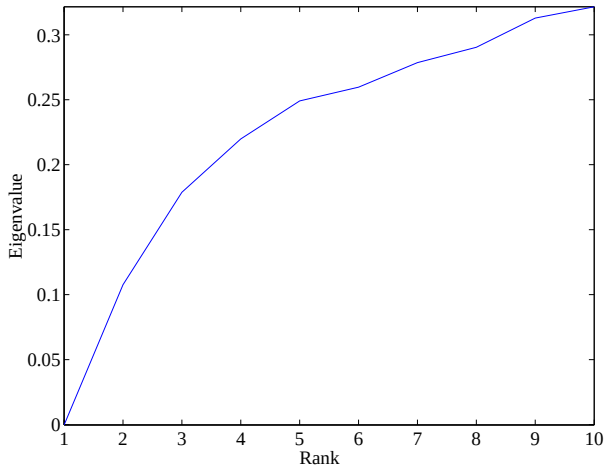


Figure 8.6: First eigenvalues obtained in equation (8.3).
Note that the eigenvalues can be interpreted as a measure of distortion . As could be expected, the first eigenvalue is 0 (and the associated map is uninformative). We keep for further study the second and third eigenvalues.

a distinction between two kinds of patterns, which can be unambiguously identified as the positive (purple-yellow) and negative (blue-green) task-related patterns. Note the discrepancy between the two types of patterns in figure 8.7 (b), and the different locations in figure 8.7 (c). The interpretation of the second map is more complex; the vertical arm in the feature distribution seems to indicate a slight shift in *activation timing* - the voxels of the arm (yellow) respond a little bit earlier than the central cluster (red). Last, the voxels at the bottom of the figure (purple cluster), are distinguished from the other by a modulation by the second condition (motion) selectively. All these characteristics can be read in figure 8.7(b), which presents the averaging of each cluster response. We display the cluster map in figure 8.7 (c). The color of the voxel corresponds to the color of the time model and the cluster in the feature space.

Additionally, a projection of the ray maps on the grey/white matter interface of the monkey is provided in figure 8.8. The three maps comprise a representation of the areas that are negative for components 1, negative for component 2, and positive for component 2.

Discussion It is probably not necessary to describe in detail those findings, which are coherent -and redundant- with those obtained with kernel PCA, section 7.3.1, State-space analysis followed by kernel PCA, section 7.3.2 and the Laplacian eigenmaps obtained from raw data, section 8.3.1, as well as SPM analysis, section 7.3.1. Let us only point out some interesting facts:

- As can be seen clearly in figure 8.7(b)(c), there are symmetric regions that respond uniquely to the second stimulus, i.e. the passive viewing moving of a static texture (*purple* regions). So far, this point had been made clear only in 7.3.1, with an overcomplete representation of the dataset.

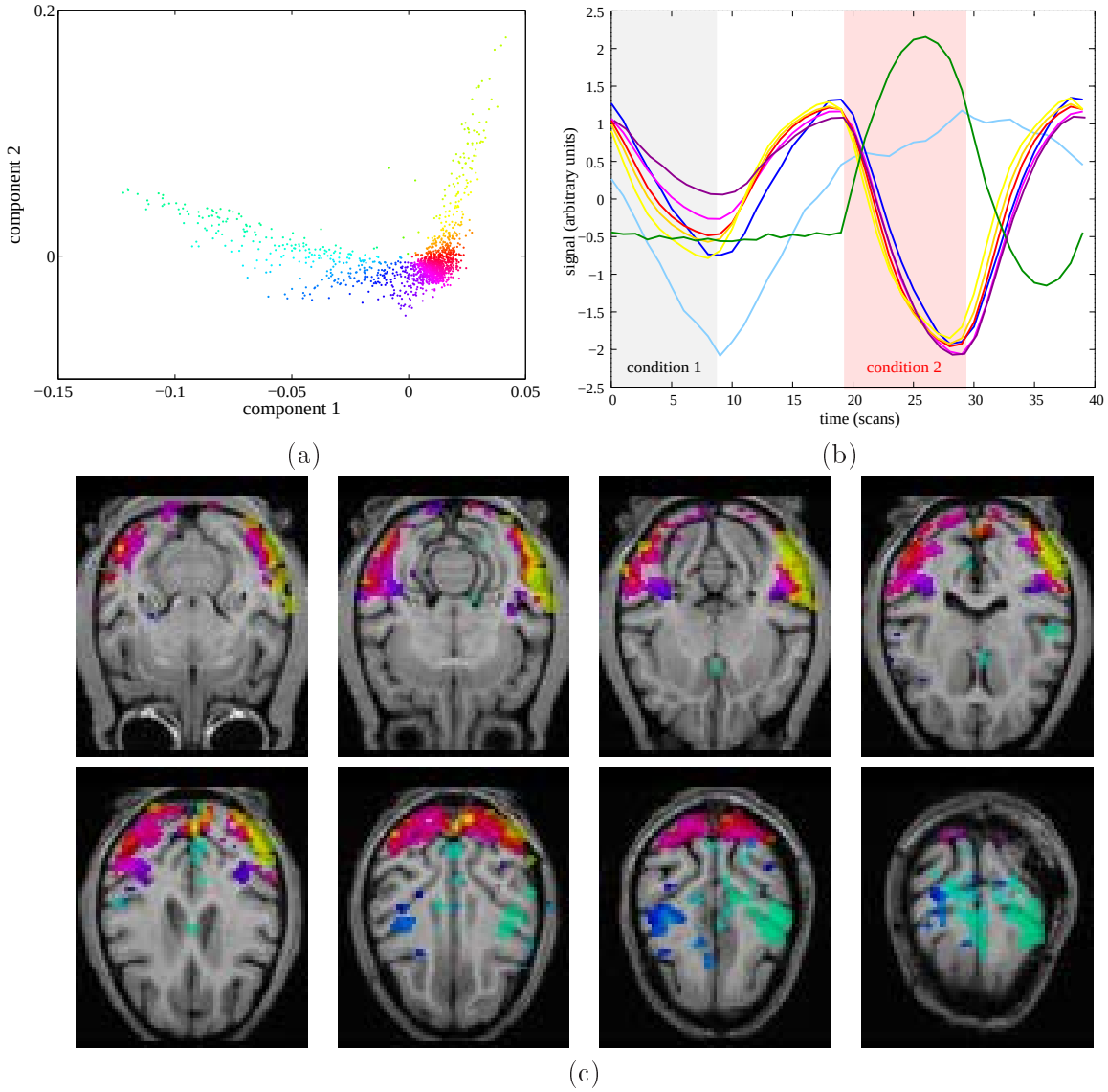
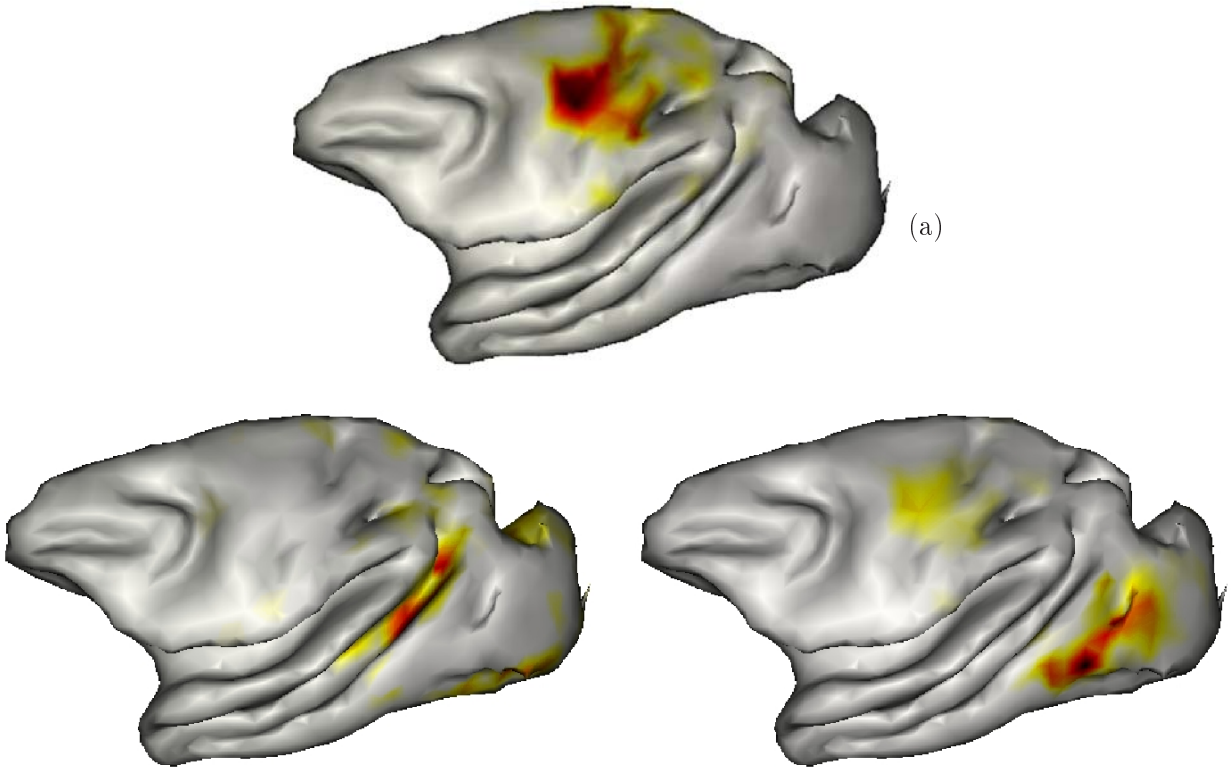


Figure 8.7: Analysis of a real dataset with the Laplacian embedding method (a) Optimal 2D representation of the signal space derived through Laplacian eigenmaps approach. The Boomerang-shaped manifold has been divided into smaller clusters by color coding by the hue the polar angle of each point- we find that is a satisfactory representation of the manifold; we keep the same color coding for subsequent interpretation. A 3D map of the cluster of points does not bring much more information. (b) Temporal models associated with the different sub-clusters. They illustrate the variability of the task-related responses within the dataset. The different experimental conditions are indicated by the background color. (c) 8 axial slices of the corresponding spatial maps; the color code is similar. Let us insist that colors do not indicate an intensity, but rather a qualitative behavior of the impulse response. Note the symmetry of the maps.



(a)

(b) (c)

Figure 8.8: Projections of the Laplacian eigenmaps on the cortical surface (grey matter-white matter interface)
(a) map 1: the negative areas are displayed in yellow/red; in the other figures, they have been identified as the red cluster; it corresponds to the parietal cortex, and the task-related pattern is shown to be negative. (b) map 2: the negative areas are displayed in yellow/red; they can be identified with the V5/MT area; in the other figures, they are also identified as the purple/black cluster. (c) map 2: the positive areas are displayed in red/yellow; in the other figure, these areas correspond to the green pattern. Interestingly, it is the only part of the spatial maps that is not symmetric.

- Interestingly, the Laplacian embedding uncovers an asymmetry between the left and right hemispheres, i.e. in the yellow pattern in figure 8.7. One can notice that this corresponds to a slight time shift in the response.

- Introducing a third dimension does not fundamentally change the setting. If we trust the Laplacian eigenmap analysis, this means that the data structure is essentially embedded in a U-shaped cluster.

One can conclude from this study that Laplacian eigenmaps yield an easy qualitative account of the task-related patterns that appear within the dataset. This method is economical in practice, since it tries to embed a maximum of information within a minimum number of dimensions. A development of interest would be to allow for an explicit formula for the mapping from the low dimensional space towards the feature space, which we have not addressed here.

8.3.3 Dependence on the neighborhood size

We investigate here the dependence of the method on the size of the neighborhood (noted k in the text) of each voxel [1]. Note that this is in fact the only parameter that has to be tuned for Laplacian eigenmaps (the influence of σ in (8.2) is quite negligible). To study the effect of k, we have done the same analysis as previously on the real dataset for $k = 5, 10, 15, 20$.

We study the impact of k *i)* in the eigenvalues of the Laplacian matrix (equation (8.3)), and *ii)* on the structure of the optimal 2-dimensional representation of the data.

Effect of k on the eigenvalues Let us emphasize once again that the eigenvalues of the Laplacian matrix can be interpreted as a measure of distortion when representing the dataset by a low dimensional map. Intuitively, one can expect that this distortion increases with k, since the neighborhood structure becomes then more complex. The empirical distribution obtained by letting $k = 5, 10, 15, 20$ is displayed in figure 8.9; it shows a discrepancy between $k = 5$ and $k = 10, 15, 20$, with $k = 5$ giving very low distortion values. This may be a hint that the structure with 5 neighbors is not sufficient to model the data properly.

Effect of k on the 2D maps As can be seen in figure 8.10, the study of the resulting cluster gives confirms this impression; while the clusters obtained for $k = 10, 15, 20$ have the familiar structure described in section 8.3.2, the cluster obtained for $k = 5$ is quite different; in fact, all the voxels that have been characterized as *negative* for the first map, are here scattered throughout the plane, which indicates that the dataset is almost shattered into several connected components, so that the global manifold is no longer apparent. Put more simply, this shows that $k = 5$ is not enough to describe this dataset.

Finally, once can conclude from this study that, provided that k is sufficient to account for the connectivity of the data manifold ($k \geq 10$), there is little impact of k on the results.

[1] Note that a neighbor matrix has to be symmetric, so that the numbers of neighbors considered for each voxel are greater than k; in fact the neighboring system considered in the topological construction of the graph is the smallest symmetric system that contains the k nearest neighbors of each voxel.

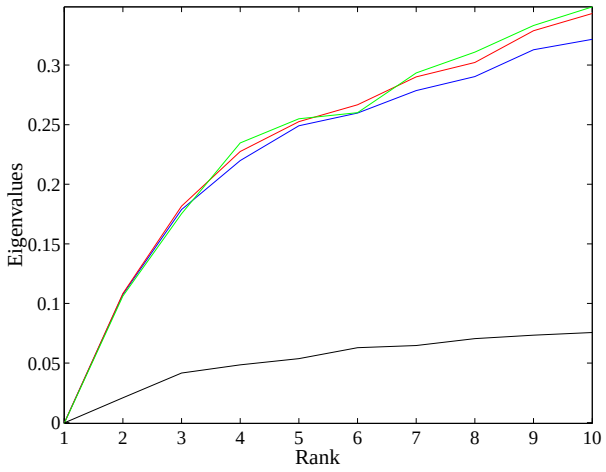


Figure 8.9: First eigenvalues obtained in equation (8.3) for different values of k $k = 5$ (black), $k = 10$ (blue), $k = 15$ (red) and $k = 20$ (green); note that the last three ones are very close to each other, in contrast with the first model.

8.3.4 Conclusion on Laplacian embedding technique

The Laplacian embedding technique yields a structured representation of datasets based on geometrical concepts. It is comparable in its result to Self-Organizing Maps. However, this representation offers several advantages:

- The optimal embedding is theoretically well defined and easy to compute in practice, even for real size datasets.
- The parameters of the method are fairly easy to tune, since only one parameter (k in the topological model) has an impact on the final result, and that a reasonable choice is not difficult to find.
- The embedding dimension can be determined from the dataset. However, we acknowledge that the Laplacian eigenvalues do not give very precise information with this respect.

It is thus a classification tool that enables a local-to-global approach of fMRI data: treating each voxel-based data as a sample of information, it builds a global framework that contains and summaries all the information of interest about the dataset. It can be used either for unsupervised/data-driven analysis or after a first temporal modeling of the data.

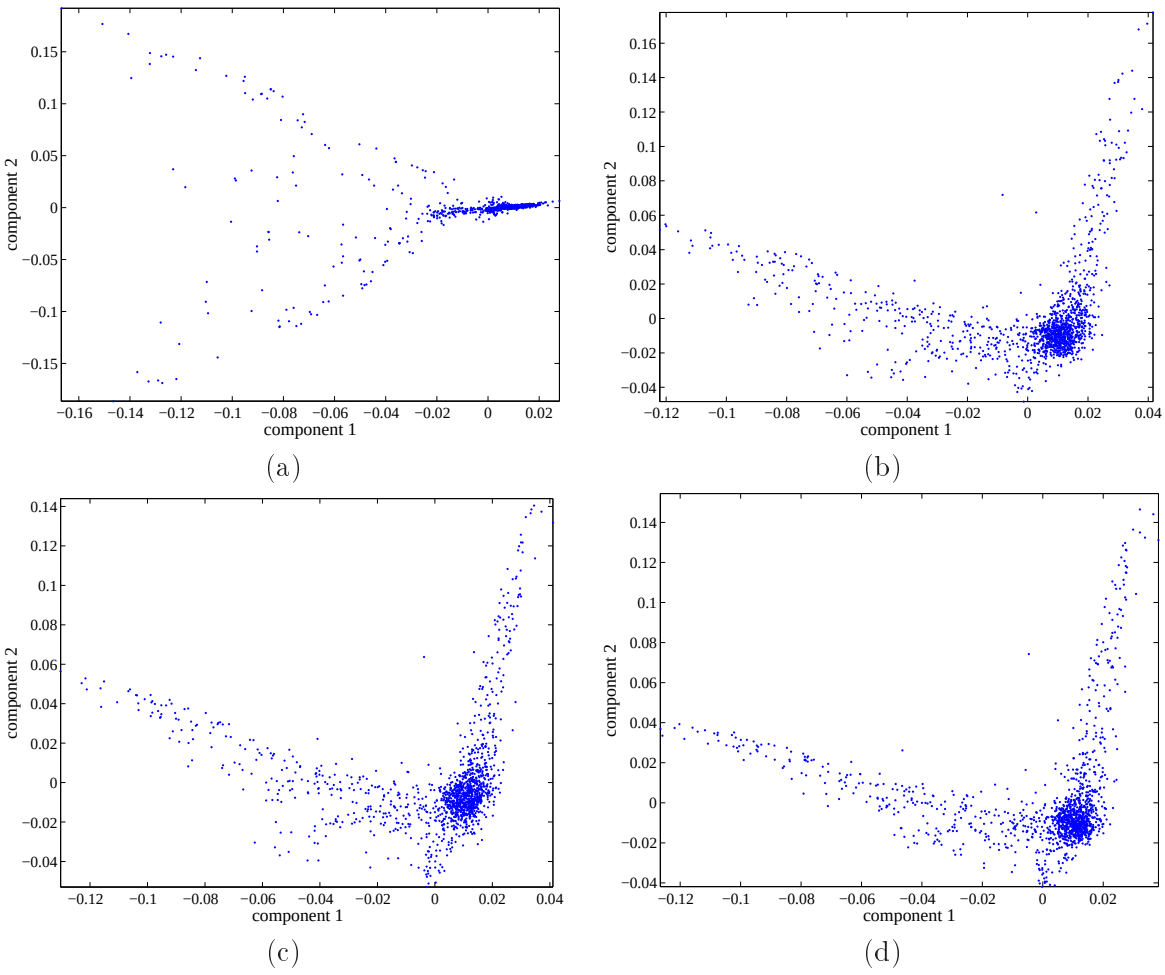


Figure 8.10: Optimal 2-D maps obtained with for different values of k $k = 5$ (a), $k = 10$(b), $k = 15$(c), $k = 20$(d). Note the scattered structure of the first map, and the similarity of the latter maps.

Chapter 9

Putting things together

This chapter presents a synthesis of our work and a discussion of some important issues. First, we conclude on the temporal modeling of the data by making the junction between the state-space models (i.e. chapter 6) and information theory (chapter 4). Second, we propose a global framework for the analysis of multi-session data; this framework uses many points developed in this thesis (temporal modeling, state-space modeling, optimal embedding through Laplacian eigenmaps). We end the chapter by discussing some fundamental aspects of fMRI data analysis (integration of anatomical priors, multi-subject studies, multi-modality integration).

9.1 Multi-session data analysis: reconciling state-space and information theory

9.1.1 Rewriting the equations

Let us consider the following situation: given a number S of repetitions of a certain dynamical system $X(t)$, endowed with a prediction model (4.37):

$$x(t) = \hat{x}(t) + \varepsilon(t) \tag{9.1}$$

$$\hat{x}(t) = \sum_{l=1}^{L} \alpha_l x(t-l) + \sum_{m=0}^{M} \sum_{c=1}^{C} \beta_m \gamma_c P_c(t-m) \tag{9.2}$$

One would like to estimate the parameters $\theta = ((\alpha), (\beta), (\gamma))$ of the model.

As in section 6.3.1, we propose to use a state-space representation of the data:

$$x(t) = \hat{x}(t) + v(t) \tag{9.3}$$

$$X(s,t) = M_s x(t) + w(s,t) \tag{9.4}$$

Where $\hat{x}(t)$ is defined as in equation (9.2), $v(t)$ is an innovation process, $w(t)$ an observation noise, M_s a $S \times 1$ mixing matrix.

Next, we drop the autoregressive term in equation (9.2), at least for two reasons. First, the hemodynamic model we consider is quite complete, so that there is no need for correcting it by an additional autoregressive term. Second, typically autoregressive patterns (oscillations, biologically rhythms) are supposed not to be reproducible for independently

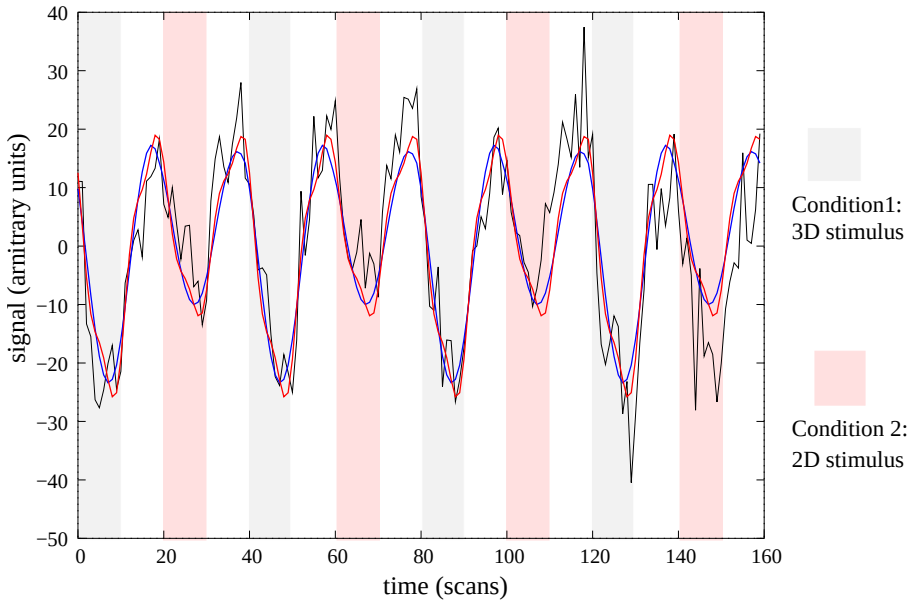


Figure 9.1: Extraction of a task-related component from the set of $S = 12$ time series displayed in figure 6.6.
The state vector (black) can be used to infer a task-related pattern (blue), but direct estimation of the task-related pattern from the data gives a better (more detailed) model of the response (red).

acquired data, and thus can be treated as noise. Since the goal of the state estimation is to derive optimally a task-related response using many data sessions, we will thus not further consider the autoregressive term in the equations.

A straightforward solution to the estimation problem would be to compute first the quantities involved in the state-space model (state vector, mixing matrix, transition matrix, noise and innovation covariance, see section 6.3.1), and then to infer the parameter θ from the state vector as in section 4.3.3. However, this is potentially suboptimal: the state estimation is not known to give a precise hemodynamic model, but rather a rough estimate based on an AR-1 scheme. As a consequence, some information of interest can be lost during the state vector estimation. In contrast, a direct estimation of θ from the multiple time series may yield a more precise model. This is illustrated in figure 9.1.

Therefore, we propose to estimate a task-related model from multi-session data by generalizing the complexity criterion developed in section 4.3.

9.1.2 The model

The equations of the model

$$\begin{aligned} \hat{x}(t) &= \sum_{m=0}^{M} \sum_{c=1}^{C} \beta_m \gamma_c P_c(t-m) && (9.5) \\ x(t) &= \hat{x}(t) + v(t) && (9.6) \\ X(s,t) &= M_s x(t) + w(s,t) && (9.7) \end{aligned}$$

are equivalent to

$$\begin{aligned} \hat{x}(t) &= \sum_{m=0}^{M} \sum_{c=1}^{C} \beta_m \gamma_c P_c(t-m) && (9.8) \\ X(s,t) &= M_s(\hat{x}(t) + v(t)) + w(t) && (9.9) \\ &= M_s \hat{x}(t) + w'(s,t) && (9.10) \end{aligned}$$

where $w'(s,t) = w(s,t) + M_s v(t)$ is a generalized residual of the model. As a consequence, the estimation of $(M_s), (\beta), (\gamma)$ is relatively simple: letting $X = (X(s,t))$ and $P = P_c(t-m)$ be the data and predictor matrices, performing the SVD of $(XP^T(PP^T)^{-1/2})$ yields a least squares solution. Letting

$$(U, \Sigma, D) = SVD(XP^T(PP^T)^{-1/2}), \tag{9.11}$$

the first column of D is nothing but the vector $\hat{x}(t)$ of $span(P_c(t-m))$ that contains most of the data variance. The first column of U provides the mixing vector. Note that this is very close to the CCA procedure, except that the data X has not been whitened (because the residuals $w'(s,t)$ of the model are assumed independent for different values of s). The amount of variance fitted by the model is the squared first eigenvalue σ_1^2. Note that the whitened predictor basis $P^T(PP^T)^{-1/2}$ can be replaced by any orthonormal basis that spans the same space, whose dimension is MC.

However, this simple procedure is not satisfactory:

- The ensuing task-related model $\hat{x}$ does not have exactly the desired form $\sum_{m=0}^{M} \sum_{c=1}^{C} \beta_m \gamma_c P_c(t-m)$. We need to apply a correction (see equation (4.49).)

- The model overfits the data: for example, a task-related model $\hat{x}$ is derived by this procedure even if there is no significant task-related component within the input data. Since our goal is to describe the generative process of the data, we need to make a distinction between the cases where the task-related pattern is plausible or not. We thus reintroduce complexity penalties as in section 4.3.

- Last, the signal space is large (its dimension being $M + C$). This induces the possibility of physiologically irrelevant patterns, and if no care is taken, of false positive. From this stems the necessity of priors on the hemodynamic model. We develop this in section 9.1.4.

9.1.3 A multi-session complexity criterion

Here we derive a criterion to control over-fitting in the derivation of $\hat{x}$. We use again the BIC-MDL heuristic and formalism (see chapter 4). As in 4.3.3, we denote the temporal regressors -i.e. a temporal basis that spans the same space as $P^T(PP^T)^{-1/2}$- by $\Omega = (\omega_i(t)), i = 1..MC$. Since the searched state-space is of dimension 1, we consider only the first column of U and D in (9.11). The first column of U yields the mixing model, while the first columns of D describes the state $\hat{X}$ in the basis Ω. The idea is to prune this description of $\hat{x}$ in order to keep only the significant coordinates. We denote these coordinates by $\delta_1, .., \delta_I$.

The amount of variance not fitted by the model is then $\sum_{s,t} X(s,t)^2 - \sigma_1^2 \sum_{i=1}^{I} \delta_i^2$. Using the notations of appendix D, this can be identified with the extensive entropy

$$H_0 = \frac{1}{2} \log \left(2\pi e \frac{\sum_{s,t} X(s,t)^2 - \sigma_1^2 \sum_{i=1}^{I} \delta_i^2}{ST} \right) \tag{9.12}$$

The parameters of the model are $\theta = ((M_s), (\delta_i))$, the number of independent parameters is $q = S + I - 1$, and the associated structural entropy is

$$H_1 = \frac{q}{2} \log ST \tag{9.13}$$

So that the multi-session complexity criterion is

$$\mathcal{C}(X, I, (\delta_i)) = STH_0 + H_1 = \frac{ST}{2} \log \left(2\pi e \frac{\sum_{s,t} X(s,t)^2 - \sigma_1^2 \sum_{i=1}^{I} \delta_i^2}{ST} \right) + \frac{S + I - 1}{2} \log ST \tag{9.14}$$

The minimization of this criterion over I is then carried out easily. The response model is then corrected by equation (4.49) (for our current datasets, this correction has mild effects). The result (red curve) displayed in figure 9.1 has been obtained by applying this multi-session criterion. One can observe that it models the hemodynamic response in greater detail than the two-steps (state estimation; parameter estimation) approach.

9.1.4 On priors

The need for more priors stems from the fact that the signal space is relatively high dimensional, so that the methods can yield many false positives. In fact, the space spanned by $P_c(t - m)$ or (ω_i) contains many *regions of no interest* (the word *region* does not refer to anatomical space but to signal space; see in figure 9.2 that some regressors of the basis are typically high-frequency, and thus physiologically not relevant). To introduce priors on the feature space, one can alternatively:

- Reduce the numbers of regressors (ω_i) by keeping only those that match a given criterion; for example, high frequency regressors can be suppressed.

- Add some spatial constraint in the model; for example, one can force the hemodynamic model to be uniform within some regions of the dataset.

S	1	2	3	4	5	7	10	12
α (%)	17.2	31.5	9.22	3.13	1.24	0.24	0.02	0.001

Table 9.1: Rate of false positives (in percent) obtained by minimum complexity description of synthetic data, as a function of the number S of sessions or realizations of the data. The rate is high for $S = 1$, due to the choice of $I = 4$ and the use of highly correlated data. Nevertheless, it drops rapidly when S increases. This means that minimum complexity description should be used with an adapted statistical test when only one or few sessions are available.

- Use a part of the data (e.g., half of the sessions) to estimate prior probabilities on the hemodynamic functions that represent the data, and use another part of the data to estimate posterior probabilities of the hemodynamic model (Bayesian approach).

The first method is easy to use. For example, taking the regressors (ω_i) from a SVD of $P_c(t-m)$, it is clear that the high frequency regressors are likely to fit noise, and can be given a low prior probability. At the risk of slightly biasing the estimate of the response, they can be canceled.

9.1.5 Minimum complexity and control of false positives

Clearly, modeling the data by a minimal complexity heuristic does not provide any control on the false positive rates of the method. This is true for the univariate model presented in 4.3 as well as the multi-session model presented in 9.1.1. For time series of length T, one can even show that the number of false positive -under white noise hypothesis- will be $\alpha \simeq 2(1-\Phi(\sqrt{\log(T)}))I$ where Φ is the gaussian cumulative density and I is the dimension of the basis that represents the activation space. For example, for $T = 130$, $\alpha = 0.0274.I$. Moreover this negative result worsens if there is autocorrelation within the data. The fact that the false positive rate increases linearly with I is an approximation, true for α small. Anyway, this shows why it is safer under general conditions to reduce the dimension of the signal space. However, this clearly shows that using a nontrivial signal space yields unacceptable false positive rates.

The solution to that problem is either the restriction of the activation space (with priors on the hemodynamic response, as in a Bayesian framework) or the use of an external criterion to deal with this problem. For example, repeatability of the activation pattern over several sessions is a potential way to overcome this problem.

We illustrate the solution on a synthetic example generated as follows: $x_{raw}(t) = b(t) + \varepsilon(t)$, where b and ϵ are a centered Brownian and white noise of equal variance, respectively. The data is then adaptatively filtered by $x_{fil}(t) = x_{raw}(t) - (x_{raw} * g_\sigma)(t)$ where g_σ is a gaussian filter of width $\sigma = 10$ scans. This process is replicated S times, S being the assumed number of sessions. Then a model is built assuming a block design with 10 scans long blocks. Similarly as in 9.1.4, the number of temporal regressors is reduced to $I = 4$ by keeping the smoothest regressors. Finally the S sessions of data are analyzed according to criterion (9.14). We study the fraction of false positives for different values of S, by simulating $N = 10^5$ such voxel-based sets of time courses. The result is given in table 9.1.

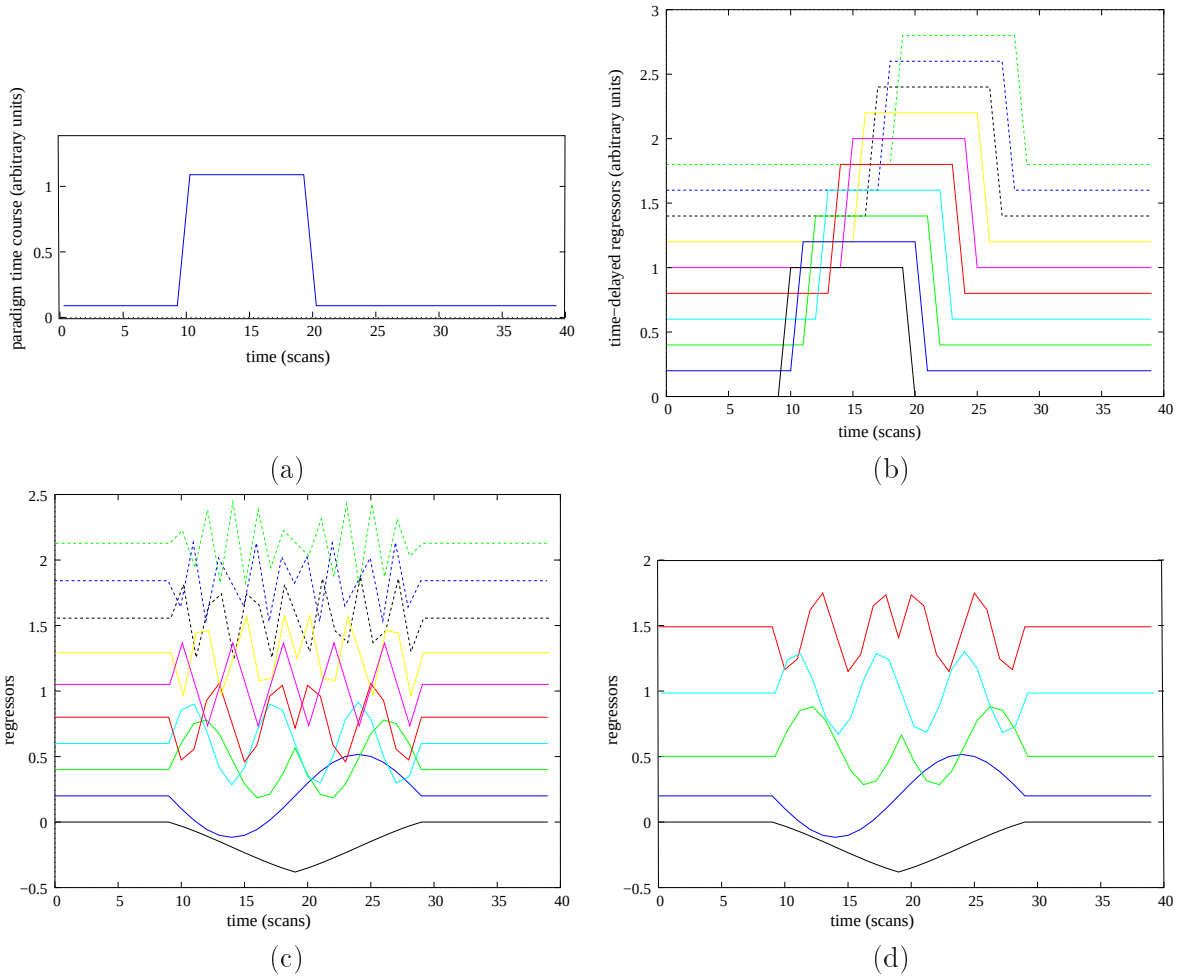


Figure 9.2: From the paradigm to temporal regressors.
This figure illustrates how an experimental paradigm can be converted to a temporal basis: (a) Basic OFF-ON-OFF sequence of a fictive stimulus condition $P_c(t)$. (b) Activation space obtained by using time-delayed versions of the experimental paradigm: $(P_c(t-m))_{m=1,..,M}$. (c) Orthonormal basis -noted $(\omega_i)_{i=1,..,I}$ in the text- constructed by SVD of $(P_c(t-m))$. Note the resemblance with a Fourier basis of $(P_c(t-m))$. (d) The first five vectors fit 99.93 % of the variance of the classical hemodynamic response function. The first four vectors fit 99.66 % of the variance. This means that the other vectors can be omitted to avoid data overfitting.

This simple simulation shows that the procedure yields low false positive rate for $S > 7$. This *miracle* results from the fact that the random occurrence of a certain pattern becomes less and less probable if one considers many realizations of the data. This observation is the reason why we present results obtained with this criterion on multi-session data. In the case of one or few sessions, an additional statistical test should be used to avoid false positives.

Let us insist on one point: the presence of false positives for S small should not be interpreted as an inconsistency in the model. Modeling by complexity minimization does not afford any control on the false positive rate. In fact, this is a feature shared with some recently proposed Bayesian approaches [87] [86] [170]. The current methodological trend in activation detection is to assess that the activation present on each voxel time series overcomes a certain level, in contrast with classical methods that reject the null hypothesis given a certain signal to noise ratio. The complexity approach is, with this respect, an asymptotic formulation of the Bayesian approach. The advantage is the computational cost and the easier specification of priors. The disadvantage is that the final description is less complete: it yields a kind of Maximum [likelihood] a Posteriori (MAP) model of the data, without confidence intervals.

9.2 Multi-session data analysis: a fully adaptable model

9.2.1 The general setting

In the preceding chapters, we have mainly considered fMRI datasets as (voxels $\times$ scans) matrices. A session dimension has now to be integrated within the model. Indeed, for the sake of SNR improvement, and given that repeatability is taken as the most convincing argument in favor of any predictive data model, fMRI datasets are often multi-session. We adapt here and generalize the signal model given in chapter 6. The general idea is the following:

- The task-related signal can be defined at the voxel level, since it represents the response of a given cortical region to the stimulation. This response should then be reproducible from session to session , though with a possible modulation in amplitude.

- By contrast, other dynamical signals within the datasets are either attributable to biological rhythms or to experimental setting, and are probably not reproduced from one session to the next; but they have probably multiple locations, so that they can be searched for within the entire dataset.

In terms of equation, this gives the following model

$$X^n(t,s) = \delta_s(n) \sum_{c=1}^{C} \sum_{l=1}^{L} \gamma_c(n)\beta_l(n)P_c(t-l) + \sum_{j=1}^{J(s)} \zeta_j(n,s)Y_j(t,s) + \varepsilon(n,t,s) \qquad (9.15)$$

where

- $P_c(t-l), c = 1..C, l = 1..L$ is the delayed time course of experimental condition c,

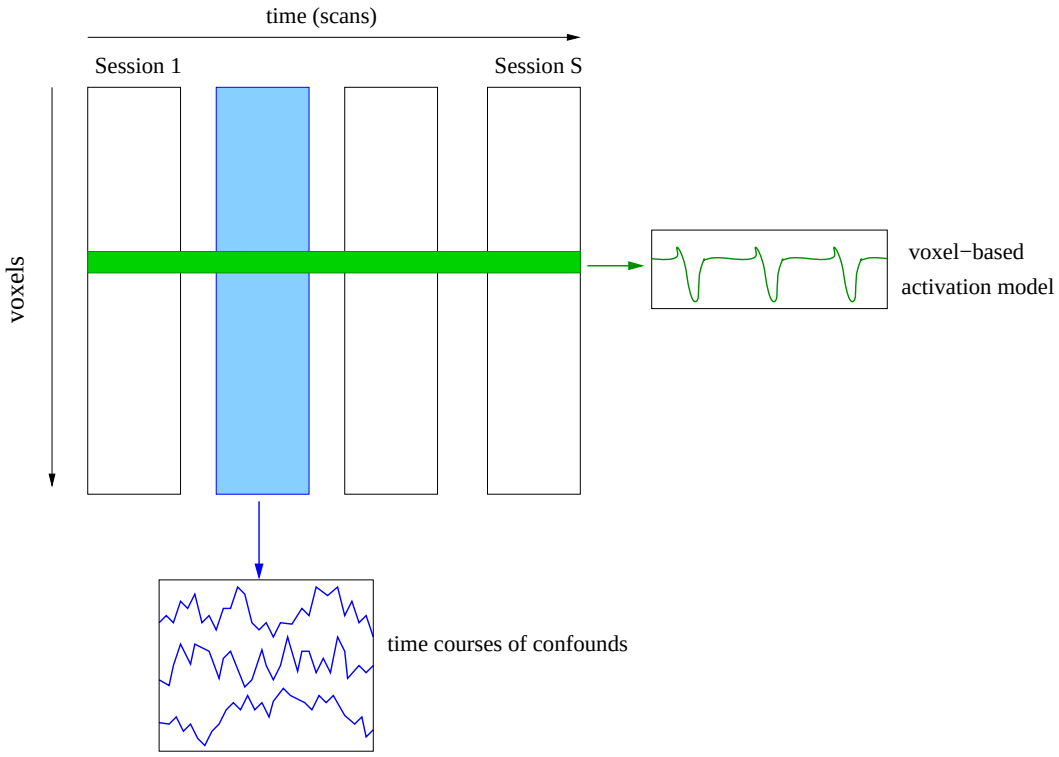


Figure 9.3: Graphical interpretation of the multi-session model.

- $(\gamma_c(n)), c = 1..C$ is the impact of condition c on the time course of the voxel n,
- $\beta_l(n), l = 1..L$ is the response filter at voxel n,
- $\delta_s(n)$ is a session effect on the response of voxels n to the stimulation
- $Y_j(t,s), j = 1,..,J(s)$ is a set of confounds during session s,
- $\zeta_j(n,s), j = 1,..,J(s)$ modulate the impact of the confounds on voxel n
- $\varepsilon(n,t,s)$ is the residual of the model

An advantage of this model with respect to previous versions ((6.31)-(6.32), (4.37)) is that the modeling of each voxel time course is simplified by the suppression of autoregressive terms. One can indeed think -and check empirically- that the autocorrelation of the voxel time courses is significantly decreased after session-wise removal of the main autocorrelated confounds of the dataset. A graphical interpretation of the model is given in figure 9.3.

We have to derive an estimation procedure. First, the multi-session voxel-wise task-related response estimation has been solved in section 9.1. There remains essentially to estimate the session-wise confounds. But this problem has received a simple solution with the state-space approach described in chapter 6. There only remains to connect both steps. Once again, we use an EM-like procedure:

- For each voxel n, estimate $(\gamma_c(n)), (\beta_l(n)), (\delta_s(n))$, given $(\zeta_j(n,s))$ and $(Y_j(t,s))$. From a theoretical viewpoint, the knowledge (and removal) of these confounds allows for the interpretation of $\varepsilon(n)$ as the true stochastic residual of the Wold model.
- For each session, estimate $(\zeta_j(n,s))$ and $(Y_j(t,s))$ given the individual voxel task-related responses. The latter are subtracted from the voxel time courses, and the state of the resulting dataset is estimated, with a purely autoregressive scheme, after adapted PCA reduction.

Though the convergence of this procedure is not guaranteed, we have obtained convergence after only one iteration on our datasets.

One might be concerned with the multiplicity of the task-related patterns obtained for a given dataset. We next propose to identify some structures within these multiple models: the Laplacian eigenmap seems perfectly suited for such a task.

9.2.2 Complete analysis of a synthetic dataset

In this section, we investigate the validity of the activity/confounds mixture model on a synthetic dataset. We address the following questions:

- Does confound estimation improve the detection of activated voxels ?
- Is our data-driven model more or less sensitive than a procedure based on the general linear model ?
- Is the estimation of the confounds accurate ?
- Does the method allow for a meaningful reconstruction/representation of the signal space ?

To simplify the analysis and the interpretation of our experiment, we use a very simple dataset: one slice of data, one session and one experimental condition of interest (the design being event-related). We proceed as follows. First, we describe the generative process for the synthetic dataset. Second, we make a detailed review of our analysis scheme. Third, we concentrate on the difficult problem of the statistical assessment in our data-driven framework -as shown in section 9.1.5, this issue remains unsolved with our usual procedure, and yields unacceptable false positive rates for mono-session data. Third, we describe our experimental results by answering the four questions above. We conclude this section with a discussion of our algorithms and the validity of this synthetic experiment.

The dataset

We have created a synthetic dataset by simulating one slice of fMRI data containing $N = 1963$ brain voxels. The length of the series is $T = 200$, and the simulated sampling time is $TR = 2s$.

The simulated paradigm comprises one condition of interest in an event-related design. Three small activation foci of 21 voxels are created; the activation time courses are obtained by convolution of the experimental condition time courses with three slightly distinct

hemodynamic filters $h_1(t), h_2(t), h_3(t)$. h_1 is simply the canonical model of SPM sampled at TR=2s [77]. $h_2(t) = h_1(t+1)$ mimics a time-shifted activation pattern, and $h_3(t) = \frac{1}{2}(h_1(t-1) + h_1(t+1))$ presents a blurred response with respect to $h_1(t)$. The activation time courses for each focus f is obtained as $a_f(t) = [h_f * P](t)$. The experimental design, and the three activation models are represented in figure 9.4 (a). The spatial organization of the foci is given in figure 9.4 (b).

Then three independent Brownian motions $B(t) = (B_1, B_2, B_3)(t)$ are simulated, and mixed into the data with a random gaussian mixture matrix. They are represented in figure 9.4 (c). Their amplitude is slightly higher than that of the task-related signal. Finally, an i.i.d. gaussian noise is added to all voxels.

Let us insist on the following features:

- The temporal model of the data seems quite realistic, if one accepts that the temporal correlation of the data can be modeled by a few sources. If yes, Brownian noise is a good candidate, since it models incremental processes, e.g. body motion. The hemodynamic model deviates mildly from the known models. Note that it can be written exactly as equation (9.15), with $\delta_s(n)$ being the characteristic function of the foci of activation. $C = 1$, $(\beta(n)) = h(n)$, $S = 1$, $J(1) = 3$, $\zeta_j(n) \sim \mathcal{N}(0,1)$ and $Y_j = B$.

- Conversely, the spatial model is not realistic, since no spatial correlation is taken into account. The reason is that we do not use it in either the General Linear Model or in our analysis framework. In particular, due to the weak number of voxels, we use uncorrected P-values.

- The hemodynamic filter space is 3-dimensional.

- The choice of 2D data is only for the sake of visualization.

Analysis scheme

The dataset is analyzed with two different methods: the G.L.M. as implemented in the SPM99 Software, and our procedure described in 9.2.1.

In the detail of the SPM analysis, we have used a model comprising:

- The specified experimental paradigm with the standard hemodynamic response function and its time derivative. This is assumed to correct for the possible shifts in the response delays.

- The data is high-passed with a cut-off period of $\tau = 80s$ (note that this cancels most of the Brownian deviation).

- The data is low-passed by convolution with the standard hrf.

- Additional correlations are not integrated within the model.

- Statistical inference consists in a F-test based on the two regressors of interest. The P-value is 10^{-3} uncorrected.

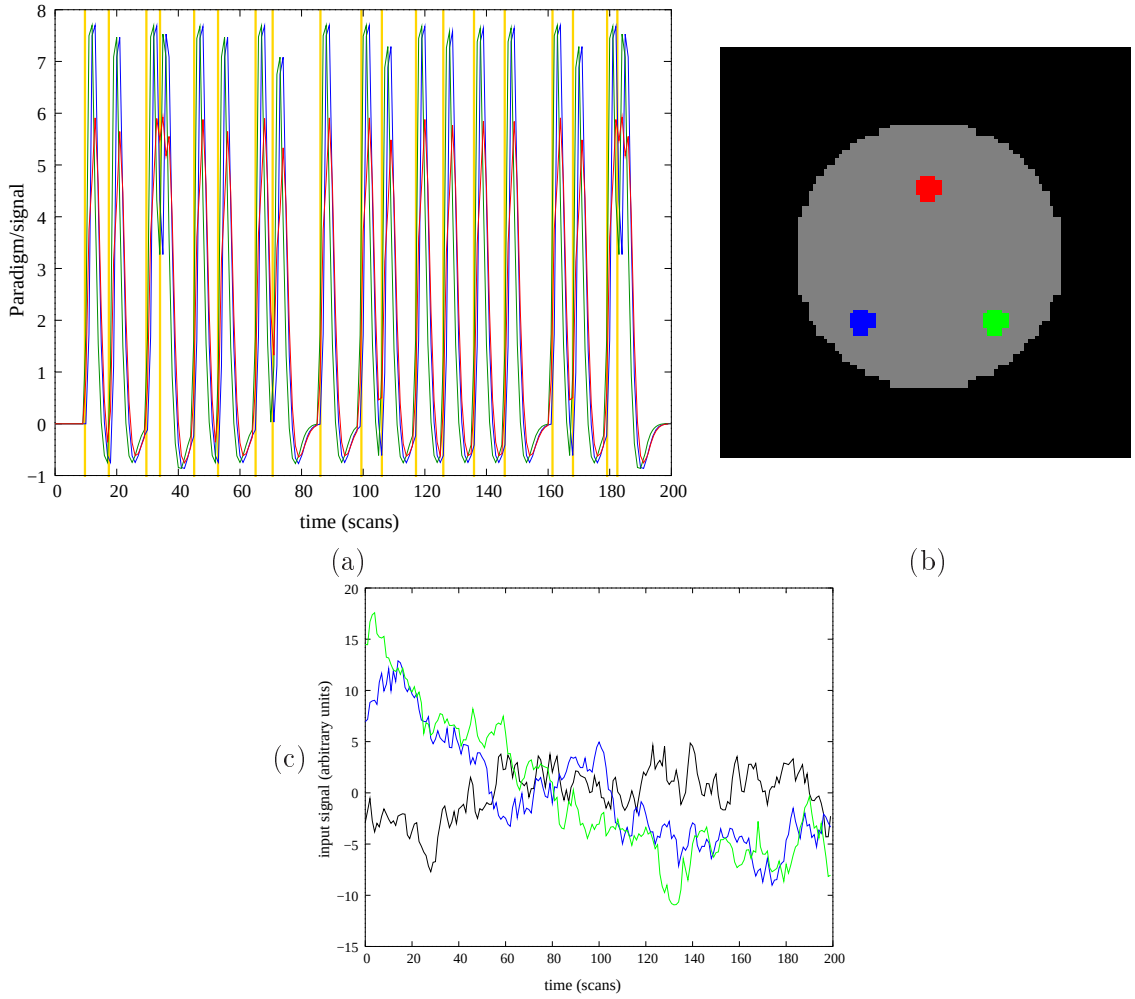


Figure 9.4: Description of the synthetic dataset built and analyzed in 9.2.2.
(a) Simulated experimental paradigm (vertical bars) and activation patterns in the synthetic dataset. (b) Spatial layout of the activations simulated in the experiment. The colors of figures (a) and (b) match. (c) Brownian confounds added to the data by random mixing.

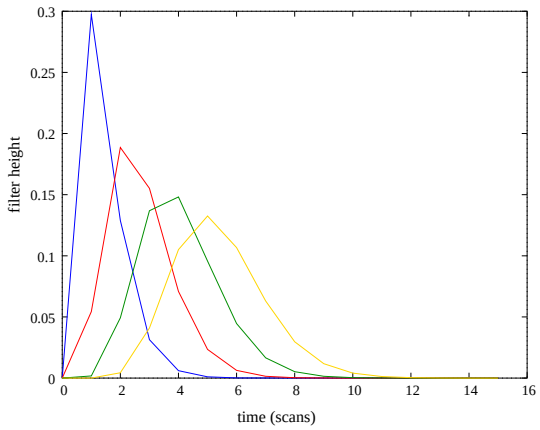


Figure 9.5: Time courses of the four filters used for activation detection in the experiment 9.2.2.
These 4 filters are obtained with gamma density functions with parameters (3,0.9), (6,0.9), (9,0.9) and (12,0.9) respectively, sampled at a 2s rate.

One can notice that the specified analysis model matches very well the hypotheses used in the generation of the data, so that SPM results can then be considered almost as a gold standard for any analysis scheme.

Our analysis procedure follows closely the scheme described in 9.2.1. Let us be more precise:

- The data is first detrended by non-linear gaussian fitting (see equation (2.2)). The width parameter is $\sigma = 10$ scans, which does not bias the activation patterns. This step is not conjoint with activation detection.

- Voxel-based activation patterns are then estimated using the complexity criterion (9.14). In order to avoid over-fitting, we use a 4-dimensional basis built with gamma functions of different latencies. This basis is represented in figure 9.5.

- The confound estimation is based upon a dimension reduction of the dataset to 40 components by PCA.

- After convergence of the joint confounds/activation estimation procedure, we obtain many voxel-based temporal patterns, many of which are not statistically significant. This is due to the fact that our analysis is based on many regressors, and only one session. Thus we add a probabilistic selection procedure based on permutation in order to keep the patterns whose amplitude is significant at $P = 10^{-3}$ uncorrected. This procedure is described next.

- Last, the surviving voxel-based signals are classified with a Laplacian eigenmap procedure.

Probabilistic selection

Our procedure for probabilistic assessment of the response amplitude is exactly the procedure described in [185]: we consider Q random permutations of the experimental paradigm (which is nothing but a binary variable). We derive the amplitude of each voxel-based response after the same complexity analysis and confounds removal. This yields an empirical voxel-based density of the amplitude of null responses. Finally we retain all the voxels that have a task-related response above $P = 10^{-3}$. We used $Q = 10^4$.

Results

We order the presentation of the results according to the four questions formulated at the beginning of the section:

a. Influence of confound estimation on the detection of activated voxels We have performed the previous analysis on the dataset with and without confounds estimation. One can notice that, at the significance P-value considered, there is no difference. Among 63 truly activated voxels, 54 and 55 are identified by the procedure, and there are 2 false positives in both cases. The map of detected voxels is presented in figure 9.6 (left).

b. Comparison with the SPM procedure We show our results together with the SPM results in figure 9.6. Both methods work well, with a good detection rate (8 false negative for our method, 2 for SPM), and reasonably few false positives (2 and 6). Once can notice that our testing method achieves the nominal rate of errors, in contrast with SPM. This is however at the expense of more false negatives. Both results could be enhanced by the use of spatial information, but in that case, one would also need to use spatially structured confounds in the simulation, which has not been done.

c. Accuracy of confound estimation The estimated confound space is of dimension 3, which is the number of Brownian components in our generative procedure. Due to the detrending/high-pass filtering procedure before the analysis, it is not possible to compare the input Brownian components and the results. Therefore, we detrend the Brownian inputs as in equation (2.2) and compare the resulting components with our confound estimations. The canonical correlations obtained between the two sets are respectively $(0.8992, 0.7915, 0.0258)$. This means that two nuisance components have been correctly identified, while the third results from a false identification. By looking at this component, we observed that it was a residual of the unfitted activation signal. Hence, it results from a slight misspecification of the activation space.

d. Resulting signal maps The Laplacian eigenvalues for the data embedding is displayed in figure 9.7 (top left). It clearly shows that the structure of the dataset is intrinsically two-dimensional. A quick look on the feature space (see figure 9.7 (top right)) shows the 3-modal structure of the feature space. The first 30 time samples of time courses associated with each of the three clusters are given in figure 9.7 (middle left), and the

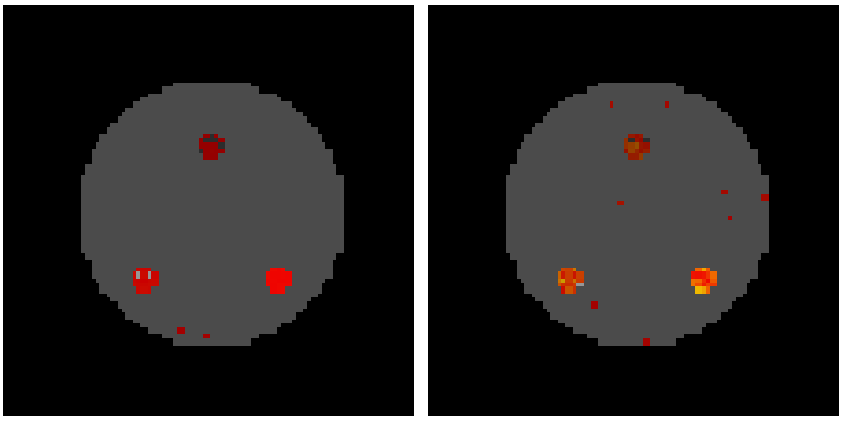


Figure 9.6: Spatial activation maps obtained from our method (left) and the standard SPM procedure (right).
The supra-threshold voxels are displayed in red, the threshold being set to a P-value of 10^{-3}, uncorrected in both cases.

spatial maps in figure 9.7 (middle right). For comparison, the first 30 time samples of the three input signals are given in figure 9.7 (bottom).

Discussion

The fact that the confound estimation does not help much for voxel detection is attributable to the choice of assessing the results through empirically derived null distribution of the data. Indeed, the empirical distribution, unlike an analytical one, adapts itself to nuisance present within the data. Moreover, this can be attributable to an over-simplistic definition of the structure of temporal correlations. For example, the autocorrelation attributable to confounds is removed by detrending/high pass filtering of the data, so that the remaining autocorrelation is weak.

Second, the comparison between SPM and our data-driven procedure gives qualitatively equally good results, with SPM being more sensitive and our method slightly more robust. It should be emphasized that SPM had been specified *the* correct linear model for this data; such a model is not available in practice ! Our main conclusion here is that the additional difficulty of estimating the temporal effects does not affect the detection power of our method.

Third, we have introduced a useful non-parametric assessment procedure, based on random permutations of the experimental paradigm. It was necessary here, because complexity analysis of mono-session data would yield over-fitting in general (see table 9.1). The principle limit of this method is the computational cost (for the data presented the computation time was 30 minutes on a PC, with $Q = 10^4$ iterations). Note that in [185], the authors use the method with fewer iterations. Finally, computation can be speeded up by considering only the voxels for which the complexity criterion has identified a component of interest. Anyway, we acknowledge the usefulness and the precision of this method.

Fourth, confound estimation has correctly identified two of the three input Brownian

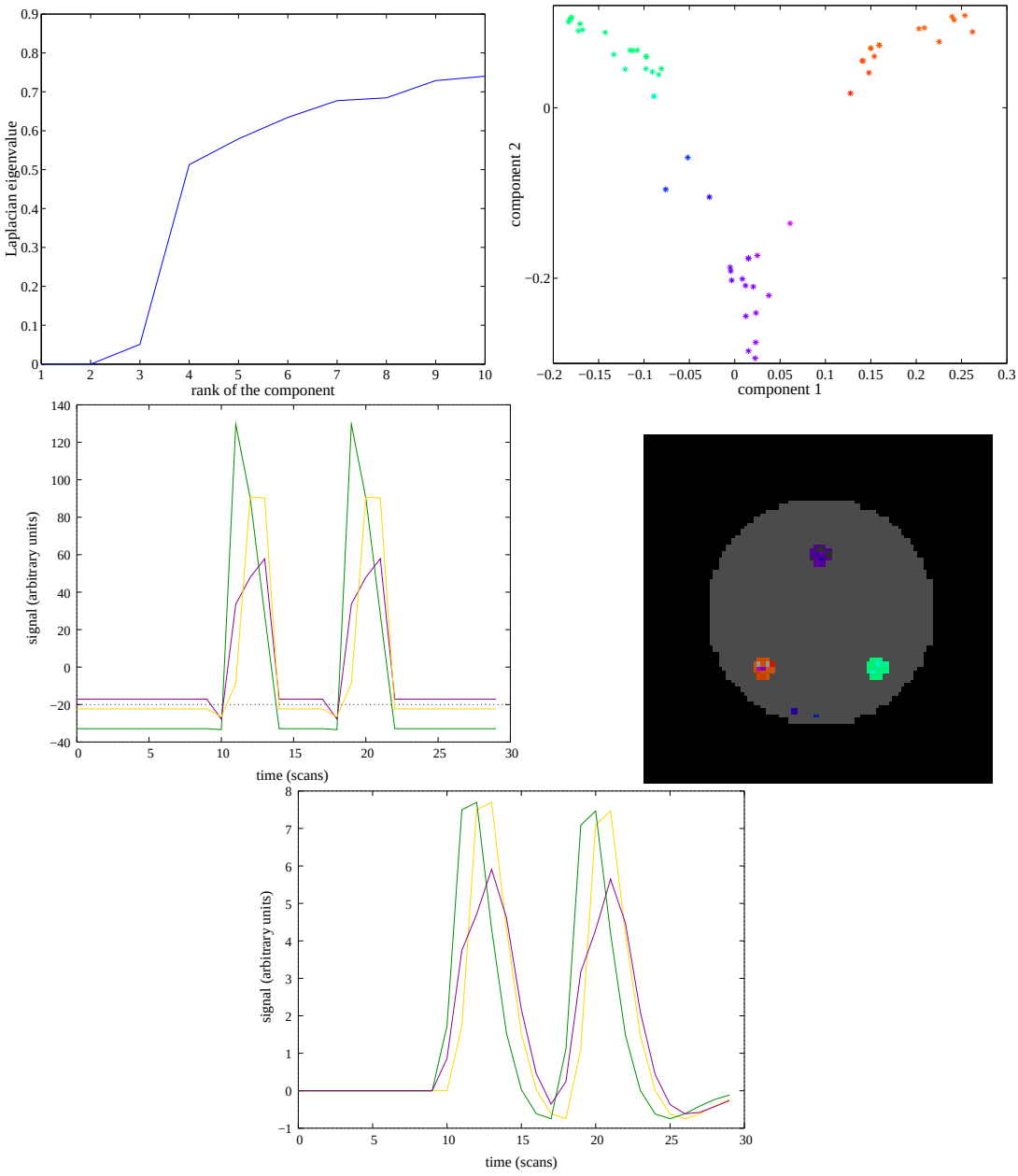


Figure 9.7: Results of the analysis of the synthetic dataset.
(left) Laplacian eigenvalues obtained in the analysis of the signal space resulting from the analysis. This clearly indicates that 2 dimensions are pertinent in the description of the dataset. (right) Resulting feature space. Note the three-modal structure of that space. This seems to fit the generative process of the data. The 30 first samples time courses (middle left) and spatial maps (middle right) indicate that the main features of the input data have been identified by the algorithm. (bottom) "True" (i.e. input) time series for the corresponding region are given for comparison with the estimated ones.

session	1	2	3	4	5	6	7	8	9	10	11
K	6	5	6	4	8	6	5	5	5	6	5

Table 9.2: Estimation of the dimension of the confound space for dataset 1. This is obtained with the procedure described in section 6.2.2.

components. While one Brownian component was not identified, our procedure has identified a fraction of the signal space that had not been modeled correctly in our hemodynamic response space. This indicates that it is always worth looking carefully at the structured parts of a dataset, even if they are not interesting *a priori*.

Last, the Laplacian embedding method gives an accurate and simplified representation of the data, solving the difficult issue of building a global model from multivariate, scattered information pieces.

9.2.3 Real dataset 1

The integrated approach has been applied to the 11 sessions of the dataset A.2; here $T = 120$, $S = 11$ and $N = 12320$. The results of this experiment comprise two parts: first, a description of the confounds in terms of session-wise models, second a description of the signal space using Laplacian eigenmaps.

The confounds Concerning the confounds, the information is twofold: first, the session-wise dimension of the confound space, and then the corresponding time courses. The dimensions are given in table 9.2. Let us notice that the session datasets are reduced to 40 components by PCA prior to confounds estimation.

A probably more instructive way to consider the effect of confound removal is to assess its effect on the autocorrelation of the voxel time series. We have computed the coefficient of order 1 of each voxel time course

$$\rho(n) = \frac{\mathbb{E}(X^n(t)X^n(t-1))}{\mathbb{E}([X^n]^2)} \tag{9.16}$$

The average value of ρ over the dataset is 0.1004 and 0.0414 before and after removal of the confounds. This is not anecdotal, since $\rho < 0.09$ with $P = 0.999$ under the white residual hypothesis. The effect of confound removal is illustrated in figure 9.8. The fact that the autocorrelation level is considerably reduced on most voxel time courses suggests that the estimation of task-related effects is less biased.

The signal space An optimal representation of the signal space is derived, similarly as in 8.3.2. The number of input voxels is higher now ($N = 1572$ instead of 1320), since the characterization of confounds allows for a less conservative detection of activated voxels. The sequence of Laplacian eigenvalues of the data embedding is given in 9.9 (left); they do not yield an evident dimension for the representation of the dataset; however the second and third eigenvalues are clearly lower than the others; a two dimensional embedding seems thus adequate; we use once again a 2-dimensional embedding. The cluster of points,

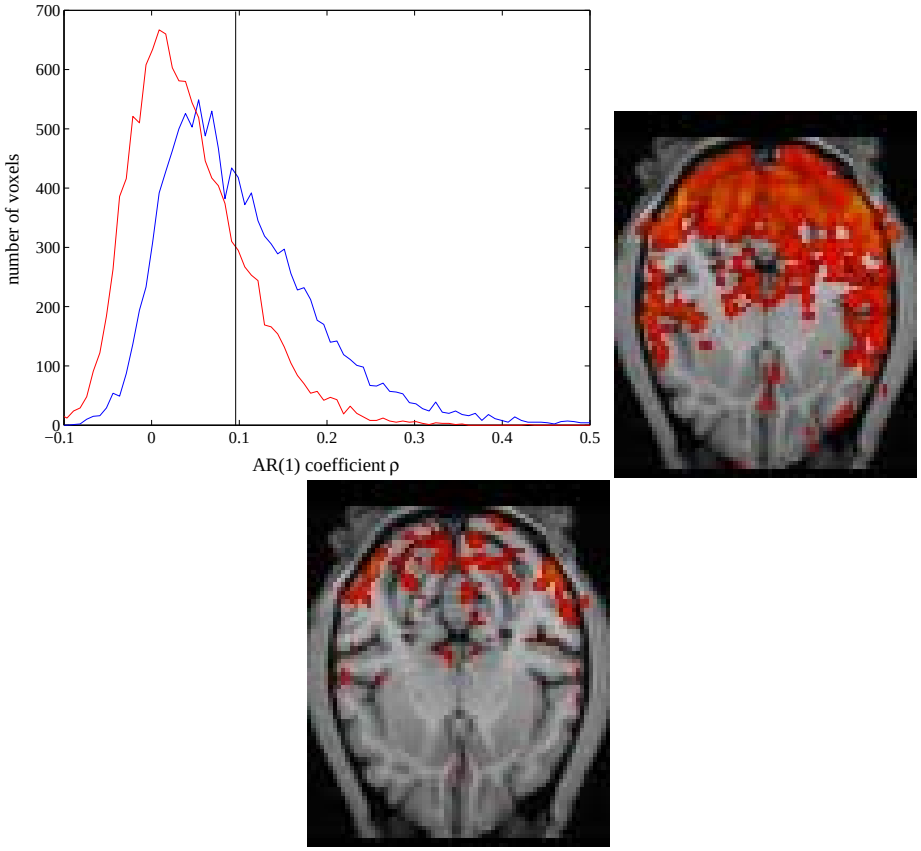


Figure 9.8: Removal of confounds and reduction of autocorrelation
(left) Histogram of the autoregressive coefficient of order 1 of the residuals throughout the dataset analyzed in 9.2.3 before (blue) and after (red) removal of the confounds. The vertical bar indicates the confidence value $\bar{\rho}$ so that $|\rho| < \bar{\rho}$ with $P = 1 - 10^{-3}$. (middle and right) Map of the ρ coefficient on an axial slice, thresholded at the significance level $\bar{\rho}$, before and after confound removal. These maps illustrate also that autoregressive patterns are anatomically organized throughout the cortex.

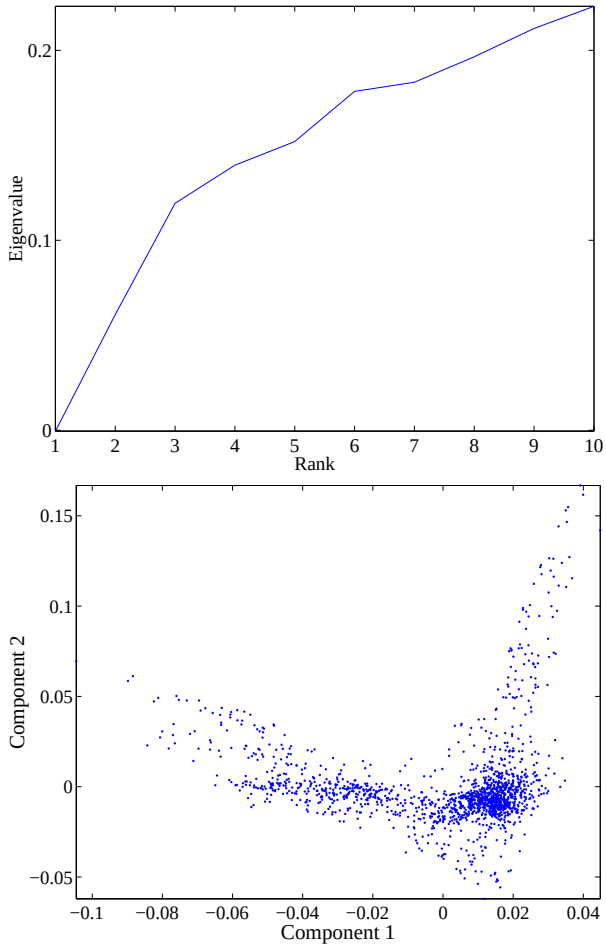


Figure 9.9: Integrated analysis of dataset A.2: results
(left) Sequence of eigenvalues obtained in equation (8.3). Note that the eigenvalues can be interpreted as a measure of distortion. As could be expected, the first eigenvalue is 0 (and the associated map is uninformative). For further study, we keep the second and third eigenvalues. (right) Optimal 2D representation obtained for the dataset. Note that the general structure is quite similar to figure 8.7 (a), though minor changes are visible.

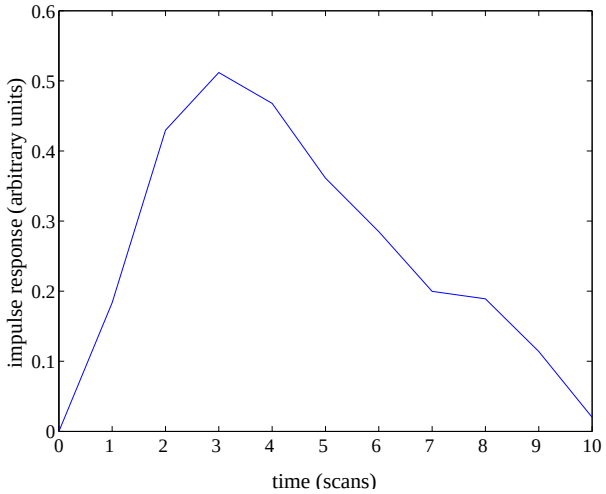


Figure 9.10: Average impulse response to the experimental task within the dataset. Note that the interscan interval is 2.976 s, so that the temporal resolution of this model is rather coarse. The response peaks roughly 9s after the beginning of stimulation. Note that this is *not* a BOLD response, but a CBV response.

session	1	2	3	4	5	6	7	8	9	10	11	12
K	7	7	4	5	7	6	6	5	6	9	7	6

Table 9.3: Estimation of the dimension of the confound space for dataset 2. This is obtained with the procedure described in section 6.2.2.

displayed in 9.9 (right), is much similar as the cluster obtained in section 8.3.2, and so is the interpretation of the feature distribution, as well as spatial maps.

Our analysis can also provide an estimation of the average impulse response in the dataset. Given the average response pattern, it suffices to find the coefficients that generated this model. It is of interest since this response is not the standard hemodynamic response, due to the use of a contrast agent. The shape of the actual response is given in figure 9.10. Of course, this model has to be taken with caution, due to the variability within the dataset.

9.2.4 Real dataset 2

We have performed the same processing on the dataset presented in A.3. The dataset is resampled and coregistered onto the anatomical model so that we have $N = 77968$. Moreover, $T = 160$ and $S = 12$. Confound estimation has been performed on each session after PCA reduction to 40 components. The dimension of the confound space for each session is given in table 9.3. It varies from 4 to 9.

More importantly, the removal of confounds has a filtering effect on the data. As an illustration, we plot the histogram of the empirical AR(1) coefficient ρ of each time course within the dataset in figure 9.11. It appears that confound removal reduces the

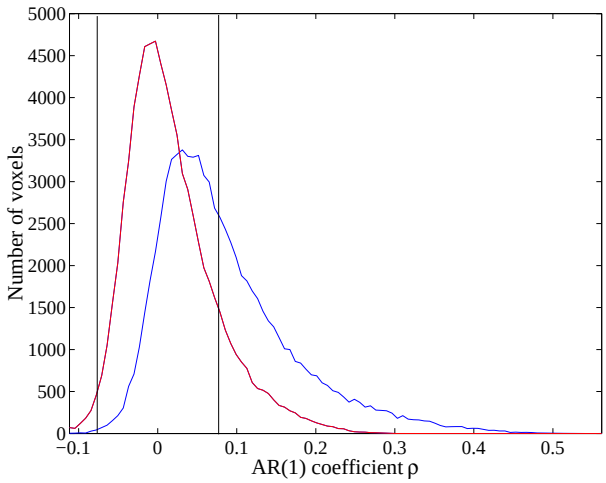


Figure 9.11: Reduction of autocorrelation by confounds removal
Histogram of the autoregressive coefficient of order 1 of the residuals throughout the dataset analyzed in 9.2.4 before (blue) and after (red) removal of the confounds. The vertical bars indicate the confidence values $\bar{\rho}$ so that $|\rho| < \bar{\rho}$ with $P = 1 - 10^3$.

value of ρ almost everywhere, shifting the global distribution towards the null hypothesis distribution (even though the tail for $\rho > 0$ remains too strong, so that the correction is not complete).

The number of *activated voxels* (according to criterion (9.14)) is $N' = 1148$ before confound removal and $N' = 1312$ after confound removal. Laplacian eigenmaps are then derived for the description of the signal space. We have chosen the model with $k = 10$ neighbors. As often with empirical data, the distribution of Laplacian eigenvalues does not give very precise information on the true dimensionality of the data -if such a thing exists; this can be checked in figure 9.12.

There is no clear-cut distinction between low and high eigenvalues. We have worked on a 3D embedding of the dataset. However, the first component of this embedding was dominated by a component of no interest, i.e. a component that belonged to the activation space, but could not be interpreted as a realistic hemodynamic response (it did not contain the fundamental frequency of each stimulus, but the first harmonic). We interpret this component as an oscillating confound present within the dataset. For this reason, we study next the embedding obtained with the second and third non-trivial components.

The 2D feature space shown in figure 9.13 (left) is structured as a dense cluster with more scattered components. To describe it more easily, we use a color encoding of the features, the color being defined by the direction in the 2D feature space. Then, we derive the average time course in the different regions of the feature space and display them in figure 9.13 (right). The eight resulting temporal patterns seem to be with respect to the delay in the response and with respect to the relative impact of the control condition (2D visual stimulation). Roughly speaking, the axis of the feature space labeled *component 2* modulates the relative impact of the experimental conditions, while the axis labeled *component 3* modulates the delay in the response: the orange and yellow curves are selec-

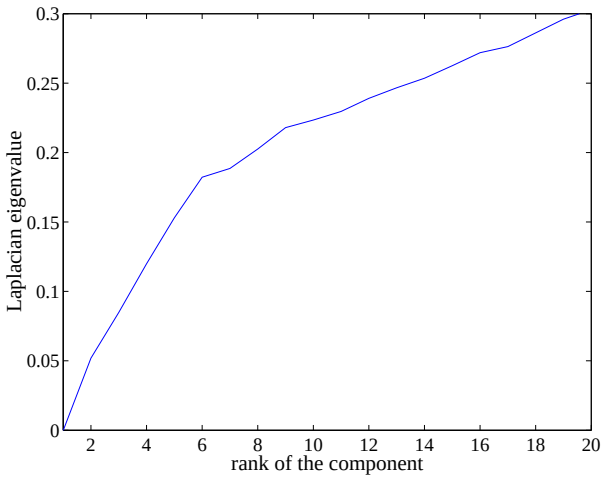


Figure 9.12: Sequence of eigenvalues obtained in equation (8.3) for the second real dataset. Note that the eigenvalues can be interpreted as a measure of distortion . As could be expected, the first eigenvalue is 0 (and the associated map is uninformative). For further study, we keep the second, third and fourth eigenvalues.

tively activated by the motion condition, while the blue and green ones show equivalent contribution of both conditions; the purple and red components show earlier responses than the green, yellow and orange patterns.

Then, the voxels selected within the signal classification process can receive a label according to their position within the feature space. This yields the maps displayed in figure 9.13 (bottom). Note that less than 2% of the voxels are color coded, which explains the scattered aspect of the map. Surprisingly, while no spatial information has been introduced in the method, the maps show an homogeneous and quite symmetrical cluster organization.

To give an account of the spatial location of the activations on the cortical surface, we display in figure 9.14 a projection of the map 2 onto the cortical surface. The maxima of this map correspond to the regions where the activity for the 2D stimulus is weaker, i.e. where the 3D-2D contrast is higher.

9.2.5 A conclusion on data modeling

We finally propose the model (9.15)- see also figure 9.3- as the most general one of this thesis for fMRI data modeling: it is based on local estimates of task-related response and global estimation of the confounds or nuisance space. The multi-session procedure for the estimation of the task-related response can also be replaced by the joint study of one voxel and its neighbors, as proposed in section 6.3.2. Otherwise, an adapted statistical procedure (as in section 9.2.2) has to be used to avoid the presence of too many false positives. This model is completed by the local-to-global step, i.e. the Laplacian embedding technique that builds a data manifold from the empirical data.

Once again, let us outline that this model results from the compromise between flexibility (let the data define the main task-related patterns present within the dataset) and

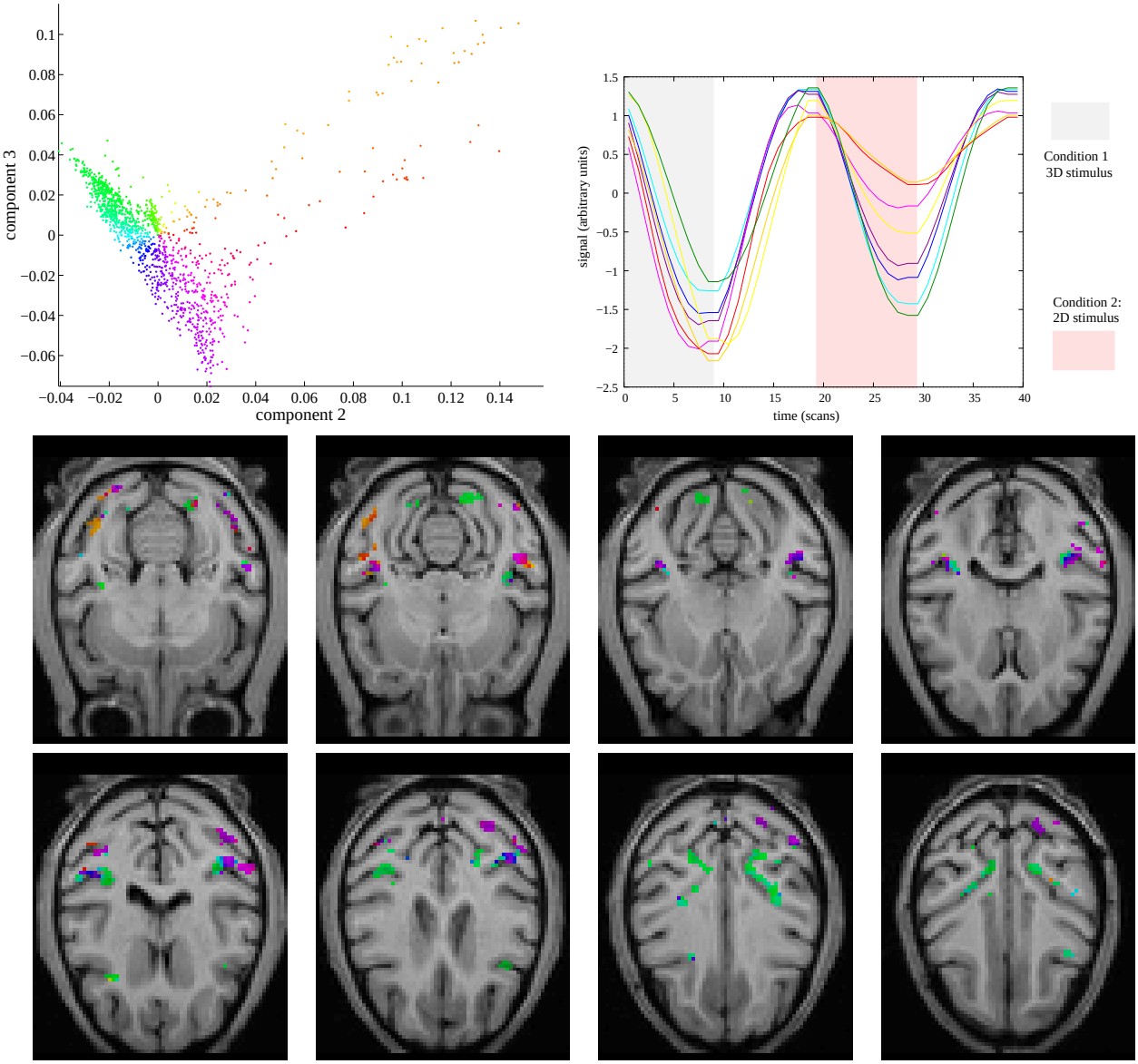


Figure 9.13: Integrated analysis of the dataset A.3: results.
(left) 2D representation obtained for the dataset described in 9.2.4 with components 2 and 3 of the dataset. (right) Time courses representing the different regions of the manifold, color coded as on the left. Note the differences in activation timing and in the relative impact of the two experimental conditions. For clarity, we limit the temporal window to one period of the stimulation, since the signals are periodic.(bottom) Spatial representation of data embedding on eight axial slices. The color code is identical to the code of the two upper figures.

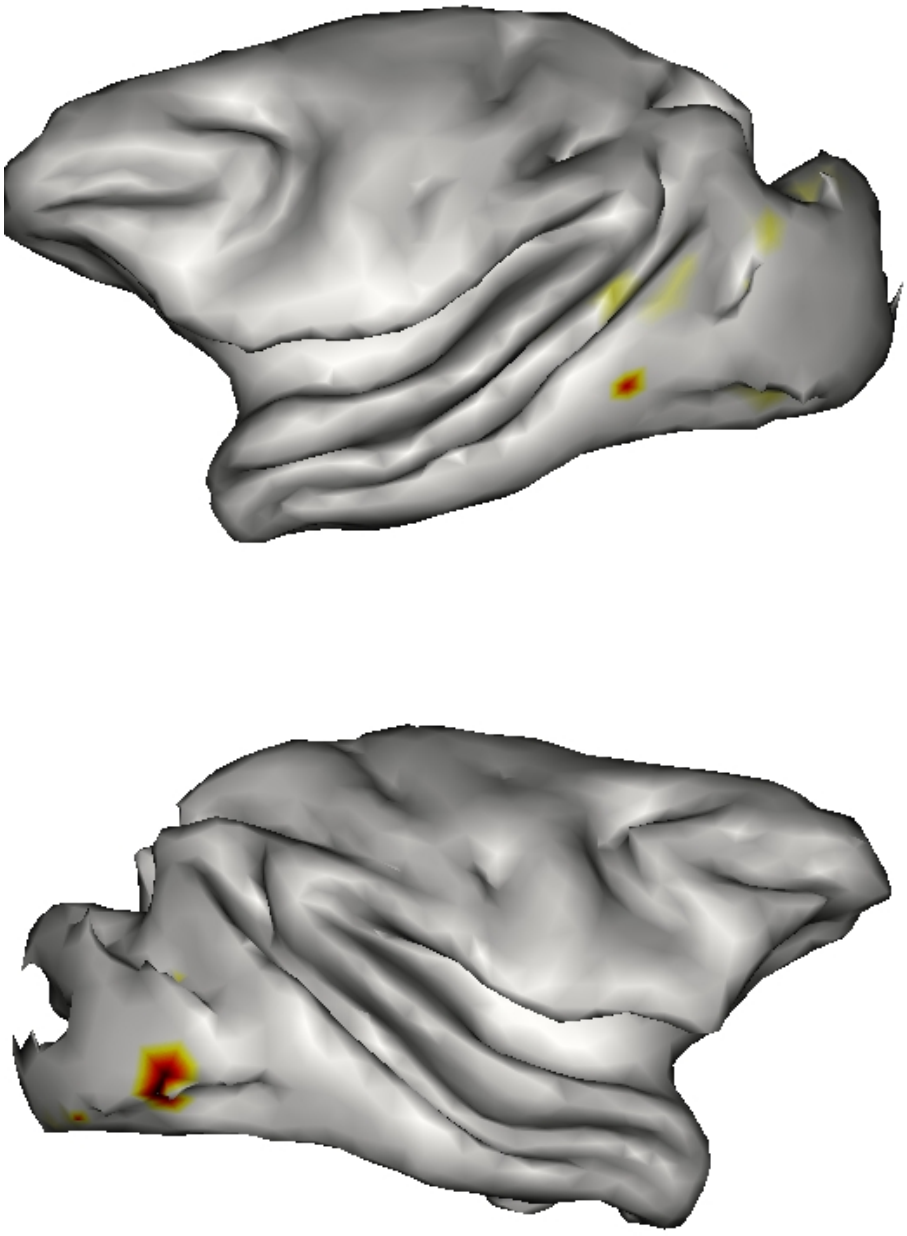


Figure 9.14: Projection of map 2 onto the left and right hemispheres of the grey-white interface.
The areas red/black correspond to the main foci of 3D stimulation selectively. Note that the maps are not symmetric, with the right hemisphere showing a stronger focus of activation. Moreover, given the difficulty of obtaining the 3D grey/white interface, some activations cannot be displayed properly on this surface.

pertinence (use the information of the experimental paradigm). To complete the picture, a study of the main confounds (non task-related structured patterns) could be interesting.

9.3 Generalization of our work

We finish this chapter with the discussion of two possible extensions of our work. The first one is the introduction of prior information in inference procedure, which is one of the goal of neuroimaging methodology; second, we discuss the extension of our estimation and exploration procedures to the case of multi-subject studies.

9.3.1 On inference

If signal estimation is one of the major concerns of fMRI data analysis, inference, i.e. the validation of biological hypotheses on the data is the ultimate goal of analysis. We thus show here how our analyses scheme can be used in inferential procedures. Since the work presented in this chapter is devoted to the joint estimation of task-related components and confounds, we can then derive a design matrix for the experiment and use it in an inferential procedure. In the case of task-related signal, a minimized model is necessary for numerical concerns, so that we limit ourselves to estimating a dataset-derived impulse response function (hrf) and then uniform task-related regressors from this model.

One can question the validity of this approach: is it relevant to include in a design matrix a set of data-driven confounds? We believe that the answer is yes. Indeed, the confounds have been defined as the non-white and non task-related components of the data, i.e. the components that deviate from the implicit assumption in noise definition: whiteness or non-predictability. The *removal* of these components has been shown in the previous section to correct -at least partially- the autocorrelation of the time series, which in the general case is far from its null distribution. We conjecture that removing these effects suppresses non-stochastic components of the signal, and thus improves the adequacy of the model. The advantage of this procedure is that, assuming that the time series autocorrelation is essentially canceled, no whitening or pre-coloring procedure is necessary for the estimation of the noise variance.

We study this procedure on the dataset A.3, by comparing two instances of the general linear model. In both cases, regressors are derived by convolution of the task function with an empirically derived impulse function. The data is high-pass filtered by the standard SPM procedure (removal of all frequencies below the stimulation fundamental frequency).

Case 1: The data is additionally low-pass filtered with an empirically derived hrf described in [218]. Some regressors are added to the design matrix: SPM-derived body motion estimates, and eye motion of the monkey, which had been recorded during the acquisition.

Case 2: We include in the design matrix (besides the task-related regressors) the regressors associated with the confounds derived in 9.2.4, whose rank per session is displayed in table 9.3. These regressors may be viewed as data-driven estimates of the confounds, in contrast with case 1, where confounds estimates are from exterior observations (monitoring).

The map of the contrast of interest (stimulus 1-stimulus2) is presented in figure 9.15. Note that the maxima for the two cases are $t = 9.69$ and $t = 12.37$ respectively. The thresholds used are identical for both maps: $t = 5.09$ corresponds to $P = 0.05$, corrected for multiple comparisons, from SPM99.

Discussion From the observation of figure 9.15, one can conclude that the differences between the resulting maps are not very important: the supra-thresholds regions are similarly located. The data-driven procedure has simply higher scores, and slightly wider regions. Once can straightforwardly conclude that this method is more sensitive. Interestingly, the increase in sensitivity does not seem to have induced false positives, at least at the threshold level considered. The difference in sensitivity is attributable to different estimations of the variance: in the *standard* case, it is likely that body motion and eye motion estimated do not include all the potential correlations of the time series. This is corrected by performing low-pass filtering of the data. The second procedure removes the main *sources* of autocorrelation of the data and takes the residual as the variance; in practice, the variance estimate seems to be smaller than in the first case. From our observation of the empirical scores distribution (symmetry of the activation patterns), this more optimistic estimation is not undue.

9.3.2 On multi-subject studies

Though we do not explicitly deal with multi-session data in this work, it seems important not to overlook completely the problem. In fact, meaningful neurophysiological conclusions can be drawn only from multi-subject studies, since one can expect that a particular subject may not execute well the task, use alternative strategies, or have some particularities that modulate the structure of the data or the response level. Technically, this problem splits into two distinct ones: *i)* Assuming that all the subjects are coregistered anatomically and *ii)* Avoiding that assumption. The first case is the most frequent one, since it allows for easier inference. However, it is plagued by the difficulty of obtaining a correct anatomical registration (an issue that we will not discuss here; some solutions for the robustness against mis-registration are given in [67]). Our contribution here is simply to suggest how one can generalize our estimation procedures to that case.

Coregistered case

We are in the following situation: one dataset $X(s,t)$ is available for each subject s, a dataset being the now familiar $N \times T$ matrix. In fact this situation can be modeled exactly by equation (9.15), where the index s stands for different subjects instead of different sessions. Each voxel is then associated with a given (eventually null) response and hemodynamic model, while some confounds are present within each dataset independently. The hypothesis of an identical hemodynamic response from subject to subject may seem too optimistic, but it seems necessary in practice for making inference on a multi-subject level. Then all the estimation procedures described in 9.2.1 applies - of course this implies that the experiment is reproduced exactly.

One might be concerned with the fact that *several* sessions are acquired for each subject, so that the dataset is indexed by voxel $n \in [1,..,N]$, subject $s \in [1,..,S]$, time

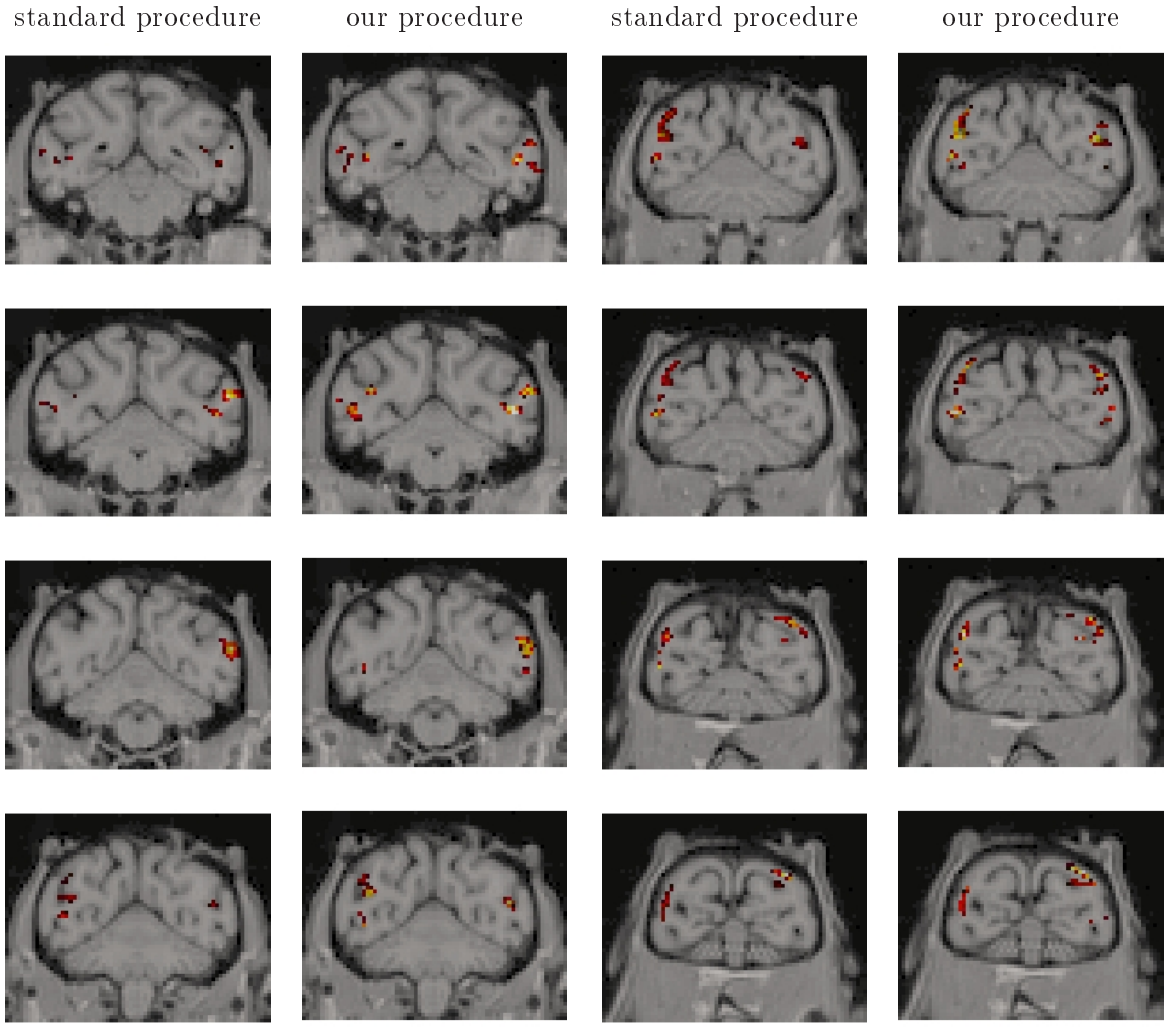


Figure 9.15: Data-driven vs hypothesis based inference
SPM t-map of the contrast stimulus1-stimulus 2 with inclusion of body motion+eye motion confounds -the *standard* procedure -(left) and data-driven confounds -*our* procedure-(right). The color-map of the activation ranges from $t = 5.09$(red) to $t = 9.7$(white).

$t \in [1, .., T]$, and session $\sigma \in [1, .., \Sigma]$. First, we can suggest that the confound space could be defined session-wise and not subject-wise, so that we may *forget* the subject effect, and treat each of the $S \times \Sigma$ sessions independently.

If the hypothesis of no subject effect -besides the hemodynamic response- is rejected, the problem should be treated hierarchically, with subject-wise confounds for example. This is an alternative to the hierarchical Bayesian model of Friston et al. [87] [81].

Non coregistered case

If, for some reason, the datasets are not aligned, estimation of response patterns becomes more difficult. Let us simply suggest the following model: let $X_1, .., X_S$ be datasets with $N(s)$ voxels, $s = 1..S$. Then

$$X_s(t) = \sum_{i=1}^{I(s)} Y_{s,i}(t) + M_s R(t) \tag{9.17}$$

where

- $Y_{s,i}(t)$ are session-wise confounds (i.e. a space of dimension $I(s)$ for each session s).
- $R(t)$ is a set of decorrelated temporal patterns which are reproducible from one session to another one (task-related or not).
- M_s is a mixing matrix that relates R to each session-wise dataset.

The estimation of all the quantities involved here can be performed by the same procedure as previously:

- Step 0: All the datasets are reduced by PCA to a given number of temporal components.
- Step 1: A set of temporal patterns common to all datasets is derived by the CCA procedure 5.1.4. This yields (M_s), $s = 1, .., S$ and R.
- Step 2: After removal of the reproducible patterns, confounds are estimated by the standard state-space procedure.

Of course, step 1 and step 2 can be iterated. This method should be quick, since it works in a low dimensional space. Once can also add some constraints within the CCA procedure (step 1) to obtain only task-related patterns. Note that this procedure is much weaker than the procedure described earlier on coregistered data. In particular, no estimation is made on a voxel basis. This is logical, since voxel-based cannot be generalized here.

These sketchy suggestions are simply intended to show that the generalization of our procedure to multi-subject studies can be done within a similar framework and/or with similar concepts.

Chapter 10

Conclusion

As a conclusion, we would like to make a quick review of the main novel technical contributions of these work and to explain in which situation they can help and, if necessary, which supplementary development would improve them. As a complement, a graphical representation of all the methods used in this thesis are presented in figure 10.1, with their links and particular function.

Temporal model: the Wold decomposition. This model aims at separating deterministic and stochastic components of the signal, and thus provides an attractive conceptual framework in signal analysis. We believe that, associated with the BIC-MDL criterion, it can provide a useful alternative to recent Bayesian models, that are computationally much more heavy and require the prior tuning of non trivial hyper-parameters. The main weakness of this model is probably the lack of physiological relevance of task-related patterns: introducing a physiologically relevant parameterization would certainly enhance its interest.

Clustering: the Information Bottleneck method. The Information Bottleneck method offers a well-grounded framework for data clustering. Besides being well-adapted to fMRI data (noise model, incorporation of priors in the feature space modeling), it probably introduces the most relevant variable to tune the bias/robustness tradeoff. For practical reasons the use of analytically derived approximations of the Kullback-Leibler divergence would bypass the problem of probability density sampling on finite grids. We recommend this method when a good temporal model of the data is available, but when the voxel-based responses to the different conditions are complex.

State-space model. The state-space model used here provides a multivariate description of the dataset that takes into account temporal structure of the components of interest within the data. It provides in a computationally economic way a rough model of the main structured patterns of the dataset. As in chapter 9, we may keep it for nuisance space estimation, since it seems to greatly enhance the whiteness of voxel-based residuals. An interesting problem is also the identification of the nuisance components given their spatial location and their structure in the frequency domain. Recalling that state-space estimation minimizes functional 6.9, it can be viewed as an alternative to bayesian temporal

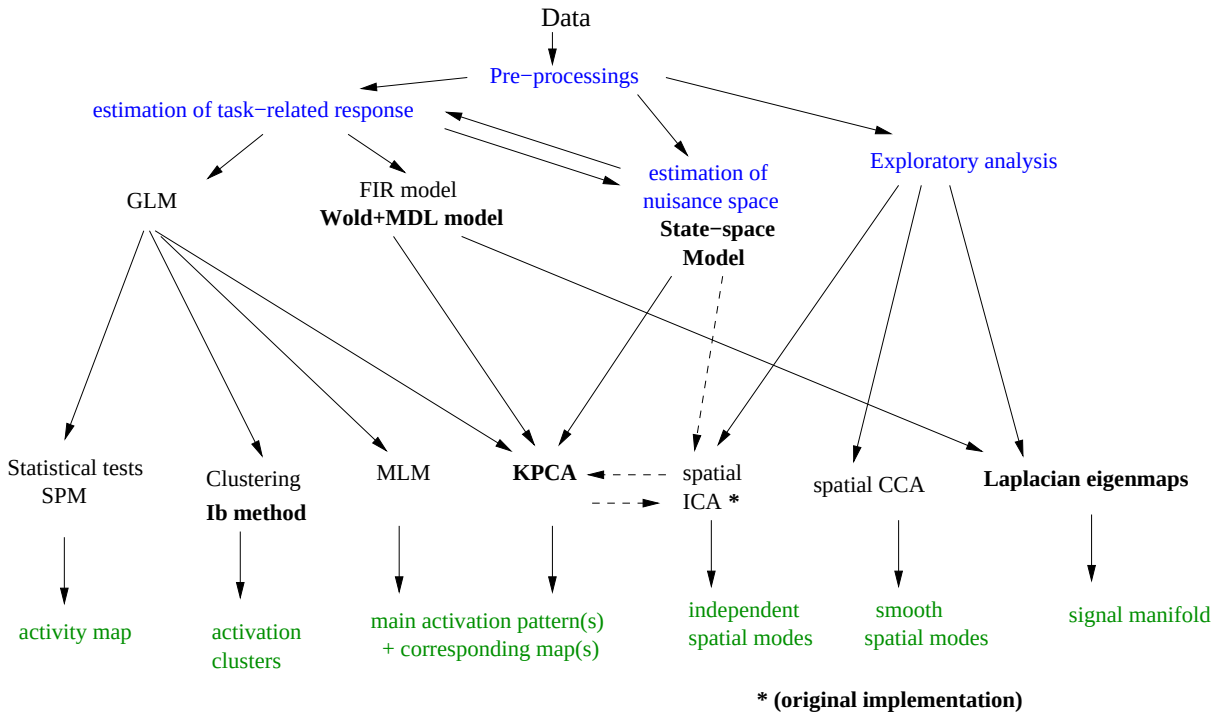


Figure 10.1: List and relationship of the methods used in this thesis.
The methods(black titles) written in bold style are original contribution of the thesis. The arrows indicate the succession of the different steps. Blue and green titles represent respectively the main steps of data analysis and the output of each method. Dashed arrows indicate steps that have not been presented in the thesis but that are worth investigating.

models, especially suitable for temporally correlated data. This thesis has thus probably not completely exhausted the potential of this framework.

Nonlinear decomposition: kernel PCA. Kernel PCA uses the potential of over-complete signal bases to account for subtle effects within the data. It is of great interest when signal bases produced by simple SVD do not fit well the non-Gaussian structure of the data. However, given the difficulty of assessing the number of components used in data decomposition and the difficulty in thresholding the heavily non-Gaussian resulting maps, we believe that it is limited to descriptive purposes-at least with non-polynomial kernels. Note also that it provides an interesting solution to the problem of state-space basis choice. The interest of this method could be enhanced by combination with ICA.

Nonlinear decomposition: Laplacian maps. Within the non linear methods, Laplacian embedding much more wisely searches for the low dimensional representation of the data that preserves its -metric or topological structure. We found that the topological point of view was better suited for fMRI data, probably due to the certainly uneven sampling of the signal space by empirical data. It gives a practical contribution to the actual trend towards functional connectivity investigation- functional connectivity being nothing but empirical time series correlation. The main strength of this method is its weak sensitivity to local perturbations of the data and to noise, and its main weakness is probably the difficulty in making explicit the embedding of the data in the natural signal space. It could still be enhanced by the introduction of spatial information (spatio-temporal clustering). Moreover, given the robustness of the method, it is a potential tool for the comparison of multiple datasets. We prefer it to Kernel PCA for data representation.

The final combination of the methods that we propose in chapter 9 is intended to propose a viable alternative to hypothesis-driven approach of fMRI data, that make up the mainstream in current neurological studies. Our main achievement is perhaps to propose a data-driven *and* pertinent -in the sense that it takes into account the experimental paradigm- approach. In practice we propose

- Purely exploratory approaches (ICA, Laplacian eigenmaps, CCA) that may be used when little prior knowledge is available on the dataset
- The semi exploratory approach described in chapter 9 which is an alternative to the GLM, especially useful for cross validation purposes.

In a near future, we will address the question of multi-subject studies.

Sur ce, qu'il me soit permis de ne pas en dire davantage.

Appendix A

Presentation of the datasets used in this work

A.1 Synthetic dataset

We have created a synthetic dataset by simulating one slice of fMRI data containing $N = 1963$ brain voxels. The length of the series is $T = 200$. The simulated paradigm comprises two alternating conditions in a block design (see figure A.1(a)). 3 small activation foci of 21 voxels are created; the activation time courses are obtained by convolution of the experimental condition time courses with the canonical hemodynamic filter of SPM sampled at TR=2s [77].

$$a_f(t) = [h * (\gamma_f(1)P_1 + \gamma_f(2)P_2)](t) \tag{A.1}$$

where $\gamma_f(1), \gamma_f(2)$ and thus $a_f(t)$ are defined for each focus of activation. The simulated time courses, $(a_f), f = 1..3$ are given in figure A.1(b) and spatial maps are presented in figure A.1(c). Next, a gaussian white noise is independently added to all voxels, so that the SNR is 0.5 in the *activated areas*.

The data is smoothed spatially as commonly done for fMRI (FWHM = 4.5mm = 1.5 voxel).

Let us insist on the following features

- The model is not realistic; for example, the non-signal part is white, which is not true for general fMRI data. Hence this simulation is not useful to test temporal models.
- However, the generative model A.1 is not unrealistic, and a question of interest could be which region of the slice responds (selectively or not) to the experimental conditions.
- The SNR is not higher than for true fMRI data, and the activated areas are relatively small, making signal activation a challenging task.
- The choice of 2D data is only for the sake of visualization

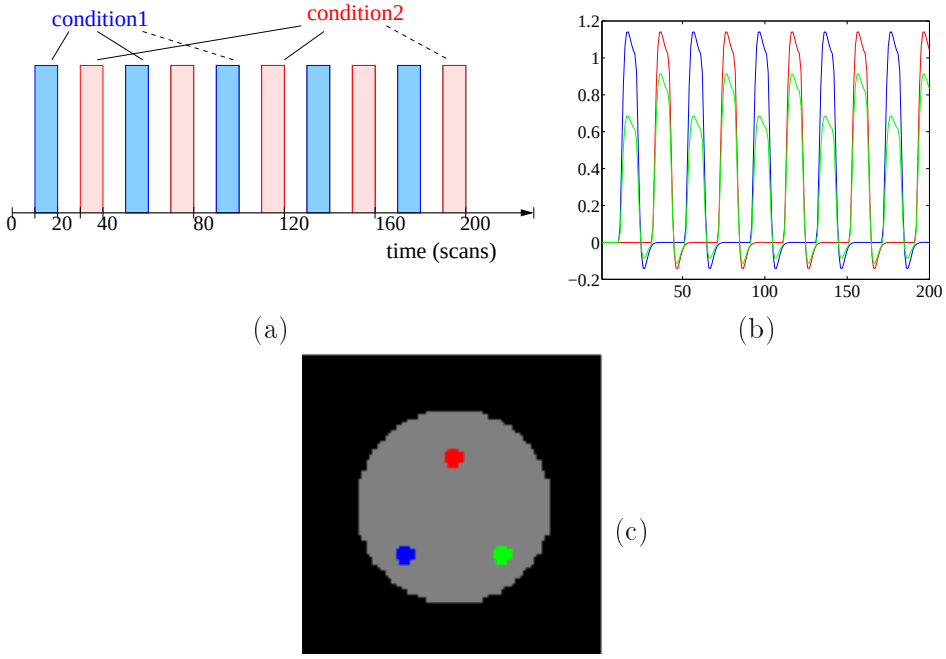


Figure A.1: Description of the synthetic dataset.
(a) Simulated experimental paradigm (two conditions, alternating block design with resting periods). (b) Synthetic activations time courses. The three patterns are obtained by convolving the canonical hrf with three different linear combination of the stimuli time courses. (c) Spatial layout of the activations simulated in the experiment. The colors are those of the time course. The colors of figures (b) and (c) match.

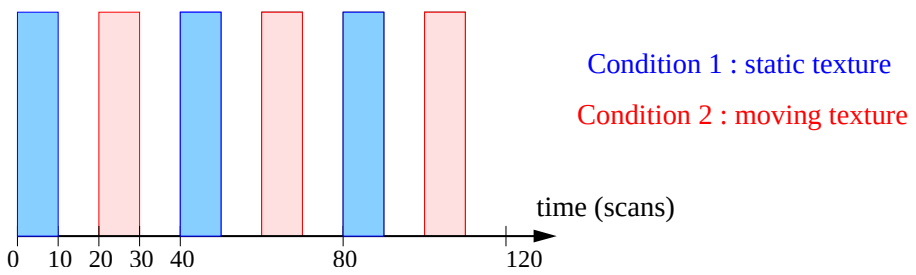


Figure A.2: Experimental paradigm used in the real data experiment described in A.2. This experiment is typically a motion localizer through the subtraction of the two alternating activation conditions.

A.2 Real dataset 1

Described in [218], this dataset belongs to a study on monkey vision : The task performed by the Rhesus monkey is the passive viewing of moving and static textures. The experimental paradigm consists in 3 repetitions of the following stimulation sequence: viewing of a static texture (random dots) during 10 scans, rest during 10 scans, viewing of a moving texture during 10 scans and rest during 10 scans, thus yielding 120 scans long sessions (see figure A.2 for a representation of the experimental paradigm). The dataset considered here comprises 11 sessions. It was acquired with a 1,5T scanner. The repetition time is TR=2.976 seconds. One volume has 64x64x32 voxels and the spatial resolution is 2x2x2 mm; it comprises the whole brain.

Before the experiment, the monkey had undergone an injection of MION (monocrystalline iron oxide nanoparticle) contrast agent of 4 mg/kg, so that the measured signal is not the BOLD contrast, but is related to the local cerebral blood volume [218]. Using this contrast agent is known to increase the contrast to noise ratio, but challenges the hypotheses used for standard models of fMRI data: it changes the sign of the activation pattern, and the shape of the response is modified, since it reflects mainly the local Cerebral Blood Volume (CBV), and not the standard BOLD effect.

In practice, the dataset is reduced to $N = 12320$ voxels by retaining only the brain voxels. For visualization of the activation maps, it is coregistered with an anatomical image through the method described in [111].

A.3 Real dataset 2

Described in [217], this dataset is devoted to another important feature of monkey vision: the ability to distinguish between the perception of motions in 2 dimensions, and motions in 3 dimensions. The experimental paradigm consists in 4 repetitions of the following stimulation sequence: viewing of a 2 dimensional motion during 10 scans, rest during 10 scans, viewing of a 3 dimensional motion during 10 scans and rest during 10 scans, thus yielding 160 scans long sessions (see figure A.3 for a representation of the experimental paradigm). The dataset considered here comprises 12 sessions. It was acquired with a 1,5T scanner. The repetition time is TR=2.368 seconds. In this case the dataset has been

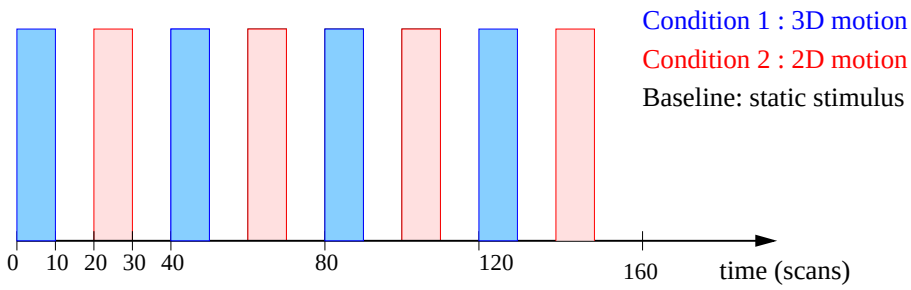


Figure A.3: Experimental paradigm used in the real data experiment described in A.3. The signal difference between the two alternating activation conditions gives the sensitivity to 3D stimulation.

pre-registered with the anatomy of the monkey with the method [111], yielding images of 1mm resolution, of size $71 \times 86 \times 56$ voxels.

Before the experiment, the monkey had undergone an injection of MION (monocrystalline iron oxide nanoparticle) contrast agent, as described in A.2.

In practice, the dataset is reduced to $N = 77968$ voxels by retaining only the brain voxels.

A.4 real dataset 3

This dataset is based on an fMRI retinotopy experiment described in [225]. Working on a 3T Bruker Medspec 30/80 Avance scanner, with a full body magnet, functional EPI images are acquired in 4 sessions; each session comprises 90 images acquisition, with $TR = 1.5s$. Each image is in turn a volume of 18 slices, with a thickness of $3mm$, with interleaved acquisition; these slices are chosen in order to encompass the occipital cortex. The first and last nine scans (without stimulation) are then discarded. A high resolution $1 \times 1 \times 1mm^3$ anatomical image of the subject is also acquired during the same session.

As usually for retinotopy ([197]) the stimulation is the display of either a rotating edge or a contracting/expanding ring; a simplified presentation of them is given in figure A.4 . The key information is the frequency ω_0 of the stimulus; here the period $T_0 = \frac{2\pi}{\omega_0}$ is 18 scans, i.e. $27s$, so that 4 complete periods are acquired for each stimulus. The subject is instructed to passively look at the display and to concentrate on a fixation red cross at the center of the image.

The 4 sessions of acquisition comprise one session of clockwise rotating wedge, one session of anti clockwise rotating wedge, one session of expanding ring and one session of contracting ring. The analysis of such datasets is usually performed by estimating the component of each voxel time course that has the same frequency as the stimulus. The amplitude of this signal can be used for testing the presence of the response, while its phase yields an estimate of the activation timing. The fact that the stimulus is performed alternatively in two opposite directions can be used to eliminate the hemodynamic offset. Finally, a value of eccentricity and polar angle can be associated to each voxel, yielding the retinotopic mapping.

The resulting maps can then be projected on an anatomical image; cortex segmentation

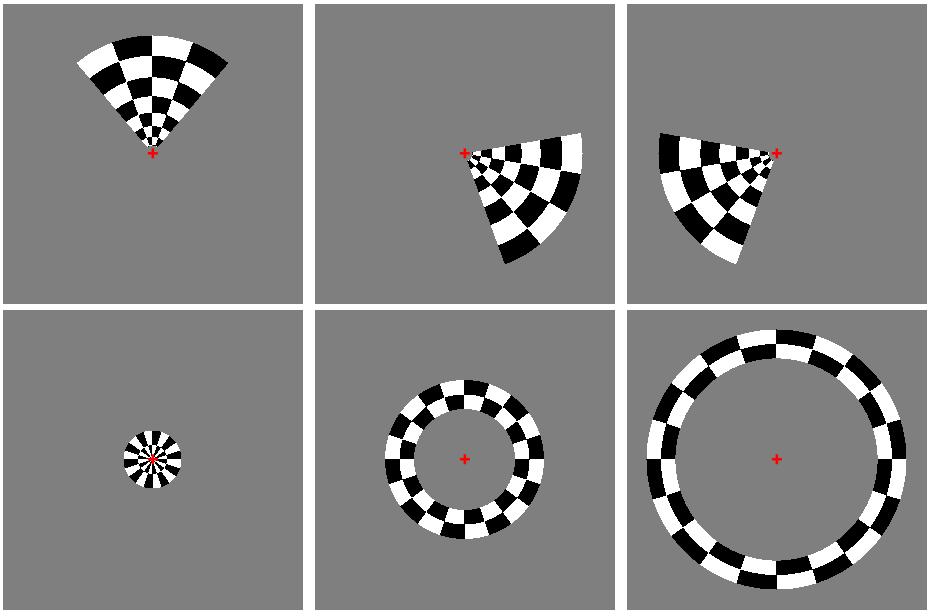


Figure A.4: Typical stimuli used in the retinotopy experiment
(top) the wedge stimulus in different position, (bottom) the ring stimulus in different positions.

and inflating are then used to derive inflated maps. In the present case, these steps have been performed with the Brainvisa [46] software.

Appendix B

Generating spatial maps from multivariate methods

Here we explain how we can give a statistical interpretation of spatial maps derived by multivariate analysis methods; note that this problem stems from many practical situations; for example, many procedures end up with the diagonalization of a matrix, whose eigenvectors are interpreted as spatial maps: this is the case for PCA, CCA, Kernel PCA, Laplacian eigenmaps, or more generally when spatial maps are manipulated as multivariate vectors, e.g. in state-space models or ICA. Note that what we present here are some practical solutions used in this work, and not systematic approaches.

In general, each map is represented by a unitary $N \times 1$ vector M:

$$\sum_{n=1}^{N} M(n)^2 = 1 \tag{B.1}$$

The most simple way to handle the problem is to consider that by (B.1), $M(n)$ results from a normalization procedure applied to a random variable. Indeed, under the hypotheses
H_1 *: All the samples are drawn randomly from the same population (i.e. from one law),*
H_2 *: they are drawn independently,*

i.e., there exists a random variable $X(n)$ such that $\forall n \in [1, N], M(n) = \frac{X(n)}{\sqrt{\sum_{i=1}^{N} X(i)^2}}$, it becomes possible to handle the map. In particular, adding the next hypothesis

H_3 *: The random distribution is gaussian with 0 mean*, the variable τ defined by

$$\tau(n) = \sqrt{N-1}\,\frac{M(n)}{\sqrt{\sum_{i\neq n} M(i)^2}} = \sqrt{N-1}\,\frac{M(n)}{\sqrt{1-M(n)^2}} = \sqrt{N-1}\,\frac{X(n)}{\sqrt{\sum_{i\neq n} X(i)^2}} \tag{B.2}$$

has a Student distribution with $N-1$ degrees of freedom. The fact that the expected value is zero stems in our case from prior assumptions or processing on the data. We will consider that it holds here. Otherwise, the mean value can be estimated.

Consequently, we can handle $\tau(n)$ as a Student variable, and make inference on its value, introducing tests and thresholds th to assess that $P(\tau(n) > th)$ is below a certain value.

An additional simplification of the problem arises for low values of N - which is not unrealistic: first the student distribution gets very close to the normal one, second $\tau(n) \simeq M(n)\sqrt{N-1}$; this may allow for quick computations.

However, this simple analysis can go wrong for several reasons:

- First, assumption (H_2) of statistically independent variables is usually false, and thus simple procedures ignore real spatial correlations.
- Second, the assumption (H_1) that all voxels are drawn from the same distribution is precisely the hypothesis that one searches to reject. While this is not necessarily a problem in univariate procedures, this becomes a problem here, because, if a fraction of them (say, the *activated* voxels) are not drawn from the null distribution, their presence biases the normalization procedure, i.e. it reduces $\tau(n)$ values.
- Last, more classically, one can reject the gaussian hypothesis for the variables; here, we will make distinction between linear and nonlinear procedures: the linear procedures yield approximately normal maps, while non-linear procedures may yield maps that are far from normal.

The first problem can be considered as solved in the literature in different manners: For example, one can warp the Student distribution to a normal one, and consider the data as the realization of a smooth gaussian field structure (see section 3.2.1; [178]); however, the validity of these models requires intensive smoothing of the data. Alternatively, applying a Bonferroni correction on P-values is probably a conservative way to solve the issue; however, the most practical solution of the problem is to use False Discovery Rates [90] instead of classical P-values; this allows for voxel-based inference.

The second problem involves the correction of τ due to the statistical heterogeneity of the voxels; however it cannot be done *a priori*.

Last, the third problem is to associate τ with a P-value; Assuming that the data is normally distributed under the null hypothesis, all linear methods can be assumed to yield normally distributed data. For this reason, we distinguish PCA, CCA and ICA from Kernel PCA, the latter being highly non-linear, while the former can reveal only small deviations from normality (even though ICA explicitly emphasizes non-normality of the data). We continue our development by separating the two cases.

B.1 When Null data is approximately Gaussian

B.1.1 Mixture modeling

A possibility is to consider mixture models (see [58], [15]) for the distribution, with one mode being the null distribution; But this solution poses some new problems: the mixture estimation may be difficult; one needs to choose the number of modes; for example, one can prefer a small number of modes to avoid over-fitting, but few modes can be suboptimal to model the non-gaussian part of the distribution.

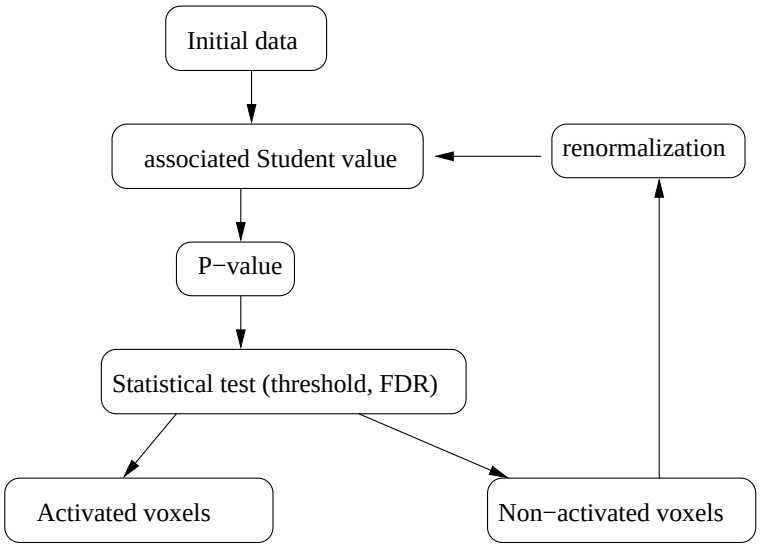


Figure B.1: An iterative scheme for the correction of spatial maps under mild deviation form normality.

Note that mixture modeling is not much different from the Information Bottleneck method presented in 5.3.2, which does note use hypotheses on the modeled distribution; but the method involves the difficult choice of the parameter β -which we have not solved- and assumes that each voxel based information $M(n)$ or $\tau(n)$ is itself a density, not a scalar.

However, we prefer the following simpler procedure.

B.1.2 A quick procedure

What we propose is an iterative procedure that alternatively recomputes the τ map, the P values and the number of voxels for which the null hypothesis is not rejected (see figure B.1). The key point is that equation B.2 is replaced by

$$\tau(n) = \sqrt{N-1}\,\frac{M(n)}{\sqrt{\sum_{i\neq n}\bigcap_{S_0} M(i)^2}} \tag{B.3}$$

Where S_0 is the null set.

The fact that this procedure converges is evident since the number of *activated voxels* always increases. In practice, A False Discovery Rate testing (set to 0.05, for example) is conservative enough to ensure that the procedure will not start to prune the null distribution. An example of this procedure is given on a synthetic example (in figure B.2), and on an ICA example taken from section 5.2.4 in figure B.3.

This procedure seems to yield valid corrections on the normalized spatial maps. It is probably a bit conservative. Once again, it is advisable when linear decomposition procedures are used -i.e. with mild deviations from normality (PCA, CCA, ICA).

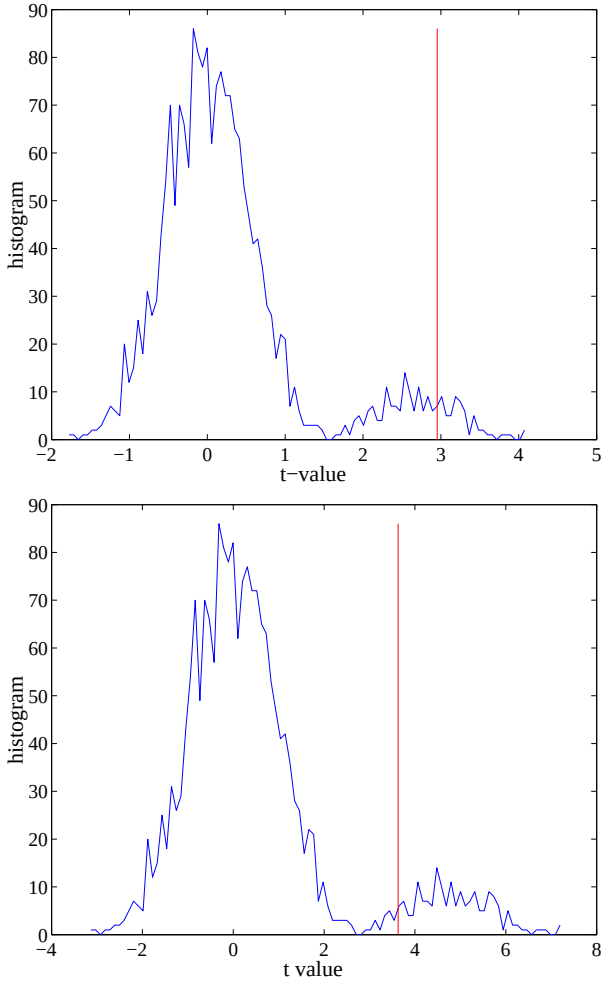


Figure B.2: Thresholding of a mixture map containing both null and activated data. The two populations are represented by 1800 and 200 samples generated from $\mathcal{N}(0,1)$ and $\mathcal{N}(5,1)$ by the FDR procedure (FDR =0.05, indicated by the vertical bar in both cases). Due to the normalization of the map, the naive approach (left) is over-conservative (only 63 voxels detected), while the iteratively corrected normalization provides a more realistic, though conservative, threshold (178 detections). Note the correction of t values.

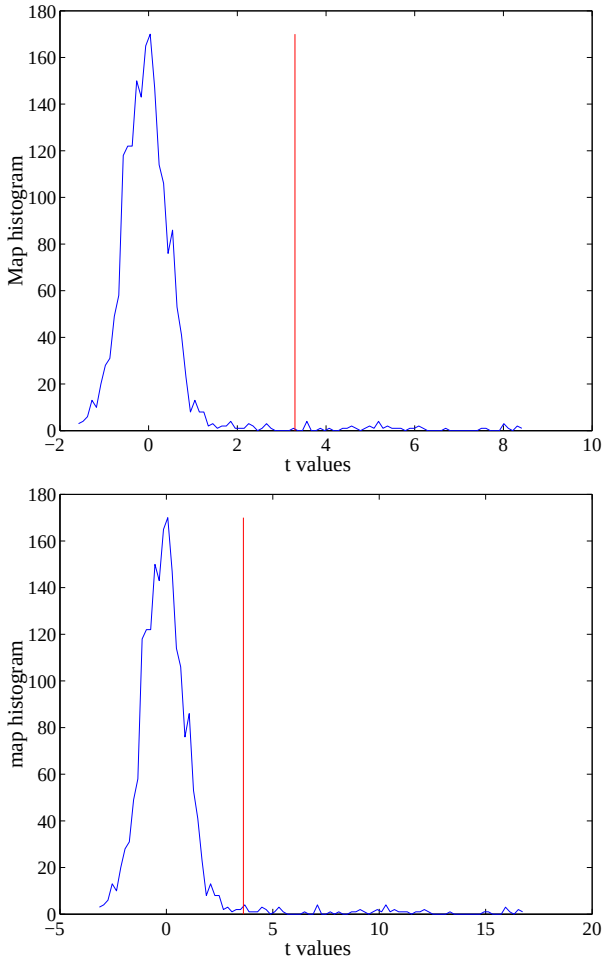


Figure B.3: Thresholding of the first map obtained in figure 5.2 by the FDR procedure FDR =0.05, indicated by the vertical bar in both cases. Due to the normalization of the map, the naive approach (left) is over-conservative (41 voxels detected), while the iteratively corrected normalization provides a more realistic, though conservative, threshold (57 detections). Note the correction of t values.

B.2 When Null data is not Gaussian

In this case the P-values under the null hypothesis are not available. There is a solution only if a simple transformation can warp the null distribution into a gaussian one. For example, in section 7.3.2, the cubic kernel makes the distribution of the data non-gaussian; but applying the function $M(n) \longrightarrow M(n)^{1/3}$ to the map approximately solves the issue: one can apply the previous method to the map, as illustrated in figure B.4.

For the more general case (section 7.3.1) where the deviation from normality cannot be corrected, there is no practical solution, apart simulating numerically the null distribution of the map to estimate the null P-values; note that we have used in figure 7.8 another way to represent spatially the data. In the case of 7.3.3, there are mild differences between PCA and KPCA, so that PCA threshold can be used for thresholding KPCA maps. This effect is due to the small dimensional input space.

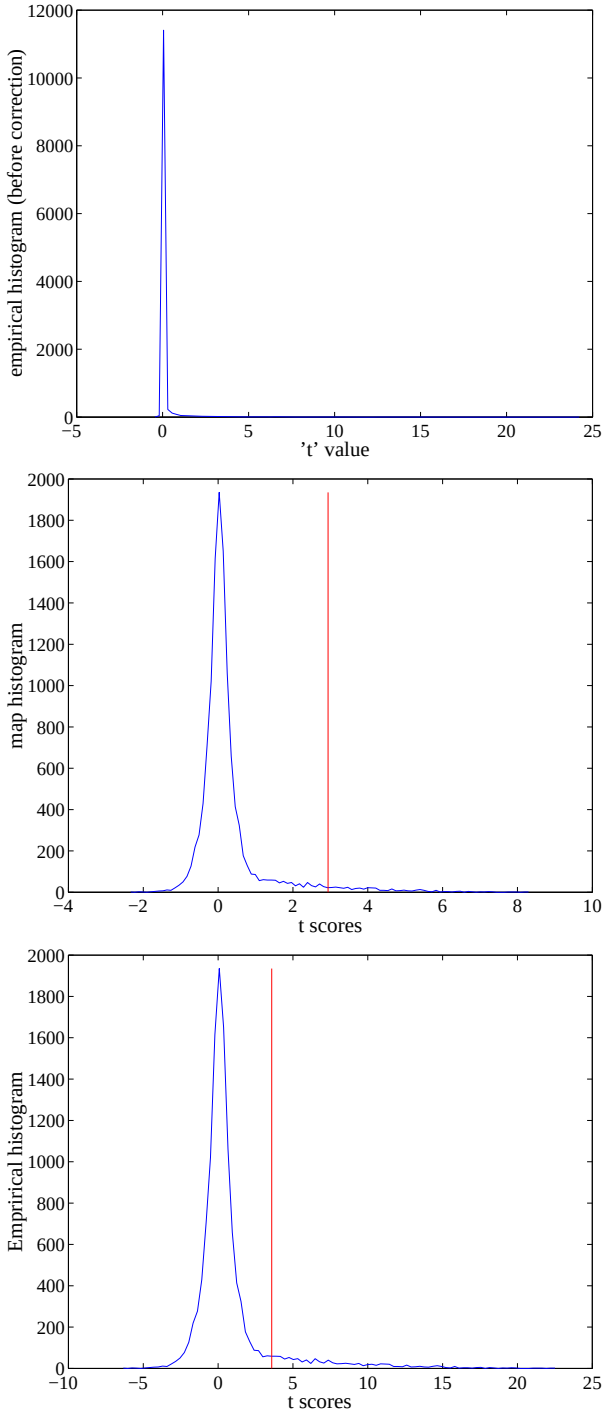


Figure B.4: Thresholding of the first map obtained in figure 7.11 by the FDR procedure FDR =0.05, indicated by the vertical bar in images (b) and (c): (a) original values, that are not student-distributed, (b) distribution after the correction $M(n) \longrightarrow M(n)^{1/3}$; the null distribution is properly more gaussian, but the t values are under-estimated, making the threshold over-conservative; (c) the iteratively corrected normalization provides a more realistic threshold. Note the correction of t values.

Appendix C

Information theory: a survivor's guide

We give here some basic definitions of information theory, which are used in this document. Proofs and complementary explanations are available in [48].

C.1 Some basic definitions

C.1.1 Entropy

Let X be a random vector of dimension $d \geq 1$, and let $x(1), .., x(N)$ be N samples or random realizations of this random vector. Let p_X be the probability density function of X. We assume that it is a function of $L^1(\mathbb{R}^d)$ that satisfies

$$p_X \geq 0 \ everywhere \tag{C.1}$$

$$\int_{\mathbb{R}^d} p_X(u)du = 1 \tag{C.2}$$

The entropy of X is the quantity

$$H(X) = -\mathbb{E}(log(p_X)) = -\int_{\mathbb{R}^d} p_X(u)\log(p_X(u))du. \tag{C.3}$$

Note that the entropy and all the quantities encountered here depend on the density and not on the random variable or vector- however one can use expressions like *entropy of a random variable* without risk of confusion. We assume that this quantity is defined for all the densities encountered in this document. This is true in particular for empirically estimated densities which are either gaussian or mixtures of gaussians (see section C.2). The entropy can be considered as the measure of the dispersion of the random variable or vector X. If the law of X is gaussian with mean μ and covariance Σ, i.e. $p_X = \mathcal{N}(\mu, \Sigma)$, then the entropy depends only on Σ:

$$H^g(X) = \frac{1}{2}\log(det(2\pi e\Sigma)) \tag{C.4}$$

The entropy of various densities can be found in [48, chapter 16].

Assuming that X is not gaussian, its negentropy is the difference between its entropy and its gaussian entropy $H^{neg}(X) = H^g(X) - H(X)$, H^g being computed from the mean μ and variance Σ of X. This quantity is always positive, since it is equal to $K[p_X|\mathcal{N}(\mu, \Sigma)]$ (see section C.1.2). It is thus a possible measure for the deviation from normality of X.

Let X and Y be two random vectors. The conditional entropy of X given Y, defined as $H(X|Y) = \int H(X|Y = y)P(Y = y)dy$, is the difference between the entropy of the joint distribution of X and Y and the entropy of Y.

$$H(X|Y) = H(X, Y) - H(Y) \tag{C.5}$$

It can be interpreted as a measure of the residual randomness of X when Y is known. The following holds : $H(X|Y) \leq H(X)$, with an equality if X and Y are independent.

C.1.2 Kullback-Leibler divergence, mutual information

If X and Y are two random variables or vectors with laws p_X and p_Y, the Kullback-Leibler divergence between their laws is defined as follows:

$$K[p_X|p_Y] = \mathbb{E}_{p_X}\left(\log \frac{p_X(u)}{p_Y(u)}\right) = \int_{\mathbb{R}^d} p_X(u) \log \frac{p_X(u)}{p_Y(u)} du \tag{C.6}$$

The Kullback-Leibler divergence is not a distance, since it is not symmetric. However, the following result holds [48, chapter 9]: given two densities p and q, $K[p|q] = 0$ only if and only if $p \equiv q$. Intuitively, one may think that $K[p|q]$ quantifies the additional *randomness* that appears when a vector of true law p is assumed to have the law q.

Given two random variables or vectors X and Y, the mutual information between X and Y is the Kullback-Leibler divergence between the product density $P_X P_Y$ and the joint density P_{XY}.

$$\begin{aligned} \mathcal{I}(X, Y) &= K[P_{XY}|P_X P_Y] && \text{(C.7)} \\ &= \int_{\mathbb{R}^d \times \mathbb{R}^d} p_{XY}(u) \log \frac{p_{XY}}{p_X p_Y}(u) du && \text{(C.8)} \\ &= H(X) + H(Y) - H(X, Y) && \text{(C.9)} \end{aligned}$$

Note that given (C.5), one has also $\mathcal{I}(X, Y) = H(X) - H(X|Y)$. Moreover, $\mathcal{I}(X, Y) \geq 0$ since it is defined as a divergence, and $\mathcal{I}(X, Y) = 0$ if and only if $p_{XY} = p_X p_Y$, which means that the random vectors X and Y are independent. Hence, the mutual information is a measure of the statistical dependence between two random variables.

C.1.3 Score vector, score function

Here we assume that the law p_X of X is differentiable. To simplify matters, we also assume that $d = 1$ (X is a variable), but the setting adapts straightforwardly to $d > 1$. The *score function* of X is the function

$$\psi_X(u) = -\frac{d}{du} \log p_X(u) = -\frac{1}{p_X(u)} \frac{dp_X(u)}{du} \tag{C.10}$$

Formally it is the derivative of the log-density of X. It can be shown that, if Y is another random variable,

$$H(X + \epsilon Y) = H(X) + \epsilon \mathbb{E}(\psi_X(X)Y) + o(\epsilon), \tag{C.11}$$

Given N samples $x(1), .., x(N)$ of the random variable X, the associated *score vector* of X is the vector $(\psi_X(x(i)))_{i=1,..,N}$. Given (C.11), the score vector can be thought of as the gradient vector of the entropy: moving each sample along the score vector locally maximizes the entropy of the resulting vector.

C.1.4 Entropy of temporal processes

The case of temporal processes $X(t), t = 1..T$ can be treated as any random vector, and the emphasis is put on the serial dependences that arise between successive samples. The key concept is the *entropy rate* of temporal processes [48, chapter 4,11]:

$$\eta(X) = \lim_{T\to\infty} H(x(1), .., x(T)) \tag{C.12}$$

$$= \lim_{T\to\infty} H(x(T)|x(1), .., x(T-1)) \tag{C.13}$$

$\eta(X)$ can be interpreted as the additional randomness of each new sample when one has an exhaustive knowledge of the past of the process. In other words, it is the intrinsic randomness of the process. If the samples are i.i.d. it is simply the marginal entropy.

In the case of gaussian processes, the entropy rate can be expressed with the spectral density f of the process:

$$\eta(X) = \frac{1}{\log(2\pi e)} + \frac{1}{4\pi}\int_{\mathbb{R}} \log(f)(u)du \tag{C.14}$$

This formula, known as Kolmogorov's formula, has been first proved by Szegö (see also [27]).

C.2 On estimation

The computation of the entropy and associated quantities requires the estimation of probability densities -unless one makes a gaussian hypothesis, but this is rejected in some methods, like ICA (see 5). Given N samples $(x(1), .., x(N))$ one needs to estimate p_X. To avoid confusion, we note the estimated density by $\mathcal{D}$, p_X being the *true* density. The standard method consists in using a mixture approximation:

$$\mathcal{D}(u) = \frac{1}{T}\sum_{t=1}^{T} g_\sigma(u - X(t)), \tag{C.15}$$

where $g - \sigma$ is the gaussian kernel of parameter σ. Note that $\mathbb{E}(\mathcal{D})(u) = (g_\sigma * p_X)(u)$, so that it should be ideal to use $\sigma = 0$. On the other hand, the estimation should not be dependent on the particular samples available, so that σ should be high enough. This is a typical bias/variance tradeoff. In practice, given e.g. the results presented in [26], $\sigma = \sqrt{var(X)}N^{\frac{1}{d+4}}$ is a good choice. The practical estimation of equation (C.15) can be made quick by the use of recursive filters [52].

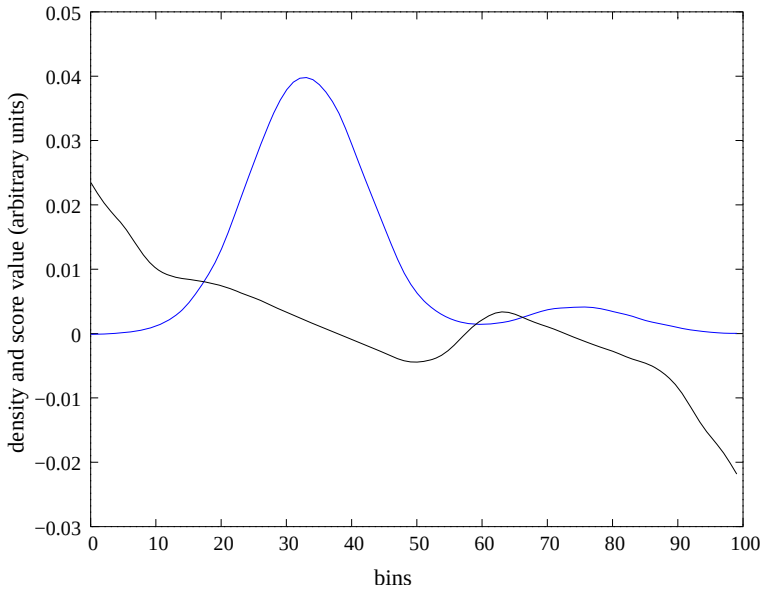


Figure C.1: Empirical estimations of a density and of the corresponding score function. A set of $N = 10^4$ samples are generated with a bimodal gaussian mixture model. The samples are then discretized on a finite grid (100 bins) and the resulting density is smoothed in order to yield the estimate (C.15). This results in the density in blue. In parallel, the score function is derived. (in black), using the same method. Note that the score function deviates from linearity in the regions where the density deviates from normality.

The estimation of the score vector follows exactly the same way. An illustration is given in figure C.1.

Let us finally point out the curse of dimensionality: Although the method presented here adapts to any dimension d, the estimation worsens rapidly for $d > 1$. This problem is known as the curse of dimensionality. However, many algorithms (e.g. ICA) require only marginal, i.e. one-dimensional densities. For high-dimensional densities, gaussian or gaussian mixtures approximations are better suited than empirically estimations.

Appendix D

Information and prediction: choosing the best model

D.1 Joint entropy of a time series

We give here a derivation of the asymptotic entropy of a time series given a prediction model as in chapter 4; this development is inspired from [20].

Let $X = (x(t)), t = 1, .., T$ be a time series, endowed with a prediction model of order q. Let $H(X)$ be the joint entropy of X. Then , under conditions that will be made precise later

$$H(X) \underset{T \to \infty}{\simeq} T\eta(X) + \frac{q}{2}\log(T) + o(\log(T)) \tag{D.1}$$

where $\eta(X)$ is the entropy rate or extensive entropy of the time series (see C.1.4). To show this result, let us first compute the joint probability of the time series, $P(X)$. We assume that the prediction model is defined by a $q-$dimensional parameter θ, and that the data has been generated by a particular $\bar{\theta} \in \mathbb{R}^q$.

$$\begin{aligned} P(X) &= \int P(\theta)P(X|\theta)d\theta & \text{(D.2)} \\ &= P(X|\bar{\theta}) \int P(\theta)\frac{P(X|\theta)}{P(X|\bar{\theta})}d\theta & \text{(D.3)} \\ &= P(X|\bar{\theta}) \int P(\theta)\exp\left(-TE(T,\theta)\right)d\theta & \text{(D.4)} \end{aligned}$$

Where $E(T,\theta) = -\frac{1}{T}\log\frac{P(X|\theta)}{P(X|\bar{\theta})}$ is the natural energy of the system. It can be estimated by its expectation given the true generative parameter $\bar{\theta}$ plus a fluctuation term:

$$\begin{aligned} E(T,\theta) &= \frac{1}{T}\int P(X|\bar{\theta})\log\frac{P(X|\theta)}{P(X|\bar{\theta})}dX + \frac{1}{T}e(X,\theta,\bar{\theta}) & \text{(D.5)} \\ &= \frac{1}{T}K[P(X|\bar{\theta})|P(X|\theta)] + \frac{1}{T}e(X,\theta,\bar{\theta}) & \text{(D.6)} \end{aligned}$$

Where $K[|]$ stands for the Kullback-Leibler divergence between the two densities given in argument; note that this divergence is computed within the T-dimensional space of

the signal X. In the sequel, we *i)* neglect the fluctuation term $\frac{1}{T}e(X,\theta,\bar{\theta})$ around the expected value, and *ii)* introduce a new hypothesis:

(Condition 1:) We assume that $\lim_{T\to\infty}\frac{1}{T}K[P(X|\bar{\theta})|P(X|\theta)]$ *exists and note this limit* $d(\theta,\bar{\theta})$*; moreover, we assume that* $d(\theta,\bar{\theta})$ *is a smooth (at least* C^2*) function of* θ*.*

In fact this condition is not restrictive at all, since most prediction models depend smoothly on their parameter. Let us assume e.g. a model $x(t)=\sum_{l=1}^{L}\theta_l^1 x(t-l)+\sum_{m=1}^{M}\theta_m^2 u(t-m)+\varepsilon(t)$, where $\theta=(\theta^1,\theta^2)$, $q=L+M$, $u(t)$ is an auxiliary explanatory variable, and ε is a gaussian i.i.d noise; then $P(X|\theta)$ is T-variate normal law with mean μ_θ and covariance Σ_θ. It is then straightforward to check that $\frac{1}{T}K[P(X|\bar{\theta})|P(X|\theta)]$ has a finite limit for $T\to\infty$.

As a consequence, $E(T,\theta)$ has a finite limit when $T\to\infty$. This implies that the integral $\int P(\theta)\exp\left(-TE(T,\theta)\right)d\theta$ essentially reduces to its value on a domain that surrounds the value of θ that minimizes $E(T,\theta)$. Neglecting the fluctuations, we will consider that the solution of the estimation problem in terms of θ is indeed $\bar{\theta}$ (which is true only asymptotically), we use a saddle-point approximation of E around its minimum:

$$E(T,\theta)\simeq d(\theta,\bar{\theta})\simeq\frac{1}{2}(\theta-\bar{\theta})A(\theta-\bar{\theta})+o((\theta-\bar{\theta})^2) \tag{D.7}$$

where A is the Hessian of $d(\theta,\bar{\theta})$ in $\bar{\theta}$. This implies that

$$\int P(\theta)\exp\left(-TE(T,\theta)\right)d\theta\simeq P(\bar{\theta})\int\exp\left(-\frac{T}{2}(\theta-\bar{\theta})A(\theta-\bar{\theta})\right)d\theta \tag{D.8}$$

The latter integral is known to be $(2\pi)^{\frac{q}{2}}|TA|^{-1/2}=(2\pi)^{\frac{q}{2}}T^{-\frac{q}{2}}|A|^{-1/2}$, where $|A|$ is the determinant of A. Finally, for $T\to\infty$,

$$P(X)\simeq P(X|\bar{\theta})P(\bar{\theta})(2\pi)^{\frac{q}{2}}T^{-\frac{q}{2}}|A|^{-1/2} \tag{D.9}$$

We can now estimate the entropy $H(X)=\mathbb{E}(-\log P(X))$ of the time series. We have

$$\log\left(P(X)\right)\simeq\log P(\bar{\theta})+\log P(X|\bar{\theta})+\log\left(|A|^{-1/2}(2\pi)^{\frac{q}{2}}\right)-\frac{q}{2}\log T \tag{D.10}$$

Thus

$$H(X)=-\mathbb{E}(\log P(\bar{\theta})+\log P(X|\bar{\theta}))-\log\left(|A|^{-1/2}(2\pi)^{\frac{q}{2}}\right)+\frac{q}{2}\log T \tag{D.11}$$

Only the first two terms are non-trivial

$$\begin{aligned}
\mathbb{E}(\log P(\bar{\theta})+\log P(X|\bar{\theta})) &= \int P(X)\left(\log P(\bar{\theta})+\log P(X|\bar{\theta})\right)dX && \text{(D.12)}\\
&= \iint P(X|\bar{\theta})P(\bar{\theta})\left(\log P(\bar{\theta})+\log P(X|\bar{\theta})\right)dXd\bar{\theta} && \text{(D.13)}\\
&= \int P(\bar{\theta})\left[\log P(\bar{\theta})+\int P(X|\bar{\theta})\log P(X|\bar{\theta})dX\right]d\bar{\theta} && \text{(D.14)}\\
&= -H(\bar{\theta})+\int P(\bar{\theta})\int P(X|\bar{\theta})\log P(X|\bar{\theta})dXd\bar{\theta} && \text{(D.15)}
\end{aligned}$$

At this point, we need a second technical hypothesis: *(Condition 2:) Let*

$$\eta(X) = -\lim_{T\to\infty} \frac{1}{T} \int P(X|\bar{\theta}) \log P(X|\bar{\theta}) dX, \tag{D.16}$$

be the entropy rate of the process; $\eta(X)$ exists and is non zero in general; we further assume that

$$H_1 = -\lim_{T\to\infty} \left(\int P(X|\bar{\theta}) \log P(X|\bar{\theta}) dX + T\eta(X) \right) \tag{D.17}$$

exists and is finite.

This condition simply means that the subextensive entropy of the true generative process is of order unity, which is true whenever there do not exist any long-range correlations within the data [20]; this holds for all models encountered in practice -for fMRI data analysis, at least after data detrending. The extensive term corresponds intuitively to the stochastic variance of the process, as defined in Wold decomposition.

Given Condition 2, we obtain

$$H(X) \simeq H(\bar{\theta}) + T\eta(X) + H_1 + \log\left(|A|^{-1/2}(2\pi)^{\frac{q}{2}}\right) - \frac{q}{2}\log T \tag{D.18}$$

The non-constant terms in the asymptotic development are thus

$$H(X) \simeq T\eta(X) + \frac{q}{2}\log T \tag{D.19}$$

The first term $T\eta(X)$ is the extensive entropy that increases linearly in the number of the samples, and is related to the variability in the observation of the variable; the second term $\frac{q}{2}\log T$ is the leading subextensive term, and can be identified as a structure term, since it is related to the dimension q of the generating process. We use it as a general complexity measure in data modeling. Last, one can notice that the conditions 1 and 2 introduced here are met in very general situations (Markov, ARMA models), so that the asymptotic formula can be viewed as a fundamental structure of many time models. However, one might be concerned by the deviation from the asymptotic model: first, data scattering around the expectation value may be important (assuming that $T \to \infty$ is optimistic for fMRI data); second, the saddle point approximation may become inaccurate whenever multiple minimizers of the energy exist; last, the constant terms in the asymptotic development may be non-negligible with respect to $\frac{q}{2}\log T$ for finite values of T.

D.2 Bayesian approach in model selection

Let assume now that different prediction models $M_i, i \in \mathcal{I}$ are possible; we follow a Bayesian approach to select the most likely one: $\forall i \in \mathcal{I}$,

$$P(M_i|X) \propto P(M_i)P(X|M_i) \tag{D.20}$$

$P(X|M_i)$ is known as the Bayes factor in the literature (this justifies the name *Bayesian Information Criterion* (BIC) applied for $H(X|M_i)$ in the literature). In many situations

each model as the same prior likelihood; thus the winning model is the one that maximizes $P(X|M_i)$. But in fact $-\mathbb{E}(log(P(X|M_i)))$ is precisely $H(X)$ given the model M_i; this means that the minimization of the criterion

$$H(X|M_i) = T\eta(X)(M_i) + \frac{q_i}{2}\log T \tag{D.21}$$

where q_i is the number of parameters in model M_i, solves the problem of model selection, at least asymptotically (for $T \to \infty$).

Appendix E

Coding and implementation

E.1 Techniques presented in this document

Essentially, the techniques presented in this thesis (see figure 10.1 for a summary) have been implemented in C++. The main exception is the General Linear Model used for statistical inference, for which we use the implementation proposed in the SPM99 software [77] in Matlab environment.

More technically, we use the Lapack library for the operations of linear algebra (matrix inversion, singular values decomposition, determinant). Among others, they are used in sections 4.3.3, 5.1.4, 6.2.2, 9.1 and 9.2.

The estimation of densities, entropies and score vectors (see appendix C) is performed mainly as described in [110]. Let us recall that these methods are based on the estimation followed by the smoothing of histograms. All the smoothing operations are done very quickly with the help of Deriche recursive filters [52].

The ICA algorithm presented in 5.2 is based on the empirical estimation for the score vector, and uses the efficient approximation of the exponential of matrices described in equation (5.29).

The algorithmically more advanced methods are the Laplacian graph algorithm described in chapter 8, and the information bottleneck algorithm described in 5.3.2.

- In the case of the Laplacian graph algorithm, the difficulties are *i)* the definition of the neighboring system, *ii)* the definition of the different connected components of the graph and *iii)* the construction of a sparse Laplacian matrix (a full matrix cannot in general be stored on a standard PC). The solution of the problem results from simple iteration of matrix/vector products. We have coded all these steps -using sparse vectors and matrices- in order to reduce memory and computation load.

- The information bottleneck algorithm relies on equation (5.37), which involves the estimation of probability densities. The latter are approximately sampled on finite grids, but the resulting computation is prohibitive as soon as the densities have to be approximated in $\mathbb{R}^2$ or $\mathbb{R}^3$. For this reason, it is preferable to code explicitly the

support of each probability density function within the grid. We have thus coded adapted routines.

Let us notice that the basic statistical -density, cumulative density- functions (Normal, Student, Fisher, Chi square) used in fMRI data analyses can be coded in C/C++ using standard routines [179]. The derivation of inverse cumulative densities follows simply, e.g. through dichotomy methods.

Moreover, we have coded the kernel PCA method in Matlab, yielding a SPM-compatible version. We are currently recoding many of the different methods -Laplacian embedding, information bottleneck- in order to yield a SPM-compatible toolbox. Some routines, e.g. the algorithm presented in appendix B, or the EM Kalman method presented in section 6.2.1, have also been coded directly in Matlab environment.

E.2 Some other technical contributions

E.2.1 Display softwares

The 2D maps presented in this document have been generated in C++ with local routines. We have indicated as often as possible the color code used.

The 3D maps presented in figures 6.11, 6.12, 8.8 and 9.14 as well as the 2D maps presented in figures 7.17 and 9.15 have been created using the Anatomist software (http://brainvisa.free.fr/).

E.2.2 Registration softwares

Let us recall that the spatial registration of the functional volumes has been performed with the INRIAlign software (http://www-sop.inria.fr/epidaure/software/INRIAlign/), while the anatomical/functional coregistration has been performed with softwares developed by Gerardo Hermosillo and Christophe Chef d'Hotel [111].

Bibliography

[1] Geoffrey K. Aguirre, Eric Zarahn, and Mark D'Esposito. A Critique of the Use of the Kolmogorov-Smirnov (KS) Statistic for the analysis of BOLD fMRI Data. *Magnetic Resonance Medicine*, 39:500–505, 1998.

[2] G.K. Aguirre, E. Zarahn, and M D'Esposito. Empirical Analyses of BOLD fMRI Statistics ii. Spatially Smoothed Data Collected under Null-Hypothesis and Experimental Conditions. *NeuroImage*, 5:199–212, 1997.

[3] Hirotugu Akaike. Use of an information theoretic quantity for statistical model identification. In *5th Hawaï Int. Conf. System Sciences*, pages 249–250, 1972.

[4] A. H. Andersen, D.M. Gash, and M.J. Avison. Principal Components Analysis of the Dynamic Response Measured by fMRI: a Generalized Linear Systems Framework. *Magnetic Resonance Imaging*, 17(6):795–815, 1999.

[5] Alexandre Andrade, Ferath Kherif, Jean-Francois Mangin, Keith Worsley, Anne-Lise Paradis, Olivier Simon, Stanislas Dehaene, Denis Le Bihan, and Jean-Baptiste Poline. Detection of fMRI Activation Using Cortical Surface Mapping. *Human Brain Mapping*, 12(2):79–93, February 2001.

[6] B.A. Ardekani, J. Kershaw, K. Kashikura, and I. Kanno. Activation Detection in Functional MRI Using Subspace Modeling and Maximum Likelihood Estimation. *IEEE Transactions on Medical Imaging*, 18(2):101–114, February 1999.

[7] H. Attias. Independent Factor Analysis. *Neural Computation*, 11:803–851, 1999.

[8] Hagai Attias and C.E. Schreiner. Blind Source Separation and Deconvolution: The Dynamic Component Analyis Algorithm. *Neural Computation*, 10:1373–1424, 1998.

[9] Francis R. Bach and Michael I. Jordan. Kernel Independent Component Analysis. Technical Report UCB/CSD-01-1166, Computer Science division (EECS), University of California, Berkeley, California 94720, 2002.

[10] Daniela Balslev, Finn A. Nielsen, et al. Cluster Analysis of Activity-Time Series in Motor Learning. *Human Brain Mapping*, 15:135–145, 2002.

[11] P.A. Bandettini, A. Jemanowicz, E.C. Wong, and J.S. Hyde. Processing Strategies for Time-Course Data Sets in Functional MRI of the Human Brain. *Magnetic Resonance Medicine*, 30:161–173, 1993.

[12] R. Baumgartner, C. Windischberger, and E. Moser. Quantification in Functional Magnetic Resonance Imaging : Fuzzy Clustering vs Correlation Analysis. *Magnetic Resonance Imaging*, 16(2):115–125, 1998.

[13] A. Baune, F.T. Sommer, M. Erb, D. Wildgruber, B. Kardatzaki, and W. Palm, G.and Grodd. Dynamical Cluster Analysis of Cortical fMRI Activation. *NeuroImage*, 9:477–489, 1999.

[14] C.F. Beckmann and S.M. Smith. Probabilistic Extensions to Independent Component Analysis in fMRI. In *Proc. of the 8th Int. Conf. on Functional Mapping of the Human Brain*, 2002.

[15] C.F. Beckmann and S.M. Smith. Probabilistic Independent Component Analysis for Functional Magnetic Resonance Imaging. Technical Report TR02CB1, Oxford Center for Functional Magnetic Resonance Imaging of the Brain (FMRIB), Department of Clinical Neurology, University of Oxford, John Radcliffe Hospital, Headley Way, Headington, Oxford, UK, 2002.

[16] Mikhail Belkin and Partha Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. Technical Report TR-2002-01, University of Chicago, Department of mathematics, Department of Computer Science, January 2002.

[17] Mikhail Belkin and Partha Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, 15:1373–1396, 2003.

[18] H. Benali, I. Buvat, J.L. Anton, M. Pelegrini, M. Di Paola, J. Bittoun, Y. Burnod, and R. Di Paola. Space-Time Statistical Model for Functional MRI Sequences. In J. Duncan and G. Gindi, editors, *IPMI*, number 1230 in Lecture Notes in Computer Science, pages 285–298. Springer–Verlag Berlin Heidelberg, 1997.

[19] Habib Benali, Melanie Pelegrini-Issac, and Frithjof Kruggel. Spatio-Temporal Covariance Model for Medical Images Sequences: Application to Functional MRI Data. In Insana and Leahy [118], pages 197–203.

[20] William Bialek, Ilya Nemenman, and Naftali Tishby. Predictability, Complexity and Learning. *Neural Computation*, 13:2409–2463, 2001.

[21] R.M. Birn, Z.S. Saad, and P.A. Bandettini. Spatial Heterogeneity of the Nonlinear Dynamics in the fMRI BOLD Response. *NeuroImage*, 14:817–826, 2001.

[22] Bharat Biswal, F. Zerrin Yetkin, Victor M. Haughton, and James S. Hyde. Functional Connectivity in the Motor Cortex of Resting Human Brain Using Echo-Planar MRI. *Magnetic Resonance Medicine*, 34:537–541, 1995.

[23] Bharat B. Biswal and John L. Ulmer. Blind Source Separation of Multiple Signal Sources of fMRI Data Sets Using Independent Component Analysis. *Journal of Computer Assisted Tomography*, 23(2):265–271, 1999.

[24] S. Bochner. *Harmonic Analysis and the Theory of Probabilty*. University of California Press, Los Angeles, California, 1955.

[25] I. Borg and P. Groenen. *Modern Multidimensional Scaling*. Spinger-Verlag, Berlin, 1997.

[26] D. Bosq. *Nonparametric Statistics for Stochastic Processes*, volume 110 of *Lecture Notes in Statistics*. Springer–Verlag, 2nd edition, 1998.

[27] P.J. Brockwell and R.A. Davis. *Time Series : Theory and Methods*. Springer Series in Statistics. Springer, 1991.

[28] Anders Brun, Hae-Jeong Park, Hans Knutsson, and Carl-Fredrik Westin. Coloring of DT-MRI Fiber Traces using Laplacian Eigenmaps. In *Eurocast*, 2003.

[29] Ed. bullmore, Chris Long, et al. Colored Noise and Computational Inference in Neurophysiological (fMRI) Time Series Analysis : Resampling Methods in Time and Wavelets Domains. *Human Brain mapping*, 12:61–78, 2001.

[30] M.A. Burock and A.M. Dale. Estimation and Detection of Event-Related fMRI Signals with Temporally Correlated Noise : A Statistically Efficient and Unbiased Approach. *Human Brain Mapping*, 11:249–260, 2000.

[31] R.B. Buxton, E.C. Wong, and L.R. Frank. Dynamics of Blood Flow and Oxygen Metabolism During Brain Activation: the Balloon Model. *Magnetic Resonance Medicine*, 39:855–864, 1998.

[32] Richard B. Buxton. The Elusive Initial Dip. *NeuroImage*, 13:953–958, 2001.

[33] Richard B. Buxton. *Introduction to Functional Magnetic Resonance Imaging.* Cambridge University Press, 2002.

[34] V.D. Calhoun, T. Adali, L.K. Hansen, J. Larsen, and J.J. Pekar. ICA of Functional MRI Data: an Overview. In *4th international symposium on Independent Component Analysis and Blind Signal Separation (ICA2003))*, Nara, Japon, 2003.

[35] V.D Calhoun, T. Adali, V.B. McGinty, J.J. Pekar, T.D. Watson, and G.D. Pearlson. fMRI Activation in a Visual-Perception Task: Network of Areas Detected Using the General Linear Model and Independent Components Analysis. *NeuroImage*, 14:1080–1088, 2001.

[36] V.D. Calhoun, T. Adali, G.D. Pearlson, and J.J. Pekar. A Method for Making Group Inferences from Functional MRI Data Using Independent Component Analysis. *Human Brain Mapping*, 14:140–151, 2001.

[37] V.D. Calhoun, T. Adali, G.D. Pearlson, and J.J. Pekar. Spatial and Temporal Independent Component Analysis of Functional MRI Data Containing a Pair of Task-Related Waveforms. *Human Brain Mapping*, 13:43–53, 2001.

[38] Vince D. Calhoun, james Pekar, Vince B. mcGinty, Tulay Adali, todd D. Watson, and Godfrey D. Pearlson. Different activation dynamics in multiple neural systems during simulated driving. *Human Brain Mapping*, 16:158–167, 2002.

[39] Jean-Francois Cardoso. Blind Signal Separation: Statistical Principles. In *Proceedings of the IEEE. Special issue on blind identification and estimation*, volume 9, pages 2009–2025, October 1998.

[40] John D. Carew, Grace Wahba, Xianhong Xie, Eric Nordheim, and Elisabeth M. Meyerand. Optimal spline smoothing of fMRI time series by generalized cross-validation. *NeuroImage*, 18:950–961, 2003.

[41] Christophe Chefd'hotel, Gerardo Hermosillo, and Olivier Faugeras. Flows of Diffeomorphisms for Multimodal Image Registration. In *International Symposium on Biomedical Imaging.* IEEE, 2002.

[42] F. Chochon, L. Cohen, P.F. van de Moortele, and S. Dehaene. Differential contributions of the left and right inferior parietal lobules to number processing. *Journal of Cognitive Neuroscience*, 11(6):617–630, November 1999.

[43] Kai-Hsiang Chuang, Ming-Jang Chiu, Chung-Chih Lin, and Chen Jyh-Horng. Model-Free Functional MRI Analysis Using Kohonen Clustering Neural Network and Fuzzy C-Means. *IEEE Transactions on Medical Imaging*, 18(12):1117–1128, 1999.

[44] Philippe Ciuciu, Guillaume Marrelec, Jean-Baptiste Poline, Jerome Idier, and Habib Benali. Robust Estimation of the Hemodynamic Response Function in Asynchronous Mutlitasks Multisessions Event-Related fMRI Paradigms. In *Proceedings of ISBI*, pages 847–850, Washington, D.C., 2002. IEEE, NIH.

[45] Mark S. Cohen. Parametric Analysis of fMRI Data Using Linear Systems Methods. *NeuroImage*, 6:93–103, 1997.

[46] Y. Cointepas, J.-F. Mangin, Line Garnero, J.-B. Poline, and H. Benali. BrainVISA: Software platform for visualization and analysis of multi-modality brain data. In *Proc. 7th HBM*, page S98, Brighton, United Kingdom, 2001.

[47] Olivier Coulon. *Analyse multi-échelle de cartes d'activations fonctionnelles cérébrales.* PhD thesis, ENST, October 1998.

[48] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* Wiley Series in Telecommunications, 1991.

[49] A.M. Dale and R.L. Buckner. Selective Averaging of Rapidly Presented Individual Trials Using fMRI. *Human Brain Mapping*, 5:329–340, 1997.

[50] P.A. De Mazière and M.M. Van Hulle. Towards a Spatio-Temporal Analysis Tool for fMRI Data: An Application to Depth-from-Motion Processing in Humans. In R.M. French and J.P. Sougné, editors, *Connectionist Models of Learning, Development and Evolution.*, pages 32–42. Springer, 2000.

[51] Valeria Della-Maggiore, Wilkin Chau, Pedro R. Peres-Neto, and Anthony R. McIntosh. An Empirical Comparison of SPM Preprocessing Parameters to the Analysis of fMRI Data. *NeuroImage*, 17(1):19–28, 2002.

[52] R. Deriche. Recursively implementing the gaussian and its derivatives. Technical Report 1893, INRIA, Unité de Recherche Sophia-Antipolis, 1993.

[53] Xavier Descombes, Frithjof Kruggel, and D.Y. von Cramon. fMRI Signal Restoration Using an Edge Preserving Spatio-temporal Markov Random Field. *NeuroImage*, 8:340–349, 1998.

[54] Jeng-Ren Duann, Tzyy-Ping Jung, Wen-Jui Kuo, Tzu-Chen Yeh, Scott Makeig, Hsieh Jen-Chuen, and Terrence J. Sejnovski. Single-Trial Variability in Event-Related BOLD Signals. *NeuroImage*, 15:823–835, 2002.

[55] J.-P. Eckmann and D. Ruelle. Ergodic Theory of Chaos and Stange Attractors. *Review of Modern Physics*, 57(3):617–656, 1985.

[56] A. Edelman. *Eigenvalues and Condition Numbers of Random Matrices.* PhD thesis, MIT, 1989.

[57] Fabrizio Esposito, Elia Formisano, Erich Seifritz, Rainer Goebel, Renato Morrone, Gioacchino Tedeschi, and Francesco Di Salle. Spatial Independent Component Analysis of Functional MRI Time-Series: To what Extent do Results Depend on the Algorithm Used ? *Human brain mapping*, 16:146–157, 2002.

[58] Brian S. Everitt and Edward T. Bullmore. Mixture Model Mapping of Brain Activation in Functional Magnetic Resonance Images. *Human Brain Mapping*, 7:1–14, 1999.

[59] M.J. Fadili, S. Ruan, D. Bloyet, and B. Mazoyer. A Multistep Unsupervised Fuzzy Clustering Analysis of fMRI Time Series. *Human Brain Mapping*, 10:160–178, 2000.

[60] M.J. Fadili, S. Ruan, D. Bloyet, and B. Mazoyer. On the Number of Clusters and the Fuzziness Index for Unsupervised FCA Applications to BOLD fMRI Time Series. *Medical Image Analysis*, 5:55–67, 2001.

[61] Pierre Faurre, Michel Clerget, and Francois Germain. *Opérateurs Rationnels Positifs.* Méthodes Mathématiques de l'Informatique. Dunod, 1979.

[62] Harald Fischer and Jürgen Hennig. Neural Network-Based Analysis of MR Time Series. *Magnetic Resonance Medicine*, 41:124–131, 1999.

[63] John W. Fisher, Alexander T. Ihler, and Paul A. Viola. Learning Informative Statistics: A Nonparametric Approach. In *Advances in Neural Information Processing Systems*, 1999.

[64] J.W. Fisher, E.R. Cosman, C. Wible, and W.M. Wells. Adaptative Entropy Rates for fMRI Time-Series Analysis. In Niessen and Viergever [167], pages 905–912.

[65] Guillaume Flandin, Ferath Kherif, Xavier Pennec, Grégoire Malandain, Nicholas Ayache, and Jean-Baptiste Poline. Improved Detection Sensitivity in Functional MRI Data Using a Brain Parcelling Technique. In Springer-Verlag, editor, *Medical Image Computing and Computer-Assisted Intervention-MICCAI2002*, volume 2488 of *Lecture Notes in Computer Science*, 2002.

[66] Guillaume Flandin, Ferath Kherif, Xavier Pennec, Denis Riviere, Nicholas Ayache, and Jean-Baptiste Poline. Parcellation of Brain Images with Anatomical and Functional Constraints for fMRI Data Analysis. In *Proceedings of the 2002 IEEE International Symposium on Biomedical Imaging*, pages 907–910, July 2002.

[67] Guillaume Flandin, Will Penny, Xavier Pennec, Nicholas Ayache, and Jean-Baptiste Poline. A multisubject anatomo-functional parcellation of the brain. In *NeuroImage (HBM'03)*, New York, USA, 2003.

[68] E. Formisano, F. Esposito, N. Kriegeskorte, G. Tedeschi, F. Di Salle, and R. Goebel. Spatial Independent Component Analysis of Functional Magnetic Resonance Imaging Time-series: Characterization of the Cortical Components. *Neurocomputing*, 49:241–254, 2002.

[69] P.T. Fox and M.E. Raichle. Stimulus Rate Dependence of Regional Cerebral Blood Flow in Human Striate Cortex, Demonstrated by Positron Emission Tomography. *Journal of Neurophysiology*, 51:1109–1120, 1991.

[70] L. Freire, A. Roche, and J.F. Mangin. What is the Best Similarity Measure for Motion Correction in fMRI Time Series? *IEEE Transactions on Medical Imaging*, 21(5):470–484, May 2002.

[71] Ola Friman, Magnus Borga, Peter Lundberg, and Hans Knutsson. Detection of Neural Activity in fMRI Using Maximum Correlation Modeling. *NeuroImage*, 15:386–395, 2002.

[72] Ola Friman, Magnus Borga, Peter Lundberg, and Hans Knutsson. Exploratory fMRI Analysis by Autocorrelation Maximization. *NeuroImage*, 16:454–464, 2002.

[73] Ola Friman, Magnus Borga, Peter Lundberg, and Hans Knutsson. Adaptative analysis of fMRI data. *NeuroImage*, 19(3):837–845, July 2003.

[74] Karl Friston, Jacquie Phillips, Dave Chawla, and Christian Buchel. Revealing Interactions among Brain Systems with Nonlinear PCA. *Human Brain Mapping*, 8:92–97, 1999.

[75] Karl Friston, Jacquie Phillips, Dave Chawla, and Christian Buchel. Nonlinear PCA: Characterizing Interactions between Modes of Brain Activity. *Philos. Trans. R. Soc. Lond. Biol. Sc.*, 355:135–146, 2000.

[76] K.J. Friston. Bayesian Estimation of Dynamical Systems: An Application to fMRI. *NeuroImage*, 16:513–530, 2002.

[77] K.J. Friston, J. Ashburner, et al. *SPM 97 course notes*. Wellcome Department of Cognitive Neurology, University College london, 1997.

[78] K.J. Friston, P. Flechter, O. Josephs, A. Holmes, M.D. Rugg, and R. Turner. Event-Related fMRI: Characterizing Differential Responses. *NeuroImage*, 7(1):30–40, 1998.

[79] K.J. Friston, C. D. Frith, R. S. J. Frackowiak, and R. Turner. Characterizing Dynamic Brain Responses with fMRI: a Multivariate Approach. *NeuroImage*, 2:166–172, 1995.

[80] K.J. Friston, C.D. Frith, P.F. Liddle, and R.S. Frackowiak. Functional Connectivity: the Principal-Component Analysis of large PET Data Sets. *Journal of Cerebral Blood Flow Metabolism*, 13(1):5–14, January 1993.

[81] K.J. Friston, D.E. Glaser, R.N.A. Henson, S. Kiebel, C. Phillips, and J. Ashburner. Classical and Bayesian Inference in Neuroimaging: Applications. *NeuroImage*, 16(2):484–512, 2002.

[82] K.J. Friston, P. Jezzard, and R. Turner. The Analysis of functional MRI Time-Series. *Human Brain Mapping*, 1:153–171, 1994.

[83] K.J. Friston, O. Josephs, G. Rees, and R. Turner. Nonlinear Event-Related Responses in fMRI. *Magnetic Resonance in Medicine*, 39(1):41–52, 1998.

[84] K.J Friston, O. Josephs, E. Zarahn, A. Holmes, S. Rouquette, and Poline J.-B. To Smooth or not to Smooth ? Bias and Efficiency in fMRI Time-Series Analysis. *NeuroImage*, 12:196–208, 2000.

[85] K.J. Friston, A. Mechelli, R. Turner, and C.J. Price. Nonlinear Responses in fMRI: The Balloon Model, Volterra Kernels and other Hemodynamics. *NeuroImage*, 12:466–477, 2000.

[86] K.J. Friston and W. Penny. Posterior Probability maps and SPMs. *NeuroImage*, 19(3):1240–1249, July 2003.

[87] K.J. Friston, W. Penny, C. Phillips, S. Kiebel, G. Hinton, and J. Ashburner. Classical and Bayesian Inference in Neuroimaging: Theory. *NeuroImage*, 16(2):465–483, June 2002.

[88] K.J. Friston, J.B. Poline, S. Strother, A.P. Holmes, C.D. frith, and R.S.J. Frackowiak. A Multivariate Analysis of PET Activation Studies. *Human Brain Mapping*, 4:140–151, 1996.

[89] K.J. Friston, K.J. Worsley, R.S.J Frackowiak, J.C. Mazziotta, and A.C. Evans. Assessing the Significance of Focal Activations Using their Spatial Extent. *Human Brain Mapping*, 1:214–220, 1994.

[90] Christopher R. Genovese, Nicole A. Lazar, and Thomas Nichols. Thresholding of Statistical Maps in Functional Neuroimaging Using the False Discovery Rate. *NeuroImage*, 15(4):870–878, 2002.

[91] Marc G. Genton. Classes of kernels for machine learning: A statistics perspective. *Journal of Machine Learning Research*, 2:299–312, 2001.

[92] Zoubin Ghahramani and Geoffrey Hinton. Parameter Estimation for Linear Dynamical Systems. Technical Report CRG-TR-96-2, University of Toronto, 1996.

[93] Gary H. Glover. Deconvolution of Impulse Response in Event-related BOLD fMRI. *NeuroImage*, 9:416–429, 1999.

[94] G.H. Golub and C.F. van Loan. *Matrix Computations*. The John Hopkins University Press, Baltimore, Maryland, second edition, 1989.

[95] C. Gössl, D.P. Auer, and L. Fahrmeir. Dynamic Models in fMRI. *Magnetic Resonance in Medicine*, 43:72–81, 2000.

[96] C. Goutte, P. Troft, E. Rostrup, A. Nielsen, and L.K. Hansen. On clustering fMRI time series. *NeuroImage*, 9(3):298–310, 1998.

[97] Cyril Goutte, Lars Kai Hansen, Mattew G. Liptrot, and Egill Rostrup. Feature space clustering for fMRI meta-analysis. Technical Report IMM-REP-1999-13, Technical University of Denmark, 1999.

[98] Cyril Goutte, Finn Arup Nielsen, and Lars Kai Hansen. Modeling the haemodynamic response in fmri using smooth fir filters. *IEEE Transactions on Medical Imaging*, 19(12):1188–1201, 2000.

[99] Peter Grassberger and Itamar Procaccia. Measuring the strangeness of strange attractors. *Physica*, 9D:189–208, 1983.

[100] K. Grill-Spector and R. Malach. fMR-adaptation : a tool for studying the functional properties of human cortical neurons. *Acta Psychologica*, 107:293–321, 2001.

[101] Hong Gu, Wolfgang Engelien, Hanhua Feng, David A. Silbersweig, Emily Stern, and Yihong Yang. Mapping transient, randomly occuring neuropsychological events using independent component analysis. *NeuroImage*, 14:1432–1443, 2001.

[102] Lars Kai Hansen. ICA of fMRI based on a convolutive mixture model. In *Ninth Annual meeting for Human Brain Mapping*, number 847, June 2003.

[103] Lars Kai Hansen, Jan Larsen, Finn Arup Nielsen, Stephen C. Strother, Egill Rostrup, Robert Savoy, Nicholas Lange, John Sidtis, Claus Svarer, and Olaf B. Paulson. Generalizable Patterns in Neuroimaging: How Many Principal Components ? *NeuroImage*, 9:534–544, 1999.

[104] Lars Kai Hansen, Finn Arup Nielsen, Stephen C. Strother, and Nicholas Lange. Consensus inference in neuroimaging. *NeuroImage*, 13(6):1212–1218, 2001.

[105] Lars Kai Hansen, Finn Årup Nielsen, Peter Toft, Matthew George Liptrot, Cyril Goutte, Stephen C. Strother, Nicholas Lange, Anders Gade, David A. Rottenberg, and Olaf B. Paulson. "lyngby" - a modeler's matlab toolbox for spatio-temporal analysis of functional neuroimages. In *NeuroImage*, volume 9, June 1999.

[106] L.K. Hansen, F.A. Nielsen, and J. Larsen. Exploring fMRI data for periodic signal components. *Artificial Intelligence in Medicine*, 25(1):35–44, 2002.

[107] Stefan Harmeling, Andreas Ziehe, and Motoaki Kawanabe. Kernel-Based Nonlinear Blind Source Separation. *Neural Computation*, 15:1089–1124, 2003.

[108] Niels Vaever Hartvig and Jens Ledet Jensen. Spatial mixture modeling of fMRI data. *Human Brain mapping*, 11:233–248, 2000.

[109] Trevor Hastie, Robert Tishbirani, and Jerom Friedman. *The Elements of Statistical Learning.* Springer Series in Statistics. Springer, 2001.

[110] Gerardo Hermosillo. *Variational Methods for Multimodal Image Matching.* PhD thesis, INRIA, The document is accessible at ftp://ftp-sop.inria.fr /robotvis/html/Papers /hermosillo:02.ps.gz, 2002.

[111] Gerardo Hermosillo, Christophe Chefd'hotel, and Olivier Faugeras. Variational methods for multimodal image matching. *ijcv*, 50(3):329–343, November 2002.

[112] P. Højen-Sørensen, O. Winther, and L.K. Hansen. Analysis of functional neuroimages using ica adaptive binary sources. *Neurocomputing*, 49:213–225, 2002.

[113] Pedro Hojen-Sorensen, Lars K. Hansen, and Carl E. Rasmussen. Bayesian modelling of fMRI time series. In *NIPS'99*, 1999.

[114] Pedro A.d.F.R. Hojen-Sorensen, Ole Winther, and Lars Kai Hansen. Mean-field approaches to independent component analysis. *Neural Computation*, 14:889–918, 2002.

[115] Barry Horwitz. The elusive concept of brain connectivity. *NeuroImage*, 19:466–470, 2003.

[116] Scott A. Huettel and Gregory McCarthy. Regional differences in the refractory period of the hemodynamic response: An event-related fMRI study. *NeuroImage*, 14:967–976, 2001.

[117] A. Hyvärinen and E. Oja. A Fast Fixed-Point Algorithm for Independent Component Analysis. *Neural Computation*, 9(7):1483–1492, 1997.

[118] M.F. Insana and R.M. Leahy, editors. volume 2082 of *Lecture Notes in Computer Science.* Springer–Verlag Berlin Heidelberg, 2001.

[119] William James. *The Principles of Psychology.* Harvard: Cambridge, MA, 1890.

[120] M. Jarmasz and R.L. Somorjai. Exploring regions of interest with cluster analysis (eroica) using a spectral peak statistic for selecting and testing the significance of fMRI activation time-series. *Artificial Intelligencein Medicine*, 25(1):45–67, 2002.

[121] P. Jezzard. *Physiological Noise: Strategies for Correction*, chapter 16, pages 171–179. Springer-Verlag, Heidelberg, 1999.

[122] Iain M. Johnstone. On the distribution of the largest principal component. Technical report, Department of Statistics, Stanford University, 2000.

[123] kota Katanoda, Yasumasa Matsuda, and Morihiro Sugishita. A spatio-temporal regression model for the analysis of functional MRI dtata. *NeuroImage*, 17:1415–1428, 2002.

[124] J. Kershaw, B.A. Ardekani, and I. Kanno. Application of Bayesian inference to fMRI data analysis. *IEEE transactions on Medical imaging*, 18(12):1138–1153, December 1999.

[125] Ferath Kherif, Guillaume Flandin, Philippe Ciuciu, Habib Benali, Olivier Simon, and Jean-Baptiste Poline. Model based spatial and temporal similarity measures between series of functional magnetic resonance images. In Takeyoshi Dohi and Ron Kikinis, editors, *Medical Image Computing and Computer-Assisted Intervention (MICCAI'02)*, volume 2489 of *Lecture Notes in Computer Science*, pages 509–516, Tokyo, 2002.

[126] Ferath Kherif, Jean-Baptiste Poline, Guillaume Flandin, Habib Benali, Olivier Simon, Stanislas Dehaene, and Keith J. Worsley. Multivariate model specification for fMRI data. *NeuroImage*, 16(4):1068–1083, August 2002.

[127] Stefan J. Kiebel, Rainer Goebel, and Friston Karl J. Anatomically informed basis functions. *NeuroImage*, 11(6):656–667, June 2000.

[128] Stefan J. Kiebel, Poline J.-B., K.J. Friston, A. P. Holmes, and K.J. Worsley. Robust smoothness estimation in statistical parametric maps using standardized residuals from the general linear model. *NeuroImage*, 10:756–766, 1999.

[129] J. Kim, J.W. Fisher, A. Tsai, C. Wible, A.S. Willsky, and W.M. Wells. Incorporating spatial priors into an information theoretic approach for fMRI data analysis. In G. Goos, J. Hartmanis, and J. van Leeuwen, editors, *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2000*, volume 1935 of *Lectures Notes in Computer Science 1935*, pages 62–72. Springer, October 2000.

[130] Vesa Kiviniemi, Juha-Heikki Kantola, Jukka Jauhiainen, Aapo Hypavärinen, and Osmo Tervonen. Independent component analysis of nondeterministic fMRI signal sources. *NeuroImage*, 19(2):253–260, June 2003.

[131] Hansjörg Klock and Joachim Buhmann. Multidimensional Scaling by Deterministic Annealing. In M. Pelillo and E.R. Hancock, editors, *Energy Minimization Methods in Computer Vision and Pattern Recognition*, volume 1223 of *Lecture Notes in Computer Science*, pages 246–260. Springer, 1997.

[132] F. Kruggel and D.Y. von Cramon. Temporal properties of the hemodynamic response in functional MRI. *Human Brain mapping*, 8(4):259–271, 1999.

[133] Frithjof Kruggel and D.Y. von Cramon. Physiollogically oriented models of the hemodynamic response in functional MRI. In A. Kuba et al., editors, *IPMI'99*, volume 1613 of *Lecture Notes in Computer Science*, pages 294–307. Springer–Verlag Berlin Heidelberg, 1999.

[134] Stephen LaConte, Jon Anderson, Muley Suraj, James Ashe, Sally Frutiger, Kelly Rehm, Lars Kai Hansen, Essa Yacoub, Xiaoping Hu, David Rottenberg, and Stephen Strother. The evaluation of preprocessing choices in single-subject BOLD fMRI using NPAIRS performance metrics. *NeuroImage*, 18(1):10–27, 2003.

[135] S. Lai, A.L. Hopkins, E.M. Haacke, D. Li, B.A. Wasserman, P. Buckley, L. Friedman, H. Meltzer, P. Hedera, and R. Friedland. Identification of vascular structures as a major source of signal contrast in high resolution 2D and 3D functional activation imaging of the motor cortex at 1.5T: Preliminary results. *Magnetic Resonance in Medicine*, 30:387–392, 1993.

[136] Angela R. Laird, Baxter P. Rogers, John D. Carew, Konstantinos Arfanakis, Chad H. Moritz, and M. Elizabeth Meyerand. Characterizing instantaneous phase relationship in whole-brain fMRI activation data. *Human Brain Mapping*, 16:71–80, 2002.

[137] Nicholas Lange and Scott L. Zeger. Non-linear fourier time serie analysis for human brain mapping by functional magnetic resonance imaging. *Appl. Statist.*, 46(1):1–29, 1997.

[138] Michael Lewicki and Terrence J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12:337–365, 2000.

[139] C.H. Liao, K.J. Worsley, J.-B. Poline, J.A.D. Aston, G. H. Duncan, and A.C. Evans. Estimating the delay of the fMRI response. *NeuroImage*, 16:593–606, 2002.

[140] Thomas T. Liu, Lawrence R. Frank, Eric C. Wong, and Richard B. Buxton. Detection power, estimation efficiency and predictability in event-related fMRI. *NeuroImage*, 13:759–773, 2001.

[141] J.J. Locascio, P.J. Jennings, Christopher I. Moore, and Suzanne Corkin. Time series analysis in the time domain and resampling for studies of functional magnetic resonance brain mapping. *Human brain mapping*, 5:168–193, 1997.

[142] N.K. Logothetis, J. Pauls, M. Augath, T. Trinath, and A. Oeltermann. Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412(6843):150–157, 2001.

[143] Gabriele Lohmann and Stefan Bohn. Using Replicator Dynamics for Analysing fMRI Data of the Human Brain. *IEEE TMI*, 21(5):485–492, 2002.

[144] Gabriele Lohmann and D. Yves von Cramon. Detecting Functionally Coherent Networks in fMRI Data of the Human Brain Using Replicator Dynamics. In Insana and Leahy [118], pages 218–224.

[145] A. Lukic, M.N. Wernick, L.K. Hansen, J. Anderson, and Strother S.C. A spatially robust ICA algorithm for multiple datasets. In *Proceedings of the 2002 IEEE International Symposium on Biomedical Imaging*, pages 839–842, July 2002.

[146] D. Malonek and A. Grinvald. Interactions between electrical activity and cortical microcirculation revealed by imaging spectroscopy: Implications for functional brain imagaging. *Science*, 272:551–554, 1996.

[147] Jonathan L. Marchini and Brian D. Ripley. A new statistical approach to detecting significant activation in functional MRI. *Neuroimage*, 12:366–380, 2000.

[148] Jonathan L. Marchini and Stephen M. Smith. On bias in the estimation of autocorrelations for fMRI voxel time-series analsysis. *NeuroImage*, 18:83–90, 2003.

[149] G. Marrelec, H. Benali, P. Ciuciu, and J.-B. Poline. Bayesian estimation of the hemodynamic response function in functional MRI. In *MAXENT 2001 (21st Int. Work. on Bayesian Inference and Maximum Entropy Methods in Science and Engineering)*, 2001.

[150] Guillaume Marrelec, Habib Benali, Philippe Ciuciu, Mélanie Pélégrini-Issac, and Jean-Baptiste Poline. Robust Bayesian Estimation of the Hemodynamic Response Function in Event-Related BOLD fMRI Using Basic Physiological Information. *Human Brain Mapping*, 19:1–17, 2003.

[151] A.R. McIntosh, F.L. Bookstein, J.V. Haxby, and C.L. Grady. Spatial pattern analysis of functional brain images using partial least squares. *NeuroImage*, 3:143–157, 1996.

[152] Martin J. McKeown. Detection of consistently task-related activations in fMRI data with hybrid independent component analysis. *NeuroImage*, 11:24–35, 2000.

[153] Martin J. McKeown, S. Makeig, et al. Analysis of fMRI data by blind separation into independant spatial components. *Human Brain Mapping*, 6:160–188, 1998.

[154] M.J. McKeown, T.P. Jung, et al. Spatially independant activity patterns in functional MRI data during the stroop color-naming task. *Proc. Natl. Acad. Sci. USA*, 95:803–810, February 1998.

[155] F.G. Meyer and G. McCarthy. Wavelet based estimation of baseline drifts in fMRI. In *Information Processing in Medical Imaging*, pages 232–238. Springer Verlag, 2001.

[156] F.M. Miezin, L. Macotta, J.M. Ollinger, S.E. Petersen, and R.L. Buckner. Characterizing the hemodynamic response: Effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. *NeuroImage*, 11:735–759, 2000.

[157] K.L. Miller, W.-M. Luh, T.T. Liu, et al. Nonlinear temporal dynamics of the cerebral blood flow response. *Human Brain Mapping*, 13:1–12, 2001.

[158] Partha P. Mitra, Seiji Ogawa, Xiaoping Hu, and Kamil Ugurbil. The Nature of Spatiotemporal Changes in Cerebral Hemodynamics As manifested in Functional Magnetic Resonance Imaging. *Magnetic Resonance in Medicine*, 37:511–518, 1997.

[159] Perry Moerland. An On-Line EM Algorithm Applied to Kernel PCA. Technical Report IDIAP-RR 00-xx, 2000, Dalle Molle Institute for Perceptual Artificial Intelligence, 2000.

[160] L. Molgedey and H.G. Schuster. Separation of independent signals using time-delayed correlations. *Physical Review Letters*, 72(23):3634–3637, 1994.

[161] U. Möller, M. Ligges, P. Georgiewa, C. Grünling, W. A. Kaiser, H. Witte, and B. Blanz. How to avoid spurious cluster validation ? A methodological investigation on simulated and fMRI data. *NeuroImage*, 17:431–446, 2002.

[162] Karsten Müller, Gabriele Lohmann, Volker Bosh, and D. Yves von Cramon. On multivariate spectral analysis of fMRI times series. *NeuroImage*, 14:347–356, 2001.

[163] F. N. Nan and R.D. Nowak. Generalized likelihood ratio detection for fMRI using complex data. *IEEE Transactions on Medical Imaging*, 18(4):320–329, April 1999.

[164] Jane Neumann, Gabriele Lohmann, Stefan Zysset, and D. Yves von Cramon. Within-subject variability of BOLD response dynamics. *NeuroImage*, 19(3):784–796, July 2003.

[165] Shing-Chung Ngan and Xiaoping Hu. Analysis of functional magnetic resonance imaging data using self-organizing mapping with spatial connectivity. *Magnetic Resonance Medicine*, 41:939–946, 1999.

[166] Thomas E. Nichols and Andrew P. Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping*, 15:1–25, 2001.

[167] W. J. Niessen and M. A. Viergever, editors. volume 2208 of *Lecture Notes in Computer Science*. Springer, 2001.

[168] S. Ogawa, T.M. Lee, A.R. Kay, and D.W. Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 87, pages 9868–9872, December 1990.

[169] S. Ogawa, D.W. Tank, R. Menon, J.M. Ellermann, S.G. Kim, H. Merkle, and K. Ugurbil. Intrinsic signal changes accompanying sensory stimulation: Functional brain mapping with magnetic resonance imaging. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 89, pages 5951–5955, 1992.

[170] Will Penny, Stefan Kiebel, and Karl Friston. Variational Bayesian inference for fMRI time series. *NeuroImage*, 19(3):727–741, July 2003.

[171] K.S. Petersen, L.K. Hansen, T. Kolenda, E. Rostrup, and S.C. Strother. On the independent components of functional neuroimages. In *ICA'2000*, pages 615–620, 2000.

[172] K.M. Petersson, T.E. Nichols, J.B. Poline, and A.P. Holmes. Statistical limitations in functional neuroimaging. 1. non-inferential methods and statistical models. *Phil. Trans. R. Soc. Lond.*, 354:1239–1260, 1999.

[173] K.M. Petersson, T.E. Nichols, J.B. Poline, and A.P. Holmes. Statistical limitations in functional neuroimaging. 2. signal detection and statistical inference. *Phil. Trans. R. Soc. Lond.*, 354:1261–1281, 1999.

[174] Josef Pfeuffer, Jeffery C. McCullough, Pierre-françois Van de Moortele, Kamil Ugurbil, and Xiaoping Hu. Spatial dependence of the nonlinear BOLD response at short stimulus duration. *NeuroImage*, 18:990–1000, 2003.

[175] Dinh Tuan Pham. Mutual Information Approach to Blind Separation of Stationary Sources. *IEEE Transactions on Information Theory*, 48(7):1935–1946, July 2002.

[176] Jayasanka Piyaratna and Jagath Rajapakse. Spatiotemporal analysis of functional images using the fixed effect model. In Insana and Leahy [118], pages 190–196.

[177] J.-B. Poline and B.M. Mazoyer. Analysis of individual brain activation maps using hierarchical description and multiscale detection. *IEEE Transactions on Medical Imaging*, 13(4):702–710, 1994.

[178] J.B. Poline, K.J. Worsley, A.C. Evans, and K.J. Friston. Combining spatial extent and peak intensity to test for activations infunctional imaging. *NeuroImage*, 5:83–96, 1997.

[179] William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. *Numerical Recipes in C*. Cambridge University Press, 1988.

[180] Itamar Procaccia. The static and dynamic invariants that characterize chaos and the relations between them in theory and experiments. *Physica Scripta*, T9:40–46, 1985.

[181] Patrick L. Purdon, Victor Solo, Robert M Weisskoff, and Emery N. Brown. Locally regularized spatiotemporal modeling and model comparison for functional MRI. *Neuroimage*, 14:912–923, 2001.

[182] Patrick L. Purdon and Robert M. Weisskoff. Effect of Temporal Autocorrelation Due to Physiological Noise and Stimulus Paradigm on Voxel-Level False-Positive Rates in fMRI. *Human Brain Mapping*, 6:239–249, 1998.

[183] Jagath C. Rajapakse, Frithjof Kruggel, Jose M. Maisog, and D. Yves von Cramon. Modeling hemodynamic response for analysis of functional MRI time-series. *Human Brain Mapping*, 6:283–300, 1998.

[184] Scott Rauch, Paul Whalen, et al. Striatal recruitment during an implicit sequence learning task as measured by functional magnetic resonance imaging. *Human Brain Mapping*, 5:124–132, 1997.

[185] Jonathan Raz, Hui Zheng, Hernando Ombao, and Bruce Turetsky. Statistical tests for fMRI based on experimental randomization. *NeuroImage*, 19(2):226–232, June 2003.

[186] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[187] J. Rissanen. Minimum description length principle. *Encyclopedia of Statistic Sciences*, 5:523–527, 1987.

[188] J.J. Rissanen. Fisher information and stochastic complexity. *IEEE Trans. Information Theory*, 42(1):40–47, January 1996.

[189] Jorma Rissanen. Universal coding, information, prediction, and estimation. *IEEE Transactions on Information Theory*, 30(4):629–636, July 1984.

[190] Sam. T. Roweis and Lawrence K. Saul. Nonlinear Dimensionality Reduction by locally Linear Embedding. *Science*, 290:2323–2326, 2000.

[191] Ziad S. Saad, Edgar A. De Yoe, and Kristina M. Ropella. Estimation of fMRI response delays. *NeuroImage*, 18:494–504, 2003.

[192] Ziad S. Saad, Kristina M. Ropella, Robert W. Cox, and Edgar A. De Yoe. Analysis and use of fMRI response delays. *Human Brain Mapping*, 13:74–93, 2001.

[193] Ziad S. Saad, Kristina M. Ropella, Edgar A. DeYoe, and Peter A. Bandettini. The spatial extent of the BOLD response. *NeuroImage*, 19(1):132–144, 2003.

[194] Bernhard Schölkkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

[195] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6:461–464, 1978.

[196] M.I. Sereno, A.M. Dale, J.B. Reppas, K.K. Kwong, J.W. Belliveau, T.J. Brady, B.R. Rosen, and R.B.H. Tootell. Borders of multiple visual areas in human revealed by functional magnetic resonance imaging. *Science*, pages 889–893, 1995.

[197] M.I. Sereno, C.T. McDonald, and J.M. Allman. Analysis of retinotopic maps in extrastriate cortex. *Cerebral Cortex*, 4:601–620, 1994.

[198] Stefano Soatto and Alessandro Chiuso. Dynamic data factorization. Technical Report 010001, Department of Computer Science, UCLA, March 2001.

[199] V. Solo, P. Purdon, R. Weisskoff, and E. Brown. A signal estimation approach to functional MRI. *IEEE Transactions on Medical Imaging*, 20(1):26–35, January 2001.

[200] James V. Stone. Blind source separation using temporal predictability. *Neural Computation*, 13:1559–1574, 2001.

[201] J.V. Stone, J. Porril, N.R. Porter, and I.D. Wilkinson. Spatiotemporal Independent Component Analysis of Event-Related fMRI Data Using Skewed Probability Density Functions. *NeuroImage*, 15(2):407–422, 2002.

[202] M. Svensén, F. Kruggel, and D.Y. von Cramon. Probabilistic modeling of single-trial fMRI data. *IEEE Transactions on Medical Imaging*, 19(1):25–35, January 2000.

[203] Jody Tanabe, David Miller, Jason Tregellas, Robert Freedman, and Francois G. Meyer. Comparison of detrending methods for optimal fMRI preprocessing. *NeuroImage*, 15:902–907, 2002.

[204] Joshua Tenenbaum, Vin De Silva, and C. Langford, John. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290:2319–2323, 2000.

[205] Bertrand Thirion and Olivier Faugeras. Activation detection and characterisation in brain fMRI sequences. application to the study of monkey vision. Technical Report RR-4213, INRIA, June 2001.

[206] Bertrand Thirion and Olivier Faugeras. Revisiting non-parametric activation detection on fMRI time series. In *MMBIA'2001*, December 2001.

[207] Bertrand Thirion and Olivier Faugeras. Dynamical components analysis of fMRI data. In *Proceedings of the 2002 IEEE International Symposium on Biomedical Imaging*, pages 915–918, July 2002.

[208] Bertrand Thirion and Olivier Faugeras. Dynamical components analysis of fMRI data through kernel PCA. *NeuroImage*, 20(1):34–49, 2003.

[209] Bertrand Thirion and Olivier Faugeras. Feature detection in fMRI data: The information bottleneck approach. In *MICCAI 2003*, November 2003. accepted.

[210] Bertrand Thirion and Olivier Faugeras. Multivariate analysis of fMRI data : A second order solution. In *Eurocast*, 2003. accepted.

[211] Bertrand Thirion, Ferath Kherif, Jean-Baptiste Poline, and Olivier Faugeras. Multivariate analysis of fMRI data: Is it worth using nonlinear methods ? In *9th International Conference on Functional Mapping of the Human Brain*, 2003.

[212] Christopher G. Thomas, Richard A. Harshman, and Ravi S. Menon. Noise reduction in BOLD-based fMRI using component analysis. *NeuroImage*, 17:1521–1537, 2002.

[213] Laurent Thoraval, Jean-Paul Armspach, and Izzie Namer. Analysis of brain functional MRI time series based on continuous wavelet transform and stimulation-response coupling distance. In Niessen and Viergever [167], pages 881–888.

[214] Naftali Tishby, Fernando C. Pereira, and William Bialek. The Information Bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

[215] A. Tsai, J.W. Fisher, C. Wible, W.M. Wells, J. Kim, and A.S. Willsky. Analysis of functional MRI data using mutual information. In *Second International Conference on medical Image Computing and Computer-Assisted Intervention*, volume 1679 of *Lecture Notes in Computer Science*, pages 473–480. Springer–Verlag, September 1999.

[216] Harri Valpola and Juha Karhunen. An unsupervised ensemble learning method for nonlinear dynamic state-space models. *Neural Computation*, 14:2647–2692, 2002.

[217] W. Vanduffel, D. Fize, H. Peuskens, K. Denys, S. Sunaert, J.T. Todd, and G. A. Orban. Extracting 3d from Motion: Differences in Human and Monkey Intraparietal Cortex. *Science*, 298:413–415, 2002.

[218] Wim Vanduffel, Denis Fize, Joseph B. Mandeville, Koen Nelissen, Paul Van Hecke, Bruce R. Rosen, Roger B.H. Tootell, and G. Orban. Visual motion processing investigated using contrast-enhanced fMRI in awake behaving monkeys. *Neuron*, 2001. in press.

[219] Alberto L. Vazquez and Douglas C. Noll. Nonlinear aspects of the BOLD response in functional MRI. *Neuroimage*, 7:108–118, 1998.

[220] Axel Wismüller and Olivier Lange. Cluster analysis of biomedical image time-series. *ijcv*, 46(2):103–128, 2002.

[221] M.W. Woolrich, B.D. Ripley, M. Brady, and S.M. Smith. Temporal autocorrelation in univariate modeling of fMRI data. *NeuroImage*, 14:1370–1386, 2001.

[222] K.J. Worsley and K.J. Friston. Analysis of fMRI time-series revisited - again. *NeuroImage*, 2:173–181, 1995.

[223] K.J. Worsley, C.H. Liao, J. Aston, V. Petre, G.H. Duncan, F. Morales, and Evans A.C. A general statistical analysis for fMRI data. *NeuroImage*, 15:1–15, 2002.

[224] K.J. Worsley, J-B. Poline, K.J. Friston, and Evans A.C. Characterizing the response of PET and fMRI data using multivariate linear models. *NeuroImage*, 6:305–319, 1997.

[225] N. Wotawa. Une Experience de Retinotopie par IRM Fonctionnelle. Technical report, DEA Mathematiques, Vision, Apprentissage, ENS-Cachan, September 2002.

[226] E. Zarahn, G.K. Aguirre, and M. D'Esposito. Empirical Analyses of Bold fMRI Statistics. 1. Spatially Unsmoothed Data Collected under Null-Hypothesis. *NeuroImage*, 5:179–197, 1997.