Probability bounds for the cross-validation estimate in the context of the statistical learning theory and statistical models applied to economics and finance

Matthieu Cornec

UNIVERSITÉ PARIS NANTERRE - PARIS 10
ÉCOLE DOCTORALE CONNAISSANCE, LANGAGE, MODÉLISATION

DOCTORAT
MATHÉMATIQUES APPLIQUÉES et APPLICATION DES MATHÉMATIQUES

présenté par
Matthieu CORNEC

sous le titre

Inégalités probabilistes pour l'estimateur de validation croisée dans le cadre de l'apprentissage statistique et Modèles statistiques appliqués à l'économie et à la finance

Directeur de Thèse
M. Patrice BERTAIL

soutenue le 4 juin 2009.

JURY

AID René, Laboratoire de Finance des Marchés d'énergies
BIAU Gérard, Université Pierre et Marie Curie - Paris VI
DOUKHAN Paul, Université Cergy Pontoise
TSYBAKOV Alexandre, CREST et Université Paris VI
VAYATIS Nicolas, Ecole Normale Supérieure de Cachan

Rapporteurs

DUDOIT Sandrine, University of California Berkeley
LUGOSI Gabor, Pompeu Fabra University

Remerciements

Je tiens à exprimer mes remerciements les plus sincères à mon directeur de thèse Patrice Bertail, pour m'avoir encadré sur ce sujet de recherche, pour sa grande gentillesse et sa disponibilité, et pour les Mathématiques que j'ai pu apprendre grâce à lui.

Je remercie très sincèrement Sandrine Dudoit et Gabor Lugosi d'avoir accepté de rapporter cette thèse. Je suis très honoré par leur lecture attentive de ce manuscrit, leurs précieux commentaires et leur intérêt pour mon travail. Je remercie également René Aid, Gérard Biau, Paul Doukhan, Sacha Tsybakov, et Nicolas Vayatis de m'avoir fait l'honneur d'accepter de faire partie de mon jury. Chacun d'entre eux a apporté au cours de ces années un soutien dans mon travail dont je suis sincèrement reconnaissant.

Le laboratoire du CREST m'a mis à disposition un bureau pendant ces deux dernières années et je tiens à remercier tous les membres pour leur accueil : Eric Gautier et Jean-François Chassagneux (notamment pour leur précieuse aide en latex), Emmanuelle Gautherat, Christian Robert, Judith Rousseau, Jean-Michel Zakoian, Romuald Elie, Hugo Harari Kermadec, Matthieu Rosenbaum, Pierre Alquier, Julian Arbel, Nicolas Chopin, Alain Monfort, Xavier Mary, Nazim Regnard, Christian-Yann Robert, Lionel Truquet, Arthur Charpentier, Stéphane Grégoir, Arnaud Porchet, Fabrice Murtin. Merci encore à Nadine Guedj pour son aide sur les problèmes d'impression de dernière minute.

J'ai suivi les séminaires du laboratoire FIME dont j'ai beaucoup appris et je tiens à remercier les organisateurs et les participants : Nizar Touzi, Alfred Galichon, Luciano Campi, Christian Gourieroux, Bertrand Villeneuve, Gilles Chemla, Olivier Feron, Nadja Oudjane, Xavier Warin, Adrien Nguyen Huu.

Je tiens également à remercier le personnel du département OSIRIS de la R&D d'EDF qui m'ont gentiment libéré du temps ces deux dernières années pour finir ma thèse. Je tiens aussi à exprimer ma gratitude envers Georges Oppenheim pour ces conseils avisés.

Je terminerai en remerciant chaleureusement évidemment mes amis et ma famille, sans eux rien n'aurait été possible.

Contents

Introduction générale

Introduction

L'objectif initial de la première partie de cette thèse est d'éclairer par la théorie une pratique communément répandue au sein des practiciens pour l'audit (ou *risk assessment* en anglais) de méthodes prédictives (ou prédicteurs) : la validation croisée (ou *cross-validation* en anglais). En effet, l'entrée "cross-validation" sur Google Scholar révèle la popularité de la méthode avec près de 200000 résultats. Cependant d'un point de vue théorique, les raisons qui justifieraient l'emploi de la validation croisée restent encore assez floues. Précisons tout de suite que la validation croisée est utilisée avec succès dans la résolution d'un grand nombre de problèmes. Sans être exhaustif, citons dans le domaine de la Statistique : la sélection de modèle, l'adaptivité et l'identification. Qui plus est, elle peut même être vue comme une application particulière des méthodes dites de rééchantillonnage.

Le second volet complémentaire de cette thèse en est la contrepartie : apporter un support théorique à des questions pratiques rencontrées au sein du monde professionnel. En effet, cette thèse a été réalisée en parallèle d'une activité salariée, d'abord au Ministère des Finances puis ensuite à EDF. Cette partie s'inscrit principalement dans la théorie des processus et son apport concerne essentiellement les applications à des données économiques et financières.

Dans la première partie de cette introduction, nous illustrons par un exemple pratique le principe de la validation croisée appliquée à l'audit de prédicteurs. Nous introduisons ensuite le cadre théorique choisi dans le cadre de cette thèse pour l'étude de la validation croisée : il s'agit de celui du Statistical Learning Theory (SLT), introduit par les travaux fondateurs de [VC71] et [VA71] dans les années 70. L'intérêt de ce cadre est celui d'une approche qui modélise peu les données (cadre non paramétrique) et qui ne suppose pas un grand nombre de données (cadre non asymptotique). Parallèlement, suivant les travaux de [DUD03], nous introduisons les notations nécessaires pour couvrir de manière générale les différentes méthodes de validation croisée notamment : leave-one-out cross-validation, k-fold cross-validation, hold-v-out cross-validation, leave-v-out cross-validation. Dans ce cadre, nous définissons le problème posé et le type de résultats souhaités. Nous rappelons ensuite l'état de l'art sur ce problème particulier : l'estimation de l'erreur de généralisation d'une méthode prédictive par validation croisée. Nous introduisons dans la première partie les principaux résultats de cette thèse concernant la validation croisée (Chapitres 1 à 3).

Le chapitre 1 s'intéresse au cas classique de prédicteurs de Vapnik-Chernovenkis dimension (VC-dimension dans la suite) finie obtenus par minimisation du risque empirique. Notons que ces hypothèses couvrent notamment la classification réalisée sur une famille d'intervalles (quand la variable explicative est dans $\mathbb{R}$), d'hyperplans (quand la variable explicative est dans $\mathbb{R}^d$), et d'ellipsoides. Dans ce cadre, nous montrons que les méthodes de validation croisée sont consistantes : la largeur de l'intervalle de confiance obtenue par validation croisée pour l'erreur de généralisation converge vers 0 quand la taille de l'échantillon tend vers l'infini. Les résultats obtenus soulignent ausi :

1. qu'il n'est pas nécessaire que le nombre d'éléments dans l'échantillon test croisse vers l'infini pour que la validation croisée soit consistante.

2. l'intérêt de la k-fold cross-validation en terme de vitesse de convergence.

Dans la pratique, des familles populaires de prédicteurs telles les k-plus proches voisins, ou les techniques de boosting sont de VC-dimension infinie. Le chapitre 2 s'intéresse ainsi à une autre classe de prédicteurs plus large que celle du chapitre 1 : les estimateurs stables. Cela couvre en particulier les algorithmes sus-cités. Dans ce cadre, nous montrons que les méthodes de validation croisée sont encore consistantes.

Dans le chapitre 3, nous exhibons un cas particulier important le *subagging* où la méthode de validation croisée permet de construire des intervalles de confiance plus étroits que la méthodologie traditionnelle issue de la minimisation du risque empirique sous l'hypothèse de VC-dimension finie. En particulier, elle permet la réalisation d'intervalles de confiances non triviaux pour de petits échantillons.

Dans la seconde partie (Ch4-Ch6), nous décrivons les modélisations proposées pour apporter des solutions pratiques aux problèmes rencontrés dans le monde professionnel.

Les chapitres 4 et 5 s'inscrivent dans la technique du filtrage de Kalman.

Le chapitre 4, en suivant les travaux de [MM03], propose un proxy mensuel du taux de croissance du P.I.B. (Produit Intérieur Brut) français qui est disponible officiellement uniquement à fréquence trimestrielle. Cela permet d'établir une grille de lecture fine de la conjoncture francaise passée de 1992 à 2004. Ce chapitre a été publié dans [COD06].

Le chapitre 5 est l'objet du travail commun réalisé avec T.Deperraz. Il décrit la méthodologie pour construire un indicateur synthétique mensuel dans les enquêtes de conjoncture dans le secteur des services en France. Ce travail a été publié dans [COR06] et l'indicateur synthétique est publié mensuellement par l'Insee dans les Informations Rapides.

Le chapitre 6 a été réalisé conjointement avec Hugo Harari Kermadec et publié dans [CHK08]. Il s'agit d'un modèle semi-paramétrique de prix spot d'électricité sur les marchés de gros ayant des applications dans la gestion du risque de la production d'électricité.

Validation croisée : un exemple pratique introductif

Supposons que nous souhaitions construire un algorithme permettant de trouver la nature -i.e. est-ce le chiffre 1 ou le chiffre 2- de l'image numérique d'un chiffre (image d'un "1",d'un "2",...). En d'autres termes, nous avons l'image (un carré de 1064 pixels sur 1064 pixels) et nous souhaitons prédire s'il s'agit de l'image du chiffre 1 ou du chiffre 2 etc...Cette question se pose par exemple à la Poste pour déchiffrer automatiquement les codes postaux écrits manuellement sur les enveloppes.

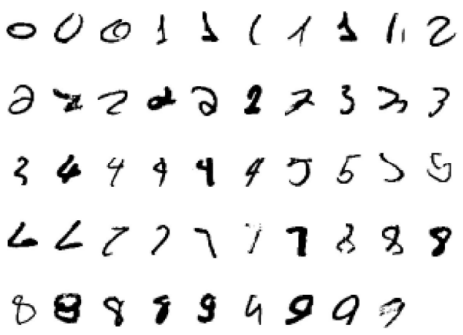


FIG. 1 – Exemples d'images numérisées de chiffres écrits à la main.

Pour ce faire, une stratégie consiste dans un premier temps à constituer une base d'images labélisées (ensemble d'apprentissage) : c'est-à-dire un nombre suffisant d'images dont on connaît la nature (i.e. le chiffre qu'elle représente, cf. graphique 1). Nous pouvons ensuite utiliser des algorithmes d'apprentissage (*Machine Learning* en anglais) qui vont "apprendre" à l'aide de cette base d'exemples. C'est-à-dire qu'ils vont construire à l'aide de ces données un prédicteur. Ce dernier sera capable de prédire la nature d'une nouvelle image. Parmi ces méthodes nous pouvons citer de manière non exhaustive : les support vector machines, les réseaux de neurones, les méthodes à noyaux, les arbres de régression.

Le problème qui se pose alors est celui de l'audit ou de la validation du prédicteur : s'agit-il d'un bon prédicteur ou non ? Mais qu'est-ce donc qu'un bon prédicteur ? Par convention, nous dirons que le prédicteur commet une erreur sur une image donnée s'il se trompe sur la nature de cette image (par exemple en classant l'image d'un 3 comme étant l'image du chiffre 1). Cette erreur nous coûte 1. S'il a bien prévu, l'erreur vaut 0. Nous dirons qu'un prédicteur est bon si son taux d'erreur moyen sur de nouvelles instances est faible. Comment se faire dès lors une idée *a priori* sur ce taux d'erreur ?

Une première possibilité est d'utiliser un grand ensemble de test constitué d'images n'ayant pas servi à l'apprentissage du prédicteur. Cependant, cela soulève un double problème : d'une part, il se peut que la constitution d'un tel ensemble soit dans la pratique très onéreuse, c'est le cas notamment des micropuces d'ADN. Ensuite, la question de la répartition des données entre ensemble d'apprentissage et ensemble de test n'est pas claire : plus l'ensemble d'apprentissage est important, plus grand est l'espoir que la méthode ait mieux appris mais plus petit sera l'ensemble de test et donc moins précise sera l'évaluation de la méthode. A l'inverse, plus l'ensemble de test est important, meilleure la validation est mais plus petit sera l'ensemble d'apprentissage et donc potentiellement moins performant sera le prédicteur.

Une seconde possibilité consiste à utiliser l'ensemble d'apprentissage comme ensemble de test : il s'agit de l'erreur de resubstitution. Cependant, le taux d'erreur moyen ainsi obtenu aura de fortes chances d'être optimiste car nous nous servons des mêmes données pour à la fois construire la méthode et pour tester la méthode. L'exemple des 1-plus proche voisin illustre même que l'erreur de resubstitution est nulle quelle que soit l'erreur de la méthode sur de futures instances.

Une troisième possibilité très populaire est la validation croisée. L'idée est justement de se prémunir contre un test trop optimiste en séparant les données d'apprentissage des données test. Prenons un cas particulier de ces procédures pour illustrer notre propos : la 2-fold cross-validation. L'échantillon est divisé en deux parties égales et disjointes. La première partie des données est utilisée pour construire notre prédicteur, la deuxième partie pour tester le prédicteur construit à l'aide de la première moitié ; nous obtenons ainsi un premier taux d'erreur. Nous répétons la procédure en intervertissant le rôle des parties : la seconde partie des données est utilisée pour construire notre prédicteur, la première partie pour tester le prédicteur construit à l'aide de la seconde moitié ; nous obtenons alors un second taux d'erreur. Le taux d'erreur est alors estimé par la moyenne des deux taux erreurs moyens ainsi construits.

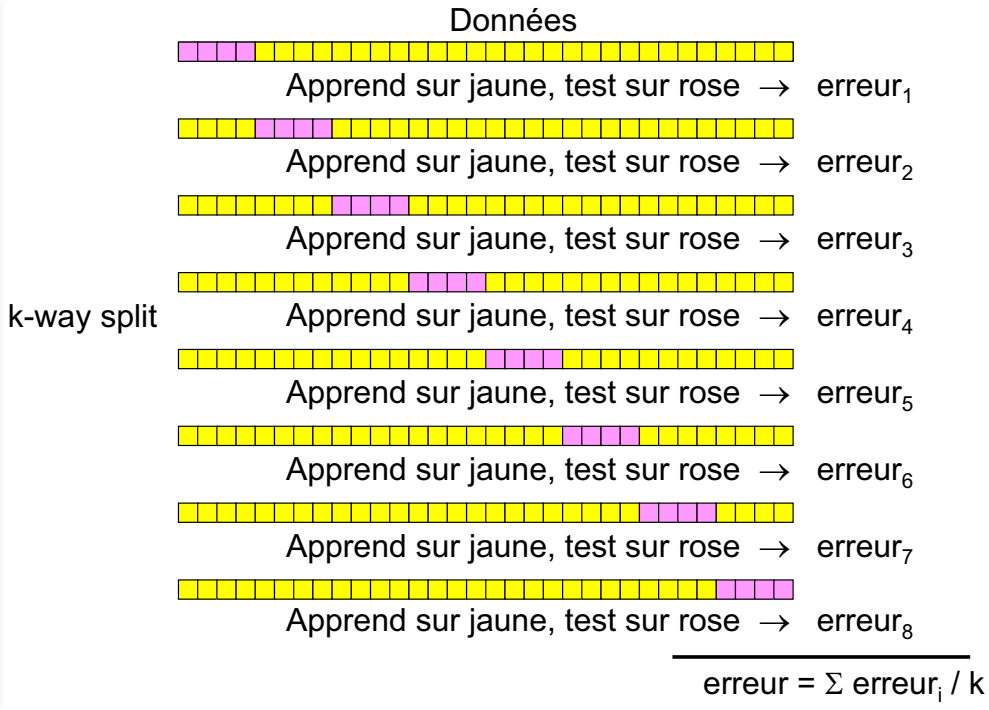


FIG. 2 – la k-fold cross-validation

Au lieu de diviser l'échantillon initial en deux parties, nous pouvons le diviser en k, il s'agit alors de k-fold cross-validation (cf. graphique 2). Ce dernier peut aussi être divisé en n parties s'il y a n données : on parle de leave-one-out cross-validation. Autre méthode plus intensive en calcul, nous répétons la procédure précédente en tirant avec remise ν données parmi n, les ν observations serviront dans l'ensemble de test tandis que les $n-\nu$ observations restantes constitueront l'ensemble d'apprentissage. Ce découpage est répété C_n^ν. On parle de leave-ν-out cross-validation.

L'intuition sur le choix du découpage peut être exprimée ainsi : si l'échantillon de test est réduit à un seul élément, nous avons bon espoir que la méthode construite avec un ensemble d'apprentissage de $n-1$ observations se comporte peu ou prou comme un prédicteur construit avec n données. En revanche, la moyenne de ces n erreurs peuvent être très dépendantes (car les prédicteurs sont contruits sur quasiment le même jeu de données) et donc fournir un très mauvais proxy de la vraie performance du prédicteur. En revanche, la leave-ν-out cross-validation avec ν grand, garantit une certaine stabilité (faible variance) car l'on moyenne sur ν observations indépendantes mais rien ne garantit a priori que l'erreur d'une méthode construite à l'aide de $n-\nu$ soit proche de celle d'une méthode construite avec n données (biais important).

Au delà de l'heuristique, se posent alors les questions suivantes auxquelles nous tenterons

d'apporter un éclairage :
-la validation croisée "marche"-t-elle ? (Ch1-2)
-faut-il diviser l'échantillon en 2 ou en 10 ? (Ch1-2)
-la validation croisée est-elle plus efficace que l'erreur de resubstitution ? (Ch 3)

La démarche va donc consister dans un cadre d'hypothèses classiques (précisément le SLT) à apporter une réponse à ces trois questions. Avant de définir ce cadre, rappelons tout d'abord que la validation-croisée s'inscrit dans les méthodes de rééchantillonnage et peut s'appliquer à la résolution d'autres problèmes que l'estimateur de l'erreur d'un predicteur. Pour une introduction détaillée, nous renvoyons à [ARL08].

Introduction au Statistical Learning Theory

L'objectif de la reconnaissance forme est de prédire le genre inconnu d'une observation. Une observation est un ensemble de mesures numériques, pouvant être représenté par un vecteur x appartenant à un espace mesurable $\mathcal{X}$ (dans notre exemple précédent, il s'agit du vecteur de pixels de l'image). La nature d'une observation est notée y et appartient à un espace mesurable $\mathcal{Y}$ (dans notre exemple, il s'agit de l'ensemble des chiffres : $0, 1, ..., 9$). Dans la pratique, nous distinguerons deux cas importants : quand $\mathcal{Y}$ est discret , il s'agit de la classification (respectivement quand $Y = \mathbb{R}^d$, il s'agit de la régression). Le but est donc de construire une fonction mesurable φ de $\mathcal{X}$ dans $\mathcal{Y}$. $\varphi(x)$ représente la prédiction de y étant donné x. Quand la vraie valeur est y, l'erreur de prédiction est mesurée par $L(y, \varphi(x))$ avec la fonction de perte $L : \mathcal{Y}^2 \to \mathbb{R}_+$. Dans le cas de la classification, la fonction de perte est typiquement $L(y, \varphi(x)) = 1_{\{y \neq \varphi(x)\}}$. Donnons maintenant deux hypothèses qui seront communes aux chapitres 1,2 et 3.

$H1$: A une même réalisation x pouvant correspondre différentes natures y, nous supposons que (x, y) est la réalisation d'une variable aléatoire (X, Y) de loi inconnue $\mathbb{P}$. La performance de φ est mesurée par le risque $R(\varphi) := \mathbb{E}_{(X,Y)} L(Y, \varphi(X))$. L'interprétation de cette grandeur probabiliste est la limite sur un échantillon test indépendant de longueur infinie du taux d'erreur moyen de φ.

$H2$: Nous supposons avoir accès à la réalisation de n variables aléatoires $\mathcal{D}_n := (X_i, Y_i)_{1 \leq i \leq n}$ indépendantes, identiquement distribuées (i.i.d.) de loi $\mathbb{P}$.

Un algorithme d'apprentissage Φ est entraîné à partir de $\mathcal{D}_n$. Donc, Φ est une application mesurable de $\mathcal{X} \times \cup_n (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Y}$. Y est prédit par $\Phi(X, \mathcal{D}_n)$. La performance de $\Phi(., \mathcal{D}_n)$ est mesurée par le risque conditionnel : $\tilde{R}_n := \mathbb{E}_{(X,Y)}[L(Y, \Phi(X, \mathcal{D}_n))|\mathcal{D}_n]$ avec $(X, Y) \sim \mathbb{P}$ indépendante de $\mathcal{D}_n$. L'interprétation de cette quantité est la limite sur un échantillon test indépendant de $\mathcal{D}_n$ de longueur infinie du taux d'erreur moyen de $\Phi(., \mathcal{D}_n)$. Remarquons qu'il s'agit d'une mesure différente du risque inconditionnel $R_n := \mathbb{E}_{\mathcal{D}_n} \tilde{R}_n$.

Nous cherchons à estimer l'erreur de généralisation $\tilde{R}_n$ sachant que $\mathbb{P}$ est inconnue. Une première approche consiste à construire le prédicteur sur l'ensemble des données et de réutiliser ces mêmes données pour tester la méthode. Il s'agit de l'erreur de resubstitution notée

$$\hat{R}_n := \frac{1}{n}\sum_{i=1}^{n} L(Y_i, \varphi(X_i, \mathcal{D}_n)).$$

qui relève exactement de la méthode du plug-in. C'est le plug-in de $\mathbb{E}_{\mathbb{P}}(L(Y, \phi(X, \mathbb{P})))$ qui s'interprète donc aussi comme une moyenne boostrap $\mathbb{E}_{\mathbb{P}_n}(L(Y_i^\star, \phi(X_i^\star, \mathbb{P}_n)))$.
Une autre méthode très populaire parmi les practiciens est la validation croisée ([HTF01]). Nous noterons par $\hat{R}_{CV}$ l'estimateur de validation croisée. Nous détaillons à présent les notations utilisées pour définir les différentes procédures de validation croisée. Pour ce faire, nous nous inspirons fortement des notations introduites dans [DUD03].

Tout d'abord, il convient de définir la séparation de l'échantillon initial entre ensemble d'apprentissage et ensemble de test. Pour ce faire, nous définissons un vecteur binaire : $V_n := (V_{n,i})_{1\leq i\leq n}$ avec V_n un vecteur de taille n tel que $V_{n,i} \in \{0,1\}$ et $\sum_{i=1}^n V_{n,i} \neq 0$. Ce dernier permet de définir un sous échantillon de $\mathcal{D}_n$ noté $\mathcal{D}_{V_n} := \{(X_i, Y_i) \in \mathcal{D}_n | V_{n,i} = 1, 1 \leq i \leq n\}$.

Ensuite, nous construisons un prédicteur à l'aide d'un sous-échantillon que nous notons $\varphi_{V_n}(.) := \Phi(., \mathcal{D}_{V_n}$.

Enfin, l'erreur empirique pondérée de φ sur le sous-échantillon D_{V_n} notée $\hat{R}_{V_n}$ est définie par :

$$\hat{R}_{V_n}(\varphi) := \frac{1}{\sum_{i=1}^n V_{n,i}} \sum_{i=1}^{n} V_{n,i} L(Y_i, \varphi(X_i)).$$

L'échantillon initial $\mathcal{D}_n$ est divisé en deux échantillons disjoints : un ensemble d'apprentissage contenant $n(1-p_n)$ observations (où p_n est le pourcentage d'observations dans l'ensemble de test tel que np_n soit entier) et un ensemble de test contenant np_n observations. Afin de définir l'ensemble d'apprentissage, nous introduisons V_n^{tr} un vecteur binaire aléatoire de taille n indépendant de $\mathcal{D}_n$. Ce dernier est appelé vecteur d'apprentissage. Nous définissons le vecteur de test noté V_n^{ts} par $1_n - V_n^{tr}$ qui représente l'échantillon de test avec $1_n := (1, ..., 1) \in \mathbb{R}^n$.

Nous pouvons maintenant définir de manière générale l'estimateur de validation croisée par :

Définition 1. *l'estimateur de validation croisée de φ_n noté $\hat{R}_{CV}$ est défini par l'espérance conditionnelle de $\hat{R}_{V_n^{ts}}(\varphi_{V_n^{tr}})$ sachant $\mathcal{D}_n$:*

$$\hat{R}_{CV} := \mathbb{E}_{V_n^{tr}} \hat{R}_{V_n^{ts}}(\varphi_{V_n^{tr}})$$

Prenons quelques exemples pour illustrer notre propos. Pour retrouver une procédure de validation, il suffit de définir la loi de V_n^{tr}. Soit k un entier naturel tel que n/k soit un entier. La k-fold cross-validation divise l'échantillon initial en k sous-échantillons. Un prédicteur est construit à l'aide de $k-1$ sous-échantillons puis tester sur l'échantillon restant. Cette procédure est répétée pour chaque sous-échantillon. Les erreurs calculées sont moyennées pour donner l'estimation k-fold.

Exemple 2 (k-fold cross-validation)**.**

$$\Pr(V_n^{tr} = (\underbrace{0,\dots,0}_{n/k \text{ observations}}, \underbrace{1,\dots,1}_{n(1-1/k) \text{ observations}})) = \frac{1}{k},$$

$$\Pr(V_n^{tr} = (\underbrace{1,\dots,1}_{n/k \text{ observations}}, \underbrace{0,\dots,0}_{n/k \text{ observations}}, \underbrace{1,\dots,1}_{n(1-2/k) \text{ observations}})) = \frac{1}{k},$$

$$\dots$$

$$\Pr(V_n^{tr} = (\underbrace{1,\dots,1}_{n(1-1/k) \text{ observations}}, \underbrace{0,\dots,0}_{n/k \text{ observations}})) = \frac{1}{k}.$$

Un autre exemple populaire est la Leave-υ-out cross-validation. Cela consiste à séparer l'échantillon initial en un échantillon d'apprentissage de taille $n - \nu$ et un échantillon test de taille ν. Cela est répété pour l'ensemble des sous-échantillons test possibles de taille ν. L'estimation de validation croisée est obtenue en moyennant l'ensemble des erreurs ainsi calculées. Nous notons par $(\xi_{n,i}^{\upsilon})_{1 \leq i \leq \binom{n}{\upsilon}}$ la famille de vecteurs binaires de taille n telle que $\sum_{i=1}^{n} \xi_{n,i}^{\upsilon} = n - \upsilon$.

Exemple 3 (Leave-υ-out cross-validation)**.**

$$\Pr(V_n^{tr} = \xi_{n,1}^{\upsilon}) = \frac{1}{\binom{n}{\upsilon}}$$

$$\Pr(V_n^{tr} = \xi_{n,2}^{\upsilon}) = \frac{1}{\binom{n}{\upsilon}}$$

$$\dots$$

$$\Pr(V_n^{tr} = \xi_{n,\binom{n}{\upsilon}}^{k}) = \frac{1}{\binom{n}{\upsilon}}.$$

Le problème est donc le suivant : l'estimateur de la validation croisée $\hat{R}_{CV}$ est-il un bon estimateur de l'erreur de généralisation ? Afin de répondre à cette question, nous pouvons imaginer plusieurs critères de proximité :

- la distance L_p entre $\hat{R}_{CV}$ et $\tilde{R}_n$: $(\mathbb{E}|\hat{R}_{CV} - \tilde{R}_n|^p)^{1/p}$.
- la distance en probabilité $\Pr(|\hat{R}_{CV} - \tilde{R}_n| \geq \varepsilon)$ avec $\varepsilon > 0$.
- une limite $\lim_{n\infty} \Pr(|\hat{R}_{CV} - \tilde{R}_n| \geq \varepsilon)$ ou des bornes $\alpha_n \leq \Pr(|\hat{R}_{CV} - \tilde{R}_n| \geq \varepsilon) \leq \beta_n$

Nous nous intéresserons principalement à la majoration non asymptotique de $\Pr(|\hat{R}_{CV} - \tilde{R}_n| \geq \varepsilon)$ (Ch1-2). Remarquons que cette dernière approche permet de répondre aussi à la question asymptotique ainsi qu'à la question en norme L_p.

Rappelons deux résultats importants sur la probabilité de déviation entre une moyenne empirique et son espérance ([HOEF63]) et sur la probabilité de déviation uniforme ([VC71]).

Théorème 4 ([HOEF63]). *Soient $X_{1,...,}X_n$ des variables indépendantes dans $[a_i, b_i]$. Alors pour tout $\varepsilon > 0$,*

$$\mathbb{P}(\sum X_i - \mathbb{E}(\sum X_i) \geq n\varepsilon) \leq e^{-\frac{2\varepsilon^2}{\sum_i (b_i - a_i)^2}} \tag{1}$$

Théorème 5 ([VC71]). *Soit $\mathcal{C}$ une classe de prédicteur de VC-dimension VC finie et L une fonction de perte bornée par 1. Alors pour tout $\varepsilon > 0$,*

$$\mathbb{P}(\sup_{\phi \in \mathcal{C}} (\widehat{R}_n(\phi) - L(\phi)) \geq \varepsilon) \leq c(n, V_{\mathcal{C}}) e^{-\frac{\varepsilon^2}{2\sigma(n)^2}} \tag{2}$$

avec $c(n, V_{\mathcal{C}}) \leq 2(2n+1)^{V_{\mathcal{C}}}$ et si $n \geq V_{\mathcal{C}}$, $c(n, V_{\mathcal{C}}) \leq 2(\frac{2ne}{V_{\mathcal{C}}})^{V_{\mathcal{C}}}$ et $\sigma(n)^2 = \frac{4}{n}$

Nous exhibons des bornes de la forme $\Pr(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon) \leq B(n, p_n, \varepsilon) + V(n, p_n, \varepsilon)$ avec $\varepsilon > 0$. Le terme $B(n, p_n, \varepsilon)$ est une borne à la "Vapnik-Chernovenkis" (cf. équation 2) alors que le terme $V(n, p_n, \varepsilon)$ est une borne à la "Hoeffding" (cf. équation 1) contrôlée par la taille de l'échantillon de test np_n.
Cette borne peut être interprétée comme une réponse quantitative à la question du compromis entre le biais et la variance. Quand le pourcentage d'observations dans l'échantillon de test p_n augmente, le terme $V(n, p_n, \varepsilon)$ diminue mais le terme $B(n, p_n, \varepsilon)$ augmente. Remarquons que cette borne est pire qu'une borne du type Vapnik-Chernovenkis et donc peut être appelée "sanity-check bound" dans l'esprit de [KEA95] . Même si ces bornes sont valables quel que soit le type de validation croisée, leur pertinence dépend grandement des valeurs de p_n le pourcentage d'éléments dans l'échantillon test ; c'est pourquoi nous classons ces bornes d'après les valeurs de p_n. Nous montrons dans ces deux premiers chapitres des résultats de la forme :

$$\Pr(|\hat{R}_{CV} - \tilde{R}_n| \geq \varepsilon) \leq B(n, p_n, \varepsilon) + V(n, p_n, \varepsilon)$$

avec $B(n, p_n, \varepsilon)$ et $V(n, p_n, \varepsilon)$ qui tendent vers 0 quand n tend vers l'infini quelque soit p_n. De plus, si p_n est fixe (c'est le cas dans la k-fold cross-validation où $p_n = 1/k$) $B(n, p_n, \varepsilon)$ et $V(n, p_n, \varepsilon)$ tendent vers 0 exponentiellement vite.

Ce résultat est rassurant. Cependant, dans le cas particulier de la minimisation du risque empirique sous des hypothèses classiques de VC-dimension finie, nous savons que $\Pr(|\hat{R}_{CV} - \tilde{R}_n| \geq \varepsilon) \leq f(n, \varepsilon)$ avec $f(n, \varepsilon) \leq V(n, p_n, \varepsilon)$. Ainsi, le résultat précédent ne nous dit pas si la validation croisée apporte un intérêt supplémentaire à la simple méthode d'erreur de resubstitution. Ces bornes seront appelées *"sanity check bounds"* dans l'esprit de [KEA95].

Ainsi, une autre question (cf. problème) d'intérêt (cf. Pest la comparaison de la validation croisée avec l'erreur de resubstitution. Il s'agit donc de comparer $\Pr(|\hat{R}_{CV} - \tilde{R}_n| \geq \varepsilon)$ et $\Pr(|\tilde{R}_n - \hat{R}_n| \geq \varepsilon)$. C'est précisément l'objet du chapitre 3. Nous prouvons un résultat plus fort dans le cas d'une méthode populaire : le subagging, méthode pour laquelle nous adaptons légèrement la validation croisée. Dans ce cas, nous obtenons alors des résultats de la forme :

$$\Pr(|\hat{R}_{CV} - \tilde{R}_n| \geq \varepsilon) \leq \min(B(n, p_n, \varepsilon), V(n, p_n, \varepsilon)) < 1$$

En particulier, pour les petits échantillons, nous avons $\min(B(n, p_n, \varepsilon), V(n, p_n, \varepsilon)) << f(n, \varepsilon)$ ce qui souligne l'intéret particulier de la validation croisée. Pour de grands échantillons, on montre que $argmin_{p_n}[\min(B(n, p_n, \varepsilon), V(n, p_n, \varepsilon))] \sim_{n\infty} f(n, \varepsilon)$.

Nous détaillons maintenant les résultats obtenus dans les différents chapitres de cette thèse.

Ch1 : Résultats sur les minimisateurs du risque empirique de dimension Vapnik-Chernovenkis finie

Hypothèses Nous notons R_{opt} l'erreur de généralisation minimale obtenue au sein d'une famille de prédicteurs $\mathcal{C}$, $R_{opt} = \inf_{\phi \in \mathcal{C}} R(\phi)$. Dans la suite, nous supposons que ϕ_n est un minimisateur du risque empirique sur la classe $\mathcal{C}$. Par simplicité, nous supposons que la borne inférieure est atteinte i.e. $\phi_n = \arg\min_{\phi \in \mathcal{C}} \widehat{R}_n(\phi)$. Remarquons que R_{opt} est un paramètre de la distribution inconnue $\mathbb{P}_{(X,Y)}$ alors que $\widetilde{R}_n$ est une variable aléatoire.
Par la suite, nous supposerons que l'échantillon d'apprentissage et l'échantillon de test sont disjoints et que le nombre d'observations dans l'échantillon d'apprentissage et dans l'échantillon de test sont respectivement $n(1 - p_n)$ et np_n. De plus, nous supposons aussi que ϕ_n est un minimisateur du risque empirique sur une classe de prédicteurs de VC-dimension $V_{\mathcal{C}}$ finie et que L est une fonction de perte majorée par 1. Nous supposons enfin que les prédicteurs sont symétriques par rapport à l'échantillon d'apprentissage, i.e. que le prédicteur ne dépend pas de l'ordre des observations dans $\mathcal{D}_n$. Finalement, la validation croisée est supposée symetrique i.e. $\Pr(V_{n,i}^{tr} = 1)$ ne dépend pas de i, ce qui exclut la hold-out cross-validation. **Nous notons ces hypothèses par $\mathcal{H}$.**

Enfin, rappelons les définitions de :

Définition 6 (Shatter coefficients)**.** *Soit $\mathcal{A}$ une famille de parties mesurables. Pour $(z_{1,\ldots,}z_n) \in \{\mathbb{R}^d\}^n$, soit $N_{\mathcal{A}}(z_{1,\ldots,}z_n)$ le nombre de parties différentes dans :*

$$\{\{z_1, \ldots, z_n\} \cap A; A \in \mathcal{A}\}.$$

Le n-shatter coefficient de $\mathcal{A}$ est

$$\mathcal{S}(\mathcal{A}, n) = \max_{(z_{1,\ldots,}z_n) \in \{\mathbb{R}^d\}^n} N_{\mathcal{A}}(z_{1,\ldots,}z_n).$$

Entre d'autres termes, le shatter coefficient est le nombre maximal de sous parties différentes de n points qui peuvent être choisies par la famille de parties $\mathcal{A}$.

Définition 7 (VC dimension)**.** *Soit $\mathcal{A}$ une famille de parties telle que $\mathcal{A} \geq 2$. Le plus grand entier $k \geq 1$ telle que $\mathcal{S}(\mathcal{A},k) = 2^k$ est notée $V_{\mathcal{C}}$, et est appelée la dimension Vapnik-Chernovenkis (ou VC-dimension) de la famille $\mathcal{A}$. Si $\mathcal{S}(\mathcal{A},n) = 2^n$ pour tout n, alors par définition $V_{\mathcal{C}} = \infty$.*

Une famille de prédicteurs $\mathcal{C}$ est dite de VC-dimension $V_{\mathcal{C}}$ finie si la dimension de la famille de parties $\{A_{\varphi,t} : \varphi \in \mathcal{C}, t \in [0,1]\}$ est égale à $V_{\mathcal{C}}$, où $A_{\varphi,t} = \{(x,y)/L(y,\varphi(x)) > t\}$.

Etat de l'art A l'exception de [BUR89] , les recherches théoriques sur la validation croisée se sont tout d'abord concentrées sur les modèles linéaires ([Li87] ;[SHAO93] ;[ZHA93]). Les premiers résultats non asymptotiques sont dus à [DEWA79] et concernent les règles k-locales et les procédures de validation croisée leave-one-out et hold-out. Plus récemment, [HOL96, HOL96bis] ont obtenu des résultats non asymptotiques pour la hold-out, $k-$fold et leave-one-out validation croisée pour des prédicteurs de VC-dimension finie dans le cas réalisable (l'erreur de généralisation est nulle). Cependant les bornes pour la $k-$fold validation croisée sont k fois plus lâches que celles pour la hold-out validation croisée. [BKL99] ont souligné quand la $k-$fold validation croisée peut obtenir de meilleurs résultats que la hold-out validation croisée dans le cas particulier du k-fold predicteur. [KR99] ont étendu ces résultats pour les prédicteurs stables dans le cas de la leave-one-out. [KEA95] ont aussi obtenu des résultats pour la hold-out validation croisée pour des minimisateurs du risque empirique sur une classe de VC-dimension finie mais sans l'hypothèse réalisable. Toutefois, ces bornes peuvent être appelées "sanity check bounds" (d'après [KR99]) ; c'est-à-dire qu'elles ne sont pas meilleures que les bornes classiques de Vapnik-Chernovenkis. [DUD04BIS] ont obtenu des résultats non asymptotiques pour la distance entre l'estimateur de validation croisée et un certain benchmark et ont prouvé des resultats asymptotiques sur la relation entre l'estimateur de validation croisée et l'erreur de généralisation. A notre connaissance, il n'y a pas de bornes pour des procédures de validation croisée intensive telle la leave-ν-out validation croisée. Cela peut être dû à l'absence d'indépendence entre les termes croisés de l'estimateur de validation croisée d'après [KMNR95] .

Résultats de la thèse Le premier résultat est consacré à la validation croisée quand l'échantillon de test est grand, dans le sens où les bornes sont d'autant meilleures si np_n est grand. Remarquons au passage que ce résultat exclut la hold-out validation croisée car cette dernière ne fait pas un usage symétrique des données.
Nous obtenons l'inégalité de concentration suivante pour la déviation absolue $|\widehat{R}_{CV} - \widetilde{R}_n|$,

Corollaire 8 (Déviation absolue pour les grands échantillons)**.** *Supposons $\mathcal{H}$. Alors, nous avons, pour tout $\varepsilon > 0$,*

$$\Pr(|\widetilde{R}_n - \widehat{R}_{CV}| \geq \varepsilon) \leq B(n, p_n, \varepsilon) + V(n, p_n, \varepsilon),$$

avec

- $B(n, p_n, \varepsilon) = 5(2n(1-p_n)+1)^{\frac{4V_{\mathcal{C}}}{1-p_n}} \exp(-\frac{n\varepsilon^2}{25})$, *une borne de type Vapnik-Chernovenkis.*
- $V(n, p_n, \varepsilon) = \exp(-\frac{2np_n\varepsilon^2}{25})$, *une borne de type Hoeffding.*

A l'aide de la précédente inégalité, nous pouvons majorer l'espérance de $|\widetilde{R}_n - \widehat{R}_{CV}|$:

Corollaire 9 (Erreur L_1 pour les grands échantillons test)**.** *Supposons $\mathcal{H}$. Alors, nous avons,*

$$\mathbb{E}|\widehat{R}_{CV} - \widetilde{R}_n| \leq 10\sqrt{\frac{V(\ln(2n(1-p_n)+1)+4)}{n(1-p_n)}} + 5\sqrt{\frac{2}{np_n}}.$$

Le précédent résultat n'est pas pertinent pour les petits échantillons test (typiquement la leave-one-out validation croisée où $np_n = 1$) puisque le terme de variance $V(n, p_n, \varepsilon)$ ne tend pas vers 0 (dans la leave-one-out validation croisée, $V(n, p_n, \varepsilon) = \exp(-2\varepsilon^2/25)$). Cependant, sous $\mathcal{H}$, la validation croisée avec de petits échantillons test demeure consistante, comme le montre le corollaire suivant :

Corollaire 10 (Déviation absolue pour les petits échantillons)**.** *Supposons $\mathcal{H}$. Alors, nous avons pour tout $\varepsilon > 0$,*

$$\Pr(|\widetilde{R}_n - \widehat{R}_{CV}| \geq \varepsilon) \leq B(n, p_n, \varepsilon) + V(n, p_n, \varepsilon),$$

- $B(n, p_n, \varepsilon) = 5(2n(1-p_n)+1)^{\frac{4V_{\mathcal{C}}}{1-p_n}} \exp(-\frac{n\varepsilon^2}{64})$
- $V(n, p_n, \varepsilon) = \frac{16}{\varepsilon}\sqrt{\frac{V_{\mathcal{C}}(\ln(2n(1-p_n)+1)+4)}{n(1-p_n)}}$.

Finalement, nous obtenons :

Corollaire 11 (Erreur L_1 pour de petits échantillons)**.** *Supposons $\mathcal{H}$. Alors, nous avons :*

$$\mathbb{E}|\widehat{R}_{CV} - \widetilde{R}_n| \leq 16\sqrt{\frac{V_{\mathcal{C}}\ln(2n(1-p_n)+1)+4}{n(1-p_n)}}\left(\ln\left(\sqrt{\frac{n(1-p_n)}{V_{\mathcal{C}}(\ln(2n(1-p_n)+1)+4)}}\right)+2\right).$$

Pour de petits échantillons test, nous obtenons aussi la consistance mais les bornes en vitesse de convergence sont moins bonnes pour le terme V : typiquement $O_n\left(\frac{1}{\varepsilon}\sqrt{\frac{\ln(n(1-p_n))}{n(1-p_n)}}\right)$ contre $O_n\left(\exp(-np_n\varepsilon^2)/8\right)$.

A ce point, nous pouvons nous interroger sur l'intérêt de moyenner encore sur les différents dossiers de la k -fold validation croisée, ce qui est couteux en temps de calcul. En fait, la borne sur le terme (V) peut être améliorée en moyennant sur les k erreurs d'apprentissage. Cette étape souligne l'intérêt de la k-fold validation croisée par rapport à une validation croisée plus simple telle la hold-out.

Proposition 12 (k-fold)**.** *Supposons $\mathcal{H}$. Alors, pour la k-fold validation croisée, nous avons pour tout $\varepsilon > 0$:*

$$\Pr(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon) \leq 2\exp(-\frac{2np_n\varepsilon^2}{25}) + 2^{\frac{1}{p_n}}\exp\left(-\frac{n\epsilon^2}{25*64(\sqrt{V_{\mathcal{C}}\ln(2(2np_n+1))}+2)}\right).$$

Ainsi, moyenner les erreurs observées pour construire l'estimateur k-fold améliore le terme $V_{\mathcal{C}}$ de

$$\min(2\exp(-\frac{32np_n\varepsilon^2}{49}), \frac{14}{\varepsilon}\sqrt{\frac{V_{\mathcal{C}}(\ln(2(1-p_n)+1)+4)}{n(1-p_n)}}).$$

à $2^{\frac{1}{p_n}} \exp\left(-\frac{n\epsilon^2}{64(\sqrt{V\ln(2(2np_n+1))}+2)}\right)$. Ce résultat est important puisqu'il montre que l'utilisation intensive des données peut être judicieuse pour améliorer les vitesses de convergence. Une autre conséquence intéressante de cette proposition est que pour une précision donnée ε, il n'est pas nécessaire que la taille de l'échantillon test tende vers l'infini pour avoir une convergence exponentielle de la validation croisée. Pour cela, il est suffisant que la taille de l'échantillon de test soit plus grand qu'un nombre fixe n_0.
A l'aide de ces bornes probabilistes, nous pouvons borner l'espérance de la différence entre l'erreur de généralisation et l'estimateur de validation croisée $\mathbb{E}_{\mathcal{D}_n}|\widehat{R}_{CV} - \widetilde{R}_n|$ est un $O_n(\sqrt{V_{\mathcal{C}}\ln(n(1-p_n))/n(1-p_n)} + \sqrt{1/np_n})$. Si nous prenons en compte $\mathbb{E}_{\mathcal{D}_n}|\widehat{R}_{CV} - \widetilde{R}_n|$, nous pouvons définir une règle de séparation : le pourcentage d'éléments p_n dans l'ensemble de test doit être proportionnel à $\frac{1}{1+V_{\mathcal{C}}^{1/3}}$, i.e. plus grande est la classe de prédicteurs, plus petit doit être l'ensemble de test dans la validation croisée.

Ch2 : Résultats sur les prédicteurs stables

Pour éviter d'avoir recours au cadre d'analyse de la VC-dimension, des concepts de stabilité ont été développés à la fin des années 90 [KEA95], [BE01], [BE02], [KUT02], et [KUNIY02]. L'objet de ce cadre d'analyse est l'algorithme d'apprentissage plutôt que la famille de prédicteurs. L'algorithme d'apprentissage est une application d'un échantillon dans un espace de prédicteurs. L'algorithme est stable en $\mathcal{D}_n$ si changer une observation dans $\mathcal{D}_n$ conduit à un petit changement dans le prédicteur ainsi contruit. L'intérêt d'une telle démarche est qu'il n'y a plus besoin d'avoir recours à la traditionnelle notion de VC-dimension et cela permet ainsi de s'intéresser à une classe plus large d'algorithmes que la minimisation empirique du risque. A titre d'exemple, cette approche a permis d'obtenir des bornes probabilistes sur l'erreur de généralisation d'algorithmes de régularisation comme le boosting ([KUNIY02]. Pour souligner l'intérêt d'une telle approche, citons une liste non exhaustive d'algorithmes satisfaisant des propriétés de stabilité : réseaux de régularisation, ERM, k-plus proches voisin, boosting.
Dans la suite, nous reprenons les notations pour la validation croisée introduite dans le chapitre 1(cf. [COR09A]). Nous considérons des notations légèrement plus compactes inspirées par la littérature sur les processus empiriques. Dans la suite, nous notons $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, et $(Z_i)_{1\leq i\leq n} := ((X_i, Y_i))_{1\leq i\leq n}$ l'échantillon initial. Pour une fonction de perte donnée L et une classe de prédicteurs $\mathcal{G}$, nous définissons une nouvelle classe $\mathcal{F}$ d'applications de $\mathcal{Z}$ dans $\mathbb{R}_+$ par $\mathcal{F} := \{\psi \in \mathbb{R}_+^{\mathcal{Z}} | \psi(Z) = L(Y, \phi(X)), \phi \in \mathcal{G}\}$. Pour un algorithme d'apprentissage Φ, nous avons la définition naturelle $\Psi(Z, \mathcal{D}_n) = L(Y, \Phi(X, \mathcal{D}_n))$. Avec ces notations, le risque conditionnel $\widetilde{R}_n$ est alors : $\widetilde{R}_n := \mathbb{E}_Z[\Psi(Z, \mathcal{D}_n) \mid \mathcal{D}_n]$ avec $Z \sim \mathbb{P}$ indépendant de $\mathcal{D}_n$. Dans la suite, nous permettrons aussi la notation $\psi(X, \mathcal{D}_n)$ au lieu de $\Psi(X, \mathcal{D}_n)$. Nous définissons comme aux chapitres précédents :

$$\psi_{V_n}(.) := \Psi(., \mathcal{D}_{V_n}).$$

Hypothèses Formellement, un algorithme d'apprentissage envoie un échantillon d'apprentissage pondéré dans un espace de prédicteurs. Ainsi, la stabilité peut être traduite comme une condition de Lipschitz avec grande probabilité.
A la suite de [KUNIY02], nous définissons une distance entre deux mesures empiriques pondérées. Une mesure empirique sur $\mathcal{Z}$ est définie par :

$$\mathbb{P}_{n,V_n} := \frac{1}{\sum_{i=1}^n V_{n,i}} \sum_{i=1}^n V_{n,i}\delta_{Z_i},$$

avec δ_{Z_i} la mesure de Dirac en $\{Z_i\}$.

La distance entre $\mathbb{P}_{n,V_n}$ et $\mathbb{P}_{n,U_n}$ deux mesures empiriques sur $\mathcal{Z}$ en lien avec les vecteurs binaires V_n et U_n. Nous ne supposons pas que leur support est égal. La distance entre elles est définie par la variation totale :

$$||\mathbb{P}_{n,U_n} - \mathbb{P}_{n,V_n}|| = \sup_{A\in\mathcal{P}(\mathcal{Z})} |(\mathbb{P}_{n,U_n} - \mathbb{P}_{n,V_n})(A)|.$$

Exemple 13. *Dans le cas de la leave-one-out (i.e. $\sum_{i=1}^n U_{n,i} = n-1$), nous avons*

$$||\mathbb{P}_{n,U_n} - \mathbb{P}_n|| = \frac{2}{n}.$$

Dans le cas de la leave-ν-out, nous avons

$$||\mathbb{P}_{n,U_n} - \mathbb{P}_n|| = \frac{2\nu}{n}.$$

Dans le cadre général, nous avons

$$||\mathbb{P}_{n,U_n} - \mathbb{P}_n|| = 2p_n$$

Enfin, nous avons besoin d'une distance d sur l'ensemble $\mathcal{F}$. Rappelons trois exemples importants. Soient $\psi_1, \psi_2 \in \mathcal{F}$. La distance uniforme est définie par : $d_\infty(\psi_1,\psi_2) = \sup_{Z\in\mathcal{Z}} |\psi_1(Z) - \psi_2(Z)|$, la distance L_1 par : $d_1(\psi_1,\psi_2) = \mathbb{P}|\psi_1 - \psi_2|$, la distance erreur par $d_e(\psi_1,\psi_2) = |\mathbb{P}(\psi_1 - \psi_2)|$. Il est important de remarquer que ce qui compte ici n'est pas directement une distance sur la classe originale des prédicteurs $\mathcal{G}$ mais une distance par rapport à la fonction de perte et la distribution $\mathbb{P}$. En particulier, pour la distance-L_1, nous ne nous soucions pas du comportement des prédicteurs φ_1 et φ_2 en dehors du support de $\mathbb{P}$. Enfin, remarquons que nous avons toujours $d_e \leq d_1 \leq d_\infty$.

Nous pouvons désormais définir les différentes notions de stabilité d'un algorithme d'apprentissage de manière notamment à couvrir les définitions de [KUNIY02]. Nous débutons avec la notion de stabilité faible. En substance, cela dit que pour n'importe quel vecteur d'apprentissage, la distance entre deux prédicteurs est contrôlée avec grande probabilité par la distance entre les vecteurs d'apprentissage.

Définition 14 (Stabilité faible)**.** *Soient $\lambda, (\delta_{n,p_n})_{n,p_n}$ des réels positifs. Un algorithme d'apprentissage Ψ est faiblement $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ stable si pour tout vecteur d'apprentissage U_n dont la somme des composantes vaut $n(1-p_n)$:*

$$\Pr(d(\psi_{U_n},\psi_n) \geq \lambda||\mathbb{P}_{n,U_n} - \mathbb{P}_n||^\alpha) \leq \delta_{n,p_n}.$$

Remarquons que dans la définition précédente Pr correspond à $\mathbb{P}^{\otimes n}$. Une notion plus forte est de considérer ψ_n construit à partir de $n-1$ observations tirées indépendamment de $\mathbb{P}$ et une additionnelle observation quelconque z. A titre de justification, remarquons que des algorithmes tels que la minimisation du risque empirique avec une VC-dimension finie ([KUNIY02]) satisfont cette propriété.

Définition 15 (Stabilité forte)**.** *Soit $z \in \mathcal{Z}$. Soit $\mathcal{D}_n = \mathcal{D}_{n-1} \cup \{z\}$ un échantillon initial. Soient $\lambda, (\delta_{n,p_n})_{n,p_n}$ des réels positifs. Un algorithme d'apprentissage Ψ est dit fortement $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ stable si pour tout vecteur d'apprentissage U_n dont la somme des composantes est égale à $n(1-p_n)$:*

$$\Pr(d(\psi_{U_n}, \psi_n) \geq \lambda ||\mathbb{P}_{n,U_n} - \mathbb{P}_n||^\alpha) \leq \delta_{n,p_n}.$$

Pour les algorithmes classiques, nous avons à l'esprit $\delta_{n,p_n} = O_n(p_n \exp(-n(1-p_n)))$.

Nous pouvons encore définir une définition plus forte de stabilité quand cette dernière est valable sur l'intégralité des vecteurs d'apprentissage de la validation croisée :

Définition 16 (Validation croisée et stabilité faible)**.** *Soit $\mathcal{D}_n = (Z_i)_{1 \leq i \leq n}$ un échantillon. Soit V_n^{tr} un vecteur d'apprentissage de distribution $\mathbb{Q}$. Soient $\lambda, (\delta_{n,p_n})_{n,p_n}$ des réels positifs. Un algorithme d'apprentissage Ψ est dit être faiblement $(\lambda, (\delta_{n,p_n})_{n,p_n}, d, \mathbb{Q})$ stable s'il est faiblement $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ stable et si :*

$$\Pr(\sup_{U_n \in support(\mathbb{Q})} \frac{d(\psi_{U_n}, \psi_n)}{||\mathbb{P}_{n,U_n} - \mathbb{P}_n||^\alpha} \geq \lambda) \leq \delta_{n,p_n}.$$

Comme précédemment, nous pouvons aussi définir la notion plus forte :

Définition 17 (Validation croisée et stabilité forte)**.** *Soit $z \in \mathcal{Z}$. Soit $\mathcal{D}_n = \mathcal{D}_{n-1} \cup \{z\}$ un échantillon. Soit V_n^{tr} un vecteur d'apprentissage de distribution $\mathbb{Q}$. Un algorithme d'apprentissage Ψ est dit être fortement $(\lambda, (\delta_{n,p_n})_{n,p_n}, d, \mathbb{Q})$ stable s'il est fortement $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ stable et si :*

$$\Pr(\sup_{U_n \in support(\mathbb{Q})} \frac{d(\psi_{U_n}, \psi_n)}{||\mathbb{P}_{n,U_n} - \mathbb{P}_n||^\alpha} \geq \lambda) \leq \delta_{n,p_n}.$$

Définition 18 (Stabilité sûre)**.** *Remarquons que si $\delta_{n,p_n} = 0$, les deux notions coincident et nous appelons alors la **stabilité sûre**.*

A titre de justification, nous citons une liste d'algorithmes d'apprentissage qui satisfont les notions de stabilité définies supra :

distance	d_∞	d_1	d_e
Faible			Lasso
Fort	Adaboost ([KUNIY02])	ERM ([KUNIY02]) k-nearest rule	Bayesian algorithm [KEA95]
Sure	Regularization networks ([BE01])		

Soit $\mathcal{D}_n$ un ensemble d'apprentissage de taille n. Soit $V_n^{tr} \sim \mathbb{Q}$ un vecteur d'apprentissage sur $\mathcal{D}_n$ tel que la validation croisée soit symétrique -i.e. $\Pr(V_{n,i}^{tr} = 1)$ est une constante indépendante de i –et le nombre d'observations dans l'échantillon test est égal à np_n. Soit d une distance parmi d_e, d_1, d_∞. Enfin, nous supposons que la fonction de perte L est majorée par 1. Nous notons par $\mathcal{H}$ ces hypothèses.

Résultats de la thèse Nous obtenons des résultats de la forme suivante :

Théorème 19 (Validation croisée et forte stabilité)**.** *Supposons $\mathcal{H}$. Soit Ψ un algorithme d'apprentissage qui est fortement $(\lambda, (\delta_{n,p_n})_{n,p_n}, d, \mathbb{Q})$ stable. Alors, pour tout $\varepsilon \geq 0$,*

$$\Pr\left(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + 2\lambda p_n\right) \leq 2\exp(-2np_n\varepsilon^2) + \delta_{n,p_n}.$$

De plus, si d est la distance uniforme d_∞, alors nous avons pour tout $\varepsilon \geq 0$:
Ainsi, si nous choisissons $\alpha = 9\lambda np_n$,

$$\Pr\left(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \delta_{n,p_n} + \lambda(2p_n)^\alpha\right) \leq 4(\exp(-\frac{\varepsilon^2}{8(18\lambda)^2 np_n^2}) + \frac{n}{9\lambda p_n}\delta'_{n,p_n}).$$

avec $\delta'_{n,p_n} = \delta_{n,p_n} + (n+1)\delta_{n+1,1/(n+1)}$.

De la même façon, pour la stabilité faible, nous obtenons :

Théorème 20 (Validation croisée et faible stabilité)**.** *Supposons $\mathcal{H}$. Soit Ψ un algorithme d'apprentissage qui est faiblement $(\lambda, (\delta_n)_n, d, \mathbb{Q})$ stable par rapport à la distance d. Alors, pour tout $\varepsilon \geq 0$,*

$$\Pr\left(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + 2\lambda p_n\right) \leq 2\exp(-2np_n\varepsilon^2) + \delta_{n,p_n}$$

Bien plus, si la distance est la distance uniforme d_∞, nous avons pour tout $\varepsilon \geq 0$:

$$\Pr\left(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \delta_{n,p_n} + 2\lambda p_n\right) \leq 4(\exp(-\frac{n\varepsilon^2}{10(9\lambda np_n)^2} + \frac{n\delta'^{1/2}_{n,p_n}}{9\lambda p_n}\exp(\frac{\varepsilon n}{4(9\lambda np_n)^2})) + n\delta'^{1/2}_{n,p_n})$$

avec $\delta'_{n,p_n} = 2\delta_{n,1/n} + \delta_{n,p_n}$

A l'aide des résultats précédents, nous obtenons des résultats en norme L_1. Dans le cas général, nous ne considérons que la notion de stabilité faible : $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ stabilité.

Théorème 21 (Erreur L_1 pour l'estimateur de la validation croisée)**.** *Supposons $\mathcal{H}$. Soit Ψ un algorithme d'apprentissage qui est faiblement $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ stable. Alors, nous avons :*

$$\mathbb{E}_{\mathcal{D}_n}|\widehat{R}_{CV} - \widetilde{R}_n| \leq 2\lambda p_n + \sqrt{\frac{2}{np_n}} + \delta_{n,p_n}$$

De plus, si Ψ est fortement $(\lambda, (\delta_n)_n, d_\infty, \mathbb{Q})$ stable, nous avons :

$$\mathbb{E}_{\mathcal{D}_n}|\widehat{R}_{CV} - \widetilde{R}_n| \leq \delta_{n,p_n} + 2\lambda p_n + 51\lambda\sqrt{n}p_n + \frac{n}{9\lambda p_n}\delta'_{n,p_n}$$

avec $\delta'_{n,p_n} = \delta_{n,p_n} + (n+1)\delta_{n+1,1/n+1}$

Rappelons nous que pour une grande partie des algorithmes d'apprentissage, nous avons $\delta_{n,p_n} = O_n(p_n \exp(-n(1-p_n))$. Nous choisissons $p_n^\star = (1/\sqrt{2}4\lambda)^{2/3}(1/n)^{1/3}$. Nous obtenons alors que $\mathbb{E}_{\mathcal{D}_n}|\widehat{R}_{CV} - \widetilde{R}_n| \leq 4(\lambda/n)^{1/3}$. Si Ψ est fortement$(\lambda, (\delta_n)_n, d_\infty, \mathbb{Q})$ stable, nous choisissons $p_n^\star = 1/n$ pour n assez grand et alors $\mathbb{E}_{\mathcal{D}_n}|\widehat{R}_{CV} - \widetilde{R}_n| = O_n(\lambda/\sqrt{n})$.
Au dela des résultats théoriques sur les concentrations, ces résultats permettent donc de construire des règles de choix, pour la taille des échantillons test/application, très utiles pour le praticien, selon le contexte considéré.

Ch3 : Résultats sur le subagging

Le Bagging est une procédure qui consiste à construire un prédicteur par rééchantillonage et une technique de combinaison. Le Bagging [bootstrap aggregating] a été introduit par [BRE96] afin de réduire la variance d'un prédicteur. A partir d'un algorithme d'apprentissage, un régresseur baggé est construit en moyennant plusieurs prédicteurs construits à l'aide d'échantillon bootstrap, un classificateur baggé est produit en votant à la majorité. Cette technique est particulièrement utile dans le cas des problèmes de grande dimension quand trouver un bon modèle en une étape est rendu difficile par la compléxité du problème. Concernant l'erreur de généralisation, cette méthode se compare favorablement avec le prédicteur original voire avec d'autres méthodes du type boosting ou randomisation. Ainsi, il est important de comprendre les raisons de ses succès et aussi occasionnellement de ses échecs. Cependant, même si cette approche a attiré beaucoup d'attention et est fréquemment appliquée, de nombreux problèmes restent ouverts sur le plan théorique. Dans ce chapitre, nous étudions une variante du bagging appelée Subagging [Subsample aggregating] qui est apparue dans [FRI00] et [BUH00]. Cette technique est plus accessible à l'analyse et a aussi des avantages non négligeables en terme de temps de calcul. L'estimateur subaggé sera noté par la suite $\Phi^B(X, \mathcal{D}_n)$ ou $\Phi_n^B(X)$.

Hypothèses Dans la suite, nous reprenons les notations pour la validation croisée introduite dans le chapitre 1 et 2. La distribution du vecteur d'apprentissage V_n^{tr} caractérise toutes les procédures de subagging :

Définition 22 (Regresseur subbagé)**.** *Le prédicteur subaggé construit à partir de φ_n et noté φ_n^B est défini par :*

$$\varphi_n^B(.) := \mathbb{E}_{V_n^{tr}} \varphi_{V_n^{tr}}(.)$$

Dans le cas des classificateurs, la règle bagging correspond au vote à la majorité. Nous supposons dans ce cas que $\mathcal{Y} = \{1, \ldots, M\}$.

Définition 23 (Classificateur subaggé.)**.** *Le classificateur subaggé dénoté φ_n^B peut être défini par :*

$$\varphi_n^B(X) := \arg \min_{k \in \{1,\ldots,M\}} \mathbb{E}_{V_n^{tr}} L(k, \Phi(X, \mathcal{D}_{V_n^{tr}}))$$

L'idée est d'adapter la validation croisée au prédicteur subaggé :

Définition 24 (Estimateur de validation croisée pour le subbagging)**.** *L'estimateur de validation croisée adapté au prédicteur subaggé φ_n^B est défini par :*

$$\widehat{R}_{CV}^{Out}(\Phi_n^B) := \mathbb{E}_{V_n^{tr}} \mathbb{P}_{n, V_n^{ts}}(\psi_{V_n^{tr}})$$

Dans la suite, la validation croisée est supposée symétrique (i.e. $\Pr(V_{n,i} = 1)$ est indépendant de i). Le nombre d'éléments dans l'échantillon de test est égal à np_n. En outre, supposons que ϕ_n appartienne à une famille de prédicteurs de VC-dimension finie. L est supposée bornée de la façon suivante : $L(Y, \varphi(X)) \leq C(h(Y, \varphi(X))$ avec C une fonction convexe -bornée elle-même par 1 sur le support de $\varphi_{v_n^{tr}}$-, et h telle que pour n'importe quel λ, $0 < \lambda < 1$, nous avons $h(y, \lambda\varphi(x_1) + (1-\lambda)\varphi(x_2)) = \lambda h(y, \varphi(x_1)) + (1-\lambda)h(y, \varphi(x_2))$. Nous supposons aussi que les prédicteurs sont symétriques en l'échantillon d'apprentissage. Nous notons ces hypothèses $\mathcal{H}$**.**

Etat de l'art [BRE96] agrége des arbres de regression pour construire des forêts aléatoires et nomme ce procédé bagging. [BUH00] considère des prédicteurs non différentiables et discontinus et prouve des effets lissants au voisinage des points de discontinuité des surfaces de décision. [GRA04] apporte de nouveaux arguments pour expliquer l'effet du bagging : les améliorations/détériorations du bagging sont expliquées par l'influence positive ou négative d'exemples très influençants. [ELI04] insiste sur le caractère stabilisant du bagging sur le prédicteur d'origine et prouve des bornes non asymptotiques sur la déviation entre l'erreur de généralisation et l'erreur leave-one-out. Trouver une réponse générale à la question $\widetilde{R}_n(\Phi_n^B) \leq \widetilde{R}_n(\Phi)$? semble difficile à obtenir dans un cadre général. En effet, à l'aide du théorème de Gauss-Markov, [GRA04] montre que comme le prédicteur baggé et non baggé sont tous non biaisés, alors la variance du prédicteur non baggé est plus petite que celle du prédicteur baggé. [BUJ00] propose des statistiques quadratiques générales pour lesquelles le prédicteur baggé augmente à la fois le biais et la variance. Ainsi, nous proposons d'estimer directement l'erreur de généralisation du prédicteur baggé à l'aide d'une validation croisée adaptée, plutôt que de déterminer les conditions générales difficilement atteignables visant à répondre à la question de l'effet du bagging.

Résultats de la thèse Nous montrons des bornes du type $\Pr(\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out} \geq \varepsilon) \leq \min(B(n, p_n, \varepsilon), V(n, p_n, \varepsilon)) < 1$ avec $\varepsilon > 0$. Le terme $B(n, p_n, \varepsilon)$ est une borne à la Vapnik-Chernovenkis alors que le terme $V(n, p_n, \varepsilon)$ est une borne à la Hoeffding contrôlée par la taille de l'échantillon de test np_n. Remarquons que ces bornes sont plus petites que 1 quel que soit la taille de l'échantillon ce qui les rend d'autant plus utile pour de petits échantillons. Nous obtenons des résultats de la forme suivante :

Théorème 25 (Déviation absolue pour la validation croisée symétrique)**.** *Supposons $\mathcal{H}$. Alors, nous avons pour tout $\varepsilon > 0$,*

$$\Pr(\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out} \geq \varepsilon) \leq \min(B_{sym}(n, p_n, \varepsilon), V_{sym}(n, p_n, \varepsilon)) < 1$$

avec
- $B_{sym}(n, p_n, \varepsilon) = (2np_n + 1)^{4V_{\mathcal{C}}/p_n} e^{-n\varepsilon^2}$
- $V_{sym}(n, p_n, \varepsilon) = \exp(-2np_n\varepsilon^2)$.

Théorème 26 (Déviation absolue pour la validation croisée symétrique)**.** *Supposons $\mathcal{H}$. Supposons aussi que ϕ_n minimise le risque empirique. Mais au lieu de minimiser $\widehat{R}_n(\phi)$, nous supposons que ϕ_n minimise $\frac{1}{n}\sum_{i=1}^n C(h(Y_i, \phi(X_i))$. Par simplicité, nous supposons que l'infimum est atteint i.e. $\phi_n = \arg\min_{\phi \in \mathcal{C}} \frac{1}{n}\sum_{i=1}^n C(h(Y_i, \phi(X_i))$. Alors, nous avons pour tout $\varepsilon > 0$,*

$$\Pr(\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out} \geq \varepsilon) \leq \min(B_{ERM}(n, p_n, \varepsilon), V_{ERM}(n, p_n, \varepsilon)) < 1$$

avec
- $B_{ERM}(n, p_n, \varepsilon) = \min((2np_n+1)^{4V_{\mathcal{C}}/p_n} \exp(-n\varepsilon^2), (2n(1-p_n)+1)^{\frac{4V_{\mathcal{C}}}{1-p_n}} \exp(-n\varepsilon^2/9))$
- $V_{ERM}(n, p_n, \varepsilon) = \exp(-2np_n\varepsilon^2)$.

Soit $\mathcal{D}_n$ un échantillon de taille n. Soit $V_n^{tr} \sim \mathbb{Q}$ un vecteur d'apprentissage indépendant de $\mathcal{D}_n$ tel que la validation croisée soit symétrique. Le nombre d'éléments dans l'échantillon test est constant et vaut np_n. Soit d une distance parmi d_e, d_1, d_∞. Enfin, nous supposons que la fonction de perte L est bornée par 1. Nous avons les résultats qui restent valides sous des procédures générales de validation croisée et des algorithmes stables. Nous notons ces hypothèses par $\mathcal{G}$.

Théorème 27 (Validation croisée et stabilité forte)**.** *Supposons $\mathcal{G}$. Soit Ψ un algorithme fortement $(\lambda, (\delta_n, p_n)_{n,p_n}, \mathbb{Q})$ stable par rapport à la distance d. Alors, pour tout $\varepsilon \geq 0$, nous avons :*

$$\Pr\left(\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out} \geq \varepsilon\right) \leq \exp(-2np_n\varepsilon^2)$$

De plus, si d est la distance uniforme d_∞, alors nous avons pour tout $\alpha > 0$:

$$\Pr\left(\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out} \geq \varepsilon\right) \leq \min(\exp(-2np_n\varepsilon^2), 2(\exp(-\frac{\varepsilon^2}{8n(8\lambda np_n+\alpha)^2}) + \frac{n}{\alpha}\delta_{n,p_n}))$$

Alors, si nous choisissons $\alpha = 8\lambda np_n$,

$$\Pr\left(\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out} \geq \varepsilon\right) \leq \min(\exp(-2np_n\varepsilon^2), 2(\exp(-\frac{\varepsilon^2}{8(16\lambda)^2np_n^2}) + \frac{n}{8\lambda p_n}\delta_{n,p_n}))$$

Le cas du subagging en classification (i.e. le vote par majorité) est aussi traité.
Demeure une question pratique importante : comment choisir p_n ? Nous donnons un début de réponse sous la forme d'une inégalité de concentration.

Travaux appliqués

Ces travaux ont été réalisés dans le cadre de mon activité salariée. Les deux premiers ont été effectués pendant mon passage à la Division Synthèse Conjoncturelle à l'Insee. Le dernier a été écrit lors de mon travail à la RetD (recherche et développement) d'EDF (Electricité de France).

Ch 4 : Analyse factorielle dynamique multifréquence appliquée à la datation de la conjoncture française

Dans cet article, nous datons de manière fine -mensuelle- les périodes de la conjoncture française de 1985. La méthode utilisée est appliquée aux enquêtes de conjoncture dans l'industrie afin de revisiter l'indicateur synthétique du climat des affaires dans l'industrie manufacturière publié par l'Insee.
La datation des cycles est rendue difficile par la multiplicité des indicateurs. Afin de pallier cette difficulté, les conjoncturistes construisent des indicateurs mensuels synthétiques. Toutefois, ces indicateurs conjoncturels mensuels présentent l'inconvénient de ne pas prendre en compte la série trimestrielle qui est déjà une synthèse particulière de grand intérêt pour les conjoncturistes : le PIB lui-même. [MM03] ont proposé une méthodologie utilisant simultanément des séries mensuelles et trimestrielles. Un indicateur synthétique mensuel est ainsi construit à partir de grands indicateurs quantitatifs de l'économie française ne se limitant pas au PIB (l'indice de la production industrielle, les dépenses des ménages en produits manufacturés, l'effectif salarié).
Au vu de cet indicateur, nous distinguons sept phases de conjoncture durant cette période. Une seule récession apparaît, de septembre 1992 à mai 1993.

- de janvier 1985 à janvier 1992 : forte croissance.
- de février 1992 à août 1992 : stabilité.
- de septembre 1992 à mai 1993 : récession.
- de juin 1993 à décembre 1994 : forte croissance.
- de janvier 1995 à avril 1997 : croissance moyenne.
- d'avril 1997 à juin 2001 : forte croissance.

- de partir de juillet 2001 à 2004 : croissance modeste.

Cette datation apparaît cohérente avec celle obtenue par [DL95] sur la période 1985-1995 qui font leur analyse à partir de l'enquête de conjoncture dans l'industrie. S'agissant de la récession de 1993, le détail des comptes trimestriels montre en effet pour l'année 1993, une baisse du PIB réel. Cette dernière est principalement présente du quatrième trimestre de 1992 au deuxième trimestre de 1993 dans les dépenses de ménages, les investissements des entreprises et des ménages. Ici, nous la datons précisément de septembre 1992 à mai 1993. De la comparaison entre indicateur commun entre les différents indicateurs quantitatifs, le

conjoncturiste peut tirer deux enseignements :

- l'avance ou le retard d'une branche par rapport à l'activité générale.
- le dynamisme relatif d'une branche par rapport à l'activité d'ensemble.

Par exemple, on constate que :

- La consommation en produits manufacturés des ménages est soit en avance soit coïncidente avec le cycle économique. La consommation en produits manufacturés des ménages présente un profil plus volatile au mois le mois, ce qui rend sa lecture plus difficile. Dès mars 1990, la consommation en produits manufacturés des ménages affiche un profil à la baisse, bien avant le déclin de l'indicateur synthétique.
- Le PIB est imprécis pour dater les cycles. La simple lecture du PIB indiquerait une récession du quatrième trimestre de 1992 jusqu'au troisième trimestre de 1993. L'indicateur du climat conjoncturel général limite cette dernière au quatrième trimestre de 1992, premier trimestre de 1993 et au deuxième trimestre de 1993.

L'approche considérée présente ainsi deux intérêts majeurs :

- elle offre une grille de lecture de l'activité passée en intégrant de façon simultanée les différentes branches de l'économie.
- elle donne un chiffrement fin (mensuel) de ces différentes périodes. La méthode de

[MM03] au PIB (trimestriel) est ensuite appliquée aux enquêtes de conjoncture dans l'industrie (mensuel) et au PIB français (trimestriel). Le but est d'extraire l'information commune entre activité économique et opinion qu'en ont les chefs d'entreprise. [DL95] ont montré l'intérêt d'une approche factorielle dynamique pour obtenir un facteur commun à partir des six soldes d'opinion émis dans l'enquête de conjoncture dans l'industrie. Le facteur commun apparaît comme un indicateur du climat général industrie. Ce facteur commun apparaît très semblable à l'ancien facteur commun de [DL95]. La comparaison entre ces deux facteurs peut s'avérer riche d'enseignements :

- nous pouvons comparer les opinions des chefs d'entreprise avec un indice conjoncturel reposant sur le PIB et faire ainsi la part entre anticipations et réalité économique.
- cette différence peut aussi être interprétée comme un renseignement complémentaire du côté des services.

Sur la période d'estimation (1978 à 2003), nous remarquons :

- de 1978 à 1985, le facteur commun industrie est avancé sur le nouveau facteur commun intégrant le PIB. Ceci peut s'interpréter comme une avance du cycle du secteur industriel sur le cycle global.

- de 1985 à 2001, les deux indicateurs sont très semblables, indiquant peut-être une synchronisation des cycles du secteur manufacturier avec le secteur services.

- de 2001 à 2003, nous remarquons de nouveau un décalage entre le cycle industriel et le cycle global.

En outre, un regard sur l'amplitude des facteurs communs semblerait indiquer :

- de 1980 à 1985, les industriels apparaissent plus pessimistes que de raison car le facteur commun industrie se trouvent en dessous du facteur commun intégrant le PIB.

- de 2001 à 2003, les industriels apparaissent plus optimistes que ne laisse suggérer la situation de l'activité économique pour les raisons opposées.

Ch5 : Un nouvel indicateur synthétique mensuel résumant le climat des affaires dans les services en France

Le nouvel indicateur synthétique mensuel présenté dans cet article résume l'information contenue dans l'enquête de conjoncture dans les services. Il est obtenu par extraction d'un signal commun à trois séries de fréquence mensuelle et trois de fréquence trimestrielle. L'approche retenue pour le construire relève du cadre de l'analyse factorielle dynamique. En effet, cette méthode est suffisamment générale pour prendre en compte des séries de fréquences différentes ainsi que des ruptures dans la fréquence des séries. L'indicateur synthétique est obtenu à travers un modèle à composantes inobservables, qui, écrit sous une forme espace-état, est estimé au moyen du filtre de Kalman. Plusieurs modèles alternatifs, dynamiques ou statiques, ont été explorés. Ces modèles fournissent des indicateurs très semblables, ce qui montre la robustesse du signal obtenu.

L'indicateur synthétique peut être décliné dans les trois sous-secteurs couverts par l'enquête de conjoncture dans les services (services aux entreprises, services aux particuliers et activités immobilières), ce qui permet d'affiner l'analyse conjoncturelle. Ainsi, son examen confirme la reprise de l'activité dans l'ensemble des services à partir de la mi-2003. Cette reprise apparaît hésitante au deuxième semestre 2004 et semble s'essouffler début 2005.

Enfin, cet indicateur peut être utilisé par le conjoncturiste pour actualiser sa prévision de production de services au mois le mois et non plus au trimestre le trimestre. En outre, combiné à l'indicateur synthétique dans l'industrie manufacturière, il apporte une information supplémentaire à la prévision du PIB.

Jusqu'à la fin des années quatre-vingt-dix, l'analyse conjoncturelle s'appuyait très largement sur les informations qualitatives et quantitatives relatives au secteur manufacturier. En particulier, elle exploitait assez peu les indicateurs concernant le secteur des services. Ce décalage, qui par le passé pouvait s'expliquer par l'abondance relative d'informations statistiques portant sur l'industrie, est en train de s'estomper. En effet, une place très importante est désormais accordée aux services dans le système statistique français, en raison notamment de leur rôle déterminant dans la compréhension des évolutions de court terme. À cet égard, [BE03] montrent que l'enquête de conjoncture dans les services est complémentaire de l'enquête dans l'industrie et permet d'améliorer la prévision du taux de croissance trimestriel du PIB. En plus des soldes d'opinion de l'enquête services, leur étude mobilise un indicateur synthétique trimestriel extrait à l'aide d'une analyse factorielle statique.

Depuis septembre 2004, l'Insee publie chaque mois les résultats de son enquête de conjoncture dans les services ainsi qu'un indicateur synthétique résumant l'information contenue dans cette enquête (cf. figure 3). Cet indicateur mensuel représente l'information commune aux principaux soldes d'opinion de l'enquête. Il a pour objectif de faciliter la lecture des résultats et d'améliorer le diagnostic conjoncturel sur le secteur des services. La construction d'un indicateur synthétique mensuel s'inscrit dans le prolongement des travaux de [BE03]. Celui-ci présente un double apport : d'une part il est de fréquence mensuelle ; d'autre part il intègre des séries de fréquence mixte, mensuelle ou trimestrielle.
Ce nouvel indicateur synthétique est calculé dans le cadre de l'analyse factorielle dynamique. En effet, cette approche est suffisamment flexible pour prendre en compte les séries de fréquence variable issues de l'enquête services. Elle a été mise en œuvre dans de nombreuses études, en particulier par [GE77, SS77, SW89]. Cette étude est également très largement fondée sur les travaux de [DL95], qui ont appliqué l'analyse factorielle dynamique à l'enquête de conjoncture de l'Insee dans l'industrie. Plus précisément, chaque solde d'opinion est représenté comme la somme de deux termes orthogonaux : le premier terme obéit à une dynamique commune à l'ensemble des séries, le second est une composante spécifique à chaque série. Chaque terme est modélisé par un processus de type ARMA (autorégressif moyenne mobile). Les paramètres du modèle sont estimés au moyen du filtre de Kalman. L'indicateur synthétique est l'espérance du facteur commun conditionnelle à l'information passée. L'application du filtre de Kalman à des séries temporelles est notamment décrite par [HA91] et par [KN89], qui l'illustrent à travers plusieurs exemples concrets.
Comme dans l'industrie, l'indicateur synthétique dans les services s'interprète comme une mesure du climat des affaires tel qu'il est perçu par les chefs d'entreprise. Il enrichit la panoplie des indicateurs de court terme ; en particulier c'est le complément naturel de l'indicateur synthétique dans l'industrie manufacturière. De plus, il présente l'avantage d'être diffusé de façon relativement précoce puisque les résultats des enquêtes de conjoncture sont disponibles avant les indicateurs quantitatifs. En outre, il peut être décliné dans les trois sous-secteurs couverts par l'enquête de conjoncture dans les services, ce qui permet d'affiner le diagnostic conjoncturel.

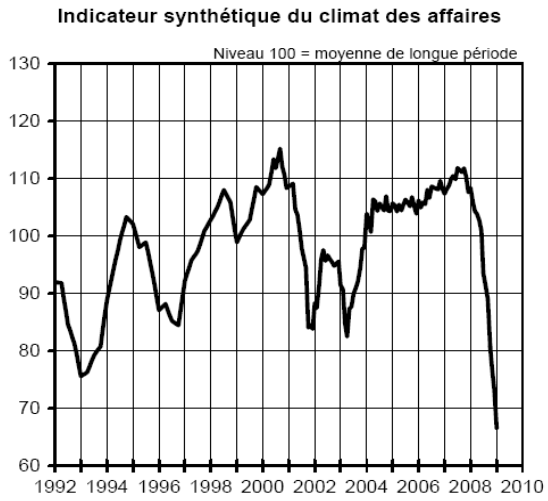


FIG. 3 – Indicateur synthétique dans les services en France.

Ch6 : Simulating spot electricity prices with regenerative blocks

Avant les années 90, la production d'électricité était un monopole dans la plupart des pays, possédé ou régulé par le gouvernement. Un brusque changement a eu lieu : les plus grands pays ont dérégulé la production et la vente d'électricité. L'une des conséquences les plus importantes de ce changement est que les prix ne sont plus réglémentés mais déterminés par la loi de l'offre et de la demande. Cependant, le comportement erratique des prix, dû notamment au caractère non stockable de l'électricité, rend caduque l'utilisation de modèles financiers classiques. Cette remarque prêche pour l'utilisation de modèle semi-paramétrique, ce qui n'est pas l'approche communément employée dans la littérature. Ainsi, au lieu de prendre des hypothèses paramétriques, nous faisons une hypothèse sur la structure de dépendance des données. Nous proposons un algorithme fondé sur les chaînes de Markov, introduit par [BC05, BC07]. Ce modèle va capturer les caractéristiques suivantes observées sur la plupart des marchés d'électriques : saisonnalités, pics de prix, retour à la moyenne, hétéroscédasticité, et non stationnarité de long terme.

Etat de l'art Une revue de littérature détaillée des modèles économétriques en finance appliqués à l'électricité peut être trouvée dans [Bunn2003].
Afin de modéliser les pics de prix, [Kaminski1997] propose une marche aléatoire avec sauts, inspiré par [Merton1976]. Cependant, cette spécification ne prend pas en compte une autre caractéristique des prix d'électricité : le retour à la moyenne. Cette prise en compte est faite par [Johnson1999]. Ce modèle capte d'importants faits stylisés tels le retour à la moyenne et les pics mais suppose toujours une volatilité déterministe. [Deng2000] étend cette classe de modèle en ajoutant des non linéarités dans la dynamique des prix comme le changement de régime et la volatilité stochastique. [Escribano2002] généralise les approches précédentes par un GARCH saisonnier. Pour séparer les sauts des larges pics, le changement de régimes a été proposé comme une alternative aux diffusions avec sauts ([Huisman2001]). Dans [Bystrom2001], un processus de prix AR-GARCH avec une composante saisonnière dans la volatilité est combiné avec des outils de la théorié des extrêmes. Enfin, d'autres modèles [Conejo2005, Cuaresma2004] ont pour but spécifique de prévoir les prix d'électricité.
Suivant [BC05, BC07], nous introduisons une approche non paramétrique, pour prendre en compte la complexité des faits stylisés des prix spot de l'électricité du marché Powernext. En outre, nous pensons que ce modèle peut être facilement adapté à d'autres marchés.

Modèle Nous présentons un modèle général pour le marché spot d'électricté (vente d'un MWH d'électricité pour le lendemain) qui reproduit les principales caractéristiques des prix d'électricité. Dans ce chapitre, nous estimons ce modèle sur les données de Powernext Energy Exchange PWX, mais cela peut facilement étre adapté à d'autres marchés. Nous décrivons le prix marché spot $(P_t)_t$ par un processus stochastique bidimensionnel journalier :

$(P_t)_t = \begin{pmatrix} P_t^{\text{peak}} \\ P_t^{\text{offpeak}} \end{pmatrix}_t$ avec P_t^{peak} le prix d'un MWH pendant les heures de pointe (8-20) et P_t^{offpeak} le prix d'un MWH pendant les heures creuses (le reste).
Nous introduisons trois facteurs : le processus de long terme Z_t, le processus de température θ_t, le processus de court terme ε_t et les quantités additionnelles suivantes : (logarithmique) courbe de prix/température $f(t,.)$ et la courbe de variance due à la température $\sigma(.)$.

L'équation du prix peut s'écrire par :

$$P_t = \exp\left(Z_t + f(t,\theta_t) + \sigma(\theta_t)\varepsilon_t\right)$$

Z_t peut être interprété comme un trend économique non stationnaire, en lien avec l'augmentation des prix des autres commodités telles le pétrole ou le gaz.
Le processus θ_t décrit la température. Les fonctions $f : \mathbb{Z}\times\mathbb{R} \to \mathbb{R}$ et $\sigma : \mathbb{R} \to \mathbb{R}_+$ dépendent de la date actuelle t et de la température θ_t. Elle décrit la relation non linéaire entre le prix et la température. La dépendance en temps t se comprend par le caractère saisonnier de la demande.
Enfin, ε_t décrit le terme résiduel des fluctuations du marché.

Dans un premier temps, nous filtrons les données afin d'extraire les résidus ε_t. Nous estimons ainsi heuristiquement Z_t, $f(t,.)$, $\sigma(.)$. Dans un deuxième temps, nous décrivons le modèle et l'appliquons aux résidus extraits.

Nous rappelons un algorithme introduit par [BC05, BC07] pour prendre en compte la dépendance dans les résidus temporels ε_t. Cette méthode est fondée sur les propriétés de régénération de ε_t qui permettent de découper la série temporelle en blocs presque indépendants.
L'hypothèse principale est que la série temporelle résiduelle est un processus de markov d'ordre 1, Harris récurrent. Il est connu depuis les travaux de [NUM78] que toute chaîne de Harris-récurrente (cf. [MT96]) peut être vue comme une comme une chaîne atomique, quitte à l'étendre : on peut en effet construire une chaîne étendue, dite split-chaine qui possède un atome et donc des propriétés de régénérations qui permettent de la découper en blocs indépendants.

Considérons une chaîne de Markov ε_t avec une transition de densité p. Un ensemble S est un petit ensemble s'il existe $\delta > 0$ et une densité ϕ de support S telle que pour tout $(x,a) \in S^2$,

$$p(x,a) \geq \delta\phi(a). \tag{3}$$

avec $p(x,A)$ la densité du noyau de transition de $P(x,A) := \Pr(X_t \in A | X_{t-1} = x)$. L'idée est de construire une nouvelle chaîne (ε, W) telle que la distribution de W sachant ε

- si $\varepsilon_t \notin S$, générer W_t comme une variable aléatoire Bernouilli $Ber(\delta)$.
- si $\varepsilon_t \in S$, générer W_t comme une variable aléatoire Bernouilli $Ber\left(\delta\phi(\varepsilon_{t+1})/p(\varepsilon_t,\varepsilon_{t+1})\right)$.

Le principe de construction est que sous la condition de minimisation,(3), $p(x,a)$ peut être vue sur S comme un mélange : $p(x,a) = (1-\delta)(p(x,a) - \delta\phi(a))/(1-\delta) + \delta\phi(a)$, qui est

constant (indépendant de x) quand le second élément est choisi (voir [BC05] et [MT96] pour les détails).
Cette construction assure que la chaîne (ε, W) est atomique, et préserve la distribution marginale d'ε (cf. [MT96]). L'atome est alors $A = S \times \{1\}$. Il est alors possible de découper la chaîne quitte à estimer p, dans une première étape.
Nous illustrons cette approche par l'exemple suivante : nous considérons l'année 2006 avec la même tendance économique que l'année 2005. Le graphique 4 donne deux simulations différentes des prix spot pour l'année 2006, toutes ayant les caractéristiques des prix spot.

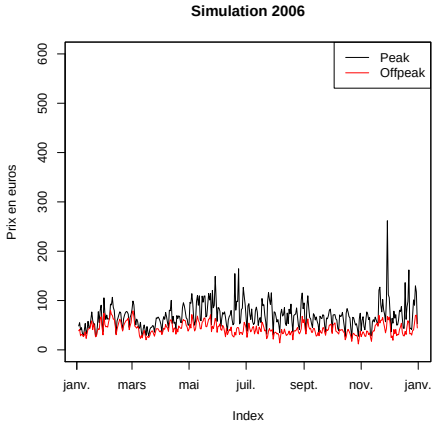


FIG. 4 – Une simulation d'une année 2006 avec la température 2005.

Part I

Concentration inequalities of the cross-validation estimator in the context of risk assessment

Chapter 1

Concentration inequalities of the cross-validation estimator of Empirical Risk Minimizer

In this article, we derive concentration inequalities for the cross-validation estimate of the generalization error for empirical risk minimizers. In the general setting, we prove sanity-check bounds in the spirit of [KR99] "*bounds showing that the worst-case error of this estimate is not much worse that of training error estimate* ". General loss functions and class of predictors with finite VC-dimension are considered. We closely follow the formalism introduced by [DUD03] to cover a large variety of cross-validation procedures including leave-one-out cross-validation, k-fold cross-validation, hold-out cross-validation (or split sample), and the leave-v-out cross-validation.

In particular, we focus on proving the consistency of the various cross-validation procedures. We point out the interest of each cross-validation procedure in terms of rate of convergence. An estimation curve with transition phases depending on the cross-validation procedure and not only on the percentage of observations in the test sample gives a simple rule on how to choose the cross-validation. An interesting consequence is that the size of the test sample is not required to grow to infinity for the consistency of the cross-validation procedure.

Keywords : Cross-validation, generalization error, concentration inequality, optimal splitting, resampling.

1.1 Introduction and motivation

Pattern recognition (or classification or discrimination) is about predicting the unknown nature of an observation: an observation is a collection of numerical measurements, represented by a vector x belonging to some measurable space $\mathcal{X}$. The unknown nature of the observation is denoted by y belonging to a measurable space $\mathcal{Y}$. In pattern recognition, the goal is

to create a measurable map $\phi : \mathcal{X} \to \mathcal{Y}$; $\phi(x)$ which represents one's prediction of y given x. The error of a prediction $\phi(x)$ when the true value is y is measured by $L(y, \phi(x))$, where the loss function $L \in \mathcal{Y}^2 \to \mathbb{R}_+$. For simplicity, we suppose $L \leq 1$. In a probabilistic setting, the distribution $\mathbb{P}$ of the random variable $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ describes the probability of encountering a particular pair in practice. The performance of ϕ, that is how the predictor can predict future data, is measured by the risk $R(\phi) := \mathbb{E}_{(X,Y)} L(Y, \phi(X))$. In practice, we have access to n independent, identically distributed (*i.i.d.*) random pairs $(X_i, Y_i)_{1 \leq i \leq n}$ sharing the same distribution as (X, Y) called the learning sample and denoted $\mathcal{D}_n$. A learning algorithm Φ is trained on the basis of $\mathcal{D}_n$. Thus, Φ is a measurable map from $\mathcal{X} \times \cup_n (\mathcal{X} \times \mathcal{Y})^n$ to $\mathcal{Y}$. Y is predicted by $\Phi(X, \mathcal{D}_n)$. The performance of $\Phi(., \mathcal{D}_n)$ is measured by the conditional risk called the generalization error denoted by $\widetilde{R}_n := \mathbb{E}_{(X,Y)}[L(Y, \Phi(X, \mathcal{D}_n)) \mid \mathcal{D}_n]$ with $(X, Y) \sim \mathbb{P}$ independent of $\mathcal{D}_n$ and with the following equivalent notation for the conditional expectation of $h(X, Y)$ given Y: $\mathbb{E}_X h(X, Y)$. In the following, if there is no ambiguity, we will also allow the notation $\phi(X, \mathcal{D}_n)$ instead of $\Phi(X, \mathcal{D}_n)$. Notice that $\widetilde{R}_n$ is a random variable measurable with respect to $\mathcal{D}_n$.

An important question is: *The distribution $\mathbb{P}$ of the generating process being unknown, can we estimate how good a predictor trained on a learning sample of size n is? In other words, can we estimate the generalization error $\widetilde{R}_n$?* This fundamental statistical problem is referred to "choice and assessment of statistical predictions" [STO74] . Many estimates have been proposed, among them the resubstitution estimate (or training estimate). The predictor is trained using the entire learning sample $\mathcal{D}_n$, and an estimate of the prediction is obtained by running the same learning process through the predictor and comparing predicted and actual responses. Thus, the resubstitution estimate $\widehat{R}_n := \frac{1}{n} \sum_{i=1}^{n} L(Y_i, \phi(X_i, \mathcal{D}_n))$ can severely underestimate the bias. It can even drop to zero for some machine learning even though the generalization error is nonzero (for example, the $1-$nearest neighbor). The difficulty arises from the fact that the learning sample is used both for training and testing. In order to get rid of this downward bias, the estimation of the generalization error based on sample reuse have been favored among practitioners. Quoting [HTF01]: *Probably the simplest and most widely used method for estimating prediction error is cross-validation.* However, the role of cross-validation estimator, denoted by $\widehat{R}_{CV}$, is far from being well understood in a general setting. In particular, the following problems remain partially solved: "Is $\widehat{R}_{CV}$ a good estimator of the generalisation error?", "How should one choose k in a k-fold cross-validation" or "Does cross-validation outperform the resubstitution error ?". The purpose of this paper is to give a partial answer to the first two questions.

We introduce our **main result** for symmetric cross-validation procedures. We divide the learning sample into two samples: the training sample and the test sample, to be defined below. We denote by p_n the percentage of elements in the test sample such that np_n is an integer. For empirical risk minimizers over a class of predictors with finite VC-dimension $V_{\mathcal{C}}$, to be defined below, we have the following concentration inequality, for all $\varepsilon > 0$:

$$\Pr\left(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon\right) \leq B(n, p_n, \varepsilon) + V(n, p_n, \varepsilon),$$

with

- $B(n, p_n, \varepsilon) = 5(2n(1-p_n)+1)^{\frac{4V_{\mathcal{C}}}{1-p_n}} \exp(-\frac{n\varepsilon^2}{64})$
- $V(n, p_n, \varepsilon) = \min\left(\exp(-\frac{np_n\varepsilon^2}{2}), \frac{16}{\varepsilon}\sqrt{\frac{V_{\mathcal{C}}(\ln(2(1-p_n)+1)+4)}{n(1-p_n)}}\right).$

The term $B(n, p_n, \varepsilon)$ is a Vapnik-Chernovenkis-type bound controlled by the size of the training sample $n(1-p_n)$ whereas the term $V(n, p_n, \varepsilon)$ is the minimum between a Hoeffding-type term controlled by the size of the test sample np_n, a polynomial term controlled by the size of the training sample. As the percentage of observations in the test sample p_n increases, the $V(n, p_n, \varepsilon)$ term decreases but the $B(n, p_n, \varepsilon)$ term increases.
The difference from the previous results on estimation of $\widetilde{R}_n$ is in the following:

- our bounds for intensive cross-validation procedures (i.e. k-fold cross-validation or leave-v-out cross-validation) are not worse than those for hold-out cross-validation.
- our inequalities not only depend on the percentage of observations in the learning sample p_n but also on the precise type of cross-validation procedure: this is why we can discriminate between k-fold cross-validation and hold-out cross-validation even if p_n is the same.
- we show that the size of the test sample does not need to grow to infinity for the cross-validation procedure to be consistent for the estimation of the generalization error.

Using these probability bounds, we can then deduce that the expectation of the difference between the generalization error and the cross-validation estimate $\mathbb{E}_{\mathcal{D}_n}|\widehat{R}_{CV} - \widetilde{R}_n|$ is of order$O_n(\sqrt{V_{\mathcal{C}}\ln(n(1-p_n))/n(1-p_n)} + \sqrt{1/np_n})$. As far as $\mathbb{E}_{\mathcal{D}_n}|\widehat{R}_{CV} - \widetilde{R}_n|$ is concerned, we can define a splitting rule: the percentage of elements p_n in the test sample should be proportional to $\frac{1}{1+V_{\mathcal{C}}^{1/3}}$, i.e. the larger the class of predictors is, the smaller the test sample in the cross-validation should be.

The paper is organized as follows. In the next section, we give a short review of literature. We detail the main cross-validation procedures and we summarize the previous results for the estimation of generalization error. In Section 3, we introduce the main notations and definitions. Finally, in Section 4, we introduce our results, in terms of concentration inequalities. In companion papers, we will show that in some cases, the cross-validation estimate can outperform the training estimate and prove that cross-validation can work out with infinite VC-dimension predictor.

1.2 Short Review of the literature on cross-validation

The cross-validation $\widehat{R}_{CV}$ includes leave-one-out cross-validation, k-fold cross-validation, hold-out cross-validation (or split sample), leave-v-out cross-validation (or Monte Carlo cross-validation or bootstrap cross-validation). In leave-one-out cross-validation, a single sample of size n is used. Each member of the sample in turn is removed, the full modeling method is applied to the remaining $n-1$ members, and the fitted model is applied to the hold-backmember. An early (1968) application of this approach to classification is that of [LM68]. [AL68] gave perhaps the first application in multiple regression and [GEI75] sketches other applications. However, this special form of cross-validation has well-known limitations, both theoretical and practical, and a number of authors have considered more general multifold cross-validation procedures [BREI84] ; [BREI92] ; [BUR89] ; [DGL96] ; [GEI75] ; [GYO02] ; [McC76] ; [PIC84] ; [RIP96] ; [SHAO93] ; [ZHA93]). The k-fold procedure divides the learning sample into k equally sized folds. Then, it produces a predictor by training on $k-1$ folds and testing on the remaining fold. This is repeated for each fold, and the observed errors are averaged to form the k-fold estimate. Leave-v-out cross-validation is a more elaborate and expensive version of cross-validation that involves leaving out all possible subsamples of v cases. In the split-sample method or hold-out, only a single subsample (the training sample) is used to estimate the generalization error, instead of k different subsamples; i.e., there is no crossing. Intuitively, there is a tradeoff between bias and variance in cross-validation procedures. Typically, we expect the leave-one-out cross-validation to have a low bias (the generalization error of a predictor trained on $n-1$ pairs should be close to the generalization error of a predictor trained on the n pairs) but a high variance. Leave-one-out cross-validation often works well for estimating generalization error for continuous loss functions such as the squared loss, but it may perform poorly for discontinuous loss functions such as the indicator loss. On the contrary, k-fold cross-validation or leave-v-out cross-validation are expected to have a higher bias but a smaller variance due to resampling.

With the exception of [BUR89], theoretical investigations of multifold cross-validation procedures have first concentrated on linear models ([Li87];[SHAO93];[ZHA93]). Results of [DGL96] and [GYO02] are discussed in Section 3. The first finite sample results are due to [DEWA79] and concern k-local rules algorithms under leave-one-out and hold-out cross-validation. More recently, [HOL96, HOL96bis] derived finite sample results for the hold-out, $k-$fold and leave-one-out cross-validations for finite VC algorithms in the realisable case (the generalization error is zero). But the bounds for $k-$fold cross-validation are k times worse than for hold-out cross-validation. [BKL99] have emphasized when $k-$fold can out perform hold-out cross-validation in a particular case of k-fold predictor. [KR99] has extended such results in the case of stable algorithms for the leave-one-out cross-validation procedure. [KEA95] also derived results for hold-out cross-validation for VC algorithms without the realisable assumption. However, the bounds obtained are "sanity check bounds" in the sense that they are not better than classical Vapnik-Chernovenkis's bounds. [DUD04BIS] derived finite sample results for the distance between the cross-validation estimate and a special benchmark and proved asymptotic results for the relation between the cross-validation risk

and the generalization error. To our knowledge, bounds for intensive cross-validation procedures are missing. This might be due to the lack of independence between the crossing terms of the cross-validated estimate [KMNR95].

1.3 Notations and definitions

We introduce here useful definitions to define the various cross-validation procedures. First, we define binary vectors, i.e. $V_n = (V_{n,i})_{1 \leq i \leq n}$ is a vector of size n, such that for all $i, V_{n,i} \in \{0,1\}$ and $\sum_i V_{n,i} \neq 0$. Consequently, knowing the binary vector, we can define the subsample associated with it: $\mathcal{D}_{V_n} := \{(X_i, Y_i) \in \mathcal{D}_n | V_{n,i} = 1, 1 \leq i \leq n\}$. The weighted empirical error of φ is denoted by $\hat{R}_{V_n}(\phi)$ and defined by:

$$\hat{R}_{V_n}(\phi) := \frac{1}{\sum_{i=1}^n V_{n,i}} \sum_{i=1}^n V_{n,i} L(Y_i, \phi(X_i)).$$

For $\hat{R}_{1_n}$, with 1_n the binary vector of size n with 1 at every coordinate, we will use the simpler notation $\hat{R}_n$. For a predictor trained on a subsample, we define:

$$\phi_{V_n}(.) := \Phi(., \mathcal{D}_{V_n}).$$

With the previous notations, notice that the predictor trained on the learning sample $\phi(., \mathcal{D}_n)$ can be denoted by $\phi_{1_n}(.)$. We will allow the simpler notation $\phi_n(.)$. The learning sample is divided into two disjoint samples: the training sample of size $n(1-p_n)$ and the test sample of size np_n, where p_n is the percentage of elements in the test sample. To represent the training sample, we define a random binary vector V_n^{tr} of size n independent of $\mathcal{D}_n$. V_n^{tr} is called the training vector. We define the test vector by $V_n^{ts} := 1_n - V_n^{tr}$ to represent the test sample.
The distribution of V_n^{tr} characterizes all the cross-validation procedures described in the previous section. Using our notations, we can now define the cross-validation estimator.

Definition 28 (Cross-validation estimator)**.** *With the previous notations, the generalized cross-validation error of ϕ_n denoted by $\widehat{R}_{CV}$ is defined by the conditionnal expectation of $\hat{R}_{V_n^{ts}}(\phi_{V_n^{tr}})$ with respect to the random vector V_n^{tr} given $\mathcal{D}_n$:*

$$\widehat{R}_{CV} := \mathbb{E}_{V_n^{tr}} \hat{R}_{V_n^{ts}}(\phi_{V_n^{tr}}).$$

We will give here some examples of distributions of V_n^{tr} to show that we retrieve cross-validation procedures described previously. Suppose n/k is a integer. The k-fold procedure divides the data into k equally sized folds. It then produces a predictor by training on k-1 folds and testing on the remaining fold. This is repeated for each fold, and the observed errors are averaged to form the k-fold estimate.

Example 29 (k-fold cross-validation)**.**

$$\Pr(V_n^{tr} = (\underbrace{0,\dots,0}_{n/k \text{ observations}}, \underbrace{1,\dots,1}_{n(1-1/k) \text{ observations}})) = \frac{1}{k},$$
$$\Pr(V_n^{tr} = (\underbrace{1,\dots,1}_{n/k \text{ observations}}, \underbrace{0,\dots,0}_{n/k \text{ observations}}, \underbrace{1,\dots,1}_{n(1-2/k) \text{ observations}})) = \frac{1}{k},$$
$$\dots$$
$$\Pr(V_n^{tr} = (\underbrace{1,\dots,1}_{n(1-1/k) \text{ observations}}, \underbrace{0,\dots,0}_{n/k \text{ observations}})) = \frac{1}{k}.$$

We provide another popular example: the leave-one-out cross-validation. In leave-one-out cross-validation, a single sample of size n is used. Each member of the sample in turn is removed, the full modeling method is applied to the remaining $n-1$ members, and the fitted model is applied to the hold-backmember.

Example 30 (leave-one-out cross-validation)**.**

$$\Pr(V_n^{tr} = (0,1,\dots,1)) = \frac{1}{n}$$
$$\Pr(V_n^{tr} = (1,0,1,\dots,1)) = \frac{1}{n}$$
$$\dots$$
$$\Pr(V_n^{tr} = (1,\dots,1,0)) = \frac{1}{n}.$$

We denote by R_{opt} the minimal generalization error attained among the class of predictors $\mathcal{C}$, $R_{opt} = \inf_{\phi\in\mathcal{C}} R(\phi)$. In the sequel, we suppose that ϕ_n is an empirical risk minimizer over the class $\mathcal{C}$. For simplicity, we suppose the infimum is attained i.e. $\phi_n = \arg\min_{\phi\in\mathcal{C}} \widehat{R}_n(\phi)$. Notice that R_{opt} is a parameter of the unknown distribution $\mathbb{P}_{(X,Y)}$ whereas $\widetilde{R}_n$ is a random variable.

At last, recall the definitions of:

Definition 31 (Shatter coefficients)**.** *Let $\mathcal{A}$ be a collection of measurable sets. For $(z_{1,\dots,}z_n) \in \{\mathbb{R}^d\}^n$, let $N_{\mathcal{A}}(z_{1,\dots,}z_n)$ be the number of differents sets in*

$$\{\{z_1,\dots,z_n\} \cap A; A \in \mathcal{A}\}.$$

The n-shatter coefficient of $\mathcal{A}$ is

$$\mathcal{S}(n,\mathcal{A}) = \max_{(z_{1,\dots,}z_n)\in\{\mathbb{R}^d\}^n} N_{\mathcal{A}}(z_{1,\dots,}z_n).$$

That is, the shatter coefficient is the maximal number of different subsets of n points that can be picked out by the class of sets $\mathcal{A}$.

Definition 32 (VC dimension)**.** *Let $\mathcal{A}$ be a collection of sets with $\mathcal{A} \geq 2$. The largest integer $k \geq 1$ for which $\mathcal{S}(k, \mathcal{A}) = 2^k$ is denoted by $V_{\mathcal{C}}$, and it is called the Vapnik-Chernovenkis dimension (or VC dimension) of the class $\mathcal{A}$. If $\mathcal{S}(n, \mathcal{A}) = 2^n$ for all n, then by definition $V_{\mathcal{C}} = \infty$.*

A class of predictors $\mathcal{C}$ is said to have a finite VC-dimension $V_{\mathcal{C}}$ if the dimension of the collection of sets $\{A_{\phi,t} : \phi \in \mathcal{C}, t \in [0,1]\}$ is equal to $V_{\mathcal{C}}$, where $A_{\phi,t} = \{(x,y)/L(y,\phi(x)) > t\}$.

1.4 Results

1.4.1 Hypotheses $\mathcal{H}$

In the sequel, we suppose that the training sample and the test sample are disjoint and that the number of observations in the training sample and in the test sample are respectively $n(1-p_n)$ and np_n. Moreover, we suppose also that the ϕ_n is an empirical risk minimizer on a sample with finite VC-dimension $V_{\mathcal{C}}$ and L a loss function bounded by 1. We also suppose that the predictors are symmetric according to the training sample, i.e. the predictor does not depend on the order of the observations in $\mathcal{D}_n$. Eventually, the cross-validation are symmetric i.e. $\Pr(V_{n,i}^{tr} = 1)$ does not depend on i, this excludes the hold-out cross-validation. **We denote these hypotheses by $\mathcal{H}$.**

We will show upper bounds of the kind $\Pr(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon) \leq B(n, p_n, \varepsilon) + V(n, p_n, \varepsilon)$ with $\varepsilon > 0$. The term $B(n, p_n, \varepsilon)$ is a Vapnik-Chernovenkis-type bound whereas the term $V(n, p_n, \varepsilon)$ is a Hoeffding-like term controlled by the size of the test sample np_n. This bound gives can be interpreted as a quantitative answer to the bias-variance trade-off question. As the percentage of observations in the test sample p_n increases, the $V(n, p_n, \varepsilon)$ term decreases but the $B(n, p_n, \varepsilon)$ term increases. Notice that this bound is worse than the Vapnik-Chernovenkis-type bound and thus can be called a "sanity-check bound" in the spirit of [KR99]. Even though these bounds are valid for almost all the cross-validation procedures, their relevance depends highly on the percentage p_n of elements in the test sample; this is why we first classify them according to p_n. At last, notice that our bounds can be refined using chaining arguments. However, this is not the purpose of this paper.

1.4.2 Cross-validation with large test samples

The first result deals with large test samples, i.e. the bounds are all the better if np_n is large. Note that this result excludes the hold-out cross-validation because it does not make a symmetric use of the data.

Proposition 33 (Large test sample)**.** *Suppose that $\mathcal{H}$ holds. Then, we have for all $\varepsilon > 0$,*

$$\Pr(\widehat{R}_{CV} - \widetilde{R}_n \geq \varepsilon) \leq B(n, p_n, \varepsilon) + V(n, p_n, \varepsilon),$$

with

- $B(n, p_n, \varepsilon) = 4(2n(1-p_n)+1)^{\frac{4V_{\mathcal{C}}}{1-p_n}} \exp(-\frac{n\epsilon^2}{25})$,
- $V(n, p_n, \varepsilon) = \exp(-\frac{2np_n\varepsilon^2}{25})$.

First, we begin with a useful lemma(for the proof, see Appendices)

Lemma 34. *Under the assumption of Proposition 33, we have for all* $\varepsilon > 0$,

$$\Pr\left(\mathbb{E}_{V_n^{tr}} \sup_{\phi\in\mathcal{C}}(\widehat{R}_{V_n^{tr}}(\phi) - R(\phi)) \geq \varepsilon\right) \leq (\mathcal{S}(2n(1-p_n), \mathcal{C}))^{\frac{4}{1-p_n}} e^{-n\varepsilon^2},$$

and symmetrically

$$\Pr\left(\mathbb{E}_{V_n^{tr}} \sup_{\phi\in\mathcal{C}}(R(\phi) - \widehat{R}_{V_n^{tr}}(\phi)) \geq \varepsilon\right) \leq (\mathcal{S}(2n(1-p_n), \mathcal{C}))^{\frac{4}{1-p_n}} e^{-n\varepsilon^2}.$$

Proof of proposition 33.
Recall that ϕ_n is based on empirical risk minimization. Moreover, for simplicity, we have supposed the infimum is attained *i.e.* $\phi_n = \arg\min_{\phi\in\mathcal{C}} \widehat{R}_n(\phi)$. Define $\bar{R}_{n(1-p)} := \mathbb{E}_{V_n^{tr}} R(\phi_{V_n^{tr}})$.

We have by splitting according to $\bar{R}_{n(1-p)}$:

$$\Pr\left(\widehat{R}_{CV} - \widetilde{R}_n \geq 5\varepsilon\right) \leq \underbrace{\Pr\left(\widehat{R}_{CV} - \bar{R}_{n(1-p)} \geq \varepsilon\right)}_{V} + \underbrace{\Pr(\bar{R}_{n(1-p)} - \widetilde{R}_n \geq 4\varepsilon)}_{B}.$$

Notice that $\mathbb{E}_{\mathcal{D}_n}(\widehat{R}_{CV} - \bar{R}_{n(1-p)}) = 0$. Intuitively, V corresponds to the variance term and is controlled in some way by the resampling plan. On the contrary, in the general setting, $\mathbb{E}_{\mathcal{D}_n}(\bar{R}_{n(1-p)} - \widetilde{R}_n) \neq 0$, and B is the bias term and measures the discrepancy between the error rate of size n and of size $n(1-p_n)$.
The first term V can be bounded via Hoeffding's inequality, as follows

$$\begin{aligned} V &= \Pr(\mathbb{E}_{V_n^{tr}}(\widehat{R}_{V_n^{ts}}(\phi_{V_n^{tr}}) - R(\phi_{V_n^{tr}})) \geq \varepsilon) \\ &\leq \inf_{s>0} e^{-s\varepsilon}\mathbb{E}e^{s\mathbb{E}_{V_n^{tr}}(\widehat{R}_{V_n^{ts}}(\phi_{V_n^{tr}}) - R(\phi_{V_n^{tr}}))} \text{ (by Chernoff's bound).} \end{aligned}$$

Then, by Jensen's inequality, we have

$$V \leq \inf_{s>0} e^{-s\varepsilon}\mathbb{E}_{\mathcal{D}_\ltimes}\mathbb{E}_{V_n^{tr}} e^{s(\widehat{R}_{V_n^{ts}}(\phi_{V_n^{tr}}) - R(\phi_{V_n^{tr}}))}.$$

Thus, for $\mathbf{v_n^{tr}}, \mathbf{v_n^{ts}}$ fixed vectors, we have by linearity of expectation and the i.i.d assumption

$$\begin{aligned} V &\leq \inf_{s>0} e^{-s\varepsilon}\mathbb{E}e^{s(\widehat{R}_{\mathbf{v_n^{ts}}}(\phi_{\mathbf{v_n^{tr}}}) - R(\phi_{\mathbf{v_n^{tr}}}))} \\ &\leq \inf_{s>0} e^{-s\varepsilon}\mathbb{E}_{\mathcal{D}_{\mathbf{v}_n^{tr}}}\mathbb{E}(e^{s(\widehat{R}_{\mathbf{v}_n^{ts}}(\phi_{\mathbf{v}_n^{ts}}) - R(\phi_{\mathbf{v}_n^{tr}}))} \mid \mathcal{D}_{\mathbf{v}_n^{tr}}). \end{aligned}$$

Finally, by lemma 1 in [Lug03] since $\mathbb{E}(\widehat{R}_{\mathbf{v}_n^{ts}}(\phi_{\mathbf{v}_n^{ts}}) - R(\phi_{\mathbf{v}_n^{tr}}) \mid \mathcal{D}_{\mathbf{v}_n^{tr}}) = 0$ and the conditional independence:

$$V \leq \inf_{s>0} e^{-s\varepsilon}\mathbb{E}e^{\frac{s^2}{8np_n}} \leq e^{-2np_n\varepsilon^2}.$$

The second term may be treated by introducing the optimal error R_{opt} which should be close to $\widetilde{R}_n$,

$$\begin{aligned} B &= \Pr\left(\bar{R}_{n(1-p)} - \widetilde{R}_n \geq 4\varepsilon\right) \\ &= \Pr\left(\mathbb{E}_{V_n^{tr}}(R(\phi_{V_n^{tr}}) - \widehat{R}_{V_n^{tr}}(\phi_{V_n^{tr}}) + \widehat{R}_{V_n^{tr}}(\phi_{V_n^{tr}}) - R_{opt}) + R_{opt} - \widetilde{R}_n \geq 4\varepsilon\right). \end{aligned}$$

Using the supremum and the fact that $\phi_{V_n^{tr}}$ is an empirical risk minimizer, we obtain:

$$\begin{aligned} B \leq \Pr\Big(&\mathbb{E}_{V_n^{tr}} \sup_{\phi\in\mathcal{C}}(R(\phi) - \widehat{R}_{V_n^{tr}}(\phi)) + \mathbb{E}_{V_n^{tr}} \inf_{\phi\in\mathcal{C}} \widehat{R}_{V_n^{tr}}(\phi) - \inf_{\phi\in\mathcal{C}} R(\phi) \\ &+ R_{opt} - \hat{R}_n + \hat{R}_n - \widetilde{R}_n \geq 4\varepsilon\Big). \end{aligned}$$

Then, since $\inf(A) - \inf(B) \leq \sup(A - B)$ and by definition of ϕ_n, we deduce

$$\begin{aligned} B \leq &\Pr\left(\mathbb{E}_{V_n^{tr}} \sup_{\phi\in\mathcal{C}}(R(\phi) - \widehat{R}_{V_n^{tr}}(\phi)) \geq \varepsilon\right) + \Pr\left(\mathbb{E}_{V_n^{tr}}(\sup_{\phi\in\mathcal{C}}(\widehat{R}_{V_n^{tr}}(\phi) - R(\phi)) \geq \varepsilon\right) \\ &+ \Pr\left(\sup_{\phi\in\mathcal{C}}(R(\phi) - \widehat{R}_n(\phi)) \geq \varepsilon\right) + \Pr\left(\sup_{\phi\in\mathcal{C}}(\widehat{R}_n(\phi) - R(\phi)) \geq \varepsilon\right). \end{aligned}$$

Thus, by Lemma 34, we get

$$B \leq 2(\mathcal{S}(2n(1-p_n), \mathcal{C}))^{\frac{4}{1-p_n}} e^{-n\varepsilon^2} + 2\mathcal{S}(2n, \mathcal{C})^4 e^{-n\varepsilon^2}.$$

Recall the following result (see e.g. [DGL96])

$$\forall n, \mathcal{S}(n, \mathcal{C}) \leq (n+1)^{V_\mathcal{C}}. \tag{1}$$

Thus, we finally obtain

$$\begin{aligned} B &\leq 2(2n(1-p_n)+1)^{\frac{4V_\mathcal{C}}{1-p_n}} e^{-n\varepsilon^2} + 2(2n+1)^{4V_\mathcal{C}} e^{-n\varepsilon^2} \\ &\leq 4(2n(1-p_n)+1)^{\frac{4V_\mathcal{C}}{1-p_n}} e^{-n\varepsilon^2}. \end{aligned}$$

□

Next, we obtain

Proposition 35 (Large test sample)**.** *Suppose that $\mathcal{H}$ holds. Then, we have, for all $\varepsilon > 0$,*

$$\Pr(\widetilde{R}_n - \widehat{R}_{CV} \geq \varepsilon) \leq (2n+1)^{4V} \exp(-n\varepsilon^2).$$

Proof
First, the following lemma holds (for the proof, see appendices),

Lemma 36. *Suppose that $\mathcal{H}$ holds, then we have $\widehat{R}_{CV} \geq \widehat{R}_n$.*

Thus,

$$\Pr\left(\widetilde{R}_n - \widehat{R}_{CV} \geq \varepsilon\right) \leq \Pr\left(\widetilde{R}_n - \widehat{R}_n \geq \varepsilon\right) \leq \mathcal{S}(2n, \mathcal{C})^4 e^{-\varepsilon^2 n} \leq (2n+1)^{4V_{\mathcal{C}}} e^{-n\varepsilon^2}.$$

□
Using the two previous results, we have a concentration inequality for the absolute error $|\widehat{R}_{CV} - \widetilde{R}_n|$,

Corollary 37 (Absolute error for large test sample)**.** *Suppose that $\mathcal{H}$ holds. Then, we have, for all $\varepsilon > 0$,*

$$\Pr(|\widetilde{R}_n - \widehat{R}_{CV}| \geq \varepsilon) \leq B(n, p_n, \varepsilon) + V(n, p_n, \varepsilon),$$

with

- $B(n, p_n, \varepsilon) = 5(2n(1-p_n)+1)^{\frac{4V_{\mathcal{C}}}{1-p_n}} \exp(-\frac{n\varepsilon^2}{25})$,
- $V(n, p_n, \varepsilon) = \exp(-\frac{2np_n\varepsilon^2}{25})$.

With the previous concentration inequality, we can bound from above the expectation of $|\widetilde{R}_n - \widehat{R}_{CV}|$:

Corollary 38 (L_1 error for large test sample)**.** *Suppose that $\mathcal{H}$ holds. Then, we have,*

$$\mathbb{E}|\widehat{R}_{CV} - \widetilde{R}_n| \leq 10\sqrt{\frac{V(\ln(2n(1-p_n)+1)+4)}{n(1-p_n)}} + 5\sqrt{\frac{2}{np_n}}.$$

Proof.
This is a direct consequence of the following lemma:

Lemma 39 ([DGL96])**.** *Let X be a nonnegative random variable. Let K, C nonnegative real such that $C \geq 1$. Suppose that for all $\varepsilon > 0$ $\mathbb{P}(X \geq \varepsilon) \leq C\exp(-K\varepsilon^2)$. Then:*

$$\mathbb{E}X \leq \sqrt{\frac{\ln(C)+2}{K}}.$$

□

1.4.3 Cross-validation with small test samples

The previous bound is not relevant for all small test samples (typically leave-one-out cross-validation) since we are not assured that the variance term converges to 0 (in leave-one-out cross-validation, $V(n, p_n, \varepsilon) = \exp(-2\varepsilon^2/25)$). However, under $\mathcal{H}$, cross-validation with small test samples works also, as stated in the next proposition.

Proposition 40 (Small test sample)**.** *Suppose that $\mathcal{H}$ holds. Then, we have, for all $\varepsilon > 0$,*

$$\Pr(\widehat{R}_{CV} - \widetilde{R}_n \geq \varepsilon) \leq B(n, p_n, \varepsilon) + V(n, p_n, \varepsilon),$$

with

- $B(n, p_n, \varepsilon) = 4(2n(1-p_n)+1)^{\frac{4V_{\mathcal{C}}}{1-p_n}} \exp(-\frac{n\varepsilon^2}{64})$,
- $V(n, p_n, \varepsilon) = \frac{1}{16\varepsilon}\left(\sqrt{\frac{V_{\mathcal{C}}(\ln(2n(1-p_n)+1)+4)}{n(1-p_n)}}\right)$.

For small test samples, we get the same conclusion but the rate of convergence for the term V is slower than for large test samples: typically $O_n\left(\frac{1}{\varepsilon}\sqrt{\frac{\ln(n(1-p_n))}{n(1-p_n)}}\right)$ against $O_n\left(\exp(-np_n\varepsilon^2)/8\right)$.

Proof.
Now, we get by splitting according to $\bar{R}_{n(1-p)}$:

$$\Pr\left(\widehat{R}_{CV} - \widetilde{R}_n \geq 8\varepsilon\right) \leq \underbrace{\Pr\left(\widehat{R}_{CV} - \bar{R}_{n(1-p)} \geq 4\varepsilon\right)}_{V} + \underbrace{\Pr\left(\bar{R}_{n(1-p)} - \widetilde{R}_n \geq 4\varepsilon\right)}_{B}.$$

First, from the proof of proposition 44, we have $B \leq 4(2n(1-p_n)+1)^{\frac{4V_{\mathcal{C}}}{1-p_n}} e^{-n\varepsilon^2}$.

Secondly, notice that $\mathbb{E}(\widehat{R}_{CV} - \bar{R}_{n(1-p)}) = 0$. To control V, we will need the following lemma (for the proof see appendices) which says that if a bounded random variable X is centered and is nonpositive with small probability then it is nonnegative with also small probability.

Lemma 41. *If $|X| \leq 1$ and $\mathbb{E}X = 0$. Then for all $\varepsilon > 0$, we get*

$$\mathbb{P}(X \geq \varepsilon) \leq \frac{\int_0^1 \mathbb{P}(X \leq -x)dx}{\varepsilon}.$$

Moreover, we have since $\widehat{R}_{CV} \geq \widehat{R}_n$ by lemma 36

$$\begin{aligned}\Pr\left(\widehat{R}_{CV} - \bar{R}_{n(1-p)} \leq -4\varepsilon\right) &\leq \Pr\left(\widehat{R}_n - \bar{R}_{n(1-p)} \leq -4\varepsilon\right) \\ &\leq \Pr\left(\widehat{R}_n - \widetilde{R}_n \leq -\varepsilon\right) + \Pr\left(\widetilde{R}_n - \bar{R}_{n(1-p)} \leq -3\varepsilon\right).\end{aligned}$$

Using lemma 34, it follows:

$$\begin{aligned}\Pr\left(\widehat{R}_{CV} - \bar{R}_{n(1-p)} \leq -4\varepsilon\right) &\leq \mathcal{S}(2n, \mathcal{C})^4 e^{-\varepsilon^2 n} + 3\mathcal{S}(2n(1-p_n), \mathcal{C})^{\frac{4V_{\mathcal{C}}}{1-p_n}} e^{-n\varepsilon^2} \\ &\leq 4(2n(1-p_n)+1)^{\frac{4V_{\mathcal{C}}}{1-p_n}} e^{-n\varepsilon^2}.\end{aligned}$$

Applying lemmas 41 and inequality 1 allows to conclude.
□

We have the following complementary but not symmetrical result:

Proposition 42 (Small test sample bis)**.** *Suppose that $\mathcal{H}$ holds. Then, we have for all $\varepsilon > 0$,*

$$\mathbb{P}(\widetilde{R}_n - \widehat{R}_{CV} \geq \varepsilon) \leq (2n+1)^{4V_{\mathcal{C}}} \exp(-n\varepsilon^2).$$

Proof.
We have since $\widehat{R}_{CV} \geq \widehat{R}_n$:

$$\Pr\left(\widetilde{R}_n - \widehat{R}_{CV} \geq \varepsilon\right) \leq \Pr\left(\widetilde{R}_n - \widehat{R}_n \geq \varepsilon\right) \leq \mathcal{S}(2n, \mathcal{C})^4 e^{-\varepsilon^2 n} \leq (2n+1)^{4V_{\mathcal{C}}} e^{-n\varepsilon^2}.$$

□
From this result, we deduce that,

Corollary 43 (Absolute error for small test sample)**.** *Suppose that $\mathcal{H}$ holds. Then, we have for all $\varepsilon > 0$,*

$$\Pr(|\widetilde{R}_n - \widehat{R}_{CV}| \geq \varepsilon) \leq B(n, p_n, \varepsilon) + V(n, p_n, \varepsilon),$$

- $B(n, p_n, \varepsilon) = 5(2n(1-p_n)+1)^{\frac{4V_{\mathcal{C}}}{1-p_n}} \exp(-\frac{n\varepsilon^2}{64})$
- $V(n, p_n, \varepsilon) = \frac{16}{\varepsilon}\sqrt{\frac{V_{\mathcal{C}}(\ln(2n(1-p_n)+1)+4)}{n(1-p_n)}}$.

Eventually, we get

Corollary 44 (L_1 error for small test sample)**.** *Suppose that $\mathcal{H}$ holds. Then, we have:*

$$\mathbb{E}|\widehat{R}_{CV} - \widetilde{R}_n| \leq 16\sqrt{\frac{V_{\mathcal{C}} \ln(2n(1-p_n)+1)+4}{n(1-p_n)}}\left(\ln\left(\sqrt{\frac{n(1-p_n)}{V_{\mathcal{C}}(\ln(2n(1-p_n)+1)+4)}}\right)+2\right).$$

Proof.

We just need lemma 39 and the following simple lemma

Lemma 45. *Let X a nonnegative random variable bounded by 1, $A > 0$ a real such that $\mathbb{P}(X \geq \varepsilon) \leq \frac{A}{\varepsilon}$, for all $\varepsilon > 0$. Then,*

$$\mathbb{E}(X) \leq A(1 - \ln(A))$$

□

Eventually, collecting the previous results, we can summarize the previous results for upper bounds in probability with the following theorem:

Theorem 46 (Absolute error for cross-validation)**.** *Suppose that $\mathcal{H}$ holds. Then, we have for all $\varepsilon > 0$,*

$$\Pr(|\widetilde{R}_n - \widehat{R}_{CV}| \geq \varepsilon) \leq B_{sym}(n, p_n, \varepsilon) + V_{sym}(n, p_n, \varepsilon),$$

with

- $B_{sym}(n, p_n, \varepsilon) = 5(2n(1-p_n)+1)^{\frac{4V_{\mathcal{C}}}{1-p_n}} \exp(-\frac{n\varepsilon^2}{64})$
- $V_{sym}(n, p_n, \varepsilon) = \min\left(\exp(-\frac{2np_n\varepsilon^2}{25}), \frac{16}{\varepsilon}\sqrt{\frac{V_{\mathcal{C}}(\ln(2(1-p_n)+1)+4)}{n(1-p_n)}}\right).$

An interesting consequence of this proposition is that the size of the test is not required to grow to infinity for the consistency of the cross-validation procedure in terms of convergence in probability.

1.4.4 k-fold cross-validation

For k-fold cross-validation, we can simply use the previous bounds together. Thus, we get

Proposition 47 (k-fold)**.** *Suppose that $\mathcal{H}$ holds. Then, we have for all $\varepsilon > 0$,*

$$\Pr(|\widetilde{R}_n - \widehat{R}_{CV}| \geq \varepsilon) \leq B_k(n, p_n, \varepsilon) + V_k(n, p_n, \varepsilon)$$

with

- $B_k(n, p_n, \varepsilon) = 5(2n(1-1/k)+1)^{\frac{4V_{\mathcal{C}}}{1-1/k}} \exp(-\frac{n\varepsilon^2}{64})$
- $V_k(n, p_n, \varepsilon) = \min\left(\exp(-\frac{2n\varepsilon^2}{25k}), \frac{16}{\varepsilon}\sqrt{\frac{V_{\mathcal{C}}(\ln(2(1-1/k)+1)+4)}{n(1-1/k)}}\right).$

Since $k \geq 2$, notice the previous bound can itself be bounded by

$$5(2n+1)^{8V_{\mathcal{C}}} \exp(-\frac{n\epsilon^2}{64}) + \min\left(2\exp(-\frac{2n\varepsilon^2}{25k}), \frac{16}{\varepsilon}\sqrt{\frac{(V_{\mathcal{C}}\ln(2n+1)+4)}{n}}\right).$$

In fact, the bound for the variance term (V) can be improved by averaging the k training errors. This step emphasizes the interest of k-fold cross-validation against simpler cross-validation.

Proposition 48 (k-fold)**.** *Suppose that $\mathcal{H}$ holds. Then, in the case of the k-fold cross-validation procedure, we have for all $\varepsilon > 0$:*

$$\Pr(\widehat{R}_{CV} - \hat{R}_{n(1-p_n)} \geq \varepsilon) \leq 2^{\frac{1}{p_n}} \exp\left(-\frac{n\epsilon^2}{64(\sqrt{V_{\mathcal{C}} \ln(2(2np_n+1))}+2)}\right).$$

Thus, averaging the observed errors to form the k-fold estimate improves the term $V_{\mathcal{C}}$ from

$$\min(2\exp(-\frac{32np_n\varepsilon^2}{49}), \frac{14}{\varepsilon}\sqrt{\frac{V_{\mathcal{C}}(\ln(2(1-p_n)+1)+4)}{n(1-p_n)}}).$$

to $2^{\frac{1}{p_n}} \exp\left(-\frac{n\epsilon^2}{64(\sqrt{V \ln(2(2np_n+1))}+2)}\right)$. This result is important since it shows why intensive use of the data can be very fruitful to improve the estimation rate. Another interesting consequence of this proposition is that, for a fixed precision ε, the size of the test is not required to grow to infinity for the exponential convergence of the cross-validation procedure. For this, it is sufficient that the size of the test sample is larger than a fixed number n_0.

Proof.

Recall that the size of the training sample is $n(1-p_n)$, and the size of the test sample is then np_n. For this proposition, we have $p_n < \frac{1}{2}$
We are interested in the behaviour of $\widehat{R}_{CV} - \bar{R}_{n(1-p)} = \mathbb{E}_{V_n^{tr}} \widehat{R}_{V_n^{ts}}(\phi_{V_n^{tr}}) - \mathbb{E}_{V_n^{tr}} R(\phi_{V_n^{tr}})$ which is a sum of $\frac{1}{p_n} = k$ terms in the case of the k-fold cross-validation.
The difficulty is that these terms are neither independent, nor even exchangeable. We have in mind to apply the results about the sum of independent random variables. For this, we need a way to introduce independence in our samples. In the same time, we do not want to lose too much information. For this, we will introduce independence by using by using the supremum. We have,

$$\begin{aligned}\Pr(\widehat{R}_{CV} - \bar{R}_{n(1-p)} \geq \varepsilon) &= \Pr(\mathbb{E}_{V_n^{tr}}(\widehat{R}_{V_n^{ts}}(\phi_{V_n^{tr}}) - R(\phi_{V_n^{tr}})) \geq \varepsilon) \\ &\leq \Pr(\mathbb{E}_{V_n^{tr}}(\sup_{\phi\in\mathcal{C}}(\widehat{R}_{V_n^{ts}}(\phi) - R(\phi)) \geq \epsilon).\end{aligned}$$

Now, we have a sum of $k = \frac{1}{p_n}$ i.i.d terms: $\mathbb{P}(\frac{1}{k}\sum Y_i \geq \epsilon)$, with $Y_i = \sup_{\phi\in\mathcal{C}}(\widehat{R}_{V_n^{ts}}(\phi) - R(\phi))$. However, we have an extra piece of information: an upper bound for the tail probability of these variables, using the concentration inequality due to [Vap98].

$$\Pr(\sup_{\phi\in\mathcal{C}}(\widehat{R}_{V_n^{ts}}(\phi) - R(\phi)) \geq \epsilon) \leq c(np_n, V_{\mathcal{C}}) e^{-\frac{\epsilon^2}{2\sigma(np_n)^2}}.$$

with $c(n, V_{\mathcal{C}}) = 2\mathcal{S}(2n, \mathcal{C}\ ) \leq 2(2n+1)^{V_{\mathcal{C}}}$ and $\sigma(n)^2 = \frac{4}{n}$.
In fact, summing independent bounded variables with exponentially small tail probability gives us a better concentration inequality than the simple sum of independent bounded variables.
To show this, we proceed in three steps:

1. the q-Hölder norms of each variable is uniformly bounded by $\sqrt{q}$,
2. the Laplace transform of Y_i is smaller than the Laplace transform of some particular normal variable,
3. using Chernoff's method, we obtain a sharp concentration inequality.

1. First step (for the proof, see appendices), we prove

   Lemma 49. *Let Y a random variable (bounded by 1) with subgaussian tail probability $\mathbb{P}(Y \geq \varepsilon) \leq ce^{-\frac{\epsilon^2}{2\sigma^2}}$ for all $\varepsilon > 0$ with $\sigma^2 > 0$ and $c \geq 2$. Then, there exists a constant γ such that, for every integer q,*

   $$(\mathbb{E}{Y_+}^q)^{\frac{1}{q}} \leq \sqrt{\gamma q},$$

   with $\gamma = (\sigma\sqrt{4\ln(c)} + \pi^{\frac{1}{4}}3^{\frac{1}{3}}2e^{-\frac{1}{2}}\sigma)^2$.

2. Second step (see exercise 4 in [Lug03]), we have

   Lemma 50. *If there exists a constant γ, such that for every integer q*

   $$(\mathbb{E}{Y_+}^q)^{\frac{1}{q}} \leq \sqrt{\gamma q}.$$

   then we have

   $$\mathbb{E}(e^{sY}) \leq \sqrt{2}e^{\frac{1}{6}}e^{\frac{s^2e\gamma}{2}}.$$

3. Third step, we have the result using Chernoff's method.

   Lemma 51. *If, for some $\alpha > 0$, $\beta > 0$, we have:*

   $$\mathbb{E}(e^{sY}) \leq \alpha e^{\frac{s^2\beta^2}{2}}$$

   then if $(Y_i)_{1\leq i\leq n}$ are i.i.d., we have:

   $$\mathbb{P}(\frac{1}{V}\sum_{i=1}^{V} Y_i > \epsilon) \leq \alpha^V e^{\frac{-V\epsilon^2}{2\beta^2}}$$

Putting lemma 49 50 51 together, we eventually get:

$$\mathbb{P}(\mathbb{E}_{V_n^{tr}}(\sup_{\phi\in\mathcal{C}} \widehat{R}_{V_n^{tr}}(\phi) - R(\phi)) \geq \varepsilon) \leq (\sqrt{2}e^{1/6})^{\frac{1}{p_n}} \exp\left(\frac{-\frac{1}{p_n}\epsilon^2}{2\sigma(np_n)^2(e^{\frac{1}{2}}\sqrt{4\ln(c(np_n, V_{\mathcal{C}}))} + \pi^{\frac{1}{4}}3^{\frac{1}{3}}2)^2}\right).$$

□

Symmetrically, we obtain:

Proposition 52 (k-fold bis)**.** *Suppose that $\mathcal{H}$ holds. Then, in the case of the k-fold cross-validation procedure, we have for all $\varepsilon > 0$*

$$\mathbb{P}(\hat{R}_{n(1-p_n)} - \widehat{R}_{CV} \geq \varepsilon) \leq 2^{\frac{1}{p_n}} \exp\left(-\frac{n\epsilon^2}{64(\sqrt{V_{\mathcal{C}} \ln(2(2np_n+1))} + 2)} \right).$$

Eventually, we have a control on the absolute deviation

Theorem 53 (Absolute error for the k-fold)**.** *Suppose that $\mathcal{H}$ holds. Then, in the case of the k-fold cross-validation procedure, we have for all $\varepsilon > 0$,*

$$\Pr(|\widetilde{R}_n - \widehat{R}_{CV}| \geq \varepsilon) \leq B_k(n, p_n, \varepsilon) + V_k(n, p_n, \varepsilon)$$

with

- $B_k(n, p_n, \varepsilon) = 5(2n(1-1/k)+1)^{\frac{4V_{\mathcal{C}}}{1-1/k}} \exp(-\frac{n\varepsilon^2}{64})$
- $V_k(n, p_n, \varepsilon) =$

$$\min(\exp(-\frac{2n/\varepsilon^2}{25k}), \frac{16}{\varepsilon}\sqrt{\frac{V_{\mathcal{C}}(\ln(2(1-1/k)+1)+4)}{n(1-1/k)}},$$
$$22^{\frac{1}{p_n}} \exp(-\frac{n\epsilon^2}{25 * 64(\sqrt{V_{\mathcal{C}} \ln(2(2np_n+1))} + 2)})).$$

1.4.5 Hold-out cross-validation

For hold-out cross-validation, the symmetric condition that for all i, $\Pr\left(i \in \mathcal{D}_{V_n^{tr}}\right)$ is independent of i is no longer valid. Indeed, in the hold-out cross-validation (or split sample), there is no crossing again.

In the next proposition, we suppose that the training sample and the test sample are disjoint and that the number of observations in the learning sample and in the test sample are still respectively $n(1-p_n)$ and np_n. Moreover, we suppose also that the predictors ϕ_n are empirical risk minimizers on a class $\mathcal{C}$ with finite $V_{\mathcal{C}}$-dimension $V_{\mathcal{C}}$ and L a loss function bounded by 1. **We denote these hypotheses by $\mathcal{G}$.**

We get the following result

Theorem 54 (Hold-out)**.** *Suppose that $\mathcal{G}$ holds. Then, we have for all $\varepsilon > 0$,*

$$\Pr(|\widetilde{R}_n - \widehat{R}_{CV}| \geq \varepsilon) \leq B_{hold}(n, p_n, \varepsilon) + V_{hold}(n, p_n, \varepsilon)$$

with

- $B_{hold}(n, p_n, \varepsilon) = 8(2n(1-p_n)+1)^{4V_{\mathcal{C}}} \exp(-\frac{2n(1-p_n)\varepsilon^2}{25})$

- $V_{hold}(n, p_n, \varepsilon) = 2\exp(-\frac{2np_n\varepsilon^2}{25})$.

Proof. We just have to follow the same steps as in proposition 102. But in the case of hold-out cross-validation, notice that

$$\begin{aligned}\Pr\left(\mathbb{E}_{V_n^{tr}} \sup_{\phi\in\mathcal{C}}(\widehat{R}_{V_n^{tr}}(\phi) - R(\phi)) \geq \varepsilon\right) &= \Pr(\sup_{\phi\in\mathcal{C}}(\widehat{R}_{\mathbf{v}_n^{tr}}(\phi) - R(\phi)) \geq \varepsilon) \\ &\leq \mathcal{S}(2n(1-p_n), \mathcal{C})^4 e^{-n(1-p_n)\varepsilon^2}\end{aligned}$$

Moreover, the lemma 41 is no longer valid, since $\mathbb{E}_{V_n^{tr}} R_{V_n^{ts}}(\phi_n) \neq \widehat{R}_n$. □

1.4.6 Discussion

We base the next discussion on upperbounds, so the following heuristic arguments are questionable if the bounds are loose.

Crossing versus non-crossing

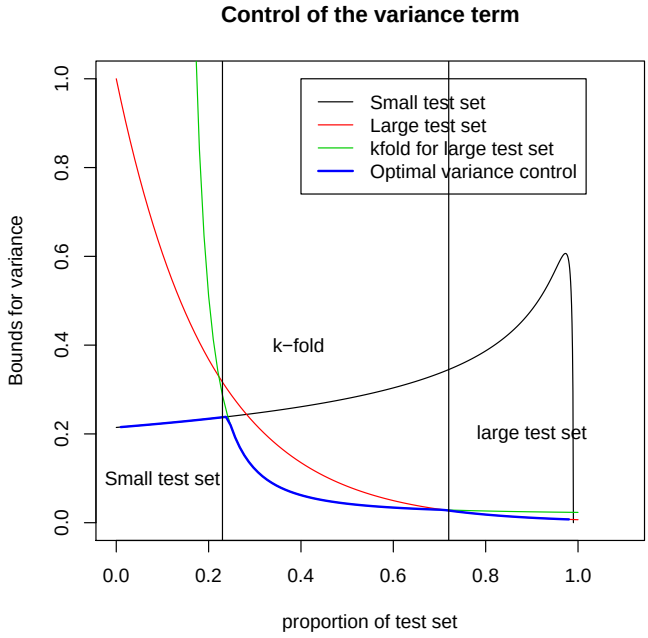
One can wonder: what is the use of averaging again over the different folds of the k-fold cross-validation, which is time consuming? As far as the expected errors are concerned, the upper bounds are the same for crossing cross-validation procedures and for hold-out cross-validation. But suppose we are given a level of precision ε, and we want to find an interval of length 2ε with maximal confidence. Then notice that $B_{sym}/B_{hold} = (2n(1-p_n)+1)^{\frac{4V_{\mathcal{C}}p_n}{1-p_n}}\exp(-np_n\varepsilon^2)$. Thus if p_n is constant, $B_{sym}/B_{hold} \to_{n\to\infty} 0$: the term B will be much greater for hold-out based on large learning size. On the contrary, if the learning size is small, then the term B is smaller for non crossing procedure for a given p_n. This might due to the absence of resampling.
Regarding the variance term $V_{hold}(n, p_n, \varepsilon)$, we need the size of the test sample to grow to infinity for the consistency of the hold-out cross-validation. On the contrary, for crossing cross-validation, the term V converges to 0 whatever the size of the test is.

k-fold cross-validation versus others

If we consider the L_1 error, the upper bounds are the same for crossing cross-validation procedures and for other cross-validation procedures. But if we look for the interval of length 2ε with maximal confidence, then notice that $V_k/V_{sym} \to_{n\to\infty} 0$ (with V_k, V_{sym} defined respectively in theorems 53, 102) if the number of elements in the training sample np_n is constant and large enough. Thus, if the learning size is large enough, the V term is much smaller for the k-fold cross-validation, thanks to the crossing.

Estimation curve

The expression of the variance term V depends on the percentage of observations p_n in the test sample and on the type of cross-validation procedure. We have thus a control of the variance term depending on p_n.



We can define the estimation curve (in probability or in L_1 norm) which gives for each cross-validation procedure and for each p_n the estimation error.

Definition 55 (Estimation curve in probability)**.** *Let $\varepsilon > 0$:*

$$\mathcal{AC} : p_n \mapsto B(n, p_n, \varepsilon) + V(n, p_n, \varepsilon).$$

with $B(n, p_n, \varepsilon)$ and $V(n, p_n, \varepsilon)$ defined in theorem 102.

This can be done with the expectation of the absolute of deviation or with the probability upper bound if the level of precision is ε.

Definition 56 (Estimation curve in L_1 norm)**.**

$$\mathcal{AC} : p_n \mapsto B(n, p_n) + V(n, p_n).$$

with $B(n, p_n)$ and $V(n, p_n)$ defined as in proposition 38.

We say that the estimation curve in probability experiences a phase transition when the convergence rate $V(n, p_n, \varepsilon)$ changes. The estimation curve experiences at least one transition phase. The transition phases just depend on the class of predictors and on the sample

size. On the contrary of the learning curve, the transition phases of the estimation curve are independent of the underlying distribution. The different transition phases define three different regions in the values of p_n the percentage of observations in the test sample. This three regions emphasize the different roles played by small test sample cross-validation, large test samples cross-validation and k-fold cross-validation.

Optimal splitting and confidence intervals

The estimation curve gives a hint for this simple but important question: how should one choose the cross-validation procedure in order to get the best estimation rate? How should one choose k in the k-fold cross-validation? The quantitative answer of theses questions is the arg min of the estimation curve $\mathcal{AC}$.
That is in probability

$$p_n^\star(\varepsilon) = \arg\min_{p_n} \mathcal{AC}(p_n, \varepsilon).$$

or in L_1 norm:

$$p_n^\star = \arg\min_{p_n} \mathcal{AC}(p_n).$$

As far as the L_1 norm is concerned, we can derive a simple expression for the choice of p_n. Indeed, if we use chaining arguments in the proof of proposition 34, that is: there exists a universal constant $c > 0$ such that $\mathbb{E}\sup_{\phi\in\mathcal{C}}(\widehat{R}_{\mathbf{W}_n^{tr}}(\phi) - R(\phi)) \leq c\sqrt{\frac{V_\mathcal{C}}{n(1-p_n)}}$ (for the proof, see e.g. [DGL96]). The proposition 38 thus becomes:

Corollary 57 (L_1 error for large test sample)**.** *Suppose that $\mathcal{H}$ holds. Then, there exists a universal constant $c > 0$ such that:*

$$\mathbb{E}|\widehat{R}_{CV} - \widetilde{R}_n| \leq c\sqrt{\frac{V_\mathcal{C}}{n(1-p_n)}} + 2\sqrt{\frac{6}{np_n}}.$$

We can then minimize the last expression in p_n. After derivation, we obtain $p_n^\star = ((\frac{c^2 V_\mathcal{C}}{2\sqrt{(6)}})^{1/3} + 1)^{-1}$. Thus, the larger the VC-dimension is, the larger the training sample should be. Since it may be difficult to find an explicit constant, one may try to solve: $\sqrt{\frac{V_\mathcal{C}(\ln(2n)+4))}{n(1-p_n)}} + 2\sqrt{\frac{6}{np_n}}$. We obtain then a computable rule $p_n^\star = ((\frac{V_\mathcal{C}(\ln(2n)+4))}{2\sqrt{(6)}})^{1/3} + 1)^{-1}$

Another interesting issue is: knowing the number of observations n and the class of predictors, we can now derive an optimal minimal $1-\alpha$-confidence interval, together with the cross-validation procedure. We look at the values (ε, p_n) such that the upperbound $B(n, p_n, \varepsilon) + V(n, p_n, \varepsilon)$ is below the threshold α. Then, we select the couple (ε^*, p_n^*) among those values for which ε is minimal. On figure 1.1, we fix a choice of $\alpha = 5\%$. We observe that, for values of n between 1000 and 10000 and for small VC-dimension, a choice of $p \simeq 10\%$, i.e. the ten-fold cross-validation, seems to be a reasonable choice.

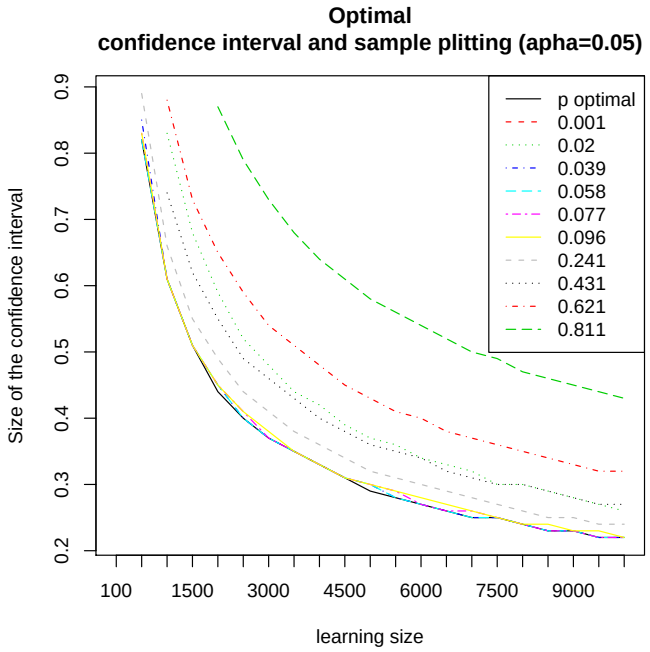


Figure 1.1: Upperbounds for cross-validation procedures with different splitting

1.5 Appendices

1.5.1 Notations and definitions

We recall the main notations and definitions.

Name	Notation	Definition
Generalisation error	$\widetilde{R}_n$	$\mathbb{E}_P[L(Y, \phi(X, \mathcal{D}_n)) \mid \mathcal{D}_n]$
Resubstitution estimate	$\widehat{R}_n$	$\frac{1}{n}\sum_{i=1}^n L(Y_i, \phi_n(X_i, \mathcal{D}_n))$
Cross-validation estimate	$\widehat{R}_{CV}$	$\mathbb{E}_{V_n^{tr}} \hat{R}_{V_n^{ts}}(\phi_{V_n^{tr}})$
Cross-validation risk	$\bar{R}_{n(1-p)}$	$\mathbb{E}_{V_n^{tr}} R(\phi_{V_n^{tr}})$
Optimal error	R_{opt}	$\inf_{\phi \in \mathcal{C}} R(\phi)$

Table 1.1: Main notations

1.5.2 Proofs

We recall three very useful results. The first one, due to [HOEF63], bounds the difference between the empirical mean and the expected value. The second one, due to [VC71] , bounds the supremum over the class of predictors of the difference between the training

error and the generalization error. The last one is called the bounded differences inequality [McD89] .

Theorem 58 ([HOEF63]). *Let $X_{1,...,}X_n$ independent random variables in $[a_i, b_i]$. Then for all $\varepsilon > 0$,*

$$\mathbb{P}(\sum X_i - \mathbb{E}(\sum X_i) \geq n\varepsilon) \leq e^{-\frac{2\varepsilon^2}{\sum_i (b_i - a_i)^2}}$$

Theorem 59 ([VC71]). *Let $\mathcal{C}$ a class of predictors with finite VC-dimension and L a loss function bounded by 1. Then for all $\varepsilon > 0$,*

$$\mathbb{P}(\sup_{\phi \in \mathcal{C}}(\widehat{R}_n(\phi) - L(\phi)) \geq \varepsilon) \leq c(n, V_{\mathcal{C}}) e^{-\frac{\varepsilon^2}{2\sigma(n)^2}}$$

with $c(n, V_{\mathcal{C}}) = 2\mathcal{S}(2n, \mathcal{C}) \leq 2(2n+1)^{V_{\mathcal{C}}}$ and if $n \geq V_{\mathcal{C}}$, $2\mathcal{S}(2n, \mathcal{C}) \leq 2(\frac{2ne}{V_{\mathcal{C}}})^{V_{\mathcal{C}}}$ and $\sigma(n)^2 = \frac{4}{n}$

Theorem 60 (McDiarmid). *Let $X_{1,...,}X_n$ be independent random variables taking values in a sample $\mathcal{X}$, and assume that $f : \mathcal{X}^n \to \mathcal{R}$ satisfies*

$$\forall i, \sup_{\substack{x_1,...,x_i,...,x_n \\ x_i'}} |f(x_1, ..., x_n) - f(x_1, ..., x_{i'}, ..., x_n)| \leq c_i.$$

Then, for all $\varepsilon > 0$,

$$\mathbb{P}(f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n) \geq \varepsilon) \leq e^{-\frac{2\varepsilon^2}{\sum_i c_i^2}}.$$

Proof of lemma 34

First, notice that

$$\mathbb{P}(\mathbb{E}_{V_n^{tr}} \sup_{\phi \in \mathcal{C}}(\widehat{R}_{V_n^{tr}}(\phi) - R(\phi)) - \mathbb{E}\mathbb{E}_{V_n^{tr}} \sup_{\phi \in \mathcal{C}}(\widehat{R}_{V_n^{tr}}(\phi) - R(\phi)) \geq \varepsilon) \leq e^{-2n\varepsilon^2},$$

using McDiarmid's inequality by setting $f(X_1, \ldots, X_n) = \mathbb{E}_{V_n^{tr}} \sup_{\phi \in \mathcal{C}}(\widehat{R}_{V_n^{tr}}(\phi) - R(\phi))$ and since for all i,

$$\sup_{\substack{x_1,\dots,x_i,\dots,x_n \\ x_i'}} |\mathbb{E}_{V_n^{tr}} \sup_{\phi\in\mathcal{C}}(\widehat{R}_{V_n^{tr}}(\phi) - R(\phi)) - \mathbb{E}_{V_n^{tr}} \sup_{\phi\in\mathcal{C}}(\widehat{R}'_{V_n^{tr}}(\phi) - R(\phi))|$$

$$= \sup_{\substack{x_1,\dots,x_i,\dots,x_n \\ x_i'}} \left| \mathbb{E}_{V_n^{tr}} \left[\sup_{\phi\in\mathcal{C}}(\widehat{R}_{V_n^{tr}}(\phi) - R(\phi)) - \sup_{\phi\in\mathcal{C}}(\widehat{R}'_{V_n^{tr}}(\phi) - R(\phi)) \right] \right|$$

$$\leq \sup_{\substack{x_1,\dots,x_i,\dots,x_n \\ x_i'}} \mathbb{E}_{V_n^{tr}} \left| \sup_{\phi\in\mathcal{C}}(\widehat{R}_{V_n^{tr}}(\phi) - R(\phi)) - \sup_{\phi\in\mathcal{C}}(\widehat{R}'_{V_n^{tr}}(\phi) - R(\phi)) \right|$$

by Jensen's inequality

$$\leq \sup_{\substack{x_1,\dots,x_i,\dots,x_n \\ x_i'}} \mathbb{E}_{V_n^{tr}} \sup_{\phi\in\mathcal{C}} |\widehat{R}_{V_n^{tr}}(\phi) - \widehat{R}'_{V_n^{tr}}(\phi)|$$

since $|\sup f - \sup g| \leq \sup |f - g|$

$$\leq \frac{1}{n}.$$

Indeed, if we note Q the number of elements in the sum $\mathbb{E}_{V_n^{tr}}$, the number of changes is lower than $\leq \frac{1}{Q}(\frac{1}{n(1-p_n)}$multiplied by the number of times i' in the learning sample) that is $\frac{1}{Q}(\frac{1}{n(1-p_n)}Q(1-p_n)) = \frac{1}{n}$
Furthermore, we have

$$\begin{aligned}\mathbb{E}\mathbb{E}_{V_n^{tr}} \sup_{\phi\in\mathcal{C}}(\widehat{R}_{V_n^{tr}}(\phi) - R(\phi)) &= \mathbb{E} \sup_{\phi\in\mathcal{C}}(\widehat{R}_{\mathbf{v}_n^{tr}}(\phi) - R(\phi)) \\ &\quad \text{with } \mathbf{v}_n^{tr} \text{ a fixed vector} \\ &\leq \sqrt{\frac{2\ln(\mathcal{S}(2n(1-p_n),\mathcal{C})}{n(1-p_n)}}.\end{aligned}$$

by Vapnik-Chernovenkis's inequality.

Thus, if we denote $\Pr\left(\mathbb{E}_{V_n^{tr}} \sup_{\phi\in\mathcal{C}}(\widehat{R}_{V_n^{tr}}(\phi) - R(\phi)) \geq \varepsilon\right)$ by P_1 it leads to

$$\begin{aligned}P_1 = \Pr\big(&\mathbb{E}_{V_n^{tr}} \sup_{\phi\in\mathcal{C}}(\widehat{R}_{V_n^{tr}}(\phi) - R(\phi)) - \mathbb{E}\mathbb{E}_{V_n^{tr}} \sup_{\phi\in\mathcal{C}}(\widehat{R}_{V_n^{tr}}(\phi) - R(\phi)) \\ &\geq \varepsilon - \mathbb{E}\mathbb{E}_{V_n^{tr}} \sup_{\phi\in\mathcal{C}}(\widehat{R}_{V_n^{tr}}(\phi) - R(\phi)).\end{aligned}$$

Then, using the two previous inequalities

$$\begin{aligned}P_1 \leq \Pr\big(&\mathbb{E}_{V_n^{tr}} \sup_{\phi\in\mathcal{C}}(\widehat{R}_{V_n^{tr}}(\phi) - R(\phi)) - \mathbb{E}\mathbb{E}_{V_n^{tr}} \sup_{\phi\in\mathcal{C}}(\widehat{R}_{V_n^{tr}}(\phi) - R(\phi)) \\ &\geq \varepsilon - \sqrt{\frac{2\ln(\mathcal{S}(2n(1-p_n),\mathcal{C})}{n(1-p_n)}}).\end{aligned}$$

Since $2(u-v)^2 \geq u^2 - 2v^2$, it follows

$$\begin{aligned} P_1 \leq & \exp(-2n(\varepsilon - \sqrt{\frac{2\ln(\mathcal{S}(2n(1-p_n),\mathcal{C})}{n(1-p_n)}})^2) \leq \exp(-n(\varepsilon^2 - \frac{4\ln(\mathcal{S}(2n(1-p_n),\mathcal{C})}{n(1-p_n)}))) \\ \leq & \mathcal{S}(2n(1-p_n),\mathcal{C})^{4/(1-p_n)} \exp(-n\varepsilon^2). \end{aligned}$$

□

Proof of lemma 34

Recall that $\widehat{R}_{CV} = \mathbb{E}_{V_n^{tr}} \widehat{R}_{V_n^{ts}}(\phi_{V_n^{tr}})$

But by definition of ϕ_n, we have $\widehat{R}_n(\phi_n) \leq \widehat{R}_n(\phi_{V_n^{tr}})$.

It follows that $\frac{1}{n}(np_n \widehat{R}_{V_n^{ts}}(\phi_n) + \sum_{i \in V_n^{tr}} L(Y_i, \phi_n(X_i)) \leq \frac{1}{n}(np_n \widehat{R}_{V_n^{ts}}(\phi_{V_n^{tr}}) \quad + \sum_{i \in V_n^{tr}} L(Y_i, \phi_{V_n^{tr}}(X_i))$.

Thus, since $\sum_{i \in V_n^{tr}} L(Y_i, \phi_n(X_i)) \geq \sum_{i \in V_n^{tr}} L(Y_i, \phi_{V_n^{tr}}(X_i))$ by definition of $\phi_{V_n^{tr}}$, we have $\widehat{R}_{V_n^{ts}}(\phi_n) \leq \widehat{R}_{V_n^{ts}}(\phi_{V_n^{tr}})$.

From this, we deduce $\widehat{R}_{CV} = \mathbb{E}_{V_n^{tr}} \widehat{R}_{V_n^{ts}}(\phi_{V_n^{tr}}) \geq \mathbb{E}_{V_n^{tr}} \widehat{R}_{V_n^{ts}}(\phi_n) = \widehat{R}_n$.

□

Proof. of lemma 41

$$\forall \varepsilon > 0, \mathbb{P}(X \geq \varepsilon) \leq \mathbb{P}(X_+ \geq \varepsilon) \leq \frac{\mathbb{E}X_+}{\varepsilon} = \frac{\mathbb{E}X_-}{\varepsilon} = \frac{\int_0^1 \mathbb{P}(X_- \geq x)dx}{\varepsilon} = \frac{\int_0^1 \mathbb{P}(X \leq -x)dx}{\varepsilon}.$$

□

Proof. of lemma 49

First, suppose that $q > 1$ and notice that

$$\begin{aligned} \mathbb{E}{Y_+}^q &= \int_0^\infty q y^{q-1} \mathbb{P}(Y_+ > y)dy \\ &= q \int_0^\infty y^{q-1} \mathbb{P}(Y > y)dy. \end{aligned}$$

We thus deduce that because of the subgaussian inequality:

$$\mathbb{E}{Y_+}^q \leq q \int_0^{\sigma\sqrt{4\ln(c)}} y^{q-1}dy + q \int_{\sigma\sqrt{4\ln(c)}}^\infty c y^{q-1} e^{-\frac{y^2}{2\sigma^2}} dy.$$

Then, with $\mathcal{N}$ a standard normal:

$$\begin{aligned} \mathbb{E}{Y_+}^q &\leq (\sigma\sqrt{4\ln(c)})^q + qc \int_{\sigma\sqrt{4\ln(c)}}^\infty y^{q-1} e^{-\frac{y^2}{2\sigma^2}} dy \\ &\leq (\sigma\sqrt{4\ln(c)})^q + qc\sqrt{2\pi}\sigma \mathbb{E}((\sigma\mathcal{N})^{q-1} 1_{(\sigma\sqrt{4\ln(c)} \leq \sigma\mathcal{N})}). \end{aligned}$$

This gives by Cauchy-Schwarz's inequality:

$$\mathbb{E}{Y_+}^q \leq (\sigma\sqrt{4\ln(c)})^q + qc\sqrt{2\pi}\sigma^q(\mathbb{E}\mathcal{N}^{2(q-1)}1_{0\leq\mathcal{N}})^{\frac{1}{2}}(\mathbb{P}(\sqrt{4\ln(c)}\leq\mathcal{N}))^{\frac{1}{2}}.$$

It leads to, since $\mathbb{E}\mathcal{N}^{2p}=\frac{(2p)!}{2^p p!}$, and $\sqrt{4\ln(c)}\geq 1$,

$$\begin{aligned}\mathbb{E}{Y_+}^q &\leq (\sigma\sqrt{4\ln(c)})^q + (2\pi)^{1/4}qc\sigma^q(\mathbb{E}\mathcal{N}^{2(q-1)})^{\frac{1}{2}}(e^{-\frac{(2)\ln(c)}{2}})^{\frac{1}{2}}\\ &\leq (\sigma\sqrt{4\ln(c)})^q + (2\pi)^{1/4}q\sigma^q(\tfrac{(2(q-1))!}{2^{(q-1)}(q-1)!})^{\frac{1}{2}}.\end{aligned}$$

We obtain, since $\sqrt{2\pi n}(\frac{n}{e})^n e^{\frac{1}{12n}} \leq n! \leq \sqrt{2\pi n}(\frac{n}{e})^n e^{\frac{1}{12n+1}}$,

$$\begin{aligned}\mathbb{E}{Y_+}^q &\leq (\sigma\sqrt{4\ln(c)})^q + (2\pi)^{1/4}q\sigma^q\left(\frac{\sqrt{2\pi 2(q-1)}(\frac{2(q-1)}{e})^{2(q-1)}e^{\frac{1}{24(q-1)+1}}}{2^{(q-1)}\sqrt{2\pi(q-1)}(\frac{(q-1)}{e})^{k(q-1)}e^{\frac{1}{12(q-1)}}}\right)^{\frac{1}{2}}\\ &\leq (\sigma\sqrt{4\ln(c)})^q + (2\pi)^{1/4}q\sigma^q(\sqrt{2}(\tfrac{2(q-1)}{e})^{(q-1)}e^{\frac{1}{24(q-1)+1}-\frac{1}{12(q-1)}})^{\frac{1}{2}}\\ &\leq (\sigma\sqrt{4\ln(c)})^q + (2\pi)^{1/4}q2^{\frac{1}{4}}\sigma^q(\tfrac{2(q-1)}{e})^{\frac{q-1}{2}}.\end{aligned}$$

Thus, since $(a+b)^{\frac{1}{q}}\leq a^{\frac{1}{q}}+b^{\frac{1}{q}}, a,b\geq 0$:

$$\begin{aligned}(\mathbb{E}{Y_+}^q)^{\frac{1}{q}} &\leq \left((\sigma\sqrt{4\ln(c)})^q + (2\pi)^{1/4}q2^{\frac{1}{4}}\sigma^q(\tfrac{2(q-1)}{e})^{\frac{q-1}{2}}\right)^{\frac{1}{q}}\\ &\leq \left((\sigma\sqrt{4\ln(c)})^q\right)^{\frac{1}{q}} + \left((2\pi)^{1/4}q2^{\frac{1}{4}}\sigma^q(\tfrac{2(q-1)}{e})^{\frac{q-1}{2}}\right)^{\frac{1}{q}},\end{aligned}$$

which gives since $q^{\frac{1}{q}}\leq 3^{\frac{1}{3}}, (\frac{2(q-1)}{e})^{\frac{q-1}{2q}} \leq (\frac{2q}{e})^{\frac{q-1}{2q}} \leq (\frac{2q}{e})^{\frac{q}{2q}}$ since $\frac{2q}{e}\geq 1$:

$$\begin{aligned}(\mathbb{E}{Y_+}^q)^{\frac{1}{q}} &\leq \sigma\sqrt{4\ln(c)} + q^{\frac{1}{q}}((2\pi)^{1/4}2^{\frac{1}{4}})^{\frac{1}{q}}\sigma(\tfrac{2(q-1)}{e})^{\frac{q-1}{2q}}\\ &\leq \sigma\sqrt{4\ln(c)} + 3^{\frac{1}{3}}2^{\frac{1}{4}}\sigma(\tfrac{2q}{e})^{\frac{1}{2}}.\end{aligned}$$

$$\begin{aligned}&\leq \sigma\sqrt{4\ln(c)} + (2\pi)^{1/4}3^{\frac{1}{3}}2^{\frac{3}{4}}e^{-\frac{1}{2}}\sigma\sqrt{q}\\ &\leq (\sigma\sqrt{4\ln(c)} + (2\pi)^{1/4}3^{\frac{1}{3}}2^{\frac{3}{4}}e^{-\frac{1}{2}}\sigma)\sqrt{q}\\ &\leq \sqrt{\gamma q}.\end{aligned}$$

with $\gamma = (\sigma\sqrt{4\ln(c)} + (2\pi)^{1/4}3^{\frac{1}{3}}2^{\frac{3}{4}}e^{-\frac{1}{2}}\sigma)^2$
For $q=1$, notice that:

$$\begin{aligned}(\mathbb{E}{Y_+}^q)^{\frac{1}{q}} &\leq \sigma\sqrt{4\ln(c)} + \tfrac{1}{2}\sigma\\ &\leq \sqrt{\gamma q}.\end{aligned}$$

□

Chapter 2

Concentration inequalities of the cross-validation estimator of stable predictors

In this article, we derive concentration inequalities for the cross-validation estimate of the generalization error for stable predictors in the context of risk assessment. The notion of stability has been first introduced by [DEWA79] and extended by [KEA95] , [BE01] and [KUNIY02] to characterize class of predictors with infinite VC dimension. In particular, this covers k-nearest neighbors rules, bayesian algorithm ([KEA95]), boosting,... General loss functions and class of predictors are considered. We use the formalism introduced by [DUD03] to cover a large variety of cross-validation procedures including leave-one-out cross-validation, k-fold cross-validation, hold-out cross-validation (or split sample), and the leave-v-out cross-validation.

In particular, we give a simple rule on how to choose the cross-validation, depending on the stability of the class of predictors. In the special case of uniform stability, an interesting consequence is that the number of elements in the test set is not required to grow to infinity for the consistency of the cross-validation procedure. In this special case, the particular interest of leave-one-out cross-validation is emphasized.

Keywords: Cross-validation, stability, generalization error, concentration inequality, optimal splitting, resampling.

2.1 Introduction and motivation

One of the main issue of pattern recognition is to create a predictor (a regressor or a classifier) which takes observable inputs in order to predict the unknown nature of an output. Formally, a predictor φ is a measurable map from some measurable space $\mathcal{X}$ to some measurable space $\mathcal{Y}$. When $\mathcal{Y}$ is a countable set (respectively $\mathbb{R}^m$), the predictor is called a classifier (respectively a regressor). The strategy of *Machine Learning* consists in building a learning algorithm Φ from both a set of examples and a class of methods. Typical class of methods are empirical risk minimization or k-nearest neighbors rules. The set of examples consists in the measurement of n observations $(x_i, y_i)_{1\leq i\leq n}$. Thus, formally, Φ is a measurable map from $\mathcal{X} \times \cup_n (\mathcal{X} \times \mathcal{Y})^n$ to $\mathcal{Y}$. One of the main issue of *Statistical Learning* is to analyze the performance of a learning algorithm in a probabilistic setting. $(x_i, y_i)_{1\leq i\leq n}$ are supposed to be observations from n independent and identically distributed (i.i.d.) random variables $(X_i, Y_i)_{1\leq i\leq n}$ with unknown distribution $\mathbb{P}$. $(X_i, Y_i)_{1\leq i\leq n}$ is denoted $\mathcal{D}_n$ in the following and called the learning set. In order to analyze the performance, it is usual to consider the conditional risk of a machine learning Φ denoted $\tilde{R}_n$, so called the generalization error. It is defined by the conditional expectation of $L(Y, \Phi(X, \mathcal{D}_n))$ given $\mathcal{D}_n$ where $(X, Y) \sim \mathbb{P}$ is a random variable independent of $\mathcal{D}_n$, i.e. $\tilde{R}_n := \mathbb{E}_{X,Y}(L(Y, \Phi(X, \mathcal{D}_n))|\mathcal{D}_n)$ with L a cost function from $\mathcal{Y}^2 \longrightarrow \mathbb{R}_+$. Notice that $\widetilde{R}_n$ is a random variable measurable with respect to $\mathcal{D}_n$.

An important question is: the distribution $\mathbb{P}$ of the generating process being unknown, can we estimate how good a predictor trained on a learning set of size n is? In other words, can we approximate the generalization error $\widetilde{R}_n$? This fundamental statistical problem is referred to "choice and assessment of statistical predictions" [STO74]. Many estimates have been proposed. Quoting [HTF01]: *Probably the simplest and most widely used method for estimating prediction error is cross-validation.*

The cross-validation procedures include leave-one-out cross-validation, k-fold cross-validation, hold-out cross validation (or split sample), leave-υ-out cross-validation (or Monte Carlo cross-validation or bootstrap cross-validation). With the exception of [BUR89], theoretical investigations of multifold cross-validation procedures have first concentrated on linear models ([Li87];[SHAO93];[ZHA93]). Results of [DGL96] and [GYO02] are discussed in Section 3. The first finite sample results are due to Wagner and Devroye [DEWA79] and concern k-local rules algorithms under leave-one-out and hold-out cross-validation. More recently, [HOL96, HOL96bis] derived finite sample results for υ-out cross-validation, $k-$fold cross-validation, and leave-one-out cross-validation for Empirical Risk Minimization (ERM) over a class of predictors with finite Vapnik-Chervonenkis-dimension (VC-dimension)in the realisable case (the generalization error is equal to zero). [BKL99] have emphasized when $k-$fold can beat υ-out cross-validation in the particular case of k-fold predictor. [KR99] has extended such results in the case of stable algorithms for the leave-one-out cross-validation procedure. [KEA95] also derived results for hold-out cross-validation for ERM, but their arguments rely on the traditional notion of VC-dimension. In the particular case of ERM over a class of predictors with finite VC-dimension but with general cross-validation procedures,

we derived derived probability upper bounds in chapter 1: we denote by p_n the percentage of elements in the test sample. In the sequel, we will denote by $\widehat{R}_{CV}$ the cross-validation estimator. For empirical risk minimizers over a class of predictors with finite VC-dimension $V_{\mathcal{C}}$, to be defined below, we obtained the following concentration inequality. For all $\varepsilon > 0$, we have

$$\Pr\left(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon\right) \leq B(n, p_n, \varepsilon) + V(n, p_n, \varepsilon),$$

with

- $B(n, p_n, \varepsilon) = 5(2n(1-p_n)+1)^{\frac{4V_{\mathcal{C}}}{1-p_n}} \exp(-\frac{n\varepsilon^2}{64})$,
- $V(n, p_n, \varepsilon) = \min\left(\exp(-\frac{2np_n\varepsilon^2}{25}), \frac{16}{\varepsilon}\sqrt{\frac{V_{\mathcal{C}}(\ln(2(1-p_n)+1)+4)}{n(1-p_n)}}\right)$.

Unfortunately, many popular predictors, including k-nearest neighbors rules, do not satisfy this property. Moreover, these bounds obtained are called "sanity check bounds" since they are not better than classical Vapnik-Chernovenkis's bounds.

To avoid the traditional analysis in the VC framework, notions of stability have been intensively worked through in the late 90's [KEA95], [BE01], [BE02], [KUT02], and [KUNIY02]. The object of stability framework is the learning algorithm rather than the space of classifiers. The learning algorithm is a map (effective procedure) from data sets to classifiers. An algorithm is stable at a learning set $\mathcal{D}_n$ if changing one point in $\mathcal{D}_n$ yields only a small change in the output hypothesis. The attraction of such an approach is that it avoids the traditional notion of VC-dimension, and allows to focus on a wider class of learning algorithms than empirical risk minimization. For example, this approach provides generalization error bounds for regularization-based learning algorithms that have been difficult to analyze within the VC framework such as boosting. As a motivation, we quote the following list of algorithms satisfying stability properties: regularization networks, ERM, k-nearest rules, boosting.

Algorithmic stability was first introduced by [DEWA79]. [**?**] argued that unstable weak learners benefit from randomization algorithms such as bagging. [KR99] considered both algorithmic stability and the weaker related notion of error stability. They proved bounds on the error of cross-validation estimates of generalization error, but their arguments rely on VC theory. [BE01, BE02] proved that an algorithm which is stable everywhere has low generalization error; their proof does not make any reference to VC-dimension. They showed that regularization networks are stable. In [KUNIY02], at least ten different notions were examined. In particular, they introduced a probabilistic notion of change-one stability called Cross-Validation stability or CV stability. This was shown to be necessary and sufficient for consistency of ERM in the Probably Approximately Correct (PAC) Model of [VAL84)].

The goal of this paper is to obtain exponential bounds to fill the chart 2.1 where possible bounds are missing (up to our knowledge).

	leave-one-out	hold-out	k-fold	ν-out
ERM with finite VC-dimension	Kearns, Holden, Cornec	Holden,Cornec	Holden,Cornec	Cornec
hypothesis stability	Devroye and W	Devroye and W	×	×
error stability with finite VC dimension	Kearns	Kearns	×	×
uniform stability	Bousquet and E.	×	×	×
strong hypothesis	Kutin and N	×	×	×
weak stability	×	×	×	×

Table 2.1: Missing bounds × to find

The goal of this article is also to show that cross-validation is still consistent for stable predictors. As a consequence, we will emphasize the role played by cross-validation: it can be a consistent estimate of the generalisation error when the training error defined by $\widehat{R}_n := \frac{1}{n}\sum_{i=1}^{n} L(Y_i, \phi(X_i, \mathcal{D}_n))$ is not. Indeed, for stable predictors, the training error can be arbitrarly poor: for example, the training error for 1-nearest neighboor is equal to zero whatever the generalisation error may be.

We introduce our **main result**[1]. Suppose that the cross-validation is symmetric –i.e. the probability of a observation to be in the training set is independent of its index- and that the number of elements in the test set is constant and equal to np_n with p_n the percentage of elements in the test set. All the bounds of the following form $\Pr\left(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \ldots\right) \leq B(n, p_n, \varepsilon) + V(n, p_n, \varepsilon)$.
Under certain stability conditions -satisfied for example by Empirical Risk Minimisers (ERM) or Adaboost-, we have for all $\varepsilon \geq 0$,

$$\Pr\left(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + 2\lambda p_n\right) \leq 2\exp(-2np_n\varepsilon^2) + \delta_{n,p_n}$$

with δ_{n,p_n} and λ a non-negative real numbers. For classical algorithms, we have in mind that $\delta_{n,p_n} = O_n(p_n \exp(-n(1-p_n))$. λ is in fact a Lipschitz coefficient with respect to the total variation and can be interpreted as a stability factor: the smaller λ is, the more stable the learning algorithm is. Furthermore, if the learning algorithm satisfies a stronger stability condition (for example Adaboost or regularization networks), we obtain

$$\Pr\left(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \delta_{n,p_n} + 2\lambda p_n\right) \leq 4(\exp(-\frac{\varepsilon^2}{8(18\lambda)^2 np_n^2}) + \frac{n}{9\lambda p_n}\delta'_{n,p_n}))$$

with $\delta'_{n,p_n} = \delta_{n,p_n} + (n+1)\delta_{n,1/n}$. For the latter, it is thus not required that the number of elements in the test set grows to infinity for the consistency of the cross-validation to hold.

Using these probability bounds, we can then deduce that the expectation between the generalization error and the cross-validation error $\mathbb{E}_{\mathcal{D}_n}|\widehat{R}_{CV} - \widetilde{R}_n|$ is of order $O_n((\lambda/n)^{1/3})$. As far as the expectation $\mathbb{E}_{\mathcal{D}_n}|\widehat{R}_{CV} - \widetilde{R}_n|$ is concerned, we can define a splitting rule in the general setting: the percentage of elements $p_n^\star$ in the test set should be proportional

[1] accurate inequalities can be found in section 2.3

to $(1/\lambda^2 n)^{1/3}$, i.e. the less stable (i.e. λ large) the learning algorithm is, the smaller the test set in the cross-validation should be. Furthermore, if the learning algorithm satisfies a stronger stability condition (for example Adaboost or regularization networks), we also have $\mathbb{E}_{\mathcal{D}_n}|\widehat{R}_{CV} - \widetilde{R}_n| = O_n(\lambda/\sqrt{n})$ and the leave-one-out cross-validation (i.e. $p_n^\star = 1/n$) is preferred for n large enough.

The paper is organized as follows. In the next section, we recall the main notations and definitions of cross-validation as introduced in chapter 1. We also introduce notations to unify the main notions of stability. Finally, in Section 3, we introduce our results in terms of probability upperbounds. We also prove that many traditionnal methods satisfy our generalized notion of stability (lasso,...,adaboost, k-nearest neighbors).

2.2 Notations and definitions

In the following, we follow the notations of cross-validation introduced in chapter 1.

2.2.1 Cross-validation

We will consider the following shorter notations inspired by the literature on empirical processes. In the sequel, we will denote $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, and $(Z_i)_{1\leq i\leq n} := ((X_i, Y_i))_{1\leq i\leq n}$ the learning set. For a given loss function L and a given class of predictors $\mathcal{G}$, we define a new class $\mathcal{F}$ of functions from $\mathcal{Z}$ to $\mathbb{R}_+$ by $\mathcal{F} := \{\psi \in \mathbb{R}_+^{\mathcal{Z}} | \psi(Z) = L(Y, \phi(X)), \phi \in \mathcal{G}\}$. For a machine learning Φ, we have the natural definition $\Psi(Z, \mathcal{D}_n) := L(Y, \Phi(X, \mathcal{D}_n))$. With these notations, the conditional risk $\widetilde{R}_n$ is the expectation of $\Psi(Z, \mathcal{D}_n)$ with respect to $\mathbb{P}$ conditionally on $\mathcal{D}_n$: $\widetilde{R}_n := \mathbb{E}_Z[\Psi(Z, \mathcal{D}_n) \mid \mathcal{D}_n]$ with $Z \sim \mathbb{P}$ independent of $\mathcal{D}_n$. In the following, if there is no ambiguity, we will also allow the following notation $\psi(X, \mathcal{D}_n)$ instead of $\Psi(X, \mathcal{D}_n)$.
To define the accurate type of cross-validation procedure, we introduce binary vectors. Let $V_n = (V_{n,i})_{1\leq i\leq n}$ be a vector of size n. V_n is a binary vector if for all $1 \leq i \leq n, V_{n,i} \in \{0, 1\}$ and if $\sum_{i=1}^n V_{n,i} \neq 0$. Consequently, we can define the subsample associated with it, $\mathcal{D}_{V_n} := \{Z_i \in \mathcal{D}_n | V_{n,i} = 1, 1 \leq i \leq n\}$. We define a weighted empirical measure on $\mathcal{Z}$

$$\mathbb{P}_{n,V_n} := \frac{1}{\sum_{i=1}^n V_{n,i}} \sum_{i=1}^n V_{n,i}\delta_{Z_i},$$

with δ_{Z_i} the Dirac measure at $\{Z_i\}$. We also define a weighted empirical error $\mathbb{P}_{n,V_n}\psi$ where $\mathbb{P}_{n,V_n}\psi$ stands for the usual notation of the expectation of ψ with respect to $\mathbb{P}_{n,V_n}$. For $\mathbb{P}_{n,1_n}$, with 1_n the binary vector of size n with 1 at every coordinate, we will use the traditional notation $\mathbb{P}_n$. For a predictor trained on a subsample, we define

$$\psi_{V_n}(.) := \Psi(., \mathcal{D}_{V_n}).$$

With the previous notations, notice that the predictor trained on the learning set $\psi(., \mathcal{D}_n)$ can be denoted by $\psi_{1_n}(.)$. We will allow the simpler notation $\psi_n(.)$. The learning set is divided into two disjoint sets: the training set of size $n(1 - p_n)$ and the test set of size np_n, where

p_n is the percentage of elements in the test set. To represent the training set, we define V_n^{tr} a random binary vector of size n independent of $\mathcal{D}_n$. V_n^{tr} is called the training vector. We define the test vector by $V_n^{ts} := 1_n - V_n^{tr}$ to represent the test set.
The distribution of V_n^{tr} characterizes all the cross-validation procedures described in the previous section (see e.g. chapter 1). Using our notations, we can now define the cross-validation estimator.

Definition 61 (Cross-validation estimator)**.** *With the previous notations, the generalized cross-validation error of ψ_n denoted $\widehat{R}_{CV}(\psi_n)$ is defined by*

$$\widehat{R}_{CV}(\psi_n) := \mathbb{E}_{V_n^{tr}} \mathbb{P}_{n,V_n^{ts}}(\psi_{V_n^{tr}}).$$

We will give here an example of distributions of V_n^{tr} to illustrate we retrieve cross-validation procedures described previously. Leave-υ-out cross-validation is an elaborate and expensive version of cross-validation. This procedure divides the data into two sets: the training set of size $n-\upsilon$ and the test set of size υ. It then produces a predictor by training on the training set and testing on the remaining test set. This is repeated for all possible subsamples of υ cases, and the observed errors are averaged to form the leave-υ-out estimate. Denote by $(\xi_{n,i}^{\upsilon})_{1 \leq i \leq \binom{n}{\upsilon}}$ the family of binary vectors of size n such that $\sum_{i=1}^{n} \xi_{n,i}^{\upsilon} = n - \upsilon$.

Example 62 (Leave-υ-out cross-validation)**.**

$$\begin{aligned}
\Pr(V_n^{tr} = \xi_{n,1}^{\upsilon}) &= \frac{1}{\binom{n}{\upsilon}} \\
\Pr(V_n^{tr} = \xi_{n,2}^{\upsilon}) &= \frac{1}{\binom{n}{\upsilon}} \\
&\dots \\
\Pr(V_n^{tr} = \xi_{n,\binom{n}{\upsilon}}^{k}) &= \frac{1}{\binom{n}{\upsilon}}.
\end{aligned}$$

For other examples, see chapter one.

2.2.2 Definitions and notations of stability

The basic idea is that an algorithm is stable at a training set $\mathcal{D}_n$ if changing one point in $\mathcal{D}_n$ yields only a small change in the output hypothesis. Formally, a learning algorithm maps a weighted training set into a predictor space. Thus, stability can be translated into a Lipschitz condition for this mapping with high probability.
To be more formal, following [KUNIY02], we define a distance between two weighted empirical errors.

Let $\mathbb{P}_{n,V_n}$ and $\mathbb{P}_{n,U_n}$ be two empirical measures on $\mathcal{Z}$ with respect to the binary vectors V_n and U_n. We do not assume their support to be equal. The distance between them is defined as their total variation, i.e. the number of points they do not have in common

$$||\mathbb{P}_{n,U_n} - \mathbb{P}_{n,V_n}|| = \sup_{A \in \mathcal{P}(\mathcal{Z})} |(\mathbb{P}_{n,U_n} - \mathbb{P}_{n,V_n})(A)|.$$

Example 63. *In the case of leave-one-out (i.e.* $\sum_{i=1}^{n} U_{n,i} = n-1$*), we have*

$$||\mathbb{P}_{n,U_n} - \mathbb{P}_n|| = \frac{2}{n}.$$

In the case of leave-ν*-out, we get*

$$||\mathbb{P}_{n,U_n} - \mathbb{P}_n|| = \frac{2\nu}{n}.$$

In the general setting, it follows that

$$||\mathbb{P}_{n,U_n} - \mathbb{P}_n|| = 2p_n.$$

At least, we need a distance d on the set $\mathcal{F}$. Let us quote three important examples. Let $\psi_1, \psi_2 \in \mathcal{F}$. The uniform distance is defined by: $d_\infty(\psi_1, \psi_2) = \sup_{Z \in \mathcal{Z}} |\psi_1(Z) - \psi_2(Z)|$, the L_1-distance by: $d_1(\psi_1, \psi_2) = \mathbb{P}|\psi_1 - \psi_2|$, the error-distance $d_e(\psi_1, \psi_2) = |\mathbb{P}(\psi_1 - \psi_2)|$. It is important to notice that what matters here is not an absolute distance between the original class of predictors $\mathcal{G}$ seen as functions but the distance with the respect to the loss or/and the distribution $\mathbb{P}$. In particular, for the L_1-distance, we do not care about the behavior of the original predictors φ_1 and φ_2 outside the support of $\mathbb{P}$. At last, notice that we always have $d_e \leq d_1 \leq d_\infty$.

We are now in position to define the different notions of stability of a learning algorithm which cover notions introduced by [KUNIY02]. We begin with the notion of weak stability. In essence, it says that for any given resampling vectors, the distance between two predictors is controlled with high probability by the distance between the resampling vectors.

Definition 64 (Weak stability)**.** *Let* $\alpha, \lambda, (\delta_{n,p_n})_{n,p_n}$ *be nonnegative real numbers. A learning algorithm* Ψ *is said to be weak* $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ *stable if for any training vector* U_n *whose sum is equal to* $n(1-p_n)$*:*

$$\Pr(d(\psi_{U_n}, \psi_n) \geq \lambda ||\mathbb{P}_{n,U_n} - \mathbb{P}_n||^\alpha) \leq \delta_{n,p_n}.$$

Notice that in the former definition Pr stands for $\mathbb{P}^{\otimes n}$. Indeed, ψ_n is trained with n observations, drawn independently from $\mathbb{P}$. A stronger notion is to consider ψ_n trained with $n-1$ observations drawn independently from $\mathbb{P}$ and an additionnal general observation z. We consider the stronger notion of strong stability. As a motivation, notice that algorithms such as Empirical Risk Minimization with finite VC dimension ([KUNIY02]) satisfies this property.

Definition 65 (Strong stability)**.** *Let* $z \in \mathcal{Z}$*. Let* $\mathcal{D}_n = \mathcal{D}_{n-1} \cup \{z\}$ *be a learning set. Let* $\lambda, (\delta_{n,p_n})_{n,p_n}$ *be nonnegative real numbers. A learning algorithm* Ψ *is said to be strong* $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ *stable if for any training vector* U_n *whose sum is equal to* $n(1-p_n)$

$$\Pr(d(\psi_{U_n}, \psi_n) \geq \lambda ||\mathbb{P}_{n,U_n} - \mathbb{P}_n||^\alpha) \leq \delta_{n,p_n}.$$

What we have in mind for classical algorithms is $\delta_{n,p_n} = O_n(p_n \exp(-n(1-p_n))$. We can state the last definition in other words. Let V_n^{tr} be a training vector with distribution $\mathbb{Q}$ such that the number of elements in the training set is constant and equal to $n(1-p_n)$. Notice then that the former definition also implies that $\sup_{U_n \in \text{support}(\mathbb{Q})} \mathbb{P}(\frac{d(\psi_{U_n},\psi_n)}{||\mathbb{P}_{n,U_n}-\mathbb{P}_n||^\alpha} \geq \lambda) \leq \delta_{n,p_n}$, where support($\mathbb{Q}$) stands for the support of $\mathbb{Q}$. The previous notion stands for any U_n having the same support of $\mathbb{Q}$. A stronger hypothesis would be that the previous probability stands uniformly over U_n in support($\mathbb{Q}$). This leads formally to the notion of cross-validation stability. As a motivation, notice that algorithms such as Lasso ([BTW07]) satisfies this property. To be more accurate, we define

Definition 66 (Cross-validation weak stability)**.** *Let $\mathcal{D}_n = (Z_i)_{1\leq i\leq n}$ a learning set. Let V_n^{tr} a training vector with distribution $\mathbb{Q}$. Let $\lambda, (\delta_{n,p_n})_{n,p_n}$ be nonnegative real numbers. A learning algorithm Ψ is said to be weak $(\lambda, (\delta_{n,p_n})_{n,p_n}, d, \mathbb{Q})$ stable if it is weak $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ stable and if:*

$$\Pr(\sup_{U_n \in support(\mathbb{Q})} \frac{d(\psi_{U_n},\psi_n)}{||\mathbb{P}_{n,U_n}-\mathbb{P}_n||^\alpha} \geq \lambda) \leq \delta_{n,p_n}.$$

As before, we also define the following stronger notion

Definition 67 (Cross-validation strong stability)**.** *Let $z \in \mathcal{Z}$. Let $\mathcal{D}_n = \mathcal{D}_{n-1} \cup \{z\}$ a learning set. Let V_n^{tr} be a cross-validation vector with distribution $\mathbb{Q}$. A learning algorithm Ψ is said to be strongly $(\lambda, (\delta_{n,p_n})_{n,p_n}, d, \mathbb{Q})$ stable if it is strong $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ stable and if:*

$$\Pr(\sup_{U_n \in support(\mathbb{Q})} \frac{d(\psi_{U_n},\psi_n)}{||\mathbb{P}_{n,U_n}-\mathbb{P}_n||^\alpha} \geq \lambda) \leq \delta_{n,p_n}.$$

Remark 68. *If the cardinal of the support of $\mathbb{Q}$ is denoted $\kappa(n)$, then a learning algorithm which is weak $(\lambda, (\delta_{n,p_n})_{n,p_n}, d, \mathbb{Q})$-stable is also strong $(\lambda, (\kappa(n)\delta_{n,p_n})_n, d, \mathbb{Q})$-stable.*

At last, we consider the special important case when $\delta_{n,p_n} = 0$. This is the case in particular for regularization networks ([BE01]).

Definition 69 (Sure stability)**.** *Notice that when $\delta_{n,p_n} = 0$, the two notions coincides and are called **sure stability**.*

As an example of strong stability, we develop the description of [FRE95] who introduced the algorithm.

Example 70 (Adaboost)**.** *We give an initial distribution p^1 and let $w^1 = p^{(1)}$ and $Z_1 = 1$. Let Φ be a learning algorithm. Let T the number of rounds.*

For each $t = 1...T$:

1. *Train the learning algorithm Φ on the learning set with distribution $p^{(t)}$. The predictor obtained is denoted by $\varphi^{(t)}$.*
2. *For each i, let $a_i^t = |\varphi^t(x_i) - y_i|$, the error of $\varphi^{(t)}$ on instance i.*

3. *Let* $\varepsilon_t = \sum_{i=1}^m p^t a_i^t$, *the error rate of* $\varphi^{(t)}$ *with respect to* $p^{'t)}$.

4. *Let* $\beta_t = \frac{\varepsilon_t}{1-\varepsilon_t}$ *and let* $\alpha_t = \ln(1/\beta_t)$

5. *reweight the data: for all* i, *let* $w_i^{(t+1)} = w_i^{(t)} \beta_t^{1-a_i^t}$.

6. *Normalize the distribution: let* $Z_{t+1} = \sum_{i=1}^m w_i^{(t+1)}$ *and* $p_i^{(t+1)} = w_i^{'t+1)}/Z_{t+1}$

The final output is $H_T(x) = \sum_{s=1}^T \alpha_s \varphi^{(s)}(x)$.
[KUNIY01] shows that under certain hypotheses, Adaboost is strongly stable: suppose the learner Φ *-*$(\lambda, 0, d_\infty)$ *stable and other regularity assumptions, then Adaboost with* T *rounds is strong* $(\lambda^*, (\delta^*_{n,p_n})_{n,p_n}, d_\infty)$ *stable for some* λ^* *and* δ^*_{n,p_n}.

We give now an example that is surely stable introduced in [BE01].

Example 71 (Regularization networks)**.** *Regularization networks are attractive for their links with Support Vector Machines and their Bayesian interpretation. This learning algorithm consists in finding a function* $\varphi : \mathbb{R}^d \to \mathbb{R}$ *in a space* H *which minimizes the following functional:*

$$A(\varphi) = \frac{1}{n} \sum_{i=1}^n (Y_i - \varphi(X_i))^2 + \lambda ||\varphi||_H^2,$$

with $||\varphi||_H$ *the* L_2 *norm in the space* H. H *is chosen to be a reproducing kernel Hilbert Space (rkhs) with kernel* k. k *is supposed to be a symmetric function* $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. *In particular, we have the following property (for a detailed introduction of rkhs, see [ATE92])*

$$|f(x)| \leq ||f||_H ||k||_H.$$

We slightly adapt the proof in [BE01] to show that a regularization network is surely stable:

Theorem 72. *If* Ψ *is a regularization network such that* $||k||_H \leq \kappa$ *and* $(y - \varphi(x))^2 \leq M$, *then* Ψ *is* $\frac{4M\kappa^2}{n\lambda}$*- surely stable with respect to the distance* d_∞.

Proof
Define $A^i(\varphi) := \frac{1}{n-1} \sum_{j \neq i} (Y_j - \varphi(X_j))^2 + \lambda ||\varphi||_H^2$ *and* $\mathcal{D}_n^i := \mathcal{D}_n \backslash \{(X_i, Y_i)\}$. $\varphi_{\mathcal{D}_n^i}$ *is the minimizer of* A^i *over* H *whereas* $\varphi_{\mathcal{D}_n}$ *is the minimizer of* A. *Denote* $g := \varphi_{\mathcal{D}_n^i} - \varphi_{\mathcal{D}_n}$.

For $t \in [0, 1]$, *we have* $A(\varphi_{\mathcal{D}_n}) - A(\varphi_{\mathcal{D}_n} + tg)$ *is equal to*

$$\frac{-2t}{n(n-1)} \sum_{j \neq i} (n-1)(\varphi_{\mathcal{D}_n}(x_j) - y_j) g(x_j) - \frac{2t}{n(n-1)} (n-1)(\varphi_{\mathcal{D}_n}(x_i) - y_i) g(x_i) - 2t\lambda < \varphi_{\mathcal{D}_n}, g >_H + t^2 B(g)$$

with $B(g)$ *the factor of* t^2.

In the same way, we get that $A^i(\varphi_{\mathcal{D}_n^i}) - A^i(\varphi_{\mathcal{D}_n^i} - tg)$ *is equal to*

$$\frac{2t}{n(n-1)} \sum_{j \neq i} (n-1)(\varphi_{\mathcal{D}_n^i}(x_j) - y_j) g(x_j) + \frac{2t}{n(n-1)} \sum_{j \neq i} (\varphi_{\mathcal{D}_n^i}(x_j) - y_j) g(x_j) + 2t\lambda < \varphi_{\mathcal{D}_n^i}, g >_H + t^2 B^n(g).$$

By definition of A and A^i, we have $A(\varphi_{\mathcal{D}_n}) - A(\varphi_{\mathcal{D}_n} + tg) \leq 0$ and $A^i(\varphi_{\mathcal{D}_n^i}) - A^i(\varphi_{\mathcal{D}_n^i} + tg) \leq 0$. Thus, we get by summing these two inequalities, dividing by $\frac{2t}{n(n-1)}$ and making $t \to 0$.

$$\sum_{j \neq i}(n-1)g^2(x_j) + \sum_{j \neq i}\left[(\varphi_{\mathcal{D}_n^i}(x_j) - y_j)g(x_j) - (\varphi_{\mathcal{D}_n}(x_i) - y_i)g(x_i)\right] + n(n-1)||g||_H^2 \leq 0$$

which leads to

$$n(n-1)||g||_H^2 \leq \sum_{j \neq i}\left[(\varphi_{\mathcal{D}_n}(x_i) - y_i)g(x_i) - (\varphi_{\mathcal{D}_n^i}(x_j) - y_j)g(x_j)\right] \leq 2(n-1)\sqrt{M}\kappa||g||_H$$

by assumptions.

Thus, we have

$$||g||_H \leq 2(n-1)\sqrt{M}\kappa||g||_H/n\lambda$$

and also, for all x, y

$$|(\varphi_{\mathcal{D}_n}(x) - y)^2 - (\varphi_{\mathcal{D}_n^i}(x) - y)^2| \leq 2\sqrt{M}|\varphi_{\mathcal{D}_n}(x) - \varphi_{\mathcal{D}_n^i}(x)| \leq 4M\kappa^2/n\lambda.$$

□

Another popular example is given by the k-nearest neighbors which are strongly stably with respect to d_1.

Example 73 (k-nearest neighbors)**.** *In the k-nearest rule, the machine learning is a function of X and of the k nearest observations to X from $(X_1, ..., X_n)$ and of the corresponding $(Y_1, ..., Y_n)$. Because there may be ties in determining the k nearest neighbors, we use an independent sequence $(Z, Z_1, ..., Z_n)$ of i.i.d uniform random variables in $[0, 1]$. X_j is nearer X_i to X if:*

1. *$||X_j - X|| < ||X_j - X||$ or*
2. *$||X_j - X|| = ||X_j - X||$ and $|Z_j - Z| < |Z_i - Z|$, or*
3. *$||X_j - X|| = ||X_j - X||$ and $Z_j = Z_i$ and $j < i$.*

The last event does not count since its has zero probability.
Denote γ_d the maximum number of distinct points in $\mathbb{R}^d$ that share the same nearest neighbor. It can be shown that $\gamma_d \leq 3^d - 1$ and other lower and upper bounds can be found in [ROG63]. Recall the following lemma from [DEWA79]: suppose $(X_1, Z_1), \ldots, (X_n, Z_n)$ is the sequence obtained from the data by omitting the $Y_1, ..., Y_n$. If, for each j, the nearest neighbor to (X_j, Z_j) is found from $(X_1, Z_1), \ldots, (X_{j-1}, Z_{j-1}), (X_{j+1}, Z_{j+1}), ..., (X_n, Z_n)$. Then no point (X_i, Y_i) can be the nearest neighbors to more than $\gamma_d + 2$ of the remaining points.
We can derive the next result following the proofs in [DEWA79].

Theorem 74. *Let $\mathcal{D}_n := ((X_1, Z_1, Y_1), \ldots, (x, z, y))$ be a learning set. Suppose Φ is a k local rule. Then we have for all $\varepsilon > 0$,*

$$\Pr(E_{X,Y,Z}|L(Y, \Phi((X, Z), \mathcal{D}_n)) - L(Y, \Phi((X, Z), \mathcal{D}_n^i))| \geq \varepsilon) \leq 6 \exp(\frac{-(n-1)\varepsilon^3}{54k(\gamma_d + 2)})$$

with $\mathcal{D}_n^i := \mathcal{D}_n \backslash \{(X_i, Y_i, Z_i)\}$ and i a fixed index.

It says that the k nearest rule satisfies strong stability property with respect d_1 and $||\mathbb{P}_{n,U_n} - \mathbb{P}_n||^\alpha$ with $\alpha < 1/3$.
Proof

Consider one local rule first.

Let m be an integer. Consider an independent identically distributed ghost sample

$$((X_{n+1}, Y_{n+1}, Z_{n+1}), \ldots, (X_{n+m}, Y_{n+m}, Z_{n+m})).$$

Denote $\mathcal{T}_{n+m} := ((X_1, Y_1, Z_1), \ldots, (X_{n+m}, Y_{n+m}, Z_{n+m}))$ and $\mathcal{T}_{n+m}^j := T_{n+m} \backslash \{(X_j, Y_j, Z_j)\}$.

- $L_1 := E_{X,Y,Z}|L(Y, \Phi((X, Z), \mathcal{D}_n)) - L(Y, \Phi((X, Z), \mathcal{D}_n^i))|$
- $L_2 := \frac{1}{m} \sum_{j=1}^m |L(Y_{n+j}, \Phi((X_{n+j}, Z_{n+j}), \mathcal{D}_n)) - L(Y_{n+j}, \Phi((X_{n+j}, Z_{n+j}), \mathcal{D}_n^i))|$
- $L_3 := \frac{1}{m} \sum_{j=1}^m |L(Y_{n+j}, \Phi((X_{n+j}, Z_{n+j}), \mathcal{D}_n)) - L(Y_{n+j}, \Phi((X_{n+j}, Z_{n+j}), \mathcal{T}_{n+m}^{n+j}))|$
- $L_4 := \frac{1}{m} \sum_{j=1}^m |L(Y_j, \Phi((X_{n+j}, Z_{n+j}), \mathcal{D}_n^i)) - L(Y_{n+j}, \Phi((X_{n+j}, Z_{n+j}), \mathcal{T}_{n+m}^{n+j}))|$.

We have

$$\Pr(L_1 \geq 3\varepsilon) \leq \Pr(L_1 - L_2 \geq \varepsilon) + \Pr(L_3 \geq \varepsilon) + \Pr(L_4 \geq \varepsilon).$$

By Hoeffding's inequality we have $\Pr(L_1 - L_2 \geq \varepsilon) \leq \exp(-2m\varepsilon^2)$.

Now we get for the second term

$$\begin{aligned}\Pr(L_3 \geq \varepsilon) &\leq \Pr(\frac{1}{m} \sum_{j=1}^m 1_{\Phi((X_{n+j}, Z_{n+j}), \mathcal{D}_n) \neq \Phi((X_{n+j}, Z_{n+j}), \mathcal{T}_{n+m}^{n+j})} \geq \varepsilon) \\ &\leq \Pr(\frac{1}{m} \sum_{j=1}^m 1_{A(n+j)} \geq \varepsilon),\end{aligned}$$

with $A(n+j)$ the event that the nearest neighbor of (X_{n+j}, Z_{n+j}) from $\mathcal{T}_{n+m}^{n+j}$ is attained in the ghost sample $\mathcal{T}_{n+m}^{n+j} \backslash \mathcal{D}_n$.

From [DEWA79], we have, if $(\gamma_d + 2)m < (n+m)\varepsilon/2$,

$$\Pr(\frac{1}{m}\sum_{j=1}^{m} 1_{A(n+j)} \geq \varepsilon) \leq 2\exp(-2m(\varepsilon/2)^2)$$

In the same way, we find that $\Pr(L_4 \geq \varepsilon) \leq 2\exp(-2m(\varepsilon/2)^2)$ *if* $(\gamma_d+2)m < (n-1+m)\varepsilon/2$

Taking $m = \frac{(n-1)\varepsilon}{\gamma_d+2}$, *we obtain*

$$\Pr(L_3 \geq \varepsilon) \leq 2\exp(\frac{-(n-1)\varepsilon^3}{2(\gamma_d+2)})$$

and $\Pr(L_3 \geq \varepsilon) \leq 2\exp(\frac{-(n-1)\varepsilon^3}{2(\gamma_d+2)})$.

For an arbitray k, *it is sufficient to replace* $(\gamma_d + 2)$ *by* $k(\gamma_d + 2)$.
□

A last popular example is given by the Lasso which is strongly stable with respect to d_1.

Example 75 (Lasso)**.** *We follow [BTW07] who defines Lasso-type methods in the following way. Let* $((X_1, Y_1), ..., (X_n, Y_n))$ *be a sample of i.i.d. pairs distributed as* $(X, Y) \in (\mathcal{X},\mathbb{R})$, *where* $\mathcal{X}$ *is a borel subset of* $\mathbb{R}^d$. *We denote by* μ *the distribution of* X *on* $\mathcal{X}$. *Let* $f(X) = \mathbb{E}(Y|X)$ *be the unknown regression function and* $\mathcal{F}_M = \{f_1, ..., f_M\}$ *be a dictionary of real-valued functions* f_j *that are defined on* $\mathcal{X}$. *We use a data dependent* l_1*-penalty. Formally, for any* $\lambda = (\lambda_1, ..., \lambda_M) \in \mathbb{R}^M$, *define* $f_\lambda(x) = \sum_{j=1}^{M} \lambda_j f_j(x)$. *Then the penalized least squares estimator of* λ *is*

$$\hat{\lambda} = \arg\min\{1/n\sum_{i=1}^{n}(Y_i - f_\lambda(X_i))^2 + pen(\lambda)\}$$

where

$$pen(\lambda) = 2\sum_{j=1}^{M}\omega_{n,j}|\lambda_j| \text{ with } \omega_{n,j} = r_{n,M}||f_j||_n$$

where $||g||_n^2 = n^{-1}\sum_{i=1}^{n} g^2(X_i)$ *for the squared empirical empirical* L_2 *norm of any function* $g : \mathcal{X} \to R$. *The tuning sequence* $r_{n,M} > 0$ *is defined by* $r_{n,M} := A\sqrt{\log(M)/n}$ *for* A *large enough. Then we have* $\hat{f}_n = f_{\hat{\lambda}}$
Define

$$M(\lambda) = \sum_{j=1}^{M} I_{\{\lambda_j \neq 0\}}$$

the number of non-zero coordinates of λ.
We recall the definition of weak sparsity in [BTW07]. Let $C_f > 0$ *be a constant depending only on* f *and*

$$\Lambda = \{\lambda \in R^M : ||f_\lambda - f||^2 \leq C_f r_{n,M}^2 M(\lambda)\}$$

where

$$||g||^2 = \int_{\mathcal{X}} g^2(x)\mu(dx)$$

If Λ is not empty, f has the weak sparsity property relative to the dictionary $\{f_1, ..., f_M\}$. We have then the following theorem

Theorem 76. *Assume the general assumptions (A1)-(A3) and consider the notations in [BTW07]. Then, for all $\lambda \in \Lambda$,*

$$\Pr(|E_{X,Y}(Y - \hat{f}_n(X))^2 - E_{X,Y}(Y - \hat{f}_{n-1}(X))^2| > 2B_1\kappa_M^{-1} r_{n,M}^2 M(\lambda)) \leq 2\pi_{n-1,M}(\lambda)$$

with $\pi_{n-1,M}(\lambda)$ a small probability defined in [BTW07].
In other words, the Lasso-type algorithm is weakly stable with respect to d_e and $||.||_1$.

Proof

According to theorem 2.1. in [BTW07], we have:

$$\Pr(E_{X,Y}|\hat{f}_n(X)) - f(X)|^2 \leq B_1\kappa_M^{-1} r_{n,M}^2 M(\lambda)) \geq 1 - \pi_{n,M}(\lambda).$$

Thus, denote $\pi := \Pr(|E_{X,Y}(Y - \hat{f}_n(X))^2 - E_{X,Y}(Y - \hat{f}_{n-1}(X))^2| > 2B_1\kappa_M^{-1} r_{n,M}^2 M(\lambda))$. *We obtain:*

$$\begin{aligned}
\pi &= \Pr(|E_X(f(X) - \hat{f}_n(X))^2 - E_{X,Y}(f(X) - \hat{f}_{n-1}(X))^2| > 2B_1\kappa_M^{-1} r_{n,M}^2 M(\lambda)) \\
&\leq \Pr(E_X(f(X) - \hat{f}_n(X))^2 > B_1\kappa_M^{-1} r_{n,M}^2 M(\lambda)) \\
&+ \Pr(E_X(f(X) - \hat{f}_{n-1}(X))^2 > B_1\kappa_M^{-1} r_{n,M}^2 M(\lambda)) \\
&\leq 2\pi_{n-1,M}(\lambda).
\end{aligned}$$

□

As seen in the following table, we retrieve with those notations the different notions of stability introduced by [DEWA79], [KEA95] and also [BE01], [KUNIY02].

stability \ distance	d_∞	d_1	d_e
Weak	weak (λ, δ) hypothesis stability [KUNIY02]	weak (λ, δ) L_1stability [KUNIY02]	weak (λ, δ) error stability [KUNIY02]
Strong	strong (λ, δ) hypothesis stability [KUNIY02][DEWA79]	strong (λ, δ) L_1stability [KUNIY02]	strong (λ, δ) error stability [KUNIY02]
Sure Stability	uniform stability [BE01]	[DEWA79]	error stability [KEA95]

To motivate this approach, we also quote a list of class of predictors satisfying the previous stability conditions.

stability distance	d_∞	d_1	d_e
Weak			Lasso
Strong	Adaboost ([KUNIY02])	-ERM ([KUNIY02]) -k-nearest rule	Bayesian algorithm [KEA95]
Uniform	Regularization networks		

Remark 77. *We omit other weaker definition of stability such as defined in [BE01], [DEWA79], and [KUNIY02]. They consider bounds on the first moment of* $\mathbb{E}_{\mathcal{D}_n} d(\psi_{U_n}, \psi_n)$ *instead of probability bounds. Under these assumptions, they obtain polynomial upper bounds on* $\Pr\left(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon\right)$*. It is would be interesting to explore the behaviour of cross-validation estimates under these hypotheses. However, this cannot be done with the techniques presented in this paper and is left to further investigation.*

The main notations and definitions are summarized in the next table:

Name	Notation	Definition
Risk or generalization error	$\widetilde{R}_n$	$E_P[L(Y, \phi(X, D_n)) \mid D_n]$
Resubstitution error	$\widehat{R}_n$	$\frac{1}{n}\sum_{i=1}^n L(Y_i, \phi_n(X_i, D_n))$
Cross-validation error	$\widehat{R}_{CV}$	$E_{V_n^{tr}} P_{n,V_n^{ts}} \psi_{V_n^{tr}}$

Table 2.2: Main notations

2.3 Results for risk assessment for stable algorithms

Our goal is now to derive upper bounds for the probability that the distance between the cross-validation estimator and the generalization error is greater than $\varepsilon \geq 0$: $\Pr\left(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon\right)$.

2.3.1 Hypotheses $\mathcal{H}$

Let $\mathcal{D}_n$ be a learning set of size n. Let $V_n^{tr} \sim \mathbb{Q}$ be a training vector independent of $\mathcal{D}_n$ such that the cross-validation is symmetric -i.e. $\Pr(V_{n,i}^{tr} = 1)$ is a constant independent of i –and the number of elements in the training set is equal to np_n. Let d be a distance among d_e, d_1, d_∞. At last, we suppose that the loss function L is bounded by 1. We derive the following general results that stands for general cross-validation procedures and stable algorithms.

2.3.2 Strong stability

We state two results according to the class of stability. We will use the definition of strong difference bounded introduced by [KUT02] and a corollary of his main theorem inspired by [McD89].

Definition 78 (Kutin[KUT02]). *Let $\Omega_1, \ldots, \Omega_n$ be probability spaces. Let $\Omega = \prod_{k=1}^{n} \Omega_k$ and let X a random variable on Ω. We say that X is strongly difference bounded by (b, c, δ) if the following holds: there is a "bad" subset $B \subset \Omega$, where $\delta = \mathbb{P}(B)$. If $\omega, \omega' \in \Omega$ differ only in k-th coordinate, and $\omega \notin B$, then*

$$|X(\omega) - X(\omega')| \leq c.$$

Furthermore, for any $\omega, \omega' \in \Omega$,

$$|X(\omega) - X(\omega')| \leq b.$$

We will need the following theorem. It says in substance that a strongly difference bounded function of independent variables is closed to its expectation with high probability.

Theorem 79 (Kutin[KUT02]). *Let $\Omega_1, \ldots, \Omega_n$ be probability spaces. Let $\Omega = \prod_{k=1}^{n} \Omega_k$ and let X a random variable on Ω, which is strongly difference bounded by (b, c, δ). Assume $b \geq c \geq 0$ and $\alpha' > 0$. Let $\mu = \mathbb{E}(X)$. Then, for any $\tau > 0, \alpha' > 0$,*

$$\Pr(X - \mu \geq \tau) \leq 2(\exp(-\frac{\tau^2}{8n(c + b\alpha')^2}) + \frac{n}{\alpha'}\delta).$$

We are now in position to derive

Theorem 80 (Cross-validation strong stability). *Suppose that $\mathcal{H}$ holds. Let Ψ a machine learning which is strong $(\lambda, (\delta_{n,p_n})_{n,p_n}, d, \mathbb{Q})$ stable. Then, for all $\varepsilon \geq 0$,*

$$\Pr\left(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \lambda(2p_n)^{\alpha}\right) \leq 2\exp(-2np_n\varepsilon^2) + \delta_{n,p_n}.$$

Furthermore, if d is the uniform distance d_∞, then we have for all $\varepsilon \geq 0$:

$$\Pr\left(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \delta_{n(1-p_n)} + \lambda(2p_n)^{\alpha}\right) \leq 4(\exp(-\frac{\varepsilon^2}{8n(5\lambda(2p_n)^{\alpha} + \alpha')^2}) + \frac{n}{\alpha'}\delta'_{n,p_n}),$$

with $\delta'_{n,p_n} = \delta_{n,p_n} + (n+1)\delta_{n+1,1/(n+1)}$.
Thus, if we choose $\alpha = 5\lambda(2np_n)^{\alpha}$,

$$\Pr\left(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \delta_{n,p_n} + \lambda(2p_n)^{\alpha}\right) \leq 4(\exp(-\frac{\varepsilon^2}{8(10\lambda)^2 n(2p_n)^{2\alpha}}) + \frac{n}{5\lambda(2p_n)^{\alpha}}\delta'_{n,p_n}).$$

Proof

1. For the general case, denote B the bad subset, i.e. $B = \{\sup_{U_n \in \text{support}(\mathbb{Q})} \frac{d(\psi_{U_n}, \psi_n)}{||\mathbb{P}_{n,U_n} - \mathbb{P}_n||_{\alpha}} \geq \lambda\}$. Since Ψ is strong $(\lambda, (\delta_{n,p_n})_{n,p_n}, d, \mathbb{Q})$ stable, we have $\Pr(B) \leq \delta_{n,p_n}$. It is sufficient to split $|\widehat{R}_{CV} - \widetilde{R}_n|$ according to a benchmark, namely $\overline{R}_{n(1-p_n)} := \mathbb{E}_{V_n^{tr}} \mathbb{P}\psi_{V_n^{tr}}$. Thus, we get

$$\Pr\left(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \lambda(2p_n)^{\alpha}\right) \leq \Pr\left(|\widehat{R}_{CV} - \overline{R}_{n(1-p_n)}| \geq \varepsilon\right) + \Pr\left(|\overline{R}_{n(1-p_n)} - \widetilde{R}_n| \geq \lambda(2p_n)^{\alpha}\right)$$

The first term can be bounded by conditional Hoeffding inequality (see chapter 1). Thus, we obtain

$$\Pr\left(|\widehat{R}_{CV} - \overline{R}_{n(1-p_n)}| \geq \varepsilon\right) \leq 2\exp(-2np_n\varepsilon^2).$$

For the second term, notice that:

$$|\overline{R}_{n(1-p_n)} - \widetilde{R}_n| = |\mathbb{E}_{V_n^{tr}}\mathbb{P}\psi_{V_n^{tr}} - \mathbb{P}\psi_n| \leq \mathbb{E}_{V_n^{tr}}|\mathbb{P}\psi_{V_n^{tr}} - \mathbb{P}\psi_n|.$$

Recall that for any $d \in \{d_e, d_1, d_\infty\}$, we have $|\mathbb{P}\psi_{V_n^{tr}} - \mathbb{P}\psi_n| \leq d(\psi_{V_n^{tr}}, \psi_n)$ and $||\mathbb{P}_{n,V_n^{tr}} - \mathbb{P}_n||_\alpha = (2p_n)^\alpha$.

Thus, since Ψ is strong $(\lambda, (\delta_n)_n)$ stable, we have

$$\begin{aligned}\Pr\left(|\overline{R}_{n(1-p_n)} - \widetilde{R}_n| \geq \lambda(2p_n)^\alpha\right) &\leq \Pr(\sup_{V_n^{tr} \in \text{support}(\mathbb{Q})} d(\psi_{V_n^{tr}}, \psi_n)/||\mathbb{P}_{n,V_n^{tr}} - \mathbb{P}_n||_\alpha \geq \lambda) \\ &= \Pr(B) \leq \delta_{n,p_n}\end{aligned}$$

2. In the particular case, when $d = d_\infty$, the most stable notion of stability, we can obtain a stronger result. For this, we recall two very useful results.

We proceed in three steps as in [BE02],[KUNIY02] by using a bounded difference inequality

- first, we show that the expectation of $\widehat{R}_{CV} - \widetilde{R}_n$ is small,
- secondly, we show that the function $\widehat{R}_{CV} - \widetilde{R}_n$ seen as a function f of $Z_1, Z_2, \ldots, Z_n$ is strongly difference bounded, i.e.: with high probability, there exists constants $c_1, \ldots, c_n$ such that we have for all i, for all $z \in \mathcal{Z}$,

$$|f(Z_1, \ldots, Z_i, \ldots, Z_n) - f(Z_1, \ldots, Z_{i-1}, z, Z_{i+1}, \ldots, Z_n)| \leq c_i,$$

- use theorem 79 with the first two points,
- at least, use arguments of symmetry to conclude.

1. **The expectation of $\widehat{R}_{CV} - \widetilde{R}_n$ is small**

Let us denote $\mathbf{v}_n^{tr}, \mathbf{v}_n^{ts}$ fixed training and test vectors.

$$\mathbb{P}^{\otimes n}(\widehat{R}_{CV} - \widetilde{R}_n) = \mathbb{P}^{\otimes n}(\mathbb{E}_{V_n^{tr}}\mathbb{P}_{n,V_n^{ts}}\psi_{V_n^{tr}} - \mathbb{P}\psi_n) = \mathbb{P}^{\otimes n}\mathbb{P}(\psi_{\mathbf{v}_n^{tr}} - \psi_n)$$

since

$$\mathbb{P}^{\otimes n}\mathbb{E}_{V_n^{tr}}\mathbb{P}_{n,V_n^{ts}}\psi_{V_n^{tr}} = \mathbb{E}_{V_n^{tr}}\mathbb{P}^{\otimes n}\mathbb{P}_{n,V_n^{ts}}\psi_{V_n^{tr}} = \mathbb{E}_{V_n^{tr}}\mathbb{P}^{\otimes n}\mathbb{P}\psi_{V_n^{tr}} = \mathbb{P}^{\otimes n}\mathbb{P}\psi_{\mathbf{v}_n^{tr}}$$

where the first equality comes from the linearity of expectation, the second from the fact that $\mathbb{P}_{n,V_n^{tr}}$ are independent of $\mathbb{P}_{n,V_n^{ts}}$, and the third from the i.i.d. nature of $(Z_i)_i$.

Recall that $\mathbb{P}(\psi_{\mathbf{v}_n^{tr}} - \psi_n) \leq d(\psi_{\mathbf{v}_n^{tr}}, \psi_n)$ where d stands indifferently for d_1, d_e or d_∞. Thus, $\mathbb{P}^{\otimes n}\mathbb{P}(\psi_{\mathbf{v}_n^{tr}} - \psi_n) \leq \mathbb{P}^{\otimes n} d(\psi_{\mathbf{v}_n^{tr}}, \psi_n)$. By conditioning according to the small values of $d(\psi_{\mathbf{v}_n^{tr}}, \psi_n)$,we obtain

$$\begin{aligned}\mathbb{P}^{\otimes n} d(\psi_{\mathbf{v}_n^{tr}}, \psi_n) &= \mathbb{P}^{\otimes n}(d(\psi_{\mathbf{v}_n^{tr}}, \psi_n)|B)\mathbb{P}^{\otimes n}(B) + \mathbb{P}^{\otimes n}(d(\psi_{\mathbf{v}_n^{tr}}, \psi_n)|B^{\complement})(1 - \mathbb{P}^{\otimes n}(B)) \\ &\leq 1 \times \delta_{n,p_n} + \lambda \mathbb{P}^{\otimes n}||\mathbb{P}_{n,\mathbf{v}_n^{tr}} - \mathbb{P}_n||_\alpha \times (1 - \delta_{n,p_n}) = \delta_{n,p_n} + \lambda(2p_n)^\alpha(1 - \delta_{n,p_n})\end{aligned}$$

Eventually, we get $\mathbb{P}^{\otimes n}(\widehat{R}_{CV} - \widetilde{R}_n) \leq \delta_{n,p_n} + \lambda(2p_n)^\alpha$.

2. $\widehat{R}_{CV} - \widetilde{R}_n$ **is difference bounded with high probability**

Denote $f(Z_1, Z_2, \ldots, Z_n) := \widehat{R}_{CV} - \widetilde{R}_n$. Let $z \in \mathcal{Z}$. Let $\mathcal{D}_{n+1} = \mathcal{D}_{n+1} \cup \{z\}$. Now denote $B = B_1 \cup B_2$ where

$$B_1 = \{\sup_{U_n \in \text{support}(\mathbb{Q})} \frac{d(\psi_{U_n}, \psi_n)}{||\mathbb{P}_{n,U_n} - \mathbb{P}_n||_\alpha} \geq \lambda\}$$

and

$$B_2 = \{\sup_{1 \leq i \leq n+1} \frac{d(\psi_{e_{n+1}^i}, \psi_{n+1})}{||\mathbb{P}_{n+1,e_{n+1}^i} - \mathbb{P}_{n+1}||_\alpha} \geq \lambda\}$$

with e_{n+1}^i the binary of size $n+1$ equal to 0 everywhere except on the i-th coordinate $e_{n+1,k}^i := 1_{(k \neq i)}$ for $1 \leq k \leq n+1$. Under our assumptions, we have

$$\Pr(B) \leq \delta_{n,p_n} + (n+1)\delta_{n+1,1/n+1}$$

.

We want to show that with high probability there exist constants c_i such that for all $i \in \{1, \ldots, n\}$, for all $z \in \mathcal{Z}$, $|f(Z_1, \ldots, Z_i, \ldots, Z_n) - f(Z_1, \ldots, Z_{i-1}, z, Z_{i+1}, \ldots, Z_n)| \leq c_i$. Notice that

$$\begin{aligned}|f(Z_1, \ldots, Z_i, \ldots, Z_n) - f(Z_1, \ldots, z, \ldots, Z_n)| &= |(\mathbb{E}_{V_n^{tr}} \mathbb{P}_{n,V_n^{ts}} \psi_{V_n^{tr}} - \mathbb{P}\psi_{e_{n+1}^{n+1}}) \\ &\quad - (\mathbb{E}_{V_n^{tr}} \mathbb{P}'_{n,V_n^{ts}} \psi'_{V_n^{tr}} - \mathbb{P}\psi_{e_{n+1}^i})| \\ &\leq |\mathbb{E}_{V_n^{tr}} \mathbb{P}_{n,V_n^{ts}} \psi_{V_n^{tr}} - \mathbb{E}_{V_n^{tr}} \mathbb{P}'_{n,V_n^{ts}} \psi'_{V_n^{tr}}| \\ &\quad + |\mathbb{P}\psi_{e_{n+1}^{n+1}} - \mathbb{P}\psi_{e_{n+1}^i}|.\end{aligned}$$

with $\mathbb{P}^{'}_{n,V_n^{tr}}$ the weighted empirical measure on the sample

$$\mathcal{E}_n = \{Z_1, \ldots, Z_{i-1}, z, Z_{i+1}, \ldots, Z_n\}$$

and $\psi^{'}_{V_n^{tr}}$ the predictor trained on $\mathcal{E}_{V_n^{tr}}$.

So, first, let us bound the second term, recall that

$$|\mathbb{P}(\psi_{e^{n+1}_{n+1}} - \psi_{e^{i}_{n+1}})| \leq d(\psi_{e^{n+1}_{n+1}}, \psi_{e^{i}_{n+1}}) \leq d(\psi_{e^{n+1}_{n+1}}, \psi_{n+1}) + d(\psi_{n+1}, \psi_{e^{i}_{n+1}})$$

with ψ_{n+1} trained on $\mathcal{D}_{n+1} = \{Z_1, \ldots, Z_{i-1}, Z_i, Z_{i+1}, \ldots, Z_n, z\}$. Thus, we have on $B^{\subset}$, $|\mathbb{P}\psi_{e^{n+1}_{n+1}} - \mathbb{P}\psi_{e^{i}_{n+1}}| \leq 2(\frac{2\lambda}{n+1})^{\alpha}$.

To upper bound the first term, notice that

$$\begin{aligned}|\mathbb{E}_{V_n^{tr}}\mathbb{P}_{n,V_n^{ts}}\psi_{V_n^{tr}} - \mathbb{E}_{V_n^{tr}}\mathbb{P}^{'}_{n,V_n^{ts}}\psi^{'}_{V_n^{tr}}| =& |\mathbb{E}_{V_n^{tr}}(\mathbb{P}_{n,V_n^{ts}}(\psi_{V_n^{tr}} - \psi^{'}_{V_n^{tr}})|V^{tr}_{n,i} = 1) \times (1 - p_n) \\ &+ \mathbb{E}_{V_n^{tr}}((\mathbb{P}_{n,V_n^{ts}} - \mathbb{P}^{'}_{n,V_n^{ts}})\psi_{V_n^{tr}}|V^{ts}_{n,i} = 1) \times p_n|.\end{aligned}$$

We always have for any ψ, $|(\mathbb{P}_{n,V_n^{ts}} - \mathbb{P}^{'}_{n,V_n^{ts}})\psi| \leq 1/np_n$ thus

$$|\mathbb{E}_{V_n^{tr}}((\mathbb{P}_{n,V_n^{ts}} - \mathbb{P}^{'}_{n,V_n^{ts}})\psi_{V_n^{tr}}|V^{ts}_{n,i} = 1) \times p_n| \leq 1/n$$

.

Until now, the previous lines hold independently of $d \in \{d_e, d_1, d_\infty\}$. We still have to bound $|\mathbb{E}_{V_n^{tr}}(\mathbb{P}_{n,V_n^{ts}}(\psi_{V_n^{tr}} - \psi^{'}_{V_n^{tr}})|V^{tr}_{n,i} = 1)|$. In the particular case of the most stable kind of stability (i.e. when $d = d_\infty$), we have

$$|\mathbb{E}_{V_n^{tr}}(\mathbb{P}_{n,V_n^{ts}}(\psi_{V_n^{tr}} - \psi^{'}_{V_n^{tr}})|V^{tr}_{n,i} = 1)| \leq \mathbb{E}_{V_n^{tr}}(d_\infty(\psi_{V_n^{tr}}, \psi^{'}_{V_n^{tr}})|V^{tr}_{n} = 1).$$

On $B^{\subset}$, we get $d_\infty(\psi_{V_n^{tr}}, \psi^{'}_{V_n^{tr}}) \leq d_\infty(\psi_{V_n^{tr}}, \psi_{n+1}) + d_\infty(\psi_{n+1}, \psi^{'}_{V_n^{tr}}) \leq 2(2\lambda p_n)^{\alpha}$.

Thus, on $B^{\subset}$, we have

$$\mathbb{E}_{V_n^{tr}}(d_\infty(\psi_{V_n^{tr}}, \psi^{'}_{V_n^{tr}})|V^{tr}_{n,i} = 1) \leq 2(2\lambda p_n)^{\alpha}.$$

Putting all together, with probability at least $1 - \delta^{'}_{n,p_n}$,

$$\sup_{1 \leq i \leq n, z \in \mathcal{Z}} |f(Z_1, \ldots, Z_i, \ldots, Z_n) - f(Z_1, \ldots, z, \ldots, Z_n)| \leq 5(2\lambda p_n)^{\alpha}.$$

3. $\widehat{R}_{CV} - \widetilde{R}_n$ **is closed to zero with high probability**

Applying theorem 79, we obtain that for all $\varepsilon \geq 0$

$$\begin{aligned}\Pr(\widehat{R}_{CV} - \widetilde{R}_n \geq \varepsilon + \delta + \lambda(2p_n)^\alpha) &\leq \Pr\left(\widehat{R}_{CV} - \widetilde{R}_n - \mathbb{E}_{\mathcal{D}_n}(\widehat{R}_{CV} - \widetilde{R}_n) \geq \varepsilon\right) \\ &\leq 2(\exp(-\frac{\varepsilon^2}{8n(5(2\lambda p_n)^\alpha + \alpha')^2}) + \frac{n}{\alpha'}\delta') \\ &\leq 2(\exp(-\frac{\varepsilon^2}{8(10\lambda)^2 n(2\lambda p_n)^{2\alpha}}) + \frac{n}{5(2\lambda p_n)^\alpha}\delta') \\ &\text{by taking } \alpha' = 5(2\lambda p_n)^\alpha.\end{aligned}$$

By symmetry, we also have $\Pr\left(\widehat{R}_{CV} - \widetilde{R}_n \leq -(\varepsilon + \delta_{n,p_n} + 2\lambda p_n)\right) \leq 2(\exp(-\frac{\tau^2}{8(10\lambda)^2 n(2\lambda p_n)^{2\alpha}}) + \frac{n}{5(2\lambda p_n)^\alpha}\delta'_{n,pn})$ which allows to conclude.
□

Theorem 81 (Strong stability)**.** *Suppose that $\mathcal{H}$ holds. Let Ψ be a machine learning which is strong $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ stable. Then, for all $\varepsilon \geq 0$, we get*

$$\Pr\left(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \lambda(2p_n)^\alpha\right) \leq 2\exp(-2np_n\varepsilon^2) + \kappa(n)\delta_n,$$

where $\kappa(n)$ is the number of training vectors in the cross-validation.
Furthermore, if the distance d is the uniform distance d_∞, then we have for any $\varepsilon \geq 0$:

$$\Pr\left(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \delta_{n,p_n} + \lambda(2p_n)^\alpha\right) \leq 4(\exp(-\frac{\varepsilon^2}{8n(5(2\lambda p_n)^\alpha + \alpha')^2}) + \frac{n}{\alpha'}\kappa(n)\delta'_{n,p_n}),$$

with $\delta'_{n,p_n} = \delta_{n,p_n} + (n+1)\delta_{n,1/n}$. Thus, if we take $\alpha = 5(2\lambda p_n)^\alpha$, we get

$$\Pr\left(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \delta_{n,p_n} + \lambda(2p_n)^\alpha\right) \leq 4(\exp(-\frac{\varepsilon^2}{8(10\lambda)^2 n(2\lambda p_n)^{2\alpha}}) + \frac{n}{5(2\lambda p_n)^\alpha}\kappa(n)\delta'_{n,p_n}).$$

Proof
For the first inequality, it is sufficient to use remarks 68.
For the second one, we can follow the previous proof, using remarks 68 and noticing that if we denote $B_{\mathbf{v}_n^{tr}} := \{d(\psi_{\mathbf{v}_n^{tr}}, \psi_n) \geq \lambda || \mathbb{P}_{n,\mathbf{v}_n^{tr}} - \mathbb{P}_n ||\}$, then, we have

$$\begin{aligned}\mathbb{P}^{\otimes n} d(\psi_{\mathbb{P}_{\mathbf{v}_n^{tr}}}, \psi_n) &= \mathbb{P}^{\otimes \mathbf{n}}(d(\psi_{\mathbf{v}_n^{tr}}, \psi_n)|B_{\mathbf{v}_n^{tr}})\mathbb{P}^{\otimes \mathbf{n}}(B_{\mathbf{v}_n^{tr}}) + \mathbb{P}^{\otimes \mathbf{n}}(d(\psi_{\mathbf{v}_n^{tr}}, \psi_n)|B^{\subset}_{\mathbf{v}_n^{tr}})(1 - \mathbb{P}^{\otimes \mathbf{n}}(B_{\mathbf{v}_n^{tr}})) \\ &\leq 1 \times \delta_{n,p_n} + \lambda \mathbb{P}^{\otimes \mathbf{n}} || \mathbb{P}_{n,\mathbf{v}_n^{tr}} - \mathbb{P}_n ||_\alpha \times (1 - \delta_{n,p_n}) = \delta_{n,p_n} + \lambda(2p_n)^\alpha(1 - \delta_{n,p_n}).\end{aligned}$$

Eventually, we get $\mathbb{P}^{\otimes n}(\widehat{R}_{CV} - \widetilde{R}_n) \leq \delta_{n,p_n} + \lambda(2p_n)^\alpha$.
□

Now, we derive results for the hold-out cross-validation which does not make a symmetrical use of the dataset. We obtain

Theorem 82 (Strong stability and hold-out)**.** *Let Ψ be a machine learning which is strong $(\lambda, (\delta_{n,p_n})_{n,p_n}, \delta)$ stable. Then the hold-out (or split sample) cross-validation satisfies for all $\varepsilon \geq 0$,*

$$\Pr\left(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \lambda(2p_n)^\alpha\right) \leq 2\exp(-2np_n\varepsilon^2) + \delta_{n,p_n}.$$

Furthermore, if the distance is the uniform distance d_∞, then we have

$$\begin{aligned}\Pr\left(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \delta_{n,p_n} + \lambda(2p_n)^\alpha\right) \leq 4(&\exp(-\frac{\varepsilon^2}{8(4\lambda(2p_n)^\alpha + 1/np_n)^2}) \\ &+ \frac{n^2}{4\lambda(2p_n)^\alpha + 1/np_n}\delta^{'}_{n,p_n}),\end{aligned}$$

with $\delta^{'}_{n,p_n} = \delta_{n,p_n} + n\delta_{n,1/n}$

Proof
For the first inequality, it is enough to use remarks 68.
For the second one, we start as previously. First, we bound in the same way the expectation.

Secondly, we show that $\widehat{R}_{CV} - \widetilde{R}_n$ is difference-bounded with high probability.
Denote $f(Z_1, Z_2, \ldots, Z_n) := \widehat{R}_{CV} - \widetilde{R}_n$. Let $z \in \mathcal{Z}$. Now denote as previously $B := B_1 \cup B_2$ with $B_1 = \{\frac{d(\psi_{\mathbf{v}_n^{tr}}, \psi_n)}{||\mathbb{P}_{n,\mathbf{v}_n^{tr}} - \mathbb{P}_n||_\alpha} \geq \lambda\}$ and $B_2 = \{\sup_i \frac{d(\psi_{e^i_{n+1}}, \psi_{n+1})}{||\mathbb{P}_{n+1,e^i_{n+1}} - \mathbb{P}_{n+1}||_\alpha} \geq \lambda\}$. Eventually, we have $\Pr(B) \leq \delta_{n,1-p_n} + n\delta_{n,1/n}$

We want to show that with high probability there exists constants c_i such that for all i, for all $z \in \mathcal{Z}$, $|f(Z_1, \ldots, Z_i, \ldots, Z_n) - f(Z_1, \ldots, z, \ldots, Z_n)| \leq c_i$. Since $V_n^{tr} = \mathbf{v}_n^{ts}$ fixed vector in the case of hold-out, notice that:

$$\begin{aligned}|f(Z_1, \ldots, Z_i, \ldots, Z_n) - f(Z_1, \ldots, z, \ldots, Z_n)| &= |\mathbb{P}_{n,\mathbf{v}_n^{ts}}\psi_{\mathbf{v}_n^{tr}} - \mathbb{P}\psi_{e^{n+1}_{n+1}}) - (\mathbb{P}^{'}_{n,\mathbf{v}_n^{ts}}\psi^{'}_{\mathbf{v}_n^{tr}} - \mathbb{P}\psi_{e^i_{n+1}})| \\ &\leq |\mathbb{P}_{n,\mathbf{v}_n^{ts}}\psi_{\mathbf{v}_n^{tr}} - \mathbb{E}_{\mathbf{v}_n^{tr}}\mathbb{P}^{'}_{n,\mathbf{v}_n^{ts}}\psi^{'}_{\mathbf{v}_n^{tr}}| \\ &\quad + |\mathbb{P}\psi_{e^{n+1}_{n+1}} - \mathbb{P}\psi_{e^i_{n+1}}|,\end{aligned}$$

with $\mathbb{P}^{'}_{n,\mathbf{v}_n^{tr}}$ the weighted empirical measures of the sample $\mathcal{E}_n = \{Z_1, \ldots, Z_{i-1}, z, Z_{i+1}, \ldots, Z_n\}$ and $\psi^{'}_{\mathbf{v}_n^{tr}}$ the predictor trained on $\mathcal{E}_{\mathbf{v}_n^{tr}}$.

So, first, let us bound the second term, recall that:

$$|\mathbb{P}(\psi_{e^{n+1}_{n+1}} - \psi_{e^i_{n+1}})| \leq d(\psi_{e^{n+1}_{n+1}}, \psi_{e^i_{n+1}}) \leq d(\psi_{e^{n+1}_{n+1}}, \psi_{n+1}) + d(\psi_{n+1}, \psi_{e^i_{n+1}})$$

.
Thus, on $B^\subset$, $|\mathbb{P}\psi_{e^{n+1}_{n+1}} - \mathbb{P}\psi_{e^i_{n+1}}| \leq 2\lambda(\frac{2}{n+1})^\alpha$.

To upper bound the first term, notice that:

$$|\mathbb{P}_{n,\mathbf{v}_n^{ts}}\psi_{\mathbf{v}_n^{tr}} - \mathbb{P}'_{n,\mathbf{v}_n^{ts}}\psi'_{\mathbf{v}_n^{tr}}| = |\mathbb{P}_{n,\mathbf{v}_n^{ts}}(\psi_{\mathbf{v}_n^{tr}} - \psi'_{\mathbf{v}_n^{tr}})1_{\{\mathbf{v}_{n,i}^{tr}=1\}} + (\mathbb{P}_{n,\mathbf{v}_n^{ts}} - \mathbb{P}'_{n,\mathbf{v}_n^{ts}})\psi_{\mathbf{v}_n^{tr}}1_{\{\mathbf{v}_{n,i}^{ts}=1\}}|$$

We always have for any ψ, $|(\mathbb{P}_{n,\mathbf{v}_n^{ts}} - \mathbb{P}'_{n,\mathbf{v}_n^{ts}})\psi| \leq 1/np_n$ thus

$$|(\mathbb{P}_{n,\mathbf{v}_n^{ts}} - \mathbb{P}'_{n,\mathbf{v}_n^{ts}})\psi_{\mathbf{v}_n^{tr}}1_{\{\mathbf{v}_{n,i}^{ts}=1\}}| \leq 1/np_n$$

.

We still have to bound $|\mathbb{P}_{n,\mathbf{v}_n^{ts}}(\psi_{\mathbf{v}_n^{tr}} - \psi'_{\mathbf{v}_n^{tr}})1_{\{\mathbf{v}_{n,i}^{tr}=1\}}|$. As in the previous proof, we have when $d = d_\infty$, $|\mathbb{P}_{n,\mathbf{v}_n^{ts}}(\psi_{\mathbf{v}_n^{tr}} - \psi'_{\mathbf{v}_n^{tr}})1_{\{\mathbf{v}_{n,i}^{tr}=1\}}| \leq d_\infty(\psi_{\mathbf{v}_n^{tr}}, \psi'_{\mathbf{v}_n^{tr}})1_{\{\mathbf{v}_{n,i}^{tr}=1\}}$
On $B^\subset$, $d_\infty(\psi_{\mathbf{v}_n^{tr}}, \psi'_{\mathbf{v}_n^{tr}}) \leq d_\infty(\psi_{\mathbf{v}_n^{tr}}, \psi_{n+1}) + d_\infty(\psi_{n+1}, \psi'_{\mathbf{v}_n^{tr}}) \leq 2\lambda(2p_n)^\alpha$. Thus, on $B^\subset$ we get $d_\infty(\psi_{\mathbf{v}_n^{tr}}, \psi'_{\mathbf{v}_n^{tr}})1_{\{\mathbf{v}_{n,i}^{tr}=1\}} \leq 2\lambda(2p_n)^\alpha$.

Putting all together, with probability at least $1 - \delta'_{n,p_n}$,

$$\begin{aligned}\sup_{i,z}|f(Z_1,\ldots,Z_i,\ldots,Z_n) - f(Z_1,\ldots,z,\ldots,Z_n)| &\leq 2\lambda(\frac{2}{n+1})^\alpha + \max((np_n)^{-1}, 2\lambda(2p_n)^\alpha) \\ &\leq 4\lambda(2p_n)^\alpha + (np_n)^{-1}\end{aligned}$$

To conclude, apply again theorem 79.
□

2.3.3 Weak stability

We now derive results that stands for general cross-validation procedures and weakly stable predictors. We recall here the interest of the notion of weak stability. For some class of machine learning, the notion of strong stability may be too demanding. That is why weak stability is introduced. As a motivation, algorithms such as Adaboost satisfies the following definition of weak stability.

We will use the definition of weak difference bounded introduced by [KUT02] and a corollary of his main theorem.

Definition 83 (Kutin[KUT02]). *Let $\Omega_1,\ldots,\Omega_n$ be probability spaces. Let $\Omega = \prod_{k=1}^n \Omega_k$ and let X a random variable on Ω. We say that X is weakly difference bounded by (b,c,δ) if the following holds: for any k,*

$$\forall^\delta(\omega, v) \in \Omega \times \Omega_k,\ \mathbb{P}(|X(\omega) - X(\omega')|) \leq c$$

where $\omega'_k = v$ and $\omega'_i = \omega_i$ for $i \neq k$. and the notation $\forall^\delta\omega, \Phi(\omega)$ means "$\Phi(\omega)$ holds for all but but a δ fraction of Ω"

$$|X(\omega) - X(\omega')| \leq c$$

Furthermore, for any $\omega, \omega' \in \Omega$*, differing only one coordinate:*

$$|X(\omega) - X(\omega')| \leq b$$

We will need the following theorem. It says in substance that a weakly difference bounded function of independent variables is closed to its expectation with probability.

Theorem 84 (Kutin[KUT02])**.** *Let* $\Omega_1, \ldots, \Omega_n$ *be probability spaces. Let* $\Omega = \prod_{k=1}^{n} \Omega_k$ *and let* X *a random variable on* Ω*.which is weakly difference bounded by* (b, c, δ)*. Assume* $b \geq c \geq 0$ *and* $\alpha > 0$*. Let* $\mu = \mathbb{E}(X)$*. Then, for any* $\varepsilon > 0$

$$\Pr(|X - \mu| \geq \varepsilon) \leq 2\exp(-\frac{\varepsilon^2}{10nc^2(1 + \frac{2\varepsilon}{15nc})^2}) + \frac{2nb\delta^{1/2}}{c}\exp(\frac{\varepsilon b}{4nc^2})) + 2n\delta^{1/2}.$$

Theorem 85 (Cross-validation Weak stability)**.** *Suppose that* $\mathcal{H}$ *holds. Let* Ψ *be a machine learning which is weak* $(\lambda, (\delta_n)_n, d, \mathbb{Q})$ *stable with respect to the distance* d*. Then, for all* $\varepsilon \geq 0$*,*

$$\Pr\left(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \lambda(2p_n)^{\alpha}\right) \leq 2\exp(-2np_n\varepsilon^2) + \delta_{n,p_n}$$

Furthermore, if the distance is the uniform distance d_∞*, we have for all* $\varepsilon \geq 0$*:*

$$\begin{aligned}\Pr\left(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \delta_{n,p_n} + \lambda(2p_n)^{\alpha}\right) \leq 4(\exp(-\frac{\varepsilon^2}{10n(5\lambda(2p_n)^{\alpha})^2(1 + \frac{2\varepsilon}{15n(5\lambda(2p_n)^{\alpha})})^2}) \\ + \frac{2n\delta'^{1/2}_{n,p_n}}{5\lambda(2p_n)^{\alpha}}\exp(\frac{\varepsilon n}{4n(5\lambda(2p_n)^{\alpha})^2})) + n\delta'^{1/2}_{n,p_n})),\end{aligned}$$

with $\delta'_{n,p_n} = 2\delta_{n,1/n} + \delta_{n,p_n}$

Proof
In the following, denote B the bad subset, i.e. $B = \cup_{v_n^{tr}} B_{v_n^{tr}}$ with $B_{v_n^{tr}} = \{d(\psi_{\mathbb{P}_{n,v_n^{tr}}}, \psi_{\mathbb{P}_n}) \geq \lambda||\mathbb{P}_{n,v_n^{tr}} - \mathbb{P}_n||\}$. Since Ψ is strong $(\lambda, (\delta_n)_{n,p_n}, d, \mathbb{Q})$ stable, we have $\mathbb{P}(B) \leq \delta_{n,p_n}$.

1. For the general case, it is again sufficient to split $|\widehat{R}_{CV} - \widetilde{R}_n|$ according to the same benchmark, namely $\bar{R}_{n(1-p)} = \mathbb{E}_{V_n^{tr}}\mathbb{P}\psi_{V_n^{tr}}$.

   Thus,

   $$\Pr\left(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \lambda(2p_n)^{\alpha}\right) \leq \Pr\left(|\widehat{R}_{CV} - \bar{R}_{n(1-p)}| \geq \varepsilon\right) + \Pr\left(|\bar{R}_{n(1-p)} - \widetilde{R}_n| \geq \lambda(2p_n)^{\alpha}\right)$$

   The first term can be bounded as previously by $2\exp(-2np_n\varepsilon^2)$.

For the second term, notice that $|\bar{R}_{n(1-p)} - \widetilde{R}_n| = |\mathbb{E}_{V_n^{tr}} \mathbb{P}\psi_{V_n^{tr}} - \mathbb{E}_{V_n^{tr}} \mathbb{P}\psi_n| \leq \mathbb{E}_{V_n^{tr}} |\mathbb{P}\psi_{V_n^{tr}} - \mathbb{P}\psi_n|$. Recall that $|\mathbb{P}\psi_{V_n^{tr}} - \mathbb{P}\psi_n| \leq d(\psi_{V_n^{tr}}, \psi_n)$ and $||\mathbb{P}_{n,V_n^{tr}} - \mathbb{P}_n||^\alpha = \lambda(2p_n)^\alpha$. Thus, since Ψ is weak $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ stable, we have

$$\begin{aligned}
\Pr\left(|\hat{R}_{n(1-p_n)} - \widetilde{R}_n| \geq \lambda(2p_n)^\alpha\right) &\leq \Pr\left(\mathbb{E}_{v_n^{tr}} d(\psi_{v_n^{tr}}, \psi_n) \geq \lambda(2p_n)^\alpha\right) \\
&\leq \Pr\left(\cup_{v_n^{tr}} \{d(\psi_{v_n^{tr}}, \psi_n) \geq \lambda ||\mathbb{P}_{n,v_n^{tr}} - \mathbb{P}_n||_\alpha\}\right) \\
&= \Pr\left(\cup_{v_n^{tr}} B_{v_n^{tr}}\right) \leq \kappa(n)\delta_{n,p_n}.
\end{aligned}$$

2. In the particular case, when $d = d_\infty$, we can also obtain a stronger result.

We proceed in three steps as in [BE02],[KUNIY02] by using a bounded difference inequality:

1. first, we show that the expectation of $\widehat{R}_{CV} - \widetilde{R}_n$ is small of the same order as for the strong stability.
2. secondly, we show that the function $\widehat{R}_{CV} - \widetilde{R}_n$ seen as a function f of $Z_1, Z_2, \ldots, Z_n$ is weakly difference bounded, i.e. there exists constants $c_1, \ldots, c_n$ such that for all i, if $Z_1, \ldots, Z_i, \ldots, Z_n, Z_{i'}$ i.i.d. random variables, we have with high probability

$$|f(Z_1, \ldots, Z_i, \ldots, Z_n) - f(Z_1, \ldots, Z_{i'}, \ldots, Z_n)| \leq c_i.$$

3. finally, we use theorem 84 with the first two points to conclude.

1. **The expectation of $\widehat{R}_{CV} - \widetilde{R}_n$ is small**

As previously, denote $\mathbf{v}_n^{tr}, \mathbf{v}_n^{ts}$ fixed vectors. We still have

$$\mathbb{P}^{\otimes n}(\widehat{R}_{CV} - \widetilde{R}_n) = \mathbb{P}^{\otimes n}(\mathbb{E}_{V_n^{tr}} \mathbb{P}_{n,V_n^{ts}} \psi_{V_n^{tr}} - \mathbb{P}\psi_n) = \mathbb{P}^{\otimes n}\mathbb{P}(\psi_{\mathbf{v}_n^{tr}} - \psi_n).$$

since $\mathbb{P}^{\otimes n}\mathbb{E}_{V_n^{tr}}\mathbb{P}_{n,V_n^{ts}}\psi_{V_n^{tr}} = \mathbb{E}_{V_n^{tr}}\mathbb{P}^{\otimes n}\mathbb{P}_{n,V_n^{ts}}\psi_{V_n^{tr}} = \mathbb{E}_{V_n^{tr}}\mathbb{P}^{\otimes n}\mathbb{P}\psi_{V_n^{tr}} = \mathbb{P}^{\otimes n}\mathbb{P}\psi_{\mathbf{v}_n^{tr}}$ where the first equality comes from the linearity of expectation, the second from the fact that $\mathbb{P}_{n,V_n^{tr}}$ are independent of $\mathbb{P}_{n,V_n^{ts}}$, and the third one from the *i.i.d.* nature of $(Z_i)_i$.

Recall that $\mathbb{P}(\psi_{\mathbf{v}_n^{tr}} - \psi_n) \leq d(\psi_{\mathbf{v}_n^{tr}}, \psi_n)$ where d stands indifferently for d_1, d_e or d_∞. Thus, $\mathbb{P}^{\otimes n}\mathbb{P}(\psi_{\mathbf{v}_n^{tr}} - \psi_n) \leq \mathbb{P}^{\otimes n} d(\psi_{\mathbf{v}_n^{tr}}, \psi_n)$. By conditioning according to the small values of $d(\psi_{\mathbf{v}_n^{tr}}, \psi_n)$, we obtain

$$\begin{aligned}
\mathbb{P}^{\otimes n} d(\psi_{\mathbf{v}_n^{tr}}, \psi_n) &= \mathbb{P}^{\otimes \mathbf{n}}(d(\psi_{\mathbf{v}_n^{tr}}, \psi_n)|B_{\mathbf{v}_n^{tr}})\mathbb{P}^{\otimes \mathbf{n}}(B_{\mathbf{v}_n^{tr}}) \\
&\quad + \mathbb{P}^{\otimes \mathbf{n}}(d(\psi_{\mathbf{v}_n^{tr}}, \psi_n)|B_{\mathbf{v}_n^{tr}}^{\subset})(1 - \mathbb{P}^{\otimes \mathbf{n}}(B_{\mathbf{v}_n^{tr}})) \\
&\leq 1 \times \delta_{n,p_n} + \lambda \mathbb{P}^{\otimes \mathbf{n}} ||\mathbb{P}_{n,\mathbf{v}_n^{tr}} - \mathbb{P}_n||_\alpha \times (1 - \delta_{n,p_n}) \leq \delta_{n,p_n} + \lambda(2p_n)^\alpha.
\end{aligned}$$

Eventually, we still have $\mathbb{P}^{\otimes n}(\widehat{R}_{CV} - \widetilde{R}_n) \leq \delta_{n,p_n} + \lambda(2p_n)^\alpha$.

2. $\widehat{R}_{CV} - \widetilde{R}_n$ **is difference bounded with high probability**
Denote $f(Z_1, Z_2, \ldots, Z_n) := \widehat{R}_{CV} - \widetilde{R}_n$.

We want to show that for all i, there exists constant c_i such

$$|f(Z_1, \ldots, Z_i, \ldots, Z_n) - f(Z_1, \ldots, Z_i^{'}, \ldots, Z_n)| \leq c_i$$

with high probability where $Z_1, \ldots, Z_i, \ldots, Z_n, Z_i^{'}$ are i.i.d. variables. Denote

$$B_i = \{\frac{d(\psi_{e^i_{n+1}}, \psi_{n+1})}{||\mathbb{P}_{n+1,e^i_{n+1}} - \mathbb{P}_{n+1}||_\alpha} \geq \lambda\}.$$

We proceed as previously where $(\cup_{v_n^{tr}} B_{v_n^{tr}}) \cup B_i \cup B_{n+1}$ will play the role of B.

$$\begin{aligned}|f(Z_1, \ldots, Z_i, \ldots, Z_n) - f(Z_1, \ldots, Z_i', \ldots, Z_n)| &= |(\mathbb{E}_{V_n^{tr}} \mathbb{P}_{n,V_n^{ts}} \psi_{V_n^{tr}} - \mathbb{P}\psi_n) - \\ &\quad (\mathbb{E}_{V_n^{tr}} \mathbb{P}^{'}_{n,V_n^{ts}} \psi^{'}_{V_n^{tr}} - \mathbb{P}\psi^{'}_n)| \\ &\leq |\mathbb{E}_{V_n^{tr}} \mathbb{P}_{n,V_n^{ts}} \psi_{V_n^{tr}} - \mathbb{E}_{V_n^{tr}} \mathbb{P}^{'}_{n,V_n^{ts}} \psi^{'}_{V_n^{tr}}| \\ &\quad + |\mathbb{P}\psi_n - \mathbb{P}\psi^{'}_n|,\end{aligned}$$

with $\mathbb{P}^{'}_n, \mathbb{P}^{'}_{n,V_n^{ts}}$ the weighted empirical measures of the sample

$$\mathcal{D}^{'}_n = \{Z_1, \ldots, Z^{'}_i, \ldots, Z_n\}$$

and $\psi^{'}_n$ the predictor built on $\mathcal{D}^{'}_n$.

So, first, let us bound the second term, recall that: $|\mathbb{P}(\psi_n - \psi^{'}_n)| \leq d(\psi_n, \psi^{'}_n) \leq d(\psi_n, \psi_{n+1}) + d(\psi_{n+1}, \psi^{'}_n)$.with ψ_{n+1} the predictor trained on the sample $\mathcal{D}_{n+1} = \{Z_1, \ldots, Z_i, \ldots, Z_n, Z_{i'}\}$. Thus, on $B^{\subset}$, we have $|\mathbb{P}\psi_n - \mathbb{P}\psi^{'}_n| \leq 2\lambda(2/n)^\alpha$.

To upper bound the first term, notice that

$$\begin{aligned}|E_{V_n^{tr}} P_{n,V_n^{ts}} \psi_{V_n^{tr}} - E_{V_n^{tr}} P^{'}_{n,V_n^{ts}} \psi^{'}_{V_n^{tr}}| &= |E_{V_n^{tr}}(P_{n,V_n^{ts}}(\psi_{V_n^{tr}} - \psi^{'}_{V_n^{tr}})|V^{tr}_{n,i} = 1) \times (1 - p_n) \\ &\quad + E_{V_n^{tr}}((P_{n,V_n^{ts}} - P^{'}_{n,V_n^{ts}})\psi_{V_n^{tr}}|V^{ts}_{n,i} = 1) \times p_n|.\end{aligned}$$

We always have for all ψ, $|(\mathbb{P}_{n,V_n^{ts}} - \mathbb{P}^{'}_{n,V_n^{ts}})\psi| \leq 1/np_n$ thus we get

$$|\mathbb{E}_{V_n^{tr}}((\mathbb{P}_{n,V_n^{ts}} - \mathbb{P}^{'}_{n,V_n^{ts}})\psi_{V_n^{tr}}, V_n^{ts} = 1) \times p_n| \leq 1/n$$

We still have to bound

$$|\mathbb{E}_{V_n^{tr}}(\mathbb{P}_{n,V_n^{ts}}(\psi_{V_n^{tr}} - \psi'_{V_n^{tr}})|V_{n,i}^{tr} = 1)| \leq \mathbb{E}_{V_n^{tr}}(d_\infty(\psi_{V_n^{tr}}, \psi'_{V_n^{tr}})|V_{n,i}^{tr} = 1)$$

On $B_{v_n^{tr}}^{\subset}$, $d_\infty(\psi_{v_n^{tr}}, \psi_{v_n^{tr}}) \leq d_\infty(\psi_{v_n^{tr}}, \psi_{n+1}) + d_\infty(\psi_{n+1}, \psi'_{v_n^{tr}}) \leq 2\lambda(2p_n)^\alpha$.
Thus, we get $\mathbb{E}_{V_n^{tr}}(d_\infty(\psi_{V_n^{tr}}, \psi'_{V_n^{tr}}), V_n^{tr} = 1) \leq 2\lambda(2p_n)^\alpha$ on $(\cup_{v_n^{tr}} B_{v_n^{tr}})^{\subset}$.

Putting all together, with probability at least $1 - 2\delta_{n+1,1/(n+1)} - \delta_{n,p_n}$,

$$|f(Z_1, \ldots, Z_i, \ldots, Z_n) - f(Z_1, \ldots, Z_{i'}, \ldots, Z_n)| \leq 5\lambda(2p_n)^\alpha.$$

3. $\widehat{R}_{CV} - \widetilde{R}_n$ **is closed to zero with high probability**

Applying theorem 84, we obtain for all $\varepsilon \geq 0$:

$$\begin{aligned}\Pr(\widehat{R}_{CV} - \widetilde{R}_n \geq \varepsilon + \delta_{n,p_n} + \lambda(2p_n)^\alpha) &\leq 2(\exp(-\frac{\varepsilon^2}{10n(5\lambda(2p_n)^\alpha)^2(1+\frac{2\varepsilon}{15n(5\lambda(2p_n)^\alpha)})^2}) \\ &\quad + \frac{2n\delta_{n,p_n}^{'1/2}}{5\lambda(2p_n)^\alpha}\exp(\frac{\varepsilon n}{4n(5\lambda(2p_n)^\alpha)^2})) + n\delta_{n,p_n}^{'1/2}) \\ &\leq 2(\exp(-\frac{\varepsilon^2}{10n(5\lambda(2p_n)^\alpha)^2(1+\frac{2\varepsilon}{15n(5\lambda(2p_n)^\alpha)})^2}) \\ &\quad + \frac{2n\delta_{n,p_n}^{'1/2}}{5\lambda(2p_n)^\alpha}\exp(\frac{\varepsilon n}{4n(5\lambda(2p_n)^\alpha)^2})) + n\delta_{n,p_n}^{'1/2}).\end{aligned}$$

By symmetry, we also upper bound $\Pr(\widehat{R}_{CV} - \widetilde{R}_n \leq -(\varepsilon + \delta_{n,p_n} + \lambda(2p_n)^\alpha))$ by the same quantity.

□

Theorem 86 (Weak stability)**.** *Suppose that $\mathcal{H}$ holds. Let Ψ be a machine learning which is weak $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$. Then for all $\varepsilon \geq 0$, we have*

$$\Pr(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \lambda(2p_n)^\alpha) \leq 2\exp(-2np_n\varepsilon^2) + \kappa(n)\delta_{n,p_n}$$

where $\kappa(n)$ is the number of elements in the cross-validation.
Furthermore, if the distance is the uniform distance d_∞, we have

$$\begin{aligned}\Pr(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \delta_{n,p_n} + \lambda(2p_n)^\alpha) &\leq 4(\exp(-\frac{\varepsilon^2}{10n(5\lambda(2p_n)^\alpha)^2(1+\frac{2\varepsilon}{15n(5\lambda(2p_n)^\alpha)})^2}) \\ &\quad + \frac{2n\delta_{n,p_n}^{'1/2}}{5\lambda(2p_n)^\alpha}\exp(\frac{\varepsilon n}{4n(5\lambda(2p_n)^\alpha)^2})) + n\delta_{n,p_n}^{'1/2}))\end{aligned}$$

with $\delta'_{n,p_n} = \delta_{n,1/n} + \kappa(n)\delta_{n,p_n}$

Proof.

For the first inequality, it is enough to use remarks 68.
For the second, it is enough to follow the previous proofs and to notice that $\Pr(\cup_{v_n^{tr}} B_{v_n^{tr}}) \leq \kappa(n)\delta_{n,p_n}$.
□

Similar results for hold-out can be derived in the spirit of proposition 82. We can now use the previous probability upper bounds to derive upper bounds for the expectation of $|\widehat{R}_{CV} - \widetilde{R}_n|$.

2.3.4 Results for the L_1 norm

For the sake of simplicity, we suppose here that $\alpha = 1$. In the general case, we just consider the weakest notion: weak $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ stability.

Theorem 87 (L_1 norm of cross-validation estimate)**.** *Suppose that $\mathcal{H}$ holds. Let Ψ be a machine learning which is weak $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ stable. Then, we have*

$$\mathbb{E}_{\mathcal{D}_n}|\widehat{R}_{CV} - \widetilde{R}_n| \leq 2\lambda p_n + \sqrt{\frac{2}{np_n}} + \delta_{n,p_n}.$$

Furthermore, if Ψ is a machine learning which is strong $(\lambda, (\delta_n)_n, d_\infty, \mathbb{Q})$ stable, we have

$$\mathbb{E}_{\mathcal{D}_n}|\widehat{R}_{CV} - \widetilde{R}_n| \leq \delta_{n,p_n} + 2\lambda p_n + 51\lambda\sqrt{n}p_n + \frac{n}{9\lambda p_n}\delta'_{n,p_n},$$

with $\delta'_{n,p_n} = \delta_{n,p_n} + (n+1)\delta_{n+1,1/n+1}$

Proof.
These inequalities are a consequence of the previous propositions and of the following lemma (for a proof, see e.g. [DGL96]):

Lemma 88. *Let X be a nonnegative random variable. Let K, C nonnegative real such that $C \geq 1$. Suppose that for all $\varepsilon > 0$, $\mathbb{P}(X \geq \varepsilon) \leq C\exp(-K\varepsilon^2)$. Then:*

$$\mathbb{E}X \leq \sqrt{\frac{\ln(C) + 2}{K}}.$$

For the second one, it is enough to follow the previous proofs and to notice that $\Pr(\cup_{V_n^{tr}} B_{V_n^{tr}}) \leq \kappa(n)\delta_{n,p_n}$
□

We deduce that

Corollary 89. *Suppose that $\mathcal{H}$ holds. If Ψ be a machine learning which is weak $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ stable, we define the splitting rule $p_n^\star = (1/\sqrt{2}4\lambda)^{2/3}(1/n)^{1/3}$. Then, we have*

$$\mathbb{E}_{\mathcal{D}_n}|\widehat{R}_{CV} - \widetilde{R}_n| \leq 4(\lambda/n)^{1/3}.$$

Furthermore, if Ψ is a machine learning which is strong $(\lambda, (\delta_n)_n, d_\infty, \mathbb{Q})$ stable, we use leave-one-out cross-validation for n large enough. And we have

$$\mathbb{E}_{\mathcal{D}_n}|\widehat{R}_{CV} - \widetilde{R}_n| = O_n(\lambda/\sqrt{n}).$$

Proof.

Recall that for a large class of learning algorithm, we have in mind that $\delta_{n,p_n} = O_n(p_n \exp(-n(1-p_n)))$. Thus $2\lambda p_n + \sqrt{\frac{2}{np_n}} + \delta_{n,p_n} \leq 4\lambda p_n + \sqrt{\frac{2}{np_n}}$. We can differentiate this last bound seen as a function of p_n. We obtain $p_n^\star = (1/\sqrt{2}4\lambda)^{2/3}(1/n)^{1/3}$. Thus, we deduce that $\mathbb{E}_{\mathcal{D}_n}|\widehat{R}_{CV} - \widetilde{R}_n| \leq 4(\lambda/n)^{1/3}$. If Ψ is a machine learning which is strong $(\lambda, (\delta_n)_n, d_\infty, \mathbb{Q})$ stable, we obtain $\delta_{n,p_n} + 2\lambda p_n + 51\lambda\sqrt{n}p_n + \frac{n}{9\lambda p_n}\delta'_{n,p_n} \leq 4\lambda p_n + 51\lambda\sqrt{n}p_n$ for n large enough since $\frac{n}{9\lambda p_n}\delta'_{n,p_n} = O_n(n^3 \exp(-n/2))$ if $p_n \leq 1/2$. Thus, $p_n^\star = 1/n$ for n large enough and $\mathbb{E}_{\mathcal{D}_n}|\widehat{R}_{CV} - \widetilde{R}_n| = O_n(\lambda/\sqrt{n})$.

□

We have obtained the following conclusions:

- Cross-validation is consistent as an estimator of the generalization error of stable algorithms.
- There is a tradeoff interpretation in the choice of the proportion of elements p_n of the test set: the smaller p_n is, the greater the term $B(n, p_n, \varepsilon)$ is controlled but the less the term $V(n, p_n, \varepsilon)$ is upper bounded.
- In the general setting, our bounds require that the sizes of the training set and the test set grow to infinity.
- In the particular case of the stability with respect to the most stable kind of stability (namely the uniform stability), we can have a stronger result: the number of elements in the test set does need to grow to infinity for the consistency of symmetric cross-validation procedures. But we lose this property with the hold-out cross-validation.
- Symmetric cross-validation out performs hold-out cross-validation for large sets.
- At last, as far as the expectation $\mathbb{E}_{\mathcal{D}_n}|\widehat{R}_{CV} - \widetilde{R}_n|$ is concerned, we can define a splitting rule in the general setting.

2.4 Appendices

2.4.1 Inequalities

We recall three very useful results. The first one, due to [HOEF63], bounds the difference between the empirical mean and the expected value. The second one, due to [VC71], bounds the supremum over the class of predictors of the difference between the training error and the generalization error. The last one is called the bounded differences inequality [McD89].

Theorem 90 (Hoeffding's inequality)**.** *Let $X_{1,...,}X_n$ independent random variables in $[a_i, b_i]$. Then for all $\varepsilon > 0$, we get*

$$\mathbb{P}(\sum X_i - \mathbb{E}(\sum X_i) \geq n\varepsilon) \leq e^{-\frac{2\varepsilon^2}{\sum_i (b_i - a_i)^2}}.$$

Theorem 91 (McDiarmid, [McD89])**.** *Let $X_{1,...,}X_n$ be independent random variables taking values in a set A, and assume that $f : A^n \to \mathcal{R}$ satisfies*

$$\forall i, \sup_{\substack{x_1,...,x_i,...,x_n \\ x_i'}} |f(x_1, ..., x_n) - f(x_1, ..., x_{i'}, ..., x_n)| \leq c_i.$$

Then for all $\varepsilon > 0$, we have

$$\mathbb{P}(f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n) \geq \varepsilon) \leq e^{-\frac{2\varepsilon^2}{\sum_i c_i^2}}.$$

Chapter 3

Estimating Subagging by cross-validation

In this article, we derive concentration inequalities for the cross-validation estimate of the generalization error for subagged estimators, both for classification and regressor. General loss functions and class of predictors with both finite and infinite VC-dimension are considered. We slightly generalize the formalism introduced by [DUD03] to cover a large variety of cross-validation procedures including leave-one-out cross-validation, k-fold cross-validation, hold-out cross-validation (or split sample), and the leave-υ-out cross-validation.

An interesting consequence is that the probability upper bound is bounded by the minimum of a Hoeffding-type bound and a Vapnik-type bounds, and thus is smaller than 1 even for small learning set. Finally, we give a simple rule on how to subbag the predictor.

Keywords: Cross-validation, generalization error, concentration inequality, optimal splitting, resampling.

3.1 Introduction and motivation

One of the main issue of pattern recognition is to create a predictor (a regressor or a classifier) which takes observable inputs in order to predict the unknown nature of an output. Typical applications range from predicting the figures of a digitalized zip code to predicting the chance of survival from clinical measurements. Formally, a predictor ϕ is a measurable map from some measurable space $\mathcal{X}$ to some measurable space $\mathcal{Y}$. When $\mathcal{Y}$ is a countable set (respectively $\mathbb{R}^m$), the predictor is called a classifier (respectively a regressor). The strategy of *Machine Learning* consists in building a learning algorithm Φ from both a set of examples and a class of methods. Typical class of methods are empirical risk minimization or k-nearest neighbors rules. The set of examples consists in the measurement of n observations $(x_i, y_i)_{1 \leq i \leq n}$. Thus, formally, Φ is a measurable map from $\mathcal{X} \times \cup_n (\mathcal{X} \times \mathcal{Y})^n$ to $\mathcal{Y}$. One of the main issue of *Statistical Learning* is to analyse the performance of a learning machine in a probabilistic setting. $(x_i, y_i)_{1 \leq i \leq n}$ are supposed to be observations

from n independent and identically distributed (i.i.d.) random variables $(X_i, Y_i)_{1\leq i\leq n}$ with distribution $\mathbb{P}$. $(X_i, Y_i)_{1\leq i\leq n}$ is denoted $\mathcal{D}_n$ in the following and called the learning set. In order to analyse the performance, it is usual to consider the conditionnal risk of a machine learning Φ denoted $\tilde{R}_n$, so called the generalization error. It is defined by the conditional expectation of $L(Y, \Phi(X, \mathcal{D}_n))$ given $\mathcal{D}_n$ where $(X, Y) \sim \mathbb{P}$ is a random variable independent of $\mathcal{D}_n$, i.e. $\tilde{R}_n := \mathbb{E}_{X,Y}(L(Y, \Phi(X, \mathcal{D}_n))|\mathcal{D}_n)$ with L a cost function from $\mathcal{Y}^2 \longrightarrow \mathbb{R}_+$. Notice that $\widetilde{R}_n$ is a random variable measurable with respect to $\mathcal{D}_n$.

Bagging, to be defined formally below, is a procedure building an estimator by a resample and combine technique. Bagging [bootstrap aggregating] was introduced by [BRE96] to reduce the variance of a predictor. From an original estimator, a bagged regressor is produced by averaging several replicates trained on bootstrap samples, a bagged classifier is produced by voting at the majority. It is one of the recent and successful computationally intensive methods for improving unstable estimation or classification schemes. It is extremely useful for large, high dimensional data set problems where finding a good model or classifier in one step is impossible because of the complexity and scale of the problem. Regarding prediction error, the method often compares favorably with the original predictor, and also, in situations with substantial noise, with other ensemble methods such as boosting or randomization. Hence it is very important to understand the reasons for its successes, and also for its occasional failures. However, even if it has attracted much attention and is frequently applied, important questions remain unanswered theoretically. In this article, we study a variant of bagging called Subagging [Subsample aggregating] that has appeared in [FRI00] and [BUH00]. It is more accessible for analysis and has also substantial computational advantages. The subagged estimator will be denoted by $\Phi^B(X, \mathcal{D}_n)$ or $\Phi_n^B(X)$ in the following.

Important questions are: *Is the generalization error of a subagged predictor lower than the original predictor, i.e* $\widetilde{R}_n(\Phi_n^B) \leq \widetilde{R}_n(\Phi)$*? The distribution* $\mathbb{P}$ *of the generating process being unknown, can we estimate the generalization error of a subagged predictor?* Our strategy is the following: after briefly emphasizing the difficulty to provide a general answer to the first question, we will concentrate on the second question. To estimate the generalization error of a subagged predictor, we propose to use an adapted cross-validation estimator denoted by $\widehat{R}_{CV}^{Out}(\Phi)$.

[BRE96] aggregates regression trees to build random forest and calls this process bagging. [BUJ02] prove that the bagged functional is always smooth in some sense. [BUJ00] also show that bagging can increase both bias and variance. [FRI00] prove that (in the limit of infinite samples) bagging reduces the variance of non-linear components of the Taylor decomposition while leaving the linear part unaffected. [BUH00] consider non-differentiable and discontinuous predictors and concentrate on the asymptotic smoothing effect of bagging on neighborhood of discontinuities of decision surfaces. [GRA04] brings new argument to explain bagging effect: bagging's improvement/deteriations are explained by the goodness/badness of highly influential examples. [ELI04] prove the effect of bagging on the stability of a learning method and derive non asymptotic bounds for the approximation error of

the bagging predictor. An interesting asymptotic result was derived in [BIA08] : asymptotically, bagging of weak predictors can produce a strong learner, namely the bayes classifier. However, a general answer to the following non-asymptotic question $\widetilde{R}_n(\Phi_n^B) \leq \widetilde{R}_n(\Phi)$? seems hard to reach in a general framework. Using Gauss-Markov theorem, [GRA04] shows that both bagged and unbagged predictor are unbiased, thus the variance of the unbagged predictor is lower than the variance of the bagged one. [BUJ00] exhibit general quadratic statistics for which the bagged predictor increase both variance and bias. Thus, we propose to estimate directly the generalization error of the subagged predictor by an adapted cross-validation procedure. The latter is inspired by [PET07], who proposed to use the left-out example of the bootstrap samples.

In the general setting, the cross-validation procedures include leave-one-out cross-validation, k-fold cross-validation, hold-out cross-validation (or split sample), leave-υ-out cross-validation (or Monte Carlo cross-validation or bootstrap cross-validation). With the exception of [BUR89], theoretical investigations of multifold cross-validation procedures have first concentrated on linear models ([Li87] ;[SHAO93] ; [ZHA93]). Results of [DGL96] and [GYO02] are discussed in Section 3. The first finite sample results are due to Wagner and Devroye [DEWA79] and concern k-local rules algorithms under leave-one-out and hold-out cross-validation. More recently, [HOL96, HOL96bis] derived finite sample results for υ-out cross-validation, $k-$fold cross-validation, and leave-one-out cross-validation for ERM over a class of predictors with finite VC-dimension in the realisable case (the generalization error is equal to zero). [BKL99] have emphasized when $k-$fold can beat υ-out cross-validation in the particular case of k-fold predictor. [KR99] has extended such results in the case of stable algorithms for the leave-one-out cross-validation procedure. [KEA95] also derived results for hold-out cross-validation for ERM, but their arguments rely on the traditional notion of VC-dimension. In the particular case of ERM over a class of predictors with finite VC-dimension but with general cross-validation procedures, [COR09A] derived probability upper bounds. [COR09B] derived upper bounds for general cross-validation estimate of the generalization error of stable predictors that do no make reference to VC-dimension. However, these bounds obtained are called "sanity check bounds" since they are not better than classical Vapnik-Chernovenkis's bounds.

We introduce our **main result** for symmetric cross-validation procedures (i.e. the probability for an observation to be in the test set is independent of its index) in the special case of empirical risk minimization (ERM). We divide the learning sample into two samples: the training sample and the test sample, to be defined below. We denote by p_n the percentage of elements in the test sample. Suppose that $\mathcal{H}$ holds, to be defined below. Suppose also that ϕ_n is an empirical risk minimizer. Then, we have for all $\varepsilon > 0$,

$$\Pr(\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out} \geq \varepsilon) \leq \min(B_{ERM}(n, p_n, \varepsilon), V_{ERM}(n, p_n, \varepsilon)) < 1,$$

with

- $B_{ERM}(n, p_n, \varepsilon) = \min((2np_n+1)^{4V_C/p_n} \exp(-n\varepsilon^2), (2n(1-p_n)+1)^{\frac{4V_C}{1-p_n}} \exp(-n\varepsilon^2/9))$

- $V_{ERM}(n, p_n, \varepsilon) = \exp(-2np_n\varepsilon^2)$.

The term $B(n, p_n, \varepsilon)$ is a Vapnik-Chernovenkis-type bound controlled by the size of the training sample $n(1-p_n)$ whereas the term $V(n, p_n, \varepsilon)$ is the minimum between a Hoeffding-type term controlled by the size of the test sample np_n, a polynomial term controlled by the size of the training sample. This bound can be interpreted as a quantitative answer to a trade-off issue. As the percentage of observations in the test sample p_n increases, the term $V(n, p_n, \varepsilon)$ decreases but the term $B(n, p_n, \varepsilon)$ increases. Other similar bounds are derived for infinite VC-dimension machine learning in the stability framework.
The main interest of the previous results is in the following

- our bounds are valid for machine learning with both finite and infinite VC-dimension. In the latter, it is sufficient that the machine learning satisfies some stablity property as introduced in chapter 2. As a motivation, we quote the following list of algorithms satisfying stability properties: regularization networks, ERM, k-nearest rules, boosting.
- our bounds are strictly less than 1 for any size of learning set. Thus it is also valid for small samples.

Using these probability bounds, we can then deduce that the expectation of the difference between the generalization error and the cross-validation estimate

$$\mathbb{E}_{\mathcal{D}_n}\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out} \leq \min(\sqrt{1/np_n}, 6\sqrt{\frac{V_{\mathcal{C}}(\ln(n(1-p_n))+2)}{n(1-p_n)}}).$$

Eventually, we define a splitting rule on how to chose the percentage of elements $p_n^\star$ in the test sample in order to get both a low generalization error together with a good approximation rate. We derive for this optimal choice of $p^\star$ a bound of the form

$$\Pr(\widetilde{R}_n(\Phi_n^{B,\star}) - \hat{R}_{CV}^{Out}(p_n^\star) \geq \varepsilon) = O_n((n+1)^{8V_{\mathcal{C}}}\exp(-2n(\varepsilon - 2\sqrt{2}V_{\mathcal{C}}^{1/2}\sqrt{\ln(n)/n})^2/(1-\exp(-2\varepsilon^2)).$$

The paper is organized as follows. We detail the main cross-validation procedures and we summarize the previous results for the estimation of generalization error. In Section 3, we introduce the main notations and definitions. Finally, in Section 4, we introduce our results, in terms of concentration inequalities.

3.2 Main notations

In the following, we follow the notations of cross-validation introduced in [COR09A].
We will consider the following shorter notations inspired by the literature on empirical processes. In the sequel, we will denote $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, and $(Z_i)_{1\leq i\leq n} := ((X_i, Y_i))_{1\leq i\leq n}$ the

learning set. For a given loss function L and a given class of predictors $\mathcal{G}$, we define a new class $\mathcal{F}$ of functions from $\mathcal{Z}$ to $\mathbb{R}_+$ by $\mathcal{F} := \{\psi \in \mathbb{R}_+^{\mathcal{Z}} | \psi(Z) = L(Y, \phi(X)), \phi \in \mathcal{G}\}$. For a machine learning Φ, we have the natural definition $\Psi(Z, \mathcal{D}_n) = L(Y, \Phi(X, \mathcal{D}_n))$. With these notations, the conditional risk $\widetilde{R}_n$ is the expectation of $\Psi(Z, \mathcal{D}_n)$ with respect to $\mathbb{P}$ conditionally on $\mathcal{D}_n$: $\widetilde{R}_n := \mathbb{E}_Z[\Psi(Z, \mathcal{D}_n) \mid \mathcal{D}_n]$ with $Z \sim \mathbb{P}$ independent of $\mathcal{D}_n$. In the following, if there is no ambiguity, we will also allow the following notation $\psi(X, \mathcal{D}_n)$ instead of $\Psi(X, \mathcal{D}_n)$.
To define the accurate type of cross-validation procedure, we introduce binary vectors. Let $V_n = (V_{n,i})_{1 \leq i \leq n}$ be a vector of size n. V_n is a binary vector if for all $1 \leq i \leq n, V_{n,i} \in \{0, 1\}$ and if $\sum_{i=1}^n V_{n,i} \neq 0$. Consequently, we can define the subsample associated with it: $\mathcal{D}_{V_n} := \{Z_i \in \mathcal{D}_n | V_{n,i} = 1, 1 \leq i \leq n\}$. We define a weighted empirical measure on $\mathcal{Z}$

$$\mathbb{P}_{n,V_n} := \frac{1}{\sum_{i=1}^n V_{n,i}} \sum_{i=1}^n V_{n,i} \delta_{Z_i},$$

with δ_{Z_i} the Dirac measure at $\{Z_i\}$. We also define a weighted empirical error $\mathbb{P}_{n,V_n}\psi$ where $\mathbb{P}_{n,V_n}\psi$ stands for the usual notation of the expectation of ψ with respect to $\mathbb{P}_{n,V_n}$. For $\mathbb{P}_{n,1_n}$, with 1_n the binary vector of size n with 1 at every coordinate, we will use the traditional notation $\mathbb{P}_n$. For a predictor trained on a subsample, we define

$$\psi_{V_n}(.) := \Psi(., \mathcal{D}_{V_n}).$$

With the previous notations, notice that the predictor trained on the learning set $\psi(., \mathcal{D}_n)$ can be denoted by $\psi_{1_n}(.)$. We will allow the simpler notation $\psi_n(.)$. The learning set is divided into two disjoint sets: the training set of size $n(1 - p_n)$ and the test set of size np_n, where p_n is the percentage of elements in the test set. To represent the training set, we define V_n^{tr} a random binary vector of size n independent of $\mathcal{D}_n$. V_n^{tr} is called the training vector. We define the test vector by $V_n^{ts} := 1_n - V_n^{tr}$ to represent the test set.

The distribution of V_n^{tr} characterizes all the subagging procedures described in the previous section. Using our notations, we can now define the bagged predictor.

Definition 92 (Subagged regressor)**.** *The subagged predictor build from ϕ_n denoted ϕ_n^B is defined by:*

$$\phi_n^B(.) := \mathbb{E}_{V_n^{tr}} \phi_{V_n^{tr}}(.).$$

In the case of classifiers, the bagging rule corresponds to the vote by majority. We suppose in this case that $\mathcal{Y} = \{1, \ldots, M\}$.

Definition 93 (Subagged classifier)**.** *Cross-validated subagged classifiers of ϕ_n^B defined by:*

$$\phi_n^B(X) := \arg \min_{k \in \{1,\ldots,M\}} \mathbb{E}_{V_n^{tr}} L(k, \Phi(X, \mathcal{D}_{V_n^{tr}}))$$

We can now define the cross-validation estimator.

Definition 94 (Cross-validated subagged estimator)**.** *Cross-validated subagged estimates of* ϕ_n^B *denoted can be defined in two different ways by:*

$$\widehat{R}_{CV}^{Out}(\Phi_n^B) := \mathbb{E}_{V_n^{tr}} \mathbb{P}_{n,V_n^{ts}}(\psi_{V_n^{tr}})$$

and

$$\widehat{R}_{CV}^{In}(\Phi_n^B) := \mathbb{E}_{V_n^{tr}} \mathbb{P}_{n,V_n^{tr}}(\psi_{V_n^{tr}})$$

Remark 95. *Recall that* $\mathbb{E}_{V_n^{tr}} \mathbb{P}_{n,V_n^{ts}}(\psi_{V_n^{tr}})$ *is the conditional expectation of* $\mathbb{P}_{n,V_n^{ts}}(\psi_{V_n^{tr}})$ *with respect to the random vector* V_n^{tr} *given* $\mathcal{D}_n$*.*

Remark 96. *The cross-validated subagged estimate differs from the usual cross-validation estimate of* $\hat{R}_{CV}^{Out}(\psi_n^B)$ *which is equal to* $\mathbb{E}_{U_n^{tr}} \mathbb{P}_{n,U_n^{ts}}(\psi_{U_n^{tr}}^B)$ *with* U_n^{tr} *the training vector as defined in chapter 1.*

We will give here a few examples of distributions of V_n^{tr} to show we retrieve subagging procedures described previously. Suppose n/k is an integer. The k-fold subagging procedure divides the data into k equally sized folds. It then produces a predictor by training on $k-1$ folds. This is repeated for each fold, and the trained predictors are averaged to form the subagged predictor.

Example 97 (k-fold cross-validation)**.**

$$\Pr(V_n^{tr} = (\underbrace{0,\dots,0}_{n/k \text{ observations}}, \underbrace{1,\dots,1}_{n(1-1/k) \text{ observations}})) = \frac{1}{k}$$

$$\Pr(V_n^{tr} = (\underbrace{1,\dots,1}_{n/k \text{ observations}}, \underbrace{0,\dots,0}_{n/k \text{ observations}}, \underbrace{1,\dots,1}_{n(1-2/k) \text{ observations}})) = \frac{1}{k}$$

$$\dots$$

$$\Pr(V_n^{tr} = (\underbrace{1,\dots,1}_{n(1-1/k) \text{ observations}}, \underbrace{0,\dots,0}_{n/k \text{ observations}})) = \frac{1}{k}.$$

We provide another popular example: the leave-one-out cross-validation. In leave-one-out cross-validation, a single sample of size n is used. Each member of the sample in turn is removed, the full modeling method is applied to the remaining $n-1$ members, and the fitted model is applied to the hold-backmember.

Example 98 (leave-one-out cross-validation)**.**

$$\Pr(V_n^{tr} = (0,1,\dots,1)) = \frac{1}{n}$$

$$\Pr(V_n^{tr} = (1,0,1,\dots,1)) = \frac{1}{n}$$

$$\dots$$

$$\Pr(V_n^{tr} = (1,\dots,1,0)) = \frac{1}{n}.$$

3.3 Results for the cross-validated subagged regressor

3.3.1 VC Framework

Notations and definition

We denote by R_{opt} the minimal generalization error attained among the class of predictors $\mathcal{C}$, $R_{opt} = \inf_{\phi \in \mathcal{C}} R(\phi)$. In the sequel, we suppose that ϕ_n belongs to some $\mathcal{C}$. Notice that R_{opt} is a parameter of the unknown distribution $\mathbb{P}_{(X,Y)}$ whereas $\widetilde{R}_n$ is a random variable.

At last, recall the definitions of:

Definition 99 (Shatter coefficients)**.** *Let $\mathcal{A}$ be a collection of measurable sets. For $(z_{1,\ldots,}z_n) \in \{\mathbb{R}^d\}^n$, let $N_{\mathcal{A}}(z_{1,\ldots,}z_n)$ be the number of differents sets in*

$$\{\{z_1, \ldots, z_n\} \cap A; A \in \mathcal{A}\}$$

The n-shatter coefficient of $\mathcal{A}$ is

$$\mathcal{S}(\mathcal{A}, n) = \max_{(z_{1,\ldots,}z_n) \in \{\mathbb{R}^d\}^n} N_{\mathcal{A}}(z_{1,\ldots,}z_n)$$

That is, the shatter coefficient is the maximal number of different subsets of n points that can be picked out by the class of sets $\mathcal{A}$.

and

Definition 100 (VC dimension)**.** *Let $\mathcal{A}$ be a collection of sets with $\mathcal{A} \geq 2$. The largest integer $k \geq 1$ for which $\mathcal{S}(\mathcal{A},k) = 2^k$ is denoted by $V_{\mathcal{C}}$, and it is called the Vapnik-Chernovenkis dimension (or VC dimension) of the class $\mathcal{A}$. If $\mathcal{S}(\mathcal{A},n) = 2^n$ for all n, then by definition $V_{\mathcal{C}} = \infty$.*

A class of predictors $\mathcal{C}$ is said to have a finite VC-dimension $V_{\mathcal{C}}$ if the dimension of the collection of sets $\{A_{\phi,t} : \phi \in \mathcal{C}, t \in [0,1]\}$ is equal to $V_{\mathcal{C}}$, where $A_{\phi,t} = \{(x,y)/L(y,\phi(x)) > t\}$.

Results

In the sequel, we suppose that the cross-validation is symmetric (i.e. $\Pr(V_{n,i} = 1)$ is independent of i) and the number of elements in the training set is constant and equal to np_n, that the training sample and the test sample are disjoint and that the number of observations in the training sample and in the test sample are respectively $n(1-p_n)$ and np_n. Moreover, we suppose also that ϕ_n belongs to a class of predictor with finite VC-dimension. Suppose also that L is bounded in the following way: $L(Y,\phi(X)) \leq C(h(Y,\phi(X))$ with C convex function -bounded itself by 1 on the support of $h(Y,\phi_{V_n^{tr}}(X))$ for simplicity-, and h such that for any $0 < \lambda < 1$, we have $h(y, \lambda\phi(x_1)+(1-\lambda)\phi(x_2) \leq \lambda h(y,\phi(x_1)+(1-\lambda)h(y,\phi(x_2)$. We will also suppose that the predictors are symmetric according to the training sample, i.e. the predictor does not depend on the order of the observations in $\mathcal{D}_n$. **We denote these hypotheses by $\mathcal{H}$.**

Remark 101. *Typical upperbounding convex cost functions are : the hinge loss* $C(x) = (1+x)_+$*, the exponential loss* $C(x) = e^x$*, the logit loss* $C(x) = \log_2(1+e^x)$*.*

We will show upper bounds of the kind $\Pr(\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out} \geq \varepsilon) \leq \min(B(n, p_n, \varepsilon), V(n, p_n, \varepsilon))$ with $\varepsilon > 0$. The term $B(n, p_n, \varepsilon)$ is a Vapnik-Chernovenkis-type bound whereas the term $V(n, p_n, \varepsilon)$ is a Hoeffding-type term controlled by the size of the test sample np_n. This bound can be interpreted as a quantitative answer to a trade-off question. As the percentage of observations in the test sample p_n increases, the $V(n, p_n, \varepsilon)$ term decreases but the $B(n, p_n, \varepsilon)$ term increases.

Theorem 102 (Absolute error for symmetric cross-validation)**.** *Suppose that* $\mathcal{H}$ *holds. Then, we have for all* $\varepsilon > 0$*,*

$$\Pr(\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out} \geq \varepsilon) \leq \min(B_{sym}(n, p_n, \varepsilon), V_{sym}(n, p_n, \varepsilon)) < 1$$

with

- $B_{sym}(n, p_n, \varepsilon) = (2np_n + 1)^{4Vc/p_n} e^{-n\varepsilon^2}$
- $V_{sym}(n, p_n, \varepsilon) = \exp(-2np_n\varepsilon^2)$.

Remark 103. *We do not require* ϕ_n *to be an empirical risk minimizer.*

Proof.

We have $\widetilde{R}_n(\Phi_n^B) = \mathbb{P}\psi_n^B = \mathbb{P}L(Y, \mathbb{E}_{V_n^{tr}}\phi_{V_n^{tr}}(X))$. Since C is a convex function -bounded itself by 1 on the support of $h(Y, \phi_{V_n^{tr}}(X))$-, and h linear in the second variable, we get

$$\widetilde{R}_n(\Phi_n^B) \leq \mathbb{P}C(h(Y, \mathbb{E}_{V_n^{tr}}\phi_{V_n^{tr}}(X)) \leq \mathbb{E}_{V_n^{tr}}\mathbb{P}C(h(Y, \phi_{V_n^{tr}}(X))$$

Then, we split according to $\mathbb{E}_{V_n^{tr}}\mathbb{P}_{n,V_n^{ts}}C(h(Y, \phi_{V_n^{tr}}(X))$:

$$\begin{aligned}\widetilde{R}_n(\Phi_n^B) &\leq \mathbb{E}_{V_n^{tr}}\mathbb{P}_{n,V_n^{ts}}C(h(Y, \phi_{V_n^{tr}}(X)) + \mathbb{E}_{V_n^{tr}}(\mathbb{P} - \mathbb{P}_{n,V_n^{ts}})C(h(Y, \phi_{V_n^{tr}}(X)) \\ &= \hat{R}_{CV}^{Out} + \mathbb{E}_{V_n^{tr}}(\mathbb{P} - \mathbb{P}_{n,V_n^{ts}})C(h(Y, \phi_{V_n^{tr}}(X)\end{aligned}$$

Thus, we obtain: $\Pr(\widetilde{R}_n(\psi_n^B) - \hat{R}_{CV}^{Out} \geq \varepsilon) \leq \Pr(\mathbb{E}_{V_n^{tr}}(\mathbb{P} - \mathbb{P}_{n,V_n^{ts}})C(h(Y, \phi_{V_n^{tr}}(X) \geq \varepsilon)$.

To prove our result, we proceed now in two steps. For this, we consider

$$\mathbb{E}_{V_n^{tr}}(\mathbb{P}_{n,V_n^{ts}}C(h(Y, \phi_{V_n^{tr}}(X)) - \mathbb{P}C(h(Y, \phi_{V_n^{tr}}(X)))$$

in two different ways

1. using conditional Hoeffding's inequality,

2. using Vapnik-Chernovenkis-type inequality to bound the supremum over a class.

1. First, by conditional Hoeffding arguments (for a proof, see e.g. chapter 1),

$$\Pr(\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out} \geq \varepsilon) \leq \exp(-2np_n\varepsilon^2).$$

2. Secondly, we derive the bound:

$$\begin{aligned}\Pr(\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out} \geq \varepsilon) &\leq \Pr(\mathbb{E}_{V_n^{tr}}(\mathbb{P} - \mathbb{P}_{n,V_n^{ts}})C(h(Y, \phi_{V_n^{tr}}(X)) \geq \varepsilon) \\ &\leq \Pr(\mathbb{E}_{V_n^{tr}} \sup_{\phi \in \mathcal{C}}(\mathbb{P} - \mathbb{P}_{n,V_n^{ts}})C(h(Y, \phi(X)) \geq \varepsilon).\end{aligned}$$

Recall a useful lemma (for the proof, see Appendices).

Lemma 104. *Under the assumptions $\mathcal{H}$, we have for all, $\varepsilon > 0$,*

$$\Pr\left(\mathbb{E}_{V_n^{tr}} \sup_{\phi \in \mathcal{C}}(\mathbb{P} - \mathbb{P}_{n,V_n^{tr}})C(h(Y, \phi(X)) \geq \varepsilon\right) \leq (\mathcal{S}(2np_n, \mathcal{C}))^{4/p_n} e^{-n\varepsilon^2}.$$

and we also have (for the proof, see e.g. [DGL96]): $\forall n, \mathcal{S}(n, \mathcal{C}) \leq (n+1)^{V_\mathcal{C}}$.

Thus, it follows that $\Pr(\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out} \geq \varepsilon) \leq (2np_n + 1)^{4V_\mathcal{C}/p_n} e^{-n\varepsilon^2}$.
Putting altogether, we get $\Pr(\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out} \geq \varepsilon) \leq \min(\exp(-2np_n\varepsilon^2), (2np_n+1)^{4V_\mathcal{C}/p_n} e^{-n\varepsilon^2})$.

□

Theorem 105 (Absolute error for symmetric cross-validation)**.** *Suppose that $\mathcal{H}$ holds. Then, we have for all $\varepsilon > 0$,*

$$\Pr(\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{In} \geq \varepsilon) \leq \min(B_{sym}(n, p_n, \varepsilon), V_{sym}(n, p_n, \varepsilon)) < 1$$

with

- $B_{sym}(n, p_n, \varepsilon) = (2n(1-p_n)+1)^{\frac{4V_\mathcal{C}}{1-p_n}} e^{-n\varepsilon^2}$
- $V_{sym}(n, p_n, \varepsilon) = \exp(-2np_n\varepsilon^2)$.

Proof.

We proceed as previously: $\widetilde{R}_n(\Phi_n^B) = \mathbb{P}\Phi_n^B = \mathbb{P}L(Y, \mathbb{E}_{V_n^{tr}}\phi_{V_n^{tr}}(X)) \leq \mathbb{P}C(h(Y, \mathbb{E}_{V_n^{tr}}\phi_{V_n^{tr}}(X)) \leq \mathbb{E}_{V_n^{tr}}\mathbb{P}C(h(Y, \phi_{V_n^{tr}}(X)$.
We then split this quantity according to $\mathbb{E}_{V_n^{tr}}\mathbb{P}_{n,V_n^{tr}}C(h(Y, \phi_{V_n^{tr}}(X)$

$$\begin{aligned}\widetilde{R}_n(\Phi_n^B) &\leq \mathbb{E}_{V_n^{tr}}\mathbb{P}_{n,V_n^{tr}}C(h(Y,\phi_{V_n^{tr}}(X)+\mathbb{E}_{V_n^{tr}}(\mathbb{P}-\mathbb{P}_{n,V_n^{ts}})C(h(Y,\phi_{V_n^{tr}}(X)\\ &= \hat{R}_{CV}^{In}+\mathbb{E}_{V_n^{tr}}(\mathbb{P}-\mathbb{P}_{n,V_n^{tr}})C(h(Y,\phi_{V_n^{tr}}(X)).\end{aligned}$$

Thus, we get

$$\begin{aligned}\Pr(\widetilde{R}_n(\Phi_n^B)-\hat{R}_{CV}^{In}\geq\varepsilon) &\leq \Pr(\mathbb{E}_{V_n^{tr}}(\mathbb{P}-\mathbb{P}_{n,V_n^{tr}})C(h(Y,\phi_{V_n^{tr}}(X))\geq\varepsilon)\\ &\leq \Pr(\mathbb{E}_{V_n^{tr}}\sup_{\phi\in\mathcal{C}}(\mathbb{P}-\mathbb{P}_{n,V_n^{tr}})C(h(Y,\phi(X))\geq\varepsilon).\end{aligned}$$

Recall two useful results (for the proof, see e.g. chapter 1)

Lemma 106. *Under the assumptions $\mathcal{H}$, we have for all $\varepsilon>0$,*

$$\Pr\left(\mathbb{E}_{V_n^{tr}}\sup_{\phi\in\mathcal{C}}(\mathbb{P}(\phi)-\mathbb{P}_{n,V_n^{tr}}(\phi))\geq\varepsilon\right)\leq(\mathcal{S}(2n(1-p_n),\mathcal{C}))^{4/(1-p_n)}e^{-n(1-p_n)\varepsilon^2}.$$

□

In the special case of empirical risk minimization, we can obtain a stronger result.

Theorem 107 (Absolute error for symmetric cross-validation)**.** *Suppose that $\mathcal{H}$ holds. Suppose also that ϕ_n is based on empirical risk minimization. But instead of minimizing $\widehat{R}_n(\phi)$, we suppose ϕ_n minimizes $\frac{1}{n}\sum_{i=1}^n C(h(Y_i,\phi(X_i))$. For simplicity, we suppose the infimum is attained i.e. $\phi_n=\arg\min_{\phi\in\mathcal{C}}\frac{1}{n}\sum_{i=1}^n C(h(Y_i,\phi(X_i))$. Then, we have for all $\varepsilon>0$,*

$$\Pr(\widetilde{R}_n(\Phi_n^B)-\hat{R}_{CV}^{Out}\geq\varepsilon)\leq\min(B_{ERM}(n,p_n,\varepsilon),V_{ERM}(n,p_n,\varepsilon))<1,$$

with

- $B_{ERM}(n,p_n,\varepsilon)=\min((2np_n+1)^{4V_{\mathcal{C}}/p_n}\exp(-n\varepsilon^2),(2n(1-p_n)+1)^{\frac{4V_{\mathcal{C}}}{1-p_n}}\exp(-n\varepsilon^2/9))$
- $V_{ERM}(n,p_n,\varepsilon)=\exp(-2np_n\varepsilon^2)$.

Remark 108.

1. *The assumption $\phi_n=\arg\min_{\phi\in\mathcal{C}}\frac{1}{n}\sum_{i=1}^n C(h(Y_i,\phi(X_i))$ is not so restrictive, since in practice in order to numerically minimizes $\frac{1}{n}\sum_{i=1}^n L(Y_i,\phi(X_i))$, one looks for C convex such that for all x,y, $L(y,\phi(x))\leq C(h(y,\phi(x))$.*
2. *Thanks to the Hoeffding's part, the bound is always smaller than 1, so it remains valid for small samples. For bigger samples, we will prefer the Vapnik-Chernovenkis's part.*

Proof.
Appying the previous result, we have $\Pr(\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out} \geq \varepsilon) \leq \min(\exp(-2np_n\varepsilon^2), (2np_n + 1)^{4V_{\mathcal{C}}/p_n} \exp(-n\varepsilon^2))$.

Recall that $\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out} \leq \mathbb{E}_{V_n^{tr}}(\mathbb{P}C(h(Y, \phi_{V_n^{tr}}(X)) - \mathbb{P}_{n,V_n^{ts}}C(h(Y, \phi_{V_n^{tr}}(X)))$.

We need the following lemma (for a proof, see chapter 1): $\mathbb{E}_{V_n^{tr}}\mathbb{P}_{n,V_n^{ts}}C(h(Y, \phi_{V_n^{tr}}(X)) \geq \mathbb{P}_n C(h(Y, \phi_n(X))$ since $\phi_n = \arg\min_{\phi \in \mathcal{C}} \frac{1}{n}\sum_{i=1}^n C(h(Y_i, \phi(X_i))$.

Denote $\psi(Z) := C(h(Y, \phi(X)))$ with $Z := (X, Y)$. We have the following natural notation $\psi_{V_n^{tr}}(Z) := C(h(Y, \phi_{V_n^{tr}}(X)))$.

We thus get

$$\Pr(\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out} \geq 3\varepsilon) \leq \Pr(\mathbb{E}_{V_n^{tr}}(\mathbb{P}\psi_{V_n^{tr}} - \mathbb{P}_{n,V_n^{ts}}\psi_{V_n^{tr}}) \geq 3\varepsilon) \leq \Pr(\mathbb{E}_{V_n^{tr}}(\mathbb{P}\psi_{V_n^{tr}} - \mathbb{P}_n\psi_n) \geq 3\varepsilon)$$

and by splitting according to $\mathbb{P}\psi_{opt}$, we have:

$$\begin{aligned}\Pr(\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out} \geq 3\varepsilon) &\leq \Pr(\mathbb{E}_{V_n^{tr}}(\mathbb{P}\psi_{V_n^{tr}} - \mathbb{P}_{n,V_n^{tr}}\psi_{V_n^{tr}} + \mathbb{P}_{n,V_n^{tr}}\psi_{V_n^{tr}} - \mathbb{P}\psi_{opt} + \mathbb{P}\psi_{opt} - \mathbb{P}_n\psi_n) \geq 3\varepsilon) \\ &\leq \Pr(\mathbb{E}_{V_n^{tr}} \sup_{\psi \in \mathcal{F}}(\mathbb{P}\psi - \mathbb{P}_{n,V_n^{tr}}\psi) \geq \varepsilon) + \Pr(\sup_{\psi \in \mathcal{F}}(\mathbb{P}_{n,V_n^{tr}}\psi - \mathbb{P}\psi) \geq \varepsilon) \\ &\quad + \Pr(\sup_{\psi \in \mathcal{F}}(\mathbb{P}\psi - \mathbb{P}_n\psi) \geq \varepsilon).\end{aligned}$$

Recall the following lemma (for the proof, see e.g.chapter 1),

Lemma 109. *Under the assumption of Proposition 33, we have for all $\varepsilon > 0$,*

$$\Pr\left(\mathbb{E}_{V_n^{tr}} \sup_{\psi \in \mathcal{F}}(\mathbb{P}_{n,V_n^{tr}}\psi - \mathbb{P}\psi) \geq \varepsilon\right) \leq (\mathcal{S}(2n(1-p_n), \mathcal{C}))^{\frac{4}{1-p_n}} e^{-n\varepsilon^2}$$

and symmetrically

$$\Pr\left(\mathbb{E}_{V_n^{tr}} \sup_{\psi \in \mathcal{F}}(\mathbb{P}\psi - \mathbb{P}_{n,V_n^{tr}}\psi) \geq \varepsilon\right) \leq (\mathcal{S}(2n(1-p_n), \mathcal{C}))^{\frac{4}{1-p_n}} e^{-n\varepsilon^2}.$$

Then, we get

$$\begin{aligned}\Pr(\widetilde{R}_n(\psi_n^B) - \hat{R}_{CV}^{Out} \geq 3\varepsilon) &\leq 2(\mathcal{S}(2n(1-p_n), \mathcal{C}))^{\frac{4}{1-p_n}} e^{-n\varepsilon^2} + (\mathcal{S}(2n, \mathcal{C}))^4 e^{-n\varepsilon^2} \\ &\leq 3(2n(1-p_n)+1)^{\frac{4V_{\mathcal{C}}}{1-p_n}} e^{-n\varepsilon^2}.\end{aligned}$$

This implies in turn that

$$\Pr(\widetilde{R}_n(\psi_n^B) - \hat{R}_{CV}^{Out} \geq \varepsilon) \leq (2n(1-p_n)+1)^{\frac{4V_{\mathcal{C}}}{1-p_n}} \exp(-n\varepsilon^2/9).$$

Putting altogether, we get

$$\Pr(\widetilde{R}_n(\psi_n^B) - \hat{R}_{CV}^{Out} \geq \varepsilon) \leq \min(\exp(-2np_n\varepsilon^2), (2np_n+1)^{4V_{\mathcal{C}}/p_n} e^{-n\varepsilon^2}, (2n(1-p_n)+1)^{\frac{4V_{\mathcal{C}}}{1-p_n}} \exp(-n\varepsilon^2/9))$$

□

Theorem 110. *Suppose that $\mathcal{H}$ holds. Suppose also and that n/k is an integer. Then, we have also for all $\varepsilon > 0$,*

$$\Pr(\widetilde{R}_n(\psi_n^B) - \hat{R}_{CV}^{Out} \geq \varepsilon) \leq \min(B_k(n, p_n, \varepsilon), V_k(n, p_n, \varepsilon))$$

with

- $B_k(n, p_n, \varepsilon) = (2n/k+1)^{4kV_{\mathcal{C}}} \exp(-n\varepsilon^2)$
- $V_k(n, p_n, \varepsilon) = \min\left(\exp(-2n/k\varepsilon^2), 2^{\frac{1}{p_n}} \exp\left(-\frac{n\epsilon^2}{64(\sqrt{V_{\mathcal{C}} \ln(2(2n/k+1))}+2)}\right)\right).$

Proof.
The proofs starts as previously. We have

$$\Pr(\hat{R}_{CV}^{Out} - \widetilde{R}_n(\psi_n^B) \geq \varepsilon) \leq \Pr(\mathbb{E}_{V_n^{tr}}(\mathbb{P}_{n,V_n^{ts}}\psi_{V_n^{tr}} - \mathbb{P}\psi_{V_n^{tr}}) \geq \varepsilon) \leq \exp(-2np_n\varepsilon^2)$$

but we also have

$$\Pr(\hat{R}_{CV}^{Out} - \widetilde{R}_n(\psi_n^B) \geq \varepsilon) \leq \Pr(\mathbb{E}_{V_n^{tr}}(\sup_{\psi\in\mathcal{F}}(\mathbb{P}_{n,V_n^{ts}}\psi - \mathbb{P}\psi) \geq \varepsilon) \leq 2^{\frac{1}{p_n}} \exp\left(-\frac{n\epsilon^2}{64(\sqrt{V_{\mathcal{C}} \ln(2(2np_n+1))}+2)}\right).$$

according to chapter 1.

□

Following the previous results, we can obtain results for the expectation of the difference $\widetilde{R}_n(\psi_n^B) - \hat{R}_{CV}^{Out}$

Theorem 111 (L_1 error)**.** *Suppose that $\mathcal{H}$ holds. Suppose also and that n/k is an integer. Then, we have also for all $\varepsilon > 0$,*

$$\mathbb{E}_{\mathcal{D}_n}\left(\widetilde{R}_n(\psi_n^B) - \hat{R}_{CV}^{Out}\right) \leq \sqrt{1/np_n}$$

Furthermore, suppose also that ϕ_n is based on empirical risk minimization. But instead of minimizing $\widehat{R}_n(\phi)$, we suppose ϕ_n minimizes $\frac{1}{n}\sum_{i=1}^n C(h(Y_i, \phi(X_i))$. For simplicity, we suppose the infimum is attained i.e. $\phi_n = \arg\min_{\phi\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^n C(h(Y_i, \phi(X_i))$. Then, we have,

$$\mathbb{E}_{\mathcal{D}_n}\left(\widetilde{R}_n(\psi_n^B) - \hat{R}_{CV}^{Out}\right) \leq \min(\sqrt{1/np_n}, 6\sqrt{\frac{V_{\mathcal{C}}(\ln(n(1-p_n))+2)}{n(1-p_n)}})$$

Proof.
We just need to apply the previous results together with the following useful lemma (for a proof, see e.g.[DGL96]):

Lemma 112. *Let X be a nonnegative random variable. Let K, C nonnegative real such that $C \geq 1$. Suppose that for all $\varepsilon > 0$, $\mathbb{P}(X \geq \varepsilon) \leq C \exp(-K\varepsilon^2)$. Then, we have*

$$\mathbb{E}X \leq \sqrt{\frac{\ln(C)+2}{K}}.$$

□

3.3.2 Stability framework

Introduction to stability

To avoid the traditional analysis in the VC framework, notions of stability have been intensively worked through in the late 90's [KEA95], [BE01], [BE02], [KUT02], and [KUNIY02]. The object of stability framework is the learning algorithm rather than the space of classifiers. The learning algorithm is a map (effective procedure) from data sets to classifiers. An algorithm is stable at a learning set $\mathcal{D}_n$ if changing one point in $\mathcal{D}_n$ yields only a small change in the output hypothesis. Several different notions of algorithmic stability are described. The attraction of such an approach is that it avoids the traditional notion of VC-dimension, and allows to focus on a wider class of learning algorithms than empirical risk minimization. For example, this approach provides generalization error bounds for regularization-based learning algorithms that have been difficult to analyze within the VC framework such as boosting. If a map is stable, exponential bounds on generalization error may be obtained. As a motivation, we quote the following list of algorithms satisfying stability properties: regularization networks, ERM, k-nearest rules, boosting.

Definitions and notations of stability

The basic idea is that an algorithm is stable at a training set $\mathcal{D}_n$ if changing one point in $\mathcal{D}_n$ yields only a small change in the output hypothesis. Formally, a learning algorithm maps a weighted training set into a predictor space. Thus, stability can be translated into a Lipschitz condition for this mapping with high probability.
To be more formal, following [COR09B], we define a distance between two weighted empirical errors:

Definition 113 (Total variation)**.** *Let $\mathbb{P}_{n,V_n}$ and $\mathbb{P}_{n,U_n}$ be two empirical measures on $\mathcal{Z}$ with respect to the binary vectors V_n and U_n. We do not assume their support to be equal. The distance between them is defined as their total variation:*

$$||\mathbb{P}_{n,U_n} - \mathbb{P}_{n,V_n}|| = \sup_{A \in \mathcal{P}(\mathcal{Z})} |(\mathbb{P}_{n,U_n} - \mathbb{P}_{n,V_n})(A)|.$$

Example 114. *In the case of leave-one-out (i.e. $\sum_{i=1}^{n} U_{n,i} = n-1$), we have:*

$$||\mathbb{P}_{n,U_n} - \mathbb{P}_n|| = \frac{2}{n}.$$

In the case of leave-ν-out, we get:

$$||\mathbb{P}_{n,U_n} - \mathbb{P}_n|| = \frac{2\nu}{n}.$$

At least, we need a distance d on the set $\mathcal{F}$. Let us quote three important examples. Let $\psi_1, \psi_2 \in \mathcal{F}$. The uniform distance is defined by: $d_\infty(\psi_1, \psi_2) = \sup_{Z \in \mathcal{Z}} |\psi_1(Z) - \psi_2(Z)|$, the L_1-distance by: $d_1(\psi_1, \psi_2) = \mathbb{P}|\psi_1 - \psi_2|$, the error-distance $d_e(\psi_1, \psi_2) = |\mathbb{P}(\psi_1 - \psi_2)|$. It is important to notice that what matters here is not an absolute distance between the original class of predictors $\mathcal{G}$ seen as functions but the distance with the respect to the loss or/and the distribution $\mathbb{P}$. In particular, for the L_1-distance, we do not care about the behavior of the original predictors ϕ_1 and ϕ_2 outside the support of $\mathbb{P}$. At last, notice that we always have $d_e \leq d_1 \leq d_\infty$.

We are now in position to define the different notions of stability of a learning algorithm which cover notions introduced by [KUNIY02]. We begin with the notion of weak stability. In essence, it says that for any given resampling vectors, the distance between two predictors is controlled with high probability by the distance between the resampling vectors. As a motivation, notice that algorithms such as Adaboost ([KUNIY02]) satisfies this property. With the previous notations, we have:

Definition 115 (Weak stability)**.** *Let $\mathcal{D}_n = (Z_i)_{1 \leq i \leq n}$ be a learning set. Let $\lambda, (\delta_{n,p_n})_{n,p_n}$ be nonnegative real numbers. A learning algorithm Ψ is said to be weak $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ stable if for any training vector U_n whose sum is equal to $n(1-p_n)$:*

$$\Pr(d(\psi_{U_n}, \psi_n) \geq \lambda ||\mathbb{P}_{n,U_n} - \mathbb{P}_n||) \leq \delta_{n,p_n}.$$

Notice that in the former definition Pr stands for $\mathbb{P}^{\otimes n}$. Indeed, ψ_n is trained with n observations, drawn independently from $\mathbb{P}$. A stronger notion is to consider ψ_n trained with $n-1$ observations drawn independently from $\mathbb{P}$ and an additionnal general observation z. We consider the stronger notion of strong stability. As a motivation, notice that algorithms such as Empirical Risk Minimization with finite VC dimension ([KUNIY02]) satisfies this property.

Definition 116 (Strong stability)**.** *Let $z \in \mathcal{Z}$. Let $\mathcal{D}_n = \mathcal{D}_{n-1} \cup \{z\}$ be a learning set. Let $\lambda, (\delta_{n,p_n})_{n,p_n}$ be nonnegative real numbers. A learning algorithm Ψ is said to be strong $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ stable if for any training vector U_n whose sum is equal to $n(1-p_n)$:*

$$\Pr(d(\psi_{U_n}, \psi_n) \geq \lambda ||\mathbb{P}_{n,U_n} - \mathbb{P}_n||) \leq \delta_{n,p_n}.$$

What we have in mind for classical algorithms is $\delta_{n,p_n} = O_n(p_n \exp(-n(1-p_n))$. We can state the last definition in other words. Let V_n^{tr} be a training vector with distribution $\mathbb{Q}$ such that the number of elements in the training set is constant and equal to $n(1-p_n)$. Notice

then that the former definition also implies that $\sup_{U_n \in \text{support}(\mathbb{Q})} \mathbb{P}(\frac{d(\psi_{U_n},\psi_n)}{||\mathbb{P}_{n,U_n}-\mathbb{P}_n||} \geq \lambda) \leq \delta_{n,p_n}$, where support($\mathbb{Q}$) stands for the support of $\mathbb{Q}$. The previous notion stands for any U_n having the same support of $\mathbb{Q}$. A stronger hypothesis would be that the previous probability stands uniformly over U_n in support($\mathbb{Q}$). This leads formally to the notion of cross-validation stability. To be more accurate:

Definition 117 (Cross-validation weak stability)**.** *Let $\mathcal{D}_n = (Z_i)_{1 \leq i \leq n}$ a learning set. Let V_n^{tr} a training vector with distribution $\mathbb{Q}$. Let $\lambda, (\delta_{n,p_n})_{n,p_n}$ be nonnegative real numbers. A learning algorithm Ψ is said to be weak $(\lambda, (\delta_{n,p_n})_{n,p_n}, d, \mathbb{Q})$ stable if it is weak $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ stable and if:*

$$\Pr(\sup_{U_n \in support(\mathbb{Q})} \frac{d(\psi_{U_n}, \psi_n)}{||\mathbb{P}_{n,U_n} - \mathbb{P}_n||} \geq \lambda) \leq \delta_{n,p_n}.$$

As before, we also define the following stronger notion:

Definition 118 (Cross-validation strong stability)**.** *Let $z \in \mathcal{Z}$. Let $\mathcal{D}_n = \mathcal{D}_{n-1} \cup \{z\}$ a learning set. Let V_n^{tr} a cross-validation vector with distribution $\mathbb{Q}$. A learning algorithm Ψ is said to be strongly $(\lambda, (\delta_{n,p_n})_{n,p_n}, d, \mathbb{Q})$ stable if it is strong $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ stable and if:*

$$\Pr(\sup_{U_n \in support(\mathbb{Q})} \frac{d(\psi_{U_n}, \psi_n)}{||\mathbb{P}_{n,U_n} - \mathbb{P}_n||} \geq \lambda) \leq \delta_{n,p_n}.$$

Remark 119. *If the cardinal of the support of $\mathbb{Q}$ is denoted $\kappa(n)$, then a learning algorithm which is weak $(\lambda, (\delta_{n,p_n})_{n,p_n}, d, \mathbb{Q})$-stable is also strong $(\lambda, (\kappa(n)\delta_{n,p_n})_n, d, \mathbb{Q})$-stable.*

As seen in the following table, we retrieve with those notations the different notions of stability introduced by [DEWA79], [KEA95] and also [BE01], [KUNIY02].

stability distance	d_∞	d_1	d_e
Weak	weak (λ, δ) hypothesis stability [KUNIY02]	weak (λ, δ) L_1stability [KUNIY02]	weak (λ, δ) error stability [KUNIY02]
Strong	strong (λ, δ) hypothesis stability [KUNIY02][DEWA79]	strong (λ, δ) L_1stability [KUNIY02]	strong (λ, δ) error stability [KUNIY02]
Sure Stability	uniform stability [BE01]	[DEWA79]	error stability [KEA95]

To motivate this approach, we also quote a list of class of predictors satisfying the previous stability conditions.

stability distance	d_∞	d_1	d_e
Weak			Lasso
Strong	Adaboost ([KUNIY02])	-ERM ([KUNIY02]) -k-nearest rule	Bayesian algorithm [KEA95]
Uniform	Regularization networks		

We recall the main notations and definitions:

Name	Notation	Definition
Risk or generalization error	$\widetilde{R}_n$	$E_P[L(Y, \phi(X, D_n)) \mid D_n]$
Resubstitution error	$\widehat{R}_n$	$\frac{1}{n}\sum_{i=1}^{n} L(Y_i, \phi_n(X_i, D_n))$
Cross-validation error	$\widehat{R}_{CV}$	$E_{V_n^{tr}} P_{n,V_n^{ts}} \psi_{V_n^{tr}}$

Table 3.1: Main notations

Main results

Let $\mathcal{D}_n$ be a learning set of size n. Let $V_n^{tr} \sim \mathbb{Q}$ be a training vector independent of $\mathcal{D}_n$ such that the cross-validation is symmetric and the number of elements in the training set is constant and equal to np_n. Let d be a distance among d_e, d_1, d_∞. At last, we suppose that the loss function L is bounded by 1. We derive the following general results that stands for general cross-validation procedures and stable algorithms.

Theorem 120 (Cross-validation Strong stability)**.** *Suppose that $\mathcal{H}$ holds. Let Ψ a machine learning which is strong $(\lambda, (\delta_n, p_n)_{n,p_n}, \mathbb{Q})$ stable with respect to the distance d. Then, for all $\varepsilon \geq 0$, we have:*

$$\Pr\left(\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out} \geq \varepsilon\right) \leq \exp(-2np_n\varepsilon^2)$$

Furthermore, if d is the uniform distance d_∞, then we have for all $\alpha > 0$:

$$\Pr\left(\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out} \geq \varepsilon\right) \leq \min(\exp(-2np_n\varepsilon^2), 2(\exp(-\frac{\varepsilon^2}{8n(8\lambda np_n + \alpha)^2}) + \frac{n}{\alpha}\delta_{n,p_n}))$$

Thus, if we choose $\alpha = 8\lambda np_n$,

$$\Pr\left(\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out} \geq \varepsilon\right) \leq \min(\exp(-2np_n\varepsilon^2), 2(\exp(-\frac{\varepsilon^2}{8(16\lambda)^2 np_n^2}) + \frac{n}{8\lambda p_n}\delta_{n,p_n}))$$

Proof.
On the one hand, we have as before by conditional Hoeffding's inequality (for a proof, see e.g. chapter 1):

$$\Pr(\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out} \geq \varepsilon) \leq \Pr(\mathbb{E}_{V_n^{tr}}(\mathbb{P}\psi_{V_n^{tr}} - \mathbb{P}_{n,V_n^{ts}}\psi_{V_n^{tr}}) \geq \varepsilon) \leq \exp(-2np_n\varepsilon^2)$$

On the other hand, notice that $\mathbb{P}^{\otimes n}\mathbb{E}_{V_n^{tr}}(\mathbb{P}\psi_{V_n^{tr}} - \mathbb{P}_{n,V_n^{ts}}\psi_{V_n^{tr}}) = 0$

Denote $f(Z_1, Z_2, \ldots, Z_n) := \mathbb{E}_{V_n^{tr}}(\mathbb{P}\psi_{V_n^{tr}} - \mathbb{P}_{n,V_n^{ts}}\psi_{V_n^{tr}})$. Let $z \in \mathcal{Z}$. Now denote:

$$B := \{\sup_{U_n \in \text{support}(\mathbb{Q})} \frac{d(\psi_{U_n}, \psi_{n+1})}{||\mathbb{P}_{n,U_n} - \mathbb{P}_{n+1}||} \geq \lambda\}$$

with ψ_{n+1} trained on $\mathcal{D}_{n+1} = \{Z_1, \ldots, Z_{i-1}, Z_i, Z_{i+1}, \ldots, Z_n, z\}$. Under our assumptions, we have $\Pr(B) \leq \delta_{n+1,p_{n+1}}$.

We want to show that with high probability there exist constants c_i such that for all $i \in \{1, \ldots, n\}$, for all $z \in \mathcal{Z}$,

$$\Delta_i := |f(Z_1, \ldots, Z_i, \ldots, Z_n) - f(Z_1, \ldots, Z_{i-1}, z, Z_{i+1}, \ldots, Z_n)| \leq c_i$$

.
Notice that:

$$\begin{aligned}|\Delta_i| &= |\mathbb{E}_{V_n^{tr}}(\mathbb{P}\psi_{V_n^{tr}} - \mathbb{P}_{n,V_n^{ts}}\psi_{V_n^{tr}}) - (\mathbb{E}_{V_n^{tr}}\mathbb{P}\psi'_{V_n^{tr}} - \mathbb{P}'_{n,V_n^{ts}}\mathbb{P}\psi'_{V_n^{tr}})| \\ &\leq |\mathbb{E}_{V_n^{tr}}\mathbb{P}(\psi_{V_n^{tr}} - \psi'_{V_n^{tr}})| + |\mathbb{E}_{V_n^{tr}}(\mathbb{P}_{n,V_n^{ts}}\psi_{V_n^{tr}} - \mathbb{P}'_{n,V_n^{ts}}\mathbb{P}\psi'_{V_n^{tr}})|\end{aligned}$$

with $\mathbb{P}'_{n,V_n^{tr}}$ the weighted empirical measure on the sample

$$\mathcal{E}_n = \{Z_1, \ldots, Z_{i-1}, z, Z_{i+1}, \ldots, Z_n\}$$

and $\psi'_{V_n^{tr}}$ the predictor trained on $\mathcal{E}_{V_n^{tr}}$.

So, first, let us bound the first term, $|\mathbb{E}_{V_n^{tr}}\mathbb{P}(\psi_{V_n^{tr}} - \psi'_{V_n^{tr}})| \leq \mathbb{E}_{V_n^{tr}}|\mathbb{P}(\psi_{V_n^{tr}} - \psi_{n+1})| + \mathbb{E}_{V_n^{tr}}|\mathbb{P}(\psi_{n+1} - \psi'_{V_n^{tr}})|$. Thus, on $B^{\subset}$, we have $|\mathbb{E}_{V_n^{tr}}\mathbb{P}(\psi_{V_n^{tr}} - \psi'_{V_n^{tr}})| \leq \frac{4\lambda}{n+1}$.

To upper bound the second term, notice that:

$$\begin{aligned}|\mathbb{E}_{V_n^{tr}}\mathbb{P}_{n,V_n^{ts}}\psi_{V_n^{tr}} - \mathbb{E}_{V_n^{tr}}\mathbb{P}'_{n,V_n^{ts}}\psi'_{V_n^{tr}}| &= |\mathbb{E}_{V_n^{tr}}(\mathbb{P}_{n,V_n^{ts}}(\psi_{V_n^{tr}} - \psi'_{V_n^{tr}})|V_{n,i}^{tr} = 1) \times (1 - p_n) \\ &\quad + \mathbb{E}_{V_n^{tr}}((\mathbb{P}_{n,V_n^{ts}} - \mathbb{P}'_{n,V_n^{ts}})\psi_{V_n^{tr}}|V_{n,i}^{ts} = 1) \times p_n|\end{aligned}$$

We always have for any ψ, $|(\mathbb{P}_{n,V_n^{ts}} - \mathbb{P}'_{n,V_n^{ts}})\psi| \leq 1/np_n$ thus $|\mathbb{E}_{V_n^{tr}}((\mathbb{P}_{n,V_n^{ts}} - \mathbb{P}'_{n,V_n^{ts}})\psi_{V_n^{tr}}, V_n^{ts} = 1) \times p_n| \leq 1/n$

We still have to bound $|\mathbb{E}_{V_n^{tr}}(\mathbb{P}_{n,V_n^{ts}}(\psi_{V_n^{tr}} - \psi'_{V_n^{tr}})|V_{n,i}^{tr} = 1)|$ which is always smaller than $\mathbb{E}_{V_n^{tr}}(d_\infty(\psi_{V_n^{tr}}, \psi'_{V_n^{tr}})|V_{n,i}^{tr} = 1)$ in the special case of the most stable kind of stability namely the uniform stability.

On $B^{\subset}$, we get $d_\infty(\psi_{V_n^{tr}}, \psi'_{V_n^{tr}}) \leq d_\infty(\psi_{V_n^{tr}}, \psi_{n+1}) + d_\infty(\psi_{n+1}, \psi'_{V_n^{tr}}) \leq 4\lambda p_n$.
Thus, on $B^{\subset}$, we derive

$$\mathbb{E}_{V_n^{tr}}(d_\infty(\psi_{V_n^{tr}}, \psi'_{V_n^{tr}})|V_{n,i}^{tr} = 1) \leq 4\lambda p_n.$$

Putting all together, with probability at least $1 - \delta_{n,p_n}$, we get

$$\sup_{1 \leq i \leq n, z \in \mathcal{Z}} |f(Z_1, \ldots, Z_i, \ldots, Z_n) - f(Z_1, \ldots, z, \ldots, Z_n)| \leq \frac{4\lambda}{n+1} + 4\lambda p_n(1 - p_n) \leq 8\lambda p_n.$$

Applying theorem 79, we obtain that for all $\varepsilon \geq 0$:

$$\Pr(\mathbb{E}_{V_n^{tr}}(\mathbb{P}\psi_{V_n^{tr}} - \mathbb{P}_{n,V_n^{ts}}\psi_{V_n^{tr}}) \geq \varepsilon) \leq 2(\exp(-\frac{\varepsilon^2}{8n(8\lambda p_n + \alpha)^2}) + \frac{n}{\alpha}\delta'_{n,p_n})$$
$$\leq 2(\exp(-\frac{\varepsilon^2}{8(16\lambda)^2 n p_n^2}) + \frac{n}{8\lambda p_n}\delta'_{n,p_n}) \text{ by taking } \alpha = 8\lambda p_n$$

□

Theorem 121 (Cross-validation Weak stability)**.** *Suppose that $\mathcal{H}$ holds. Let Ψ be a machine learning which is weak $(\lambda, (\delta_{n,p_n})_{n,p_n}, \mathbb{Q})$ stable with respect to the distance d. Then, for all $\varepsilon \geq 0$, we have*

$$\Pr\left(\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out} \geq \varepsilon\right) \leq \exp(-2np_n\varepsilon^2).$$

Furthermore, if the distance is the uniform distance d_∞, we have for all $\varepsilon \geq 0$:

$$\Pr\left(\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out} \geq \varepsilon\right) \leq \min(\exp(-2np_n\varepsilon^2), 2(\exp(-\frac{n\varepsilon^2}{10(9\lambda np_n)^2} + \frac{n\delta_{n(1-p_n)}^{1/2}}{9\lambda p_n}\exp(\frac{\varepsilon n}{4(9\lambda np_n)^2})) + n\delta_{n,p_n}^{1/2}).$$

Proof.
Denote $f(Z_1, Z_2, \ldots, Z_n) := \hat{R}_{CV}^{Out} - \widetilde{R}_n$ and $B := \{\sup_{U_n \in \text{support}(\mathbb{Q})} \frac{d(\psi_{U_n}, \psi_{n+1})}{||\mathbb{P}_{n,U_n} - \mathbb{P}_{n+1}||} \geq \lambda\}$ with ψ_{n+1} trained on $\mathcal{D}_{n+1} = \{Z_1, \ldots, Z_{i-1}, Z_i, Z_{i+1}, \ldots, Z_n, Z_i'\}$.
We want to show that for all i, there exists constant c_i such $|\Delta_i| := |f(Z_1, \ldots, Z_i, \ldots, Z_n) - f(Z_1, \ldots, Z_i', \ldots, Z_n)| \leq c_i$ with high probability where $Z_1, \ldots, Z_i, \ldots, Z_n, Z_i'$ are i.i.d. variables.

$$|\Delta_i| = |\mathbb{E}_{V_n^{tr}}(\mathbb{P}\psi_{V_n^{tr}} - \mathbb{P}_{n,V_n^{ts}}\psi_{V_n^{tr}}) - (\mathbb{E}_{V_n^{tr}}\mathbb{P}\psi'_{V_n^{tr}} - \mathbb{P}'_{n,V_n^{ts}}\mathbb{P}\psi'_{V_n^{tr}})|$$
$$\leq \mathbb{E}_{V_n^{tr}}|\mathbb{P}(\psi_{V_n^{tr}} - \psi'_{V_n^{tr}})| + \mathbb{E}_{V_n^{tr}}|(\mathbb{P}_{n,V_n^{ts}}\psi_{V_n^{tr}} - \mathbb{P}'_{n,V_n^{ts}}\mathbb{P}\psi'_{V_n^{tr}})|.$$

with $\mathbb{P}'_n, \mathbb{P}'_{n,V_n^{ts}}$ the weighted empirical measures of the sample $\mathcal{D}'_n = \{Z_1, \ldots, Z_i', \ldots, Z_n\}$ and ψ'_n the predictor built on $\mathcal{D}'_n$.

So, first, let us bound the first term, $|\mathbb{E}_{V_n^{tr}}\mathbb{P}(\psi_{V_n^{tr}} - \psi'_{V_n^{tr}})| \leq \mathbb{E}_{V_n^{tr}}|\mathbb{P}(\psi_{V_n^{tr}} - \psi_{n+1})| + \mathbb{E}_{V_n^{tr}}|\mathbb{P}(\psi_{n+1} - \psi'_{V_n^{tr}})|$ Thus, on $B^\subset$, we have $|\mathbb{E}_{V_n^{tr}}\mathbb{P}(\psi_{V_n^{tr}} - \psi'_{V_n^{tr}})| \leq \frac{4\lambda}{n+1}$.
To upper bound the second term, notice that:

$$|\mathbb{E}_{V_n^{tr}}P_{n,V_n^{ts}}\psi_{V_n^{tr}} - \mathbb{E}_{V_n^{tr}}P'_{n,V_n^{ts}}\psi'_{V_n^{tr}}| = |\mathbb{E}_{V_n^{tr}}(P_{n,V_n^{ts}}(\psi_{V_n^{tr}} - \psi'_{V_n^{tr}}), V_n^{tr} = 1) \times (1 - p_n)$$
$$+\mathbb{E}_{V_n^{tr}}((P_{n,V_n^{ts}} - P'_{n,V_n^{ts}})\psi_{V_n^{tr}}, V_n^{ts} = 1) \times p_n|.$$

We always have for all ψ, $|(\mathbb{P}_{n,V_n^{ts}} - \mathbb{P}'_{n,V_n^{ts}})\psi| \leq 1/np_n$ thus we get

$$|\mathbb{E}_{V_n^{tr}}((\mathbb{P}_{n,V_n^{ts}} - \mathbb{P}'_{n,V_n^{ts}})\psi_{V_n^{tr}}, V_n^{ts} = 1) \times p_n| \leq 1/n.$$

We still have to bound $|\mathbb{E}_{V_n^{tr}}(\mathbb{P}_{n,V_n^{ts}}(\psi_{V_n^{tr}} - \psi'_{V_n^{tr}}), V_n^{tr} = 1)| \leq \mathbb{E}_{V_n^{tr}}(d_\infty(\psi_{V_n^{tr}}, \psi'_{V_n^{tr}}), V_n^{tr} = 1)$ in the special of the uniform stability.

On $B^\subset$, we derive $d_\infty(\psi_{V_n^{tr}}, \psi'_{V_n^{tr}}) \leq d_\infty(\psi_{V_n^{tr}}, \psi_{n+1}) + d_\infty(\psi_{n+1}, \psi'_{V_n^{tr}}) \leq 4\lambda p_n$, thus on $B^\subset$

$$\mathbb{E}_{V_n^{tr}}(d_\infty(\psi_{V_n^{tr}}, \psi'_{V_n^{tr}}), V_n^{tr} = 1) \leq 4\lambda p_n.$$

Putting all together, with probability at least $1 - \delta_{n,p_n}$,

$$|f(Z_1, \ldots, Z_i, \ldots, Z_n) - f(Z_1, \ldots, Z_{i'}, \ldots, Z_n)| \leq 8\lambda p_n.$$

□

Following the previous results, we can obtain results for the expectation of the difference $\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out}$.

Theorem 122. *In the case of classification, we can bound the excess risk by*

$$\mathbb{E}_{\mathcal{D}_n}(\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out}) \leq \sqrt{1/np_n}$$

Furthermore, if d is the uniform distance d_∞, then we have for all $\alpha > 0$:

$$\mathbb{E}_{\mathcal{D}_n}(\widetilde{R}_n(\Phi_n^B) - \hat{R}_{CV}^{Out}) \leq \min(\sqrt{1/np_n}, \sqrt{16^3 n}\lambda p_n + \frac{n}{4\lambda p_n}\delta_{n,p_n})$$

Similar results can be derived in the context of the weak stability.
Proof
It is sufficient to apply the previous probability upper bounds together with the lemma 112.

□

3.4 Results for the cross-validated subagged classification

In the case of subagging of classifiers (i.e. the majority vote), we can obtain the following results:

Theorem 123. *For any subbaged classifier, we can bound the excess risk.*

$$\Pr(\widetilde{R}_n(\Phi_n^B) - \frac{1}{2}\hat{R}_{CV}^{Out} \geq \varepsilon) \leq \exp(-8np_n\varepsilon^2/9)$$

and also

$$\Pr(\widetilde{R}_n(\Phi_n^B) - l\hat{R}_{CV}^{Maj} \geq \varepsilon) \leq l\exp(-2np_n\varepsilon^2/9)$$

where N denotes the total number of training vectors in the cross-validation and l denotes $[(N-1)/2] + 1$ that is the strict majority of the subbaged classifiers and $\hat{R}_{CV}^{Maj}$ the cross-validated estimate of this majority.
Furthermore, in the particular case of binary classification we also have

$$\Pr(\widetilde{R}_n(\Phi_n^B) - (\hat{R}_{CV}^{Out}/2 - 1/2)) \leq -\varepsilon) \leq \exp(-2np_n\varepsilon^2/9)$$

and

$$\Pr(\widetilde{R}_n(\Phi_n^B) - (l\hat{R}_{CV}^{Maj} - l + 1) \leq -\varepsilon) \leq l\exp(-2np_n\varepsilon^2)$$

Proof.
We consider a ghost sample i.i.d. of size m: $(X_1^{'}, Y_1^{'}), ..., (X_m^{'}, Y_m^{'})$. Denote $\eta_i := L(Y_i^{'}, \phi_n^B(X_i^{'}))$. Then $e_m^B := \frac{1}{m}\sum_{i=1}^m \eta_i$ corresponds to the average number of mistakes of Φ_n^B on the ghost sample. In the same way, $e_m^{V_n^{tr}} := \frac{1}{m}\sum_{i=1}^m L(Y_i^{'}, \phi_{V_n^{tr}}(X_i^{'}))$ (respectively $e_m^a := \mathbb{E}_{V_n^{tr}}[\frac{1}{m}\sum_{i=1}^m L(Y_i^{'}, \phi_{V_n^{tr}}(X_i^{'}))]$) is the average number of the mistakes of $\phi_{V_n^{tr}}$ (respectively the weighted average number of mistakes of the family of predictors $\phi_{V_n^{tr}}$).

Denote by

1. $L_1 := \widetilde{R}_n(\Phi_n^B) - \frac{1}{2}\hat{R}_{CV}^{Out}$
2. $L_2 := \widetilde{R}_n(\Phi_n^B) - e_m^B$
3. $L_3 := e_m^B - e_m^a/2$
4. $L_4 := \frac{1}{2}[e_m^a - E_{X,Y}\mathbb{E}_{V_n^{tr}}L(Y, \phi_{V_n^{tr}}(X))]$
5. $L_5 := \frac{1}{2}[E_{X,Y}\mathbb{E}_{V_n^{tr}}L(Y, \phi_{V_n^{tr}}(X)) - \hat{R}_{CV}^{Out}]$

We have

$$\Pr(L_1 \geq 3\varepsilon) \leq \Pr(L_2 \geq \varepsilon) + \Pr(L_3 \geq 0) + \Pr(L_4 \geq \varepsilon) + \Pr(L_5 \geq \varepsilon)$$

By Hoeffding's inequality, we have:

$$\Pr(L_2 \geq \varepsilon) \leq \exp(-2m\varepsilon^2).$$

and also $\Pr(L_4 \geq \varepsilon) \leq \exp(-2m(2\varepsilon)^2)$

By conditionnal Hoeffding's inequality (for a proof, see e.g. [COR09A]), we deduce

$$\Pr(L_5 \geq \varepsilon) \leq \exp(-2np_n(2\varepsilon)^2)$$

By conditionnal Hoeffding's inequality, we also have

$$\Pr(e_m^a - E_{X,Y}\mathbb{E}_{V_n^{tr}}L(Y, \phi_{V_n^{tr}}(X)) \geq \varepsilon) \leq \exp(-2m\varepsilon^2).$$

since for fixed v_n^{tr} $\Pr(\frac{1}{m}\sum_{i=1}^m L(Y_i^{'}, \phi_{v_n^{tr}}(X_i^{'})) - E_{X,Y}L(Y, \phi_{v_n^{tr}}(X)) \geq \varepsilon) \leq \exp(-2m\varepsilon^2)$

We suppose here that $\Pr(V_n^{tr} = v_n)$ are rational numbers whose smallest multiplicator is denoted by N. Thus e_m^a can be seen as a simple average number of mistakes of a family of predictors $(\phi_j)_{1\leq j\leq N}$ on the ghost sample.

First notice, that if e_m^a is small then e_m^B must be small either. Indeed,we have

$$e_m^a = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{m}\sum_{i=1}^{m} L(Y_i^{'}, \phi_j(X_i^{'})) = \frac{1}{N}\frac{1}{m}\sum_{1\leq j\leq N, 1\leq i\leq m}^{N} \epsilon_{i,j}$$

with $\epsilon_{i,j} := L(Y_i^{'}, \phi_j(X_i^{'})) \in \{0,1\}$. We thus deduce that the total number of mistakes on the ghost sample of the family of predictors $(\phi_j)_{1\leq j\leq N}$ is equal to Nme_m^a. Notice that if the number of mistakes of the family $(\phi_j)_{1\leq j\leq N}$ on the i-th observation is less that $\lfloor (N-1)/2 \rfloor$ (i.e. $\sum_{j=1}^{N}\epsilon_{i,j} \leq \lfloor (N-1)/2 \rfloor$) then it means that a strict majority of predictors have classified correctly $Y_i^{'}$, which in turns tells us that a strict majority of predictors have the same output $Y_i^{'} = \phi_j(X_i^{'})$. We thus have $\phi_n^B(X_i^{'}) = Y_i^{'}$ which implies $\eta_j = L(Y_i^{'}, \phi_n^B(X_i^{'})) = 0$.

Denoting by $\kappa = me_m^B$ the number of mistakes of the subbaged classifier on the ghost sample, we necessarly have

$$\sum_{i=1}^{m}\sum_{j=1}^{N}\epsilon_{i,j} \geq \kappa(\lfloor (N-1)/2 \rfloor + 1) = \kappa(\lfloor (N+1)/2 \rfloor).$$

It follows that

$$e_m^B \leq \frac{N}{\lfloor (N+1)/2 \rfloor} e_m^a < e_m^a/2.$$

Thus $\Pr(L_3 \geq 0) = 0$

We conclude $\Pr(\widetilde{R}_n(\Phi_n^B) - \frac{1}{2}\hat{R}_{CV}^{Out} \geq 3\varepsilon) \leq \exp(-2np_n(2\varepsilon)^2) + \exp(-2m(2\varepsilon)^2) + \exp(-2m\varepsilon^2)$. If we let $m \to \infty$,

$$\Pr(\widetilde{R}_n(\Phi_n^B) - \frac{1}{2}\hat{R}_{CV}^{Out} \geq \varepsilon) \leq \exp(-8np_n\varepsilon^2/9)$$

Notice that in the particular case of the binary classification, we have by symmetry, $1 - e_m^B \leq \frac{N}{\lfloor (N+1)/2 \rfloor}(1 - e_m^a)$, which gives

$$\frac{N}{\lfloor N/2+1 \rfloor} e_m^a - (1 - \frac{N}{\lfloor N/2+1 \rfloor}) \leq e_m^B$$

and eventually $e_m^B \geq \frac{N}{\lfloor N/2+1 \rfloor} e_m^a - 1/2 \geq e_m^a - 1/2$

Thus, for binary classification, we can even obtain an probability upper bound for $\Pr(|\widetilde{R}_n(\Phi_n^B) - \frac{1}{2}\hat{R}_{CV}^{Out}| \geq \varepsilon)$ not only for $\Pr(\widetilde{R}_n(\Phi_n^B) - \frac{1}{2}\hat{R}_{CV}^{Out} \geq \varepsilon)$. Indeed, denote by

1. $L_1^{'} := \widetilde{R}_n(\Phi_n^B) - \frac{N}{\lfloor N/2+1 \rfloor}(\hat{R}_{CV}^{Out} - 1/2)$
2. $L_2^{'} := \widetilde{R}_n(\Phi_n^B) - e_m^B$
3. $L_3^{'} := e_m^B - (\frac{N}{\lfloor N/2+1 \rfloor} e_m^a - 1/2)$
4. $L_4^{'} := (\frac{N}{\lfloor N/2+1 \rfloor} e_m^a - 1/2) - (\frac{N}{\lfloor N/2+1 \rfloor} E_{X,Y} \mathbb{E}_{V_n^{tr}} L(Y, \phi_{V_n^{tr}}(X)) - 1/2)$
5. $L_5^{'} := (\frac{N}{\lfloor N/2+1 \rfloor} E_{X,Y} \mathbb{E}_{V_n^{tr}} L(Y, \phi_{V_n^{tr}}(X)) - 1/2) - (\frac{N}{\lfloor N/2+1 \rfloor} \hat{R}_{CV}^{Out} - 1/2)$

We get

$$\begin{aligned}\Pr(L_1^{'} \leq -3\varepsilon) &\leq \Pr(L_2^{'} \leq -\varepsilon) + \Pr(L_3^{'} < 0) + \Pr(L_4^{'} \leq -\varepsilon) + \Pr(L_5^{'} \leq -\varepsilon) \\ &\leq \exp(-2m\varepsilon^2) + 0 + \exp(-2m(\frac{N}{\lfloor N/2+1 \rfloor}\varepsilon)^2) + \exp(-2np_n(\frac{N}{\lfloor N/2+1 \rfloor}\varepsilon)^2)\end{aligned}$$

Taking $m \to \infty$, and noticing that $N/\lfloor N/2+1 \rfloor > 1$

$$\begin{aligned}\Pr(\widetilde{R}_n(\Phi_n^B) - (\hat{R}_{CV}^{Out}/2 - 1/2)) \leq -\varepsilon) &\leq \Pr(\widetilde{R}_n(\Phi_n^B) - (\hat{R}_{CV}^{Out}/2 - 1/2) \leq -\varepsilon) \\ &\leq \Pr(L_1^{'} \leq -\varepsilon) \leq \exp(-2np_n\varepsilon^2/9)\end{aligned}$$

For binary classification, we can eventually obtain that

$$\Pr(|\widetilde{R}_n(\Phi_n^B) - \frac{1}{2}(\hat{R}_{CV}^{Out} - 1/2)| \geq \varepsilon) \leq \exp(-8np_n\varepsilon^2/9) + \exp(-2np_n\varepsilon^2/9) \leq 2\exp(-2np_n\varepsilon^2/9)$$

Denote by $\epsilon_j := \frac{1}{m}\sum_{i=1}^m \epsilon_{i,j}$ the average number of mistakes by predictors j on the ghost sample. We can order them by increasing order: $\epsilon_{(1)}, ..., \epsilon_{(N)}$. Let $l := \lfloor N/2+1 \rfloor$ be the strict majority. An interesting case is when we know that a strict majority of classifiers are very good. Denote by

$$e_m^G := \frac{1}{l}\sum_{j=1}^{l} \epsilon_{(j)}$$

their global average error of the first l best classifiers on the ghost sample.

In the same way, denote by $\mu_j := E_{X,Y} L(Y, \phi_j(X))$ the risk of the j-th classifier. We introduce now a cross-validation estimate of the average risk $\frac{1}{l}\sum_{j=1}^{l} \mu_{(j)}$ of the l best classifiers: $\hat{R}_{CV}^{Maj}$. For this, recall that each ϕ_j corresponds to some $\phi_{v_n^{tr}}$ thus we can define an out sample error for the predictor j : $\hat{r}_j := \mathbb{P}_{n,v_n^{ts}}(L(Y, \phi_j(X))$. And we define $\hat{R}_{CV}^{Maj} := \frac{1}{l}\sum_{j=1}^{l} \hat{r}_{(j)}$

1. $R_1 := \widetilde{R}_n(\Phi_n^B) - l\hat{R}_{CV}^{Maj}$
2. $R_2 := \widetilde{R}_n(\Phi_n^B) - e_m^B$
3. $R_3 := e_m^B - le_m^G$
4. $R_4 := l(e_m^G - \frac{1}{l}\sum_{j=1}^{l}\mu_{(j)})$
5. $R_5 := l(\frac{1}{l}\sum_{j=1}^{l}\mu_{(j)} - \hat{R}_{CV}^{Maj})$

We have

$$\Pr(R_1 \geq 3\varepsilon) \leq \Pr(R_2 \geq \varepsilon) + \Pr(R_3 > 0) + \Pr(R_4 \geq \varepsilon) + \Pr(R_5 \geq \varepsilon)$$

By Hoeffding's inequality, we have:

$$\Pr(R_2 \geq \varepsilon) \leq \exp(-2m\varepsilon^2).$$

We also derive

$\Pr(R_4 \geq \varepsilon) = \Pr(e_m^G - \frac{1}{l}\sum_{j=1}^{l}\mu_{(j)} \geq \varepsilon/l) = \Pr(\sum_{j=1}^{l}\epsilon_{(j)} - \sum_{j=1}^{l}\mu_{(j)} \geq \varepsilon)$
There exist permutations σ and $\sigma^{'}$ such that $\epsilon_{(j)} = \epsilon_{\sigma(j)}$ and $\mu_{(j)} = \mu_{\sigma'(j)}$. Thus, we get

$$\begin{aligned}\Pr(R_4 \geq \varepsilon) &\leq \Pr(\sum_{j=1}^{l}\epsilon_{\sigma(j)} - \mu_{\sigma'(j)} \geq \varepsilon) \\ &\leq \Pr(\sum_{j=1}^{l}\epsilon_{\sigma'(j)} - \mu_{\sigma'(j)} \geq \varepsilon)\end{aligned}$$

by definition of $\epsilon_{(j)}$. It follows that

$$\begin{aligned}\Pr(R_4 \geq \varepsilon) &\leq \sum_{j=1}^{l}\Pr(\epsilon_{\sigma'(j)} - \mu_{\sigma'(j)} \geq \varepsilon) \\ &\leq l\exp(-2m\varepsilon^2).\end{aligned}$$

In the same way, we deduce $\Pr(R_5 \geq \varepsilon) \leq l\exp(-2np_n\varepsilon^2)$.
By conditional Hoeffding's inequality (for a proof, see e.g. [COR09A]), we deduce $\Pr(L_5 \geq \varepsilon) \leq \exp(-2np_n(2\varepsilon)^2)$ and also for a fixed v_n^{tr}

$$\Pr(|e_m^{v_n^{tr}} - E_{X,Y}L(Y, \phi_{v_n^{tr}}(X))| \geq \varepsilon) \leq 2\exp(-2m\varepsilon^2).$$

By conditional Hoeffding's inequality (for a proof, see e.g. [COR09A]), we also have

$$\Pr(|e_m^a - E_{X,Y}\mathbb{E}_{V_n^{tr}} L(Y, \phi_{V_n^{tr}}(X))| \geq \varepsilon) \leq 2\exp(-2m\varepsilon^2).$$

Notice that if all the l best classifiers classify correctly the i-th observation (i.e. $\epsilon_{i,(j)} = 0$ for all $j \in \{1, ..., M\}$), then the subbaged classification classifies also correctly. Thus $\eta_i = 0$. Let κ be the number of mistakes of the subbaged classifier on the ghost sample and let x the number of observations correctly classified by all the l classifiers. Then we obtain that the number of correctly classified observations by the subagging is greater that x, i.e. $m - \kappa \geq x$. On the other hand, there is at least one predictor that makes a mistake on each of the remaining $m - x$ observations. Thus $m - x$ is less that the total number of mistakes made by the l best classifiers

$$(m - x) \leq m l e_m^G.$$

From which, it follows that

$$e_m^B \leq l e_m^G.$$

Thus $\Pr(R_3 > 0) = 0$.

Putting altogether, we have

$$\Pr(\widetilde{R}_n(\Phi_n^B) - l\hat{R}_{CV}^{Maj} \geq 3\varepsilon) \leq \exp(-2m\varepsilon^2) + l\exp(-2m\varepsilon^2) + l\exp(-2np_n\varepsilon^2).$$

If we let $m \to \infty$, $\Pr(\widetilde{R}_n(\Phi_n^B) - l\hat{R}_{CV}^{Maj} \geq \varepsilon) \leq l\exp(-2np_n\varepsilon^2/9)$.

Once again, in the particular case of binary classification, we have by symmetry $1 - e_m^B \leq l(1 - e_m^G)$ which leads to

$$e_m^B \geq 1 - l(1 - e_m^G).$$

In the same way, we have a symmetrical result for binary classification:

$$\begin{aligned}\Pr(\widetilde{R}_n(\Phi_n^B) - (l\hat{R}_{CV}^{Maj} - l + 1) \leq -3\varepsilon) &\leq \exp(-2m\varepsilon^2) + l\exp(-2m\varepsilon^2) + l\exp(-2np_n\varepsilon^2) \\ &\leq l\exp(-2np_n\varepsilon^2).\end{aligned}$$

which gives $\Pr(|\widetilde{R}_n(\Phi_n^B) - (l\hat{R}_{CV}^{Maj} - l + 1)| \geq \varepsilon) \leq 2l\exp(-2np_n\varepsilon^2/9)$.

□

In the case of subagging of classifiers (i.e. the majority vote) whose VC dimension is finite, we can obtain a stronger result:

Theorem 124. *Suppose $\mathcal{H}$ holds and that the machine learning is based on empirical risk minimization. We can bound the excess risk.*

$$\Pr(\widetilde{R}_n(\Phi_n^B) - \frac{1}{2}\hat{R}_{CV}^{Out} \geq \varepsilon) \leq \min(\exp(-8np_n\varepsilon^2/9), (2n(1-p_n)+1)^{4V_C/(1-p_n)}e^{-4n(1-p_n)\varepsilon^2}).$$

and also

$$\Pr(\widetilde{R}_n(\Phi_n^B) - l\hat{R}_{CV}^{Maj} \geq \varepsilon) \leq l \exp(-2np_n\varepsilon^2/9)$$

with the $l := [(N-1)/2] + 1$ *the strict majority of the subagged classifiers and* $\hat{R}_{CV}^{Maj}$ *the cross-validated estimate of this majority.*
Furthermore, in the particular case of binary classification we also have

$$\Pr(\widetilde{R}_n(\Phi_n^B) - (\hat{R}_{CV}^{Out}/2 - 1/2)) \leq -\varepsilon) \leq \min(\exp(-2np_n\varepsilon^2/9), (2n(1-p_n)+1)^{4V_{\mathcal{C}}/(1-p_n)} e^{-4n(1-p_n)\varepsilon^2})$$

and

$$\Pr(\widetilde{R}_n(\Phi_n^B) - (l\hat{R}_{CV}^{Maj} - l + 1) \leq -\varepsilon) \leq l \exp(-2np_n\varepsilon^2)$$

Proof.

We use again the lemma (for a proof, see chapter 1): $\hat{R}_{CV}^{Out} \geq \mathbb{P}_n L(Y, \phi_n(X))$ since

$$\phi_n = \arg\min_{\phi \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, \phi(X_i)).$$

Following the last proof, we can bound L_5 in another way.

$$\begin{aligned} \Pr(L_5 \geq 3\varepsilon) &\leq \Pr(\mathbb{E}_{V_n^{tr}}[E_{X,Y} L(Y, \phi_{V_n^{tr}}(X)) - \mathbb{P}_n L(Y, \phi_n(X))] \geq 6\varepsilon) \\ &\leq \Pr(\mathbb{E}_{V_n^{tr}}[E_{X,Y} L(Y, \phi_{V_n^{tr}}(X)) - \mathbb{P}_n L(Y, \phi_n(X))] \geq 6\varepsilon) \end{aligned}$$

Then as in proof, we split according to $\mathbb{P}L(Y, \phi_{opt}(X))$ and we obtain by lemma 112

$$\Pr(L_5 \geq \varepsilon) \leq (2n(1-p_n)+1)^{4V_{\mathcal{C}}/(1-p_n)} e^{-n(1-p_n)(2\varepsilon)^2}$$

□

3.5 Results for the subagged predictor selection

The remaining important question is: in practice, how should we choose p_n? We give a hint for this question.
First, suppose that the final user wants to have an accuracy equal to a certain level η.

Then we need to provide him a rule to chose an optimal $p_n^\star$ and to upper bound the probability of excess risk $\Pr(\widetilde{R}_n(\phi_n^{B,p_n^\star}) - \hat{R}_{CV}^{Out}(p_n^\star) \geq \eta)$. Previous bounds tell us that for any fixed p_n, $\Pr(\widetilde{R}_n(\phi_n^B) - \hat{R}_{CV}^{Out}(p_n) \geq \varepsilon) \leq \min(B(n, p_n, \varepsilon), V(n, p_n, \varepsilon))$. Notice that $\min(B(n, p_n, \varepsilon), V(n, p_n, \varepsilon))$ seen as a function of ε is a continuous non-increasing function. Thus, we can define an inverse denoted by f.

The previous probability bound becomes for any p_n: $\Pr(\widetilde{R}_n(\phi_n^B) - \hat{R}_{CV}^{Out}(p_n) \geq f(n, p_n, \delta) \leq \delta$. For each k, define $\delta_{n,k}$ by $f(n, k/n, \delta_{n,k}) = \eta$, i.e. $\delta_{n,k} = \min(B(n, k/n, \eta), V(n, k/n, \eta))$. Denote $k_n^\star := \arg\min_{k\in\{1\ldots n-1\}} \hat{R}_{CV}^{Out}(k/n) + f(n, k/n, \delta_{n,k})$ and denote by $p_n^\star := k_n^\star/n$. Thus, we obtain:

Theorem 125 (Subbaging selection)**.** *Suppose that $\mathcal{H}$ holds. Suppose also that ϕ_n is based on empirical risk minimization. But instead of minimizing $\widehat{R}_n(\phi)$, we suppose ϕ_n minimizes $\frac{1}{n}\sum_{i=1}^n C(h(Y_i, \phi(X_i))$. For simplicity, we suppose the infimum is attained i.e. $\phi_n = \arg\min_{\phi\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^n C(h(Y_i, \phi(X_i))$. In this context, we have:*

- *if $\delta \geq \delta_n$*

$$f(n, p_n, \delta) = \sqrt{\frac{\ln(1/\delta)}{2np_n}}$$

- *and if $\delta < \delta_n$,*

$$f(n, p_n, \delta) = 3\sqrt{\frac{4V_{\mathcal{C}}\ln(2n(1-p_n)+1)/(1-p_n) + \ln(1/\delta)}{n}}$$

with $\delta_n := (2n(1-p_n)+1)^{-\frac{4p_nV_{\mathcal{C}}}{(1-p_n)(1/9-2p_n)}}$.
Furthermore, we have for all $\varepsilon > 0$:

$$\Pr(\widetilde{R}_n(\phi_n^{B,p_n^\star}) - \hat{R}_{CV}^{Out}(p_n^\star) \geq \varepsilon) = O_n((n+1)^{8V_{\mathcal{C}}}\exp\left(-\frac{2n(\varepsilon - 2\sqrt{2}V_{\mathcal{C}}^{1/2}\sqrt{\ln(n)/n})^2}{1-\exp(-2\varepsilon^2)}\right)).$$

Proof
We have:

$$\begin{aligned}\Pr(\widetilde{R}_n(\phi_n^{B,p_n^\star}) - \hat{R}_{CV}^{Out}(p_n^\star) \geq \eta) &= \Pr(\widetilde{R}_n(\phi_n^{B,p_n^\star}) - \hat{R}_{CV}^{Out}(p_n^\star) \geq f(n, p_n^\star, \delta_{n,k_n^\star}))\\ &\leq \sum_{k\in\{1\ldots n-1\}} \Pr(\widetilde{R}_n(\phi_n^{B,p_k}) \geq \hat{R}_{CV}^{Out}(p_k) + f(n, k/n, \delta_{n,k})).\end{aligned}$$

It follows that:

$$\begin{aligned}\Pr(\widetilde{R}_n(\phi_n^{B,p_n^\star}) - \hat{R}_{CV}^{Out}(p_n^\star) \geq \eta) &\leq \sum_{k\in\{1\ldots n-1\}} \Pr(\widetilde{R}_n(\phi_n^{B,p_k}) - \hat{R}_{CV}^{Out}(p_k) \geq \eta)\\ &\leq \sum_{k\in\{1\ldots n-1\}} \min(B(n, k/n, \eta), V(n, k/n, \eta)).\end{aligned}$$

Thus, using previous bounds we get:

$$\begin{aligned}\Pr(\widetilde{R}_n(\phi_n^{B,p_n^\star}) - \hat{R}_{CV}^{Out}(p_n^\star) \geq \eta) \leq & \min_{k_0 \in \{1 \ldots n-1\}} (\sum_{k=1}^{k_0-1} (2n(1-k/n)+1)^{4V_{\mathcal{C}}/(1-k/n)} \exp(-2n\eta^2) \\ & + \sum_{k=k_0}^{n-1} \exp(-2k\eta^2)) \\ \leq & \min_{k_0 \in \{1 \ldots n-1\}} (k_0(2n+1)^{4V_{\mathcal{C}}/(1-k_0/n)} \exp(-2n\eta^2) \\ & + \exp(-2k_0\eta^2) \frac{1-(\exp(-2k\eta^2))^{n-k_0}}{1-\exp(-2\eta^2)}) \\ \leq & \min_{k_0 \in \{1 \ldots n-1\}} ((2n+1)^{4V_{\mathcal{C}}/(1-k_0/n)} \alpha^n + \frac{\alpha^{k_0}}{1-\alpha}) \text{ with } \alpha := \exp(-2\eta^2)\end{aligned}$$

We look for k_0 in $\{(1-z_n)n, 0 < z_n < 1 \text{ and } z_n \to_{n\infty} 0\}$

$$\Pr(\widetilde{R}_n(\phi_n^{B,p_n^\star}) - \hat{R}_{CV}^{Out}(p_n^\star) \geq \eta) \leq \min_{z_n} \left((2n+1)^{4V_{\mathcal{C}}/z_n} \alpha^n + \frac{\alpha^{(1-z_n)n}}{1-\alpha} \right)$$

We look for z_n such that $(2n+1)^{4V_{\mathcal{C}}/z_n} \sim_{n\infty} \frac{\alpha^{-z_n n}}{1-\alpha}$
Let us even find z_n such that $(2n+1)^{4V_{\mathcal{C}}/z_n} = \frac{\alpha^{-z_n n}}{1-\alpha}$. It is thus equivalent to: $-n\ln(\alpha)z_n^2 - \ln(1-\alpha)z_n - 4V_{\mathcal{C}}\ln(2n+1) = 0$
We have $\Delta = \ln(1-\alpha)^2 - 16V_{\mathcal{C}}\ln(2n+1)n\ln(\alpha) > 0$ since $|\alpha| < 1$
Since $0 < z_n < 1$, we have necesseraly z_n the non negative root of the previous equation which leads to:

$$\begin{aligned}z_n &= \frac{\ln(1-\alpha) + \sqrt{\ln(1-\alpha)^2 - 16V_{\mathcal{C}}\ln(2n+1)n\ln(\alpha)}}{-2n\ln(\alpha)} \\ &\sim \frac{4V_{\mathcal{C}}^{1/2}}{\ln(1/\alpha)^{1/2}} \sqrt{\frac{\ln(n)}{n}} \\ &\sim \frac{2\sqrt{2}V_{\mathcal{C}}^{1/2}}{\eta} \sqrt{\frac{\ln(n)}{n}}\end{aligned}$$

We can inject z_n in $(2n+1)^{4V_{\mathcal{C}}/z_n}\alpha^n + \frac{\alpha^{(1-z_n)n}}{1-\alpha}$ and we find that

$$\Pr(\widetilde{R}_n(\phi_n^{B,p_n^\star}) - \hat{R}_{CV}^{Out}(p_n^\star) \geq \eta) = O_n((n+1)^{8V_{\mathcal{C}}} \exp(-2n(\eta - 2\sqrt{2}V_{\mathcal{C}}^{1/2}\sqrt{\ln(n)/n})^2)/(1-\exp(-2\eta^2))$$

Let us now find the expression of f the inverse of $\min_\varepsilon(B(n,p_n,\varepsilon), V(n,p_n,\varepsilon))$ with

- $B(n,p_n,\varepsilon) = \min((2n(1-p_n)+1)^{\frac{4V_{\mathcal{C}}}{1-p_n}} \exp(-n\varepsilon^2/9))$
- $V(n,p_n,\varepsilon) = \exp(-2np_n\varepsilon^2)$.

In the case of ERM algorithm,

$$\exp(-2np_n\varepsilon^2) \leq (2n(1-p_n)+1)^{\frac{4V_{\mathcal{C}}}{1-p_n}} \exp(-n\varepsilon^2/9)$$

if and only if $-2np_n\varepsilon^2 \leq \frac{4V_{\mathcal{C}}}{1-p_n}\ln(2n(1-p_n)+1) - n\varepsilon^2/9$ which is equivalent to

$$n(1/9-2p_n)\varepsilon^2 \leq \frac{4V_{\mathcal{C}}\ln(2n(1-p_n)+1)}{1-p_n}$$

and also $\varepsilon \leq \sqrt{\frac{4V_{\mathcal{C}}\ln(2n(1-p_n)+1)}{n(1-p_n)(1/9-2p_n)}} := \varepsilon_n$.

Thus if $\varepsilon \leq \varepsilon_n$, it follows that $\min(B(n,p_n,\varepsilon),V(n,p_n,\varepsilon)) = \exp(-2np_n\varepsilon^2)$, thus if $\delta = \exp(-2np_n\varepsilon^2)$ we deduce that $\varepsilon = \sqrt{\frac{\ln(1/\delta)}{np_n}}$. If $\varepsilon > \varepsilon_n$, $\min(B(n,p_n,\varepsilon),V(n,p_n,\varepsilon)) = (2n(1-p_n)+1)^{\frac{4V_{\mathcal{C}}}{1-p_n}} \exp(-n\varepsilon^2/9)$. Thus if $\delta = (2n(1-p_n)+1)^{\frac{4V_{\mathcal{C}}}{1-p_n}} \exp(-n\varepsilon^2/9)$, we then deduce that $\varepsilon = 3\sqrt{\frac{4V_{\mathcal{C}}\ln(2n(1-p_n)+1)/(1-p_n)+\ln(1/\delta)}{n}}$. Denote $\delta_n = \exp(-2np_n\varepsilon_n^2) = \exp(-\frac{4p_nV_{\mathcal{C}}\ln(2n(1-p_n)+1)}{(1-p_n)(1/9-2p_n)}) = (2n(1-p_n)+1)^{-\frac{4p_nV_{\mathcal{C}}}{(1-p_n)(1/9-2p_n)}}$.

In conclusion, if $\delta \geq \delta_n$, we have:

$$f(n,p_n,\delta) = \sqrt{\frac{\ln(1/\delta)}{2np_n}}$$

and if $\delta < \delta_n$,

$$f(n,p_n,\delta) = 3\sqrt{\frac{4V_{\mathcal{C}}\ln(2n(1-p_n)+1)/(1-p_n)+\ln(1/\delta)}{n}} \leq 6\sqrt{\frac{V_{\mathcal{C}}\ln(2n+1)+\ln(1/\delta)}{n(1-p_n)}}.$$

□

In summary, the probability of the deviation between the out-of-bag cross-validation estimate and the generalization error is bounded by the minimum of a Hoeffding-type bound and a Vapnik-Chernovenkis-type bounds, and thus it is smaller than 1 even for small learning sets. Finally, we also give a simple rule on how to subbag the predictor. However, in the case of classification, we show that subagging strong learners can give a strong learner. It would be more interesting to answer the following question : can we obtain a similar result with the subagging of weak learners ?

3.6 Appendices

We will use the definition of strong difference bounded introduced by [KUT02] and a corollary of his main theorem inspired by [McD89].

Definition 126 (Kutin[KUT02]). *Let $\Omega_1, \ldots, \Omega_n$ be probability spaces. Let $\Omega = \prod_{k=1}^{n} \Omega_k$ and let X a random variable on Ω. We say that X is strongly difference bounded by (b, c, δ) if the following holds: there is a "bad" subset $B \subset \Omega$, where $\delta = \mathbb{P}(B)$. If $\omega, \omega' \in \Omega$ differ only in k-th coordinate, and $\omega \notin B$, then*

$$|X(\omega) - X(\omega')| \leq c$$

Furthermore, for any $\omega, \omega' \in \Omega$,

$$|X(\omega) - X(\omega')| \leq b$$

We will need the following theorem. It says in substance that a strongly difference bounded function of independent variables is closed to its expectation with high probability.

Theorem 127 (Kutin[KUT02]). *Let $\Omega_1, \ldots, \Omega_n$ be probability spaces. Let $\Omega = \prod_{k=1}^{n} \Omega_k$ and let X a random variable on Ω, which is strongly difference bounded by (b, c, δ). Assume $b \geq c \geq 0$ and $\alpha > 0$. Let $\mu = \mathbb{E}(X)$. Then, for any $\tau > 0$,*

$$\Pr(X - \mu \geq \tau) \leq 2(\exp(-\frac{\tau^2}{8n(c + b\alpha)^2}) + \frac{n}{\alpha}\delta)$$

We will use the definition of weak difference bounded introduced by [KUT02] and a corollary of his main theorem.

Definition 128 (Kutin). *Let $\Omega_1, \ldots, \Omega_n$ be probability spaces. Let $\Omega = \prod_{k=1}^{n} \Omega_k$ and let X a random variable on Ω. We say that X is weakly difference bounded by (b, c, δ) if the following holds: for any k,*

$$\forall^{\delta}(\omega, v) \in \Omega \times \Omega_k,\ \mathbb{P}(|X(\omega) - X(\omega^{'})|) \leq c$$

where $\omega_k^{'} = v$ and $\omega_i^{'} = \omega_i$ for $i \neq k$. and the notation $\forall^{\delta}\omega, \Phi(\omega)$ means "$\Phi(\omega)$ holds for all but but a δ fraction of Ω"

$$|X(\omega) - X(\omega')| \leq c$$

Furthermore, for any $\omega, \omega' \in \Omega$, differing only one coordinate:

$$|X(\omega) - X(\omega')| \leq b$$

We will need the following theorem. It says in substance that a weakly difference bounded function of independent variables is closed to its expectation with probability.

Theorem 129 (Kutin)**.** *Let $\Omega_1, \ldots, \Omega_n$ be probability spaces. Let $\Omega = \prod_{k=1}^{n} \Omega_k$ and let X a random variable on Ω.which is weakly difference bounded by (b, c, δ). Assume $b \geq c \geq 0$ and $\alpha > 0$. Let $\mu = \mathbb{E}(X)$. Then, for any $\varepsilon > 0$*

$$\Pr(|X - \mu| \geq \varepsilon) \leq 2\exp(-\frac{\varepsilon^2}{10nc^2(1 + \frac{2\varepsilon}{15nc})^2}) + \frac{2nb\delta^{1/2}}{c}\exp(\frac{\varepsilon b}{4nc^2})) + 2n\delta^{1/2}$$

Deuxième partie

Statistical models applied in economics and finance

Chapitre 4

Analyse factorielle dynamique multifréquence appliquée à la datation de la conjoncture française

Analyse factorielle dynamique multifréquence appliquée à la datation de la conjoncture française

Matthieu Cornec(*)

Dans cet article, nous datons de manière fine – mensuelle – les périodes de la conjoncture française de 1985 à 2004. Puis la méthode utilisée est appliquée aux enquêtes de conjoncture dans l'industrie afin de revisiter l'indicateur synthétique du climat des affaires dans l'industrie manufacturière publié par l'Insee.

La datation des cycles est rendue difficile par la multiplicité des indicateurs. Afin de pallier cette difficulté, les conjoncturistes construisent des indicateurs mensuels synthétiques. Toutefois, ces indicateurs conjoncturels mensuels présentent l'inconvénient de ne pas prendre en compte la série trimestrielle qui est déjà une synthèse particulière de grand intérêt pour les conjoncturistes : le PIB lui-même.

Murasawa et Mariano ont proposé une méthodologie utilisant simultanément des séries mensuelles et trimestrielles.

Un indicateur synthétique mensuel (I^*_{eco}) est ainsi construit à partir de grands indicateurs quantitatifs de l'économie française ne se limitant pas au PIB (l'indice de la production industrielle, les dépenses des ménages en produits manufacturés, l'effectif salarié).

Au vu de cet indicateur, nous distinguons sept phases de conjoncture durant cette période. Une seule récession apparaît, de septembre 1992 à mai 1993 :

– de janvier 1985 à janvier 1992 : forte croissance ;

– de février 1992 à août 1992 : stabilité ;

– de septembre 1992 à mai 1993 : récession ;

– de juin 1993 à décembre 1994 : forte croissance ;

– de janvier 1995 à avril 1997 : croissance moyenne ;

– d'avril 1997 à juin 2001 : forte croissance ;

– à partir de juillet 2001 : croissance modeste.

Cette datation apparaît cohérente avec celle obtenue par Doz et Lenglart sur la période 1985-1995, lesquels font leur analyse à partir de l'enquête de conjoncture dans l'industrie. S'agissant de la récession de 1993, le détail des comptes trimestriels montre en effet pour l'année 1993, une baisse du PIB réel. Cette dernière est principalement présente du quatrième trimestre de 1992 au deuxième trimestre de 1993 dans les dépenses de ménages, les investissements des entreprises et des ménages. Ici, nous la datons précisément de septembre 1992 à mai 1993.

(*) Division synthèse conjoncturelle, Insee.

Je remercie Xavier Bonnet, Éric Dubois, Stéphane Grégoir et Dominique Ladiray pour leurs précieux commentaires sur une version antérieure de cet article. Les erreurs qui pourraient subsister relèvent de ma seule responsabilité.

De la comparaison entre l'indicateur commun I^*_{eco} et les différents indicateurs quantitatifs, le conjoncturiste peut tirer deux enseignements :

– l'avance ou le retard d'une branche par rapport à l'activité générale ;

– le dynamisme relatif d'une branche par rapport à l'activité d'ensemble.

Par exemple :

– la consommation en produits manufacturés des ménages est soit en avance soit coïncidente avec le cycle économique. La consommation en produits manufacturés des ménages présente un profil plus volatil au mois le mois ce qui rend sa lecture plus difficile. Dès mars 1990, la consommation en produits manufacturés des ménages affiche un profil à la baisse, bien avant le déclin de l'indicateur synthétique $I^*_{eco-cumul}$;

– le PIB est imprécis pour dater les cycles. La simple lecture du PIB indiquerait une récession du quatrième trimestre de 1992 jusqu'au troisième trimestre de 1993. L'indicateur du climat conjoncturel général $I^*_{eco-cumul}$ limite cette dernière aux quatrième trimestre de 1992, premier trimestre de 1993 et au deuxième trimestre de 1993.

L'approche considérée présente ainsi deux intérêts majeurs :

– elle offre une grille de lecture de l'activité passée en intégrant de façon simultanée les différentes branches de l'économie ;

– elle donne un chiffrement fin (mensuel) de ces différentes périodes.

La méthode de Murasawa et de Mariano au PIB (trimestriel) est ensuite appliquée aux enquêtes de conjoncture dans l'industrie (mensuel) et au PIB français (trimestriel). Le but est d'extraire l'information commune entre activité économique et opinion qu'en ont les chefs d'entreprise. Doz et Lenglart ont montré l'intérêt d'une approche factorielle dynamique pour obtenir un facteur commun à partir des six soldes d'opinion émis dans l'enquête de conjoncture dans l'industrie. Le facteur commun apparaît comme un indicateur du climat général dans l'industrie.

Ce facteur commun apparaît très semblable à l'ancien facteur commun de Doz et Lenglart.

La comparaison entre ces deux facteurs apporte les enseignements suivants :

– nous pouvons comparer les opinions des chefs d'entreprise avec un indice conjoncturel reposant sur le PIB et faire ainsi la part entre anticipations et réalité économique ;

– cette différence peut aussi être interprétée comme un renseignement complémentaire du côté des services.

Sur la période d'estimation (1978 à 2004), nous remarquons :

– de 1978 à 1985, le facteur commun industrie est avancé sur le nouveau facteur commun intégrant le PIB. Ceci peut s'interpréter comme une avance du cycle du secteur industriel sur le cycle global ;

– de 1985 à 2001, les deux indicateurs sont très semblables, indiquant peut-être une synchronisation des cycles du secteur manufacturier avec le secteur des services ;

– de 2001 à 2003, nous remarquons de nouveau un décalage entre le cycle industriel et le cycle global.

En outre, un regard sur l'amplitude des facteurs communs semblerait indiquer :

– de 1980 à 1985, les industriels apparaissent plus pessimistes que de raison car le facteur commun industrie se trouvent en dessous du facteur commun intégrant le PIB ;

– de 2001 à 2004, les industriels apparaissent plus optimistes que ne laisse suggérer la situation de l'activité économique pour les raisons opposées.

Dans cet article, nous cherchons à dater de manière fine – mensuelle – les périodes de la conjoncture française de 1985 à 2004.

La datation des cycles est rendue difficile par la multiplicité des indicateurs. Afin de pallier cette difficulté, les conjoncturistes construisent des indicateurs *mensuels* synthétiques. Par exemple, Doz et Lenglart (1999) définissent un indicateur synthétique du climat des affaires en France à partir des soldes d'opinion de l'enquête industrie. Stock et Watson (1989) élaborent un indicateur coïncident à partir de grands indicateurs quantitatifs de l'économie américaine : nombre d'emplois hors agriculture, revenu des ménages avant les transferts, indice de la production industrielle, dépenses de produits manufacturés et ventes des secteurs manufacturiers et commerce.

Toutefois, ces indicateurs conjoncturels *mensuels* présentent l'inconvénient de ne pas prendre en compte la série *trimestrielle* qui est déjà une synthèse particulière de grand intérêt pour les conjoncturistes : le PIB lui-même.

Murasawa et Mariano (2003) ont proposé une méthodologie utilisant simultanément des séries mensuelles et trimestrielles.

Nous rappelons la méthodologie employée. Il s'agit d'une représentation espace-état où les séries trimestrielles sont considérées comme des séries mensuelles avec valeurs manquantes.

Tout d'abord, cette approche appliquée par Murasawa et Mariano à l'économie américaine est adaptée à l'économie française. Un indicateur synthétique mensuel (I^*_{eco}) est ainsi construit à partir de grands indicateurs quantitatifs de l'économie française ne se limitant pas au PIB. Au vu de cet indicateur, nous distinguons sept phases de conjoncture durant cette période. Une seule récession apparaît, de septembre 1992 à mai 1993.

Ensuite, nous revisitons l'indicateur synthétique du climat des affaires dans l'industrie manufacturière en incorporant le PIB. La comparaison de l'indicateur synthétique, actuellement utilisé, construit avec les six soldes d'opinion de l'enquête mensuelle de conjoncture et d'un nouvel indice (I^*_{en}) élaboré à partir de ces mêmes six soldes d'opinion et du PIB trimestriel permet d'analyser la composante commune entre les enquêtes et l'activité économique.

Rappel des méthodes existantes

Doz et Lenglart (1999) synthétisent l'information commune contenue dans les différentes questions dans l'enquête de conjoncture en construisant un indice mensuel du climat dans l'industrie manufacturière. L'analyse statistique s'effectue dans le cadre des modèles factoriels dynamiques. Un petit nombre de variables temporelles inobservées expliquent linéairement les variables observées. Ces variables cachées appelées facteurs expliquent la corrélation entre séries temporelles observées.

$$\begin{cases} y_{1,t} = \lambda_{1,1} F_{1,t} + ... + \lambda_{1,p} F_{p,t} + u_{1,t} \\ \quad ... = ... \\ y_{n,t} = \lambda_{n,1} F_{1,t} + ... + \lambda_{n,p} F_{p,t} + u_{n,t} \end{cases}$$

où les séries sont supposées stationnaires et vérifient

$$\begin{cases} \mathbf{E}\,(F_{i,t}\ F_{j,t'}) = 0 \quad \forall i \neq j,\ \forall t,t' \\ \mathbf{E}\,(u_{i,t}\ u_{j,t'}) = 0 \quad \forall i \neq j,\ \forall t,t' \\ \mathbf{E}\,(F_{i,t}\ u_{j,t'}) = 0 \quad \forall i, j, t, t' \end{cases}$$

L'estimation de ces modèles s'effectue de deux manières.

La première méthode utilise la technique de l'analyse factorielle statique. Elle présente l'avantage d'être simple à mettre en place dans le cas gaussien et fournit très souvent (Doz et Lenglart, 1999) une très bonne première approximation de la seconde méthode – dynamique –. En outre, elle fournit un test sur le nombre de facteurs cachés (Doz et Lenglart, 1996). Toutefois, elle n'est pas *a priori* optimale pour les séries temporelles étant limitée aux échantillons i.i.d., car elle suppose en outre :

$$\begin{cases} \mathbf{E}(F_{i,t}\ F_{i,t'}) = 0\ \ \forall t \neq t',\ \forall i \\ \mathbf{E}(u_{i,t}\ u_{i,t'}) = 0\ \ \forall t \neq t',\ \ \forall i \end{cases}$$

La seconde est la représentation espace-état intégrant la dynamique des facteurs et des résidus. L'estimation est réalisée par la méthode du filtrage de Kalman. Cette modélisation est appropriée car elle prend en compte la dynamique des séries temporelles. En outre, elle offre l'avantage de pouvoir traiter des séries de fréquence mixte. Ainsi, Doz et Lenglart (1996) extraient un facteur commun à partir des enquêtes mensuelles et trimestrielles de conjoncture dans l'industrie en France.

Ces méthodes ont pour objectif le suivi conjoncturel en temps réel de la conjoncture. Pour la datation des cycles, elles présentent l'inconvénient de ne pas prendre en compte la principale série d'intérêt comme mesure de l'activité globale de l'économie : le PIB lui-même.

En intégrant les différents secteurs de l'économie, le *National Bureau of Economic Research* (NBER) s'éloigne d'une définition qui ne considèrerait que le PIB. Le NBER adopte une démarche plus générale : "une récession économique est un déclin significatif de l'activité économique dans les différentes branches d'une durée supérieure à quelques mois. Cette baisse significative devrait normalement être présente dans le PIB, l'emploi, la production industrielle, les ventes des secteurs manufacturier et commerce". Ainsi, un déclin dans l'activité manufacturière se transmettant au PIB, mais qui ne serait pas simultané avec une baisse dans les services ne peut être considéré comme une récession à part entière.

Selon Murasawa et Mariano, il apparaît que la modélisation en facteurs dynamiques peut donner un sens quantitatif à cette définition : le facteur commun estimé devient négatif pendant plusieurs mois consécutifs. Remarquons que cette définition est une convention et que la méthode proposée n'est pas unique.

En suivant leur méthodologie (Mariano et Murasawa, 2003 et annexes), nous considérons la représentation espace-état d'un modèle factoriel. Les séries trimestrielles sont considérées comme des séries mensuelles avec valeurs manquantes. Une dynamique commune est postulée au niveau mensuel puis traduite au niveau trimestriel. Les valeurs sont remplacées par des zéros et une nouvelle représentation espace-état est déduite permettant l'estimation du modèle par maximum de vraisemblance. La maximisation numérique est réalisée par la méthode de Newton.

Résultats : facteur commun PIB et indicateurs quantitatifs

Description des données

De manière analogue, nous cherchons les indicateurs quantitatifs de l'économie française représentant les grandes branches. D'autres critères de choix interviennent : la disponibilité des données, leur fréquence. Pour décrire la branche industrielle, nous choisissons l'indice de la production industrielle (IPI), disponible en cvs-cjo[1] depuis 1978[2].

Du côté demande, la consommation des ménages est décrite par les dépenses des ménages en produits manufacturés, disponible depuis janvier 1985.

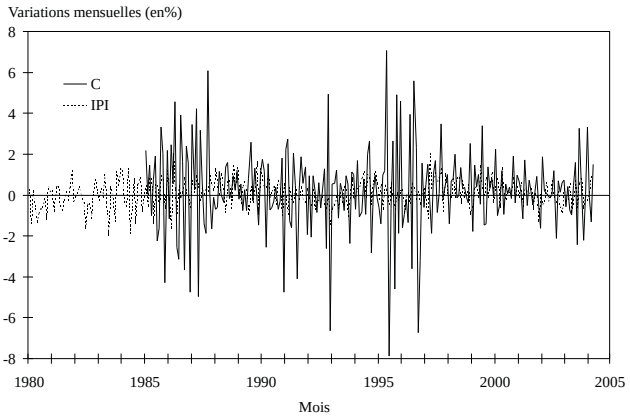
Enfin, sur le marché du travail, l'effectif salarié est retenu, en fréquence trimestrielle depuis le premier trimestre de 1978.

En résumé, nous utilisons les données suivantes :

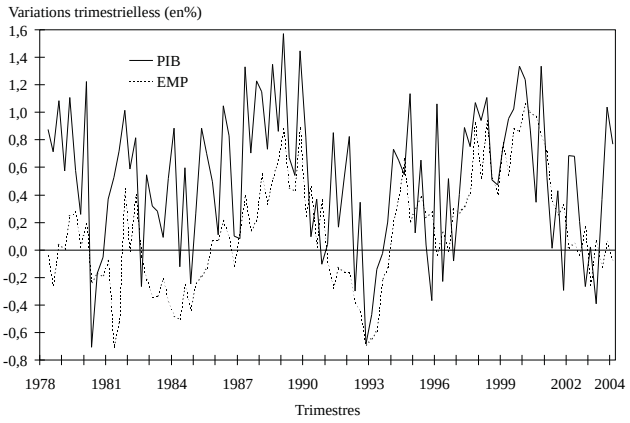
Tableau 1 : description statistique

Indicateur	Description
	Trimestrielle
PIB	Produit Intérieur Brut (en milliards d'euros base 1995, c.v.s.-c.j.o.)
EMP	Effectifs salariés trimestriels : Ensemble des secteurs (c.v.s.)
	Mensuelle
IPI	Indice de Production Industrielle (base 1995 , c.v.s.)
C	Consommation mensuelle des ménages - Dépenses de consommation des ménages en produits manufacturés (volumes aux prix de 1995, c.v.s.-c.j.o.)

Graphique 1 : données mensuelles françaises retenues



Graphique 2 : données françaises trimestrielles retenues



La description statistique de ces données est la suivante :

Tableau 2 : description des séries en %

Indicateur	Moyenne	E.T.	Min	Max	KPSS
		Trimestrielles			
$\Delta \ln PIB$	0,53	0,52	-0,68	1,57	0,27
$\Delta \ln EMP$	0,25	0,41	-0,69	1,06	0,51
		Mensuelles			
$\Delta \ln C$	0,19	2,05	-7,88	7,11	0,07
$\Delta \ln IPI$	0,13	1,21	-1,54	1,97	0,02

On s'aperçoit que la valeur moyenne des séries trimestrielles est supérieure à celle des séries mensuelles. En revanche, les écarts type des séries mensuelles sont supérieurs à ceux des séries trimestrielles. Le test de Kwiatkowski, Phillips, Schmidt et Shin (KPSS)(3) permet d'accepter l'hypothèse nulle de stationnarité.

Résultats

Pour déterminer p et q, nous regardons l'autocorrélogramme et l'autocorrégramme partiel des différentes séries. Nous choisissons *a priori* un auto-régressif d'ordre $p = 1$ et $q = 2$. Le facteur commun étant défini à un coefficient d'échelle prêt, nous fixons $\beta_1 = 1$ pour des raisons d'identification.

À l'instar de Murasawa et Mariano (2003), nous simplifions l'estimation de deux façons :
– nous enlevons aux séries leur moyenne et nous fixons $\mu = 0$;
– nous initialisons le filtre de Kalman de la manière décrite en annexe 1 (3).

Nous obtenons :

Les coefficients d'élasticité entre les différents indicateurs sont ainsi estimés :

Une variation du PIB de 1 point "correspond" à une variation de 1,5 point des dépenses de produits manufacturés aux ménages.

La version lissée du facteur $I^*_{eco-cumul}$ cumulé est représentée ainsi : $\mu_1 + \mathbf{E}(\Sigma^t_{i=1} F_t | I_T)$ avec μ_1 la moyenne trimestrielle du PIB divisé par trois. Ce calcul est simple car

$$\mathbf{E}(\Sigma^t_{i=1} F_i | I_T) = \Sigma^t_{i=1} \mathbf{E}(F_i | I_T) = \Sigma^t_{i=1} \hat{s}_{t|T}$$

Or les $\hat{s}_{t|T}$ ont justement été calculés dans le cadre de l'algorithme de de Jong décrit en annexe.

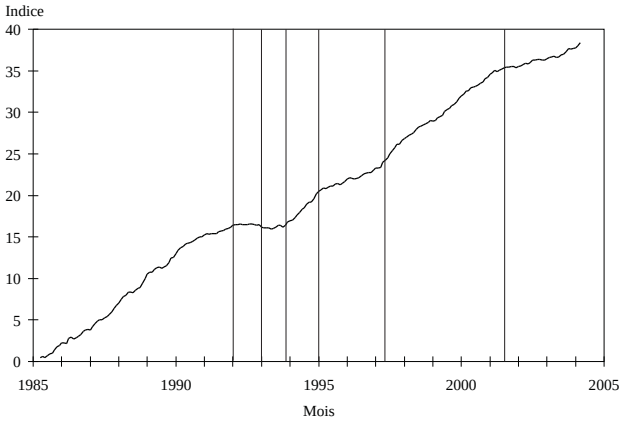
Tableau 3 : résultats numériques (écart type)

Paramètre	Δ ln *PIB*	Δ ln *EMP*	Δ ln *DET*	Δ ln *IPI*
β	1	0,77	1,44	2,71
	(0,19)	(0,39)	(0,43)	(0,17)
ϕ_f		0,48		
		(0,18)		
σ^2_1		0,02		
		(0,01)		
$\phi_{u,1}$	0,41	0,14	-0,55	-0,99
	(0,27)	(0,09)	(0,04)	(0,06)
$\phi_{u,2}$	-0,46	-0,3	-0,35	-0,76
	(0,14)	(0,16)	(0,06)	(0,1)
$\Sigma_{2,2}$	0,07	0,06	3,00	0,07
	(0,04)	(0,01)	(0,28)	(0,04)

Interprétation des résultats

Murasawa et Mariano (2003) montrent que pour les États-Unis, les évolutions du facteur commun semblent en mesure de caractériser la phase conjoncturelle :
– les parties décroissantes de la courbe correspondent aux périodes de récession ;
– les parties croissantes représentent les périodes de croissance ;
– plus la pente est décroissante (respectivement croissante), plus la récession (respectivement croissance) est forte.

Graphique 3 : facteur mensuel $I^*_{eco-cumul}$ cumulé



En transposant l'analyse pour la France, nous pouvons distinguer sept périodes de l'économie française de 1985 à 2003 au vu du graphique 3 :
– de janvier 1985 à janvier 1992 : forte croissance ;
– de février 1992 à août 1992 : stabilité ;
– de septembre 1992 à mai 1993 : récession ;
– de juin 1993 à décembre 1994 : forte croissance ;
– de janvier 1995 à avril 1997 : croissance moyenne ;
– d'avril 1997 à juin 2001 : forte croissance ;
– à partir de juillet 2001 : croissance modeste.

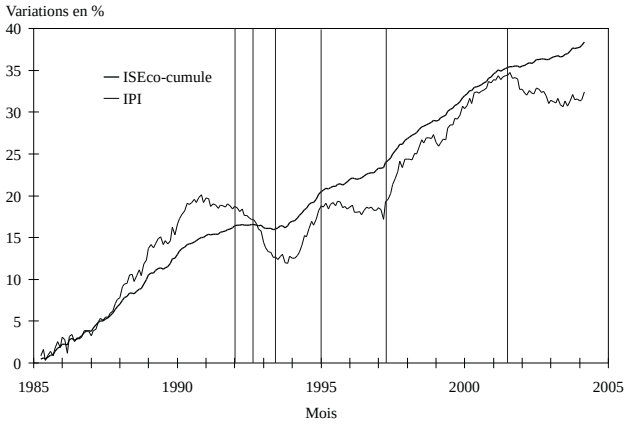
Cette datation apparaît cohérente avec celle obtenue par Doz et Lenglart (1996), sur la période 1985-1995, qui font leur analyse à partir de l'enquête de conjoncture dans l'industrie. S'agissant de la récession de 1993, le détail des comptes trimestriels montre en effet, pour l'année 1993, une baisse du PIB réel. Cette dernière est principalement présente du quatrième trimestre de 1992 au deuxième trimestre de 1993 dans les dépenses des ménages, les investissements des entreprises et ceux des ménages. Ici, nous la datons précisément de septembre 1992 à mai 1993.

Même si l'analyse permet de dégager l'existence d'un indicateur commun I^*_{eco} entre les différents indicateurs quantitatifs, il n'en demeure pas moins que ces derniers apportent un éclairage propre. En effet, à l'aide de cette grille de lecture, il est possible d'isoler la contribution propre à chaque grandeur économique et de la situer par rapport au facteur

conjoncturel commun. De cette comparaison, le conjoncturiste peut tirer deux enseignements :
– l'avance ou le retard d'une branche par rapport à l'activité générale ;
– le dynamisme relatif d'une branche par rapport à l'activité d'ensemble.

Ainsi,
– les retournements de l'indice de la production industrielle apparaissent coïncidents sur la période récente. L'IPI présente des signes de faiblesse dès la fin de 1990 (novembre), c'est-à-dire en avance par rapport au climat général représenté par l'indicateur $I^*_{eco-cumul}$. De février 1995 à mars 1997, l'IPI est atone alors que l'indicateur synthétique $I^*_{eco-cumul}$ continue de progresser. De mars à octobre 1997, l'IPI est en avance sur la conjoncture globale et présente un dynamisme supérieur. D'août 2001 à mars 2004, l'IPI faiblit tandis l'indicateur $I^*_{eco-cumul}$ connaît toujours une progression modeste (cf. graphique 4) ;
– l'emploi n'est pas toujours en phase avec le cycle conjoncturel.

Graphique 4 : indicateur cumulé et IPI

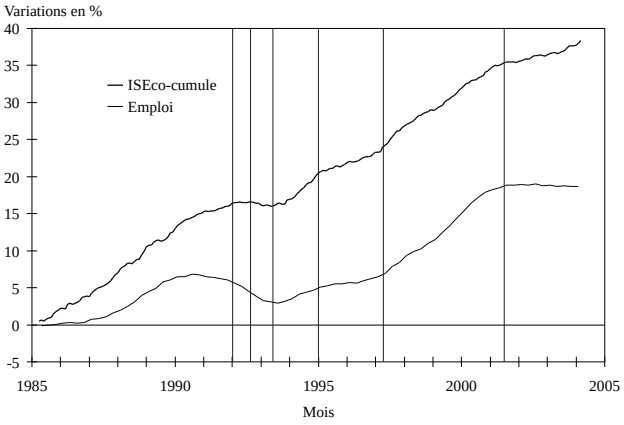


En novembre 1990, l'emploi commence déjà à faiblir alors que l'indicateur $I^*_{eco-cumul}$ est encore dans une phase de progression modeste. À partir de novembre 2001, le nombre d'emplois cesse de progresser à la différence de l'indicateur synthétique $I^*_{eco-cumul}$ (cf. graphique 5) ;

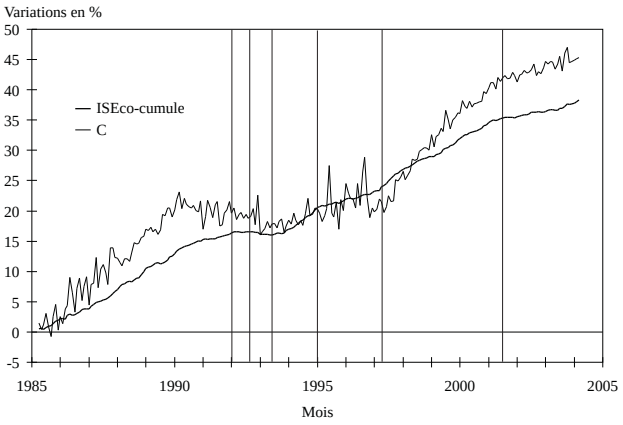
– la consommation en produits manufacturés des ménages est, soit en avance, soit coïncidente avec le cycle économique. La consommation en produits manufacturés des ménages présente un profil plus volatil au mois le mois, ce qui rend sa lecture plus difficile. Dès mars 1990, la consommation en produits manufacturés des ménages affiche un profil à la baisse, bien avant le déclin de l'indicateur synthétique $I^*_{eco-cumul}$ (cf. graphique 6) ;

– le PIB est imprécis pour dater les cycles (cf. graphique 7) La simple lecture du PIB indiquerait une récession du quatrième trimestre de 1992 jusqu'au troisième trimestre de 1993. L'indicateur

Graphique 5 : indicateur cumulé et emploi

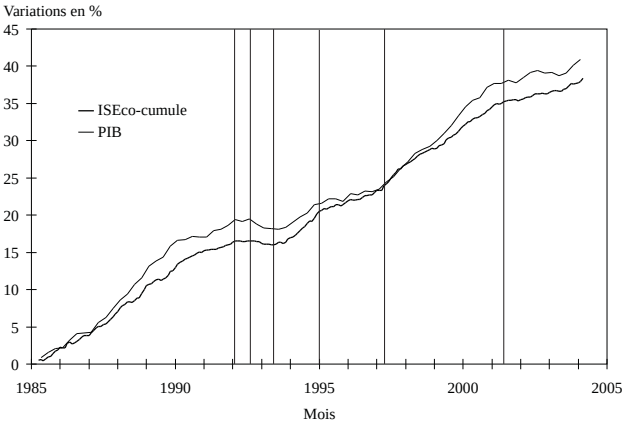


Graphique 6 : indicateur cumulé et consommation en produits manufacturés



du climat conjoncturel général $I^*_{eco-cumul}$ limite cette dernière aux quatrième trimestre de 1992, premier trimestre de 1993 et au deuxième trimestre de 1993.

Graphique 7 : indicateur cumulé et PIB

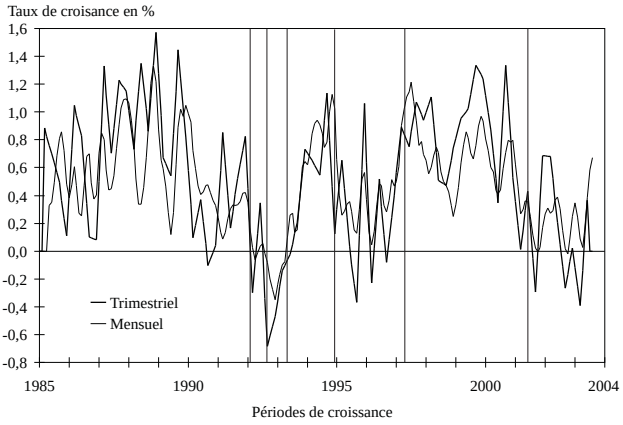


L'approche considérée présente ainsi deux intérêts majeurs :
– elle offre une grille de lecture de l'activité passée en intégrant de façon simultanée les différentes branches de l'économie ;

– elle donne un chiffrement fin (mensuel) de ces différentes périodes.

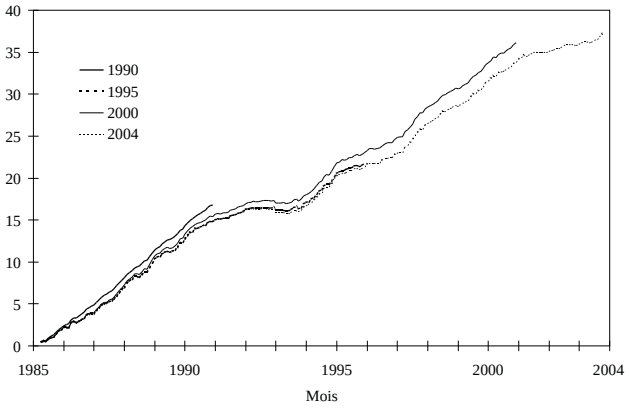
Remarque 1 : *l'indice mensuel obtenu I^*_{eco} augmenté du tiers de la valeur moyenne du taux de croissance trimestriel du PIB peut être assimilé à un taux de croissance mensuel latent du PIB* (cf. graphique 8).

Graphique 8 : PIB mensuel et PIB trimestriel



Remarque 2 : *une question se pose, celle de la stabilité d'une telle datation. Que se passe-t-il lorsque l'on effectue l'estimation sur une période de temps différente ? Les périodes du cycle se trouvent-elles changées ? Le graphique 9 montre l'indicateur $I^*_{eco-cumul}$ estimé respectivement en 1990, 1995, 2000 et 2004. Les périodes de la datation apparaissent stables quelle que soit la période d'estimation. Enfin, sur le graphique 9, un décalage en niveau apparaît, ce qui est fréquent dans une modélisation des taux de croissance sans force de rappel*.

Graphique 9 : stabilité de la datation à différentes dates

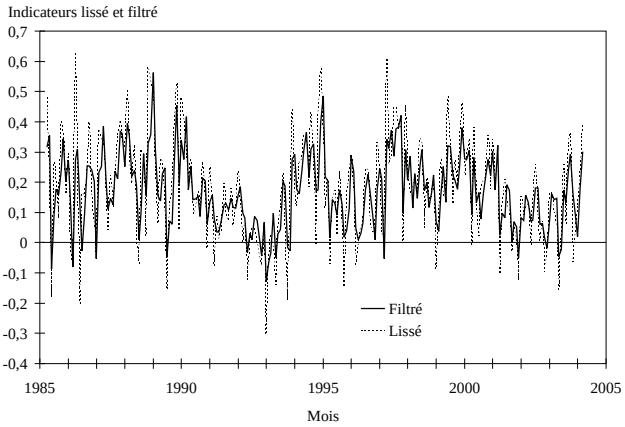


Facteur lissé ou facteur filtré ?

Jusqu'à présent, nous avons présenté l'indicateur lissé I^*_{eco} . Nous avons utilisé cet indicateur dans le cadre d'une relecture de l'activité économique, où nous disposions des données sur toute la période d'estimation. En effet, I^*_{eco} prend en compte à la date t l'information disponible jusqu'à la date T,$(T>t)$, notamment l'information disponible juste après la date t. Ce facteur correspond à $\mathbf{E}(F_t \mid I_T)$. Dans le cadre d'une lecture en temps réel de l'activité économique, notre indicateur I^*_{eco} coincide avec le facteur filtré $I^*_{eco-filtre}$. Ce facteur correspond à $\mathbf{E}(F_t \mid I_t)$ car en temps réel $t=T$. Cependant, rétrospectivement $(t<T)$, le facteur filtré se distingue du facteur lissé I^*_{eco} car le premier n'est construit qu'à l'aide de l'information passée I_t alors que le second l'est en prenant en compte l'information passée et future I_T. Quand nous effectuons l'exercice de datation en temps réel, nous utilisons en réalité le facteur filtré $I^*_{eco-filtre}$ et non le facteur lissé I^*_{eco}.Pour juger de la pertinence du facteur filtré $I^*_{eco-filtre}$ pour le conjoncturiste, il convient donc de le comparer au facteur lissé I^*_{eco} de 1985 à 2004. Les deux signaux sont très cohérents entre eux (cf. graphique 10). Toutefois, il apparaît que le facteur filtré est légèrement en retard sur le facteur lissé pendant les périodes de transition. Ainsi, le facteur filtré est un indicateur mensuel légèrement retardé (au plus d'un ou deux mois) de l'activité économique globale.

Remarque 3 : *toutefois, notons que cette approche aurait dû être effectuée avec les données dont disposait le conjoncturiste à l'époque considérée et non avec les données révisées dont nous disposons actuellement. Ainsi, faute d'une base de données en temps réel, il est difficile de conclure sur l'apport conjoncturel réel de l'indicateur filtré.*

Graphique 10 : apport conjoncturel de l'indicateur $I^*_{eco-filtre}$

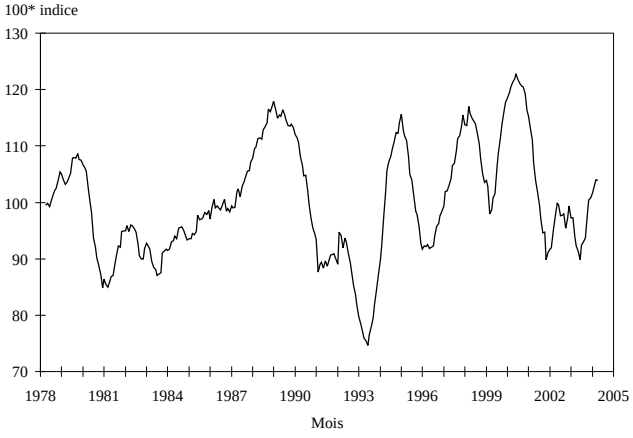


Résultats : facteur commun PIB et enquêtes manufacturières

Nous appliquons ici la méthode de Murasawa et de Mariano au PIB (trimestriel) et aux enquêtes de conjoncture dans l'industrie (mensuel). Le but est d'extraire l'information commune entre l'activité économique et l'opinion qu'en ont les chefs d'entreprise. Il serait souhaitable d'intégrer aussi les séries de la partie précédente, ce qui n'a pu être réalisé pour des problèmes numériques compte tenu du nombre de séries.

Doz et Lenglart (1999) ont montré l'intérêt d'une approche factorielle dynamique pour obtenir un facteur commun à partir des six soldes d'opinion émis dans l'enquête de conjoncture dans l'industrie. Le facteur commun est dans un premier temps obtenu par analyse factorielle statique, ce qui permet en outre de tester le nombre de facteurs communs. Le modèle est dans un second temps estimé par un filtre de Kalman en adoptant la représentation espace-état adéquate. Le facteur commun apparaît comme un indicateur du climat général dans l'industrie. Il offre l'avantage d'être moins volatil au mois le mois que les séries des enquêtes et a donc une plus grande lisibilité (cf. graphique 11).

Graphique 11 : facteur commun "à la Doz Lenglart 1999"

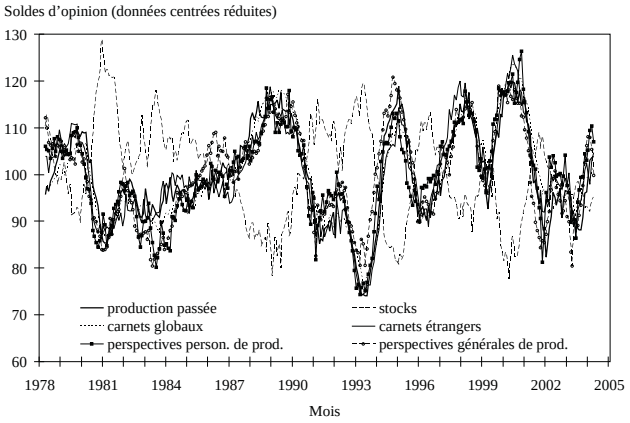


Description des données

De même que Doz et Lenglart (1999), nous considérons ici les soldes d'opinion relatifs à six questions posées dans l'enquête de conjoncture dans l'industrie. Ces questions sont : opinion sur la tendance passée de la production personnelle (TPPA), opinion sur la tendance prévue de la production personnelle (TPPRE), opinion sur la demande et les carnets de commande globaux (OSCD), opinion sur la demande et les carnets de commandes en provenance de l'étranger (OSCDE), opinion sur les niveaux des stocks (OSSK) et opinion sur les perspectives générales d'activité (PGP).

Rappelons que le principe de cette enquête de conjoncture consiste à interroger un panel d'entreprises en leur demandant de répondre de manière qualitative (amélioration, même niveau, ou détérioration) à une série de questions. La répartition des trois types de réponses est calculée en pourcentage et l'information relative à chaque question est ensuite représentée sous la forme d'un solde d'opinions (pourcentage d'entreprises jugeant qu'il y a amélioration moins le pourcentage de celles jugeant qu'il y a détérioration). L'observation au mois le mois de ces soldes permet de suivre l'évolution des opinions des industriels sur ces questions.

Graphique 12 : enquêtes de conjoncture dans l'industrie



Rappelons que le test de KPSS permet de conclure à la stationnarité de ces séries.

Ces enquêtes sont cvs-cjo et l'échantillon considéré s'étend d'avril 1978 à décembre 2003.

La description statistique des séries temporelles est la suivante :

Tableau 4 : description des séries en %

Indicateur	Moyenne	S.D.	Min.	Max.	KPSS
		Trimestrielle			
$\Delta \ln PIB$	0,50	0,51	-0,71	1,57	0,27
		Mensuelles			
TPPA	3,90	15,54	-36,1	34,91	0,21
TPPRE	3,56	11,45	-25,65	34,01	0,67
OSCD	-17,66	19,74	-64,51	28,14	0,67
OSCDE	-12,31	19,21	-61,57	35,7	0,47
OSSK	14,7	2,05	7,98	37,67	0,93
PGP	-58,15	2,05	25,57	42,69	0,40

Résultats d'estimation

Nous choisissons de préférence une dynamique de type AR(2) sur le facteur commun, et de type AR(1) sur le résidu propre à chaque série. Ce choix s'inspire de la dynamique choisie par Doz et Lenglart (1999). Toutefois, après estimation, il n'est pas apparu nécessaire de conserver la dynamique MA(1) sur le facteur commun.

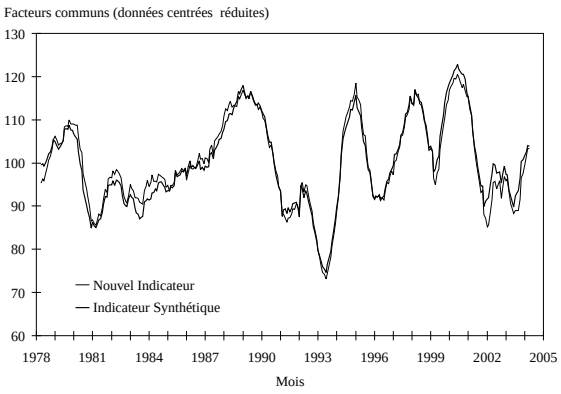
Les résultats d'estimation (cf. tableau 5) apparaissent cohérents avec les résultats de Doz et Lenglart (1999) pour ce qui concerne les coefficients autorégressifs du facteur commun. En revanche, les coefficients autorégressifs ϕ_u des résidus $u_{i,t}$ se rapprochent de 1. Ceci s'explique par le fait que les autocorrélations des séries des enquêtes sont nulles à partir d'un grand retard.

Interprétation des résultats

Comme précédemment, le nouvel indice I^*_{en} est obtenu en utilisant la version lissée $\mathbf{E}(F_t \mid I_T)$.

Ce facteur commun apparaît très semblable à l'ancien facteur commun de Doz et Lenglart (1999).

Graphique 13 : comparaison du facteur commun "à la Doz et Lenglart" avec le nouvel indicateur I^*_{en}



La comparaison entre ces deux facteurs apporte les enseignements suivants :
– nous pouvons comparer les opinions des chefs d'entreprise avec un indice conjoncturel reposant sur le PIB et faire ainsi la part entre anticipations et réalité économique ;
– cette différence peut aussi être interprétée comme un renseignement complémentaire du côté des services.

Sur la période d'estimation (1978 à 2004), nous remarquons :
– de 1978 à 1985, le facteur commun industrie est avancé sur le nouveau facteur commun intégrant le PIB. Ceci peut s'interpréter comme une avance du cycle du secteur industriel sur le cycle global ;
– de 1985 à 2001, les deux indicateurs sont très semblables, indiquant peut-être une synchronisation des cycles du secteur manufacturier avec le secteur des services ;
– de 2001 à 2003, nous remarquons de nouveau un décalage entre le cycle industriel et le cycle global.

En outre, un regard sur l'amplitude des facteurs communs semblerait indiquer :
– de 1980 à 1985, les industriels apparaissent plus pessimistes que de raison car le facteur commun industrie se trouve en dessous du facteur commun intégrant le PIB ;
– de 2001 à 2003, les industriels apparaissent plus optimistes que ne laisse suggérer la situation de l'activité économique pour les raisons opposées.

La méthode d'estimation utilisée ne "prête" peut-être pas assez d'importance à la donnée du PIB. Ceci s'explique par le nombre de séries utilisées (à savoir 6 contre 1) et par le faible nombre de valeurs du PIB considéré comme série mensuelle. Ainsi, cela fournirait une explication à la ressemblance entre le facteur commun industrie et le facteur commun avec le PIB intégré.

Tableau 5 : résultats numériques (écart type)

	$\Delta \ln PIB$	TPPA	TPPRE	OSCD	OSCDE	OSSK	PGP
β	1,00	129,01	87,99	121,23	146,56	-57,65	179,29
		(14,12)	(11,41)	(15,03)	(19,99)	(10,73)	(27,77)
$\phi_{f,1}$				1,79			
				(0,06)			
$\phi_{f,2}$				-0,80			
				(0,06)			
σ_1^2				119^{e-4}			
				(4^{e-5})			
ϕ_u	-0,61	0,30	0,83	0,99	0,98	0,99	0,96
	(0,22)	(0,11)	(0,03)	(0,01)	(0,01)	(0,01)	(0,01)
$\Sigma_{2,2}$	0,23	6,23	9,7	6,9	12,9	6,7	30,7
	$(4,90^{e-2})$	(0,77)	(0,82)	(0,67)	(1,21)	(0,55)	(2,62)

Conclusion

Dans cet article, nous avons considéré l'apport d'un modèle factoriel dynamique multi-fréquence (mensuel et trimestriel). Nous avons construit un indicateur mensuel de l'activité économique intégrant les données trimestrielles du PIB et de l'emploi en France, ainsi que les données mensuelles de l'indice de la production industrielle et des dépenses de produits manufacturés des ménages. Cet indicateur mensuel est apparu pertinent pour dater de façon précise (mensuelle) les différentes phases du cycle économique français de 1985 à 2003. Dans un deuxième temps, nous avons appliqué cette méthodologie à l'indicateur synthétique industrie en intégrant le PIB trimestriel. Cette approche a permis d'apporter une information complémentaire sur le rapport entre opinions des industriels et activité économique. Toutefois, les deux indicateurs restent très similaires sur l'ensemble de la période d'estimation, ce qui suggère que le modèle statistique utilisé converge, par construction, vers le facteur commun industrie et n'accorde pas assez d'importance au PIB lui-même. L'analyse factorielle statique réalisée avec les six soldes d'opinion de l'enquête industrie et le PIB mensuel construit avec les indicateurs quantitatifs suggère l'existence *a priori* de deux facteurs communs. Ce deuxième facteur commun pourrait être interprété comme la composante services du climat des affaires. Une modélisation à deux facteurs latents serait dès lors souhaitable.

Notes

(1) Corrigé des variations saisonnières et des jours ouvrables.

(2) Rétropolé depuis 1978.

(3) La valeur critique au-dessus de laquelle on rejette l'hypothèse nulle est 0,739 au de seuil de 1%.

Bibliographie

Brockwell et Davis (1991). *Time Series. Theory and Methods,* Springer-Verlag

Doz C. et Lenglart F. (1995). " Une grille de lecture pour l'enquête mensuelle de l'industrie ", *Note de conjoncture*, Insee.

Doz C. et Lenglart F. (1996). "Factor Analysis and Unobserved Component Models : an Application to the Study of French Business Surveys", *Document de travail*, Insee, juillet.

Doz C. et Lenglart F. (1999). " Analyse factorielle dynamique: test du nombre de facteurs, estimation et application à l'enquête de conjoncture dans l'industrie ", *Annales d'Économie et Statistiques,* n°54, pp. 91-127.

Gourieroux C. et Monfort A. (1990). *Séries temporelles et modèles dynamiques*, Economica.

Hamilton J.-D. (1991). *Time Series Analysis*, Princeton University Press.

Mariano R. et Murasawa Y. (2003). "A New Coincident Index of Business Cycles Based on Monthly and Quarterly Series", *Journal of Applied Econometrics,* vol. 18, pp. 427-443.

Stock J.-H. et Watson M.-W. (1989). "New Indexes of Coincident and Leading Economic Indicators", *NBER Macroeconomics Annual*.

Annexe 1 : méthodologie de Murasawa et Mariano

Modèle à un facteur

Les étapes de l'analyse d'un modèle factoriel multi-fréquence utilisée par Murasawa et Mariano (Mariano R. *et alii*, 2003) sont rappelées.

Soient N_1 séries trimestrielles représentées par la série multivariée $\{Y_{1,t}\}_t$ de dimension N_1 et de N_2 séries mensuelles représentées par $\{Y_{2,t}\}_t$de dimension N_2. La fréquence considérée est mensuelle. Ainsi, les données trimestrielles sont vues comme des séries mensuelles avec valeurs observables toutes les trois périodes.

Deux hypothèses successives sont effectuées sur le modèle :

– *première hypothèse* :

$\{Y_{1,t}\}_t$ est la moyenne géométrique d'une grandeur mensuelle latente inobservable $\{Y^*_{1,t}\}_t$. C'est-à-dire que :

$$(1)\quad \ln Y_{1,t}=\frac{1}{3}(\ln Y^*_{1,t-1}+\ln Y^*_{1,t}+\ln Y^*_{1,t-2})$$

L'équation (1) est différenciée une fois car les $Y_{1,t}$ sont supposés intégrés d'ordre 1. Après calculs, on obtient (exactement) :

$$(2)\quad y_{1,t}=\frac{1}{3}y^*_{1,t}+\frac{2}{3}y^*_{1,t-1}+y^*_{1,t-2}+\frac{2}{3}y^*_{1,t-3}+\frac{1}{3}y^*_{1,t-4}$$

Ainsi, cette variable latente $y^*_{1,t}$ peut être assimilée au taux de croissance du PIB mensuel. Remarquons toutefois que l'équation (1) n'est pas la relation habituelle entre le PIB en niveau trimestriel et le PIB en niveau mensuel. Elle est choisie afin d'obtenir la représentation espace-état linéaire (2) ;

– *deuxième hy*pothèse :

$y^*_{1,t}$ suit un modèle à un facteur. Plus précisément,

$$\begin{pmatrix} y^*_{1,t} \\ y_{2,t} \end{pmatrix}=\begin{pmatrix} \mu^*_1 \\ \mu_2 \end{pmatrix}+\beta f_t+u_t$$

$$\phi_f(L)f_t=\nu_{1,t}$$

$$\phi_u(L)u_t=\nu_{2,t}$$

$$\begin{pmatrix} v_{1,t} \\ v_{2,t} \end{pmatrix}=N\left(0,\begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \Sigma_{2,2} \end{bmatrix}\right)$$

Rappelons que $y^*_{1,t}\in IR^{N_1}$, $y_{2,t}\in IR^{N_2}$, $\beta\in IR^N$, $f_t\in IR$, $u_t\in IR^N$. β est le vecteur des *loadings,* $(f_t)_t$ est le processus stationnaire du facteur commun, $(u_t)_t$ le processus stationnaire des facteurs spécifiques. ϕ_f est un polynôme à coefficients réels de degré p. $\phi_u$est un polynôme à coefficients dans les matrices carrée de taille N de degré q. Pour des raisons d'identification, β_1 est fixé à 1 et seules les matrices diagonales pour $\Sigma_{2,2}$ et pour les coefficients de ϕ_u sont considérées.

Transformation

Comme les $y^{*}_{1,t}$ ne sont pas observables, la relation (2) est utilisée :

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \beta_1 (\frac{1}{3} f_t + \frac{2}{3} f_{t-1} + f_{t-2} + \frac{2}{3} f_{t-3} + \frac{1}{3} f_{t-4} \\ \beta_2 f_t \end{pmatrix} + \begin{pmatrix} \frac{1}{3} u_{1,t} + \frac{2}{3} u_{1,t-1} + u_{t-2} + \frac{2}{3} u_{1,t-3} + \frac{1}{3} u_{1,t-4} \\ u_{2,t} \end{pmatrix}$$

avec $\mu_1 = 3\mu^{*}_1$, pour obtenir le modèle à un facteur suivant :

$$s_t = F s_{t-1} + G v_t$$

$$y_t = \mu + H s_t$$

Ecriture détaillée du modèle ($p<5$ *et* $q < 5$*)*

$$s_t = \begin{pmatrix} f_t \\ . \\ f_{t-4} \\ u_t \\ . \\ u_{t-4} \end{pmatrix}_{[5(N+1),1]}$$

$$v_t = \begin{pmatrix} v_{1,t} \\ v_{2,t} \end{pmatrix}_{[5N,1]}$$

$$F = \begin{pmatrix} \phi_{f,1} & . & \phi_{f,p} & o'_{5-p} & & & & \\ 1 & . & 0 & 0 & & & & \\ & & & . & O_{5\times 5N} & & & \\ 0 & & 1 & 0 & & & & \\ & & & & \Phi_{u,1} & . & \Phi_{u,q} & O_{N\times(5-q)N} \\ & O_{5N\times 5} & & & I_N & & 0 & O_{N\times N} \\ & . & & & & . & & \\ & & & & 0 & & I_N & O_{N\times N} \end{pmatrix}_{[5(N+1),5(N+1)]}$$

$$G = \begin{pmatrix} 1 & o'_N \\ 0 & o'_N \\ . & . \\ o_N & I_N \\ o_N & O_{N\times N} \\ . & . \end{pmatrix}_{[5(N+1),N+1]}$$

$$H = \begin{pmatrix} \frac{\beta_1}{3} & \frac{2\beta_1}{3} & \beta_1 & \frac{2\beta_1}{3} & \frac{\beta_1}{3} & \frac{1}{3} I_{N_1} & O_{N_1\times N_2} & \frac{2}{3} I_{N_1} & O_{N_1\times N_2}\ldots \\ \beta_2 & & O_{N_1\times 4} & & O_{N_2\times N_1} & I_{N_2} & & O_{N_2\times N}\ \ldots \end{pmatrix}_{[N,5(N+1)]}$$

Modèle final

Il est pour l'instant impossible d'estimer le modèle car seulement une observation sur 3 est observable pour les $y_{1,t}$.

Il suffit alors de donner aux valeurs manquantes des $y_{1,t}$ la valeur z_t, où z_t est une réalisation d'une variable normale standard indépendante des y_t, z_{t-1}. Nous créons ainsi une nouvelle variable ($y^+_{1,t} = y_{1,t}$ si $y_{1,t}$ est observable, z_t sinon), qui suit donc :

$$s_t = F\, s_{t-1} + G\, v_t$$

$$y_t^+ = \mu_t + H_t\, s_t + w_t$$

avec

$$\mu_{1,t} = \begin{cases} \mu_1 \text{ si } y_{1,t} \text{ est observable} \\ 0 \text{ sinon} \end{cases}$$

$$\mu_t = \begin{pmatrix} \mu_{1,t} \\ \mu_2 \end{pmatrix}$$

$$H_{1,t} = \begin{cases} H_1 \text{ si } y_{1,t} \text{ est observable} \\ 0 \text{ sinon} \end{cases}$$

$$w_{1,t} = \begin{cases} 0 \text{ si } y_{1,t} \text{ est observable} \\ z_t \text{ sinon} \end{cases}$$

$$H_t = \begin{pmatrix} H_{1,t} \\ H_2 \end{pmatrix} \text{ et } w_t = \begin{pmatrix} w_{1,t} \\ 0 \end{pmatrix}$$

Remarquons que l'estimateur du maximum de vraisemblance de $f(y_t^+, \theta)$ est asymptotiquement équivalent à l'estimateur du maximum de vraisemblance de $f(y_t, \theta)$ sous les hypothèses de Murasawa et Mariano (cf. Mariano R. *et alii*, 2003).

En outre, comme (en posant $y_t = (y_1, y_2, \ldots, y_t), y_{2,t} = (y_{2,1}, y_{2,2} \ldots, y_{2,t})$)

$$f(y_t^+, \theta) = f(y_{1,1}, y_{1,3}, \ldots, y_{1,t-3}, y_{1,t}, y_{2,t}, \theta) \prod_{t \in A} f(z_t)$$

les réalisations de z_t sont remplacées par 0 sans changer la valeur de l'argument maximum.

Estimation

L'estimation se fait par maximisation de la vraisemblance calculée par un filtre de Kalman.

Calcul de la vraisemblance par le filtrage de Kalman

Pour des raisons de calcul, la vraisemblance est calculée récursivement en remarquant que :

$$f(y_T^+, \theta) = f(y_0, \theta) \prod_{t=1}^{T} f(y_t^+ \mid y_{t-1}^+, \theta)$$

Le filtre de Kalman détaille la récurrence (cf. Gourieroux *et alii*, 1990). Les notations sont les suivantes :

$$\hat{s}_{t|\,t-1} = \mathbf{E}(s_t \mid y_t^+)$$

$$P_{t|\,t-1} = \mathrm{var}(s_t \mid y_t^+)$$

Initialisation

Nous initialisons le filtre de la façon suivante :

(3) $\hat{s}_{1|\,0} = 0$

$$P_{1|\,0} = G \Sigma_{vv} G'$$

ce qui est en temps de calcul moins coûteux et asymptotiquement équivalent à l'estimateur exact du maximum de vraisemblance (cf Mariano R. *et alii*, 2003) défini par l'initialisation du filtre de Kalman pour un facteur stationnaire :

(4) $\hat{s}_{1|0} = 0$

$$vec(P_{1|0}) = (I_{(5+5N)^2} - F \otimes F)^{-1} vec(G \Sigma_{vv} G')$$

Mise à jour

$$\hat{s}_{t|t} = \hat{s}_{t|t-1} + B_t (y_t^+ - \mu_t - H_t \hat{s}_{t|t-1})$$

$$P_{t|t} = P_{t|t-1} - B_t H_t P_{t|t-1}$$

avec B_t la matrice de gain ($t \geq 1$) :

$$B_t = P_{t|t-1} H_t' (H_t P_{t|t-1} H_t' + \Sigma_{ww,t})^{-1}$$

et $$\Sigma_{ww,t} = \begin{cases} 0_{[N \times N]} & \text{si } y_{1,t} \text{ est observable} \\ \begin{bmatrix} I_{N1} & 0_{N_1 \times N1} \\ 0_{N_1 \times N_2} & 0_{N_2 \times N_2} \end{bmatrix} & \text{sinon} \end{cases}$$

Prévision

$$\hat{s}_{t|t-1} = F \hat{s}_{t-1|t-1}$$

$$P_{t|t-1} = F P_{t-1|t-1} F' + G \Sigma_{vv} G'$$

Au cours du filtre, il est possible de calculer la vraisemblance en remarquant que :

$$\mu_{t|t-1} = \mu_t + H_t \hat{s}_{t|t-1}$$

$$\Sigma_{t|t-1} = H_t P_{t|t-1} H_t' + \Sigma_{ww,t}$$

avec, par définition,

$$\mu_{t|t-1} = \mathbf{E}(y_t^+ | y_{t-1}^+)$$

$$\Sigma_{t|t-1} = Var(y_t^+ | y_{t-1}^+)$$

De sorte que la vraisemblance s'écrit :

$$\ln L = -\frac{NT}{2} \ln 2\pi - \frac{1}{2} \Sigma_{t=1}^{T} \ln \det \Sigma_{t|t-1} - \frac{1}{2} \Sigma_{t=1}^{T} (y_t^+ - \mu_{t|t-1})' \Sigma_{t|t-1}^{-1} (y_t^+ - \mu_{t|t-1})$$

Maximisation par la méthode de Newton

À la fin du filtre de Kalman, on a une fonction à $(N_1 - 1) + N_2 + p + (N_1 + N_2) * q + 1 + N_1 + N_2$ variables.

Le maximum de vraisemblance est cherché par la méthode de Newton. Deux difficultés apparaissent alors :

– existence de maximums locaux ;

– choix du seuil de tolérance.

Lissage

Plutôt que $\hat{s}_{t|t-1}$ nous préférons exposer les résultats obtenus pour la version lissée de l'espérance conditionnelle, à savoir $\hat{s}_{t|T}$. L'algorithme proposé par de Jong est :

$$\begin{cases} r_{T+1} = 0 \\ r_t = H_t' (H_t P_{t|t-1} H_t' + \Sigma_{ww,t})^{-1} (y_t^+ - \mu_t - H_t \hat{s}_{t|t-1}) + (I - H_t' B_t') F' r_{t+1} \\ \hat{s}_{t|T} = \hat{s}_{t|t-1} + P_{t|t-1} r_t \end{cases}$$

Choix de p et q

Reste à déterminer les ordres des coefficients autorégressifs *p* et *q*. Une idée *a priori* de ces ordres peut être obtenue en étudiant séparément les séries temporelles et en les identifiant à un *ARMA* (*p, q*) qui utilise la méthodologie de Box et Jenkins.

A posteriori, les critères employés sont des critères d'information de type Aikaike et des critères de blancheur des résidus.

Annexe 2 : Estimation avec un *ARMA* (2, 1) sur le facteur commun

Modèle

Le nouveau modèle s'écrit :

$$s_t = \begin{pmatrix} f_t \\ . \\ f_{t-4} \\ v_{1,t} \\ u_t \\ . \\ u_{t-4} \end{pmatrix}_{[5(N+1),1,1]} \qquad v_t = \begin{pmatrix} v_{1,t} \\ v_{2,t} \end{pmatrix}_{[5N,1]}$$

$$F = \begin{pmatrix} \phi_{f,1} & . & \phi_{f,p} & o'_{5-p} & -\theta & & & & \\ 1 & & 0 & 0 & 0 & & & & \\ & . & & . & . & O_{6\times 5N} & & & \\ 0 & & 1 & 0 & 0 & & & & \\ 0 & . & & 0 & 0 & & & & \\ & & & & & \Phi_{u,1} & . & \Phi_{u,q} & O_{N\times(5-q)N} \\ & O_{5N\times 6} & & & & I_N & & 0 & O_{N\times N} \\ & & & & & & . & & . \\ & & & & & 0 & & I_N & O_{N\times N} \end{pmatrix}_{[5(N+1)+1,\,5(N+1)+1]}$$

$$G = \begin{pmatrix} 1 & o'_N \\ o_4 & o'_{N\times 4} \\ 1 & o'_N \\ 0 & o'_N \\ . & . \\ o_N & I_N \\ o_N & O_{N\times N} \\ . & . \end{pmatrix}_{[5(N+1)+1,\,N+1]}$$

$$H = \begin{pmatrix} \frac{\beta_1}{3} & \frac{2\beta_1}{3} & \beta_1 & \frac{2\beta_1}{3} & \frac{\beta_1}{3} & 0 & \frac{1}{3}I_{N_1} & O_{N_1\times N_2} & \frac{2}{3}I_{N_1} & O_{N_1\times N_2}\cdots \\ \beta_2 & & O_{N_1\times 4} & & & 0 & O_{N_2\times N_1} & I_{N_2} & & O_{N_2\times N}\ \cdots \end{pmatrix}_{[N,\,5(N+1)+1]}$$

Résultats des estimations

Tableau 6 : résultats numériques (écart type)

	$\Delta \ln PIB$	TPPA	TPPRE	OSCD	OSCDE	OSSK	PGP
β	1,00	129,01 (14,29)	87,99 (11,39)	121,25 (14,69)	146,57 (18,96)	-57,64 (10,80)	179,29 (29,35)
$\phi_{f,1}$				1,88 (0,03)			
$\phi_{f,2}$				-0,96 (0,03)			
θ				0,47 (0,08)			
σ_1^2				$1,89^{e-4}$ ($0,54^{e-5}$)			
ϕ_u	-0,61 (0,22)	0,33 (0,11)	0,83 (0,03)	0,99 (0,01)	0,98 (0,01)	0,99 (0,01)	0,96 (0,02)
$\Sigma_{2,2}$	0,24 (0,05)	6,48 (0,78)	9,35 (0,80)	6,26 (0,67)	11,57 (1,18)	6,71 (0,56)	31,12 (2,72)

Chapitre 5

Un nouvel indicateur synthétique mensuel résumant le climat des affaires dans les services en France

Un nouvel indicateur synthétique mensuel résumant le climat des affaires dans les services en France

Matthieu Cornec* et Thierry Deperraz**

Le nouvel indicateur synthétique mensuel présenté dans cet article constitue un résumé de l'information contenue dans l'enquête de conjoncture dans les services. Il est obtenu par extraction d'un signal commun à trois séries de fréquence mensuelle et trois de fréquence trimestrielle. L'approche retenue pour le construire relève du cadre de l'analyse factorielle dynamique. L'indicateur synthétique est le résultat de l'estimation d'un modèle à composantes inobservables.

L'indicateur synthétique peut être appliqué aux trois sous-secteurs couverts par l'enquête de conjoncture dans les services (services aux entreprises, services aux particuliers et activités immobilières). Son examen confirme la reprise de l'activité dans l'ensemble des services à partir de la mi-2003. Cette reprise apparaît hésitante au deuxième semestre 2004 et semble s'essouffler début 2005.

Cet indicateur peut être utilisé par le conjoncturiste pour actualiser sa prévision de la production trimestrielle de services au mois le mois et non plus seulement au trimestre le trimestre. En outre, il contient une information spécifique par rapport à l'indicateur synthétique du climat des affaires dans l'industrie manufacturière et contribue ainsi à la prévision du Pib.

* *Insee, Division Synthèse conjoncturelle.*
** *Insee, Division Enquêtes de conjoncture.*
Nous remercions Karine Berger, Xavier Bonnet, Hélène Erkel-Rousse, Fabrice Lenglart, Philippe Scherrer ainsi que deux relecteurs anonymes pour leurs précieux commentaires. Les erreurs qui pourraient subsister relèvent de notre seule responsabilité. La rédaction de cet article a été achevée en mai 2006.

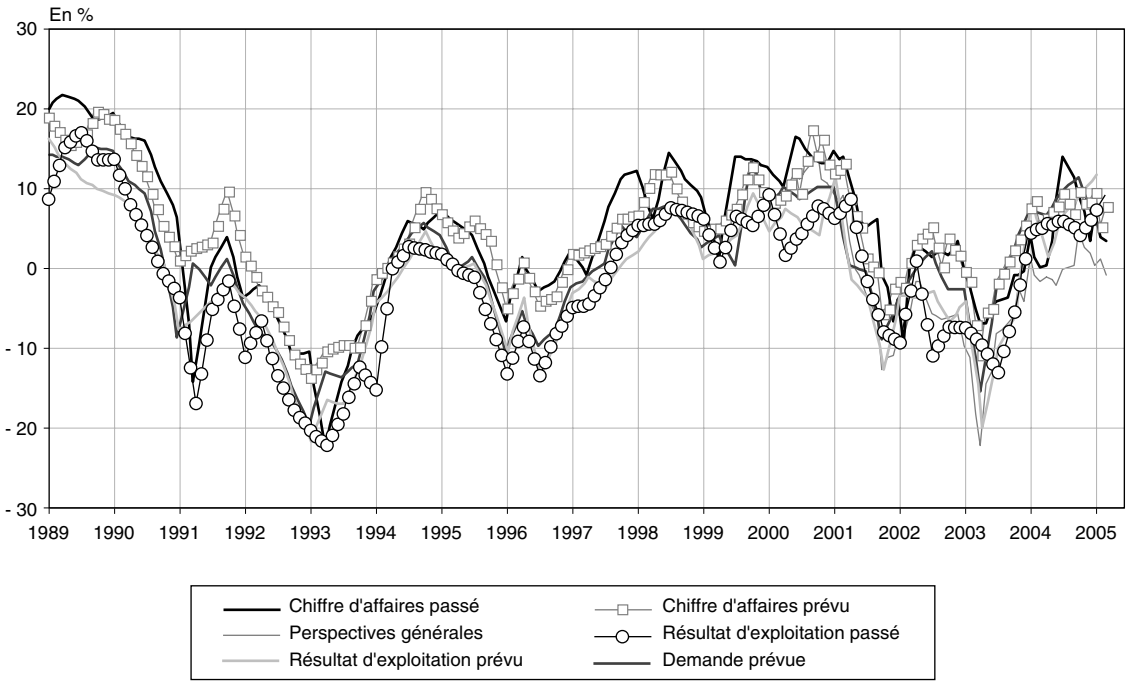
Jusqu'à la fin des années 1990, l'analyse conjoncturelle s'appuyait très largement sur les informations qualitatives et quantitatives relatives au secteur manufacturier. En particulier, elle exploitait assez peu les indicateurs concernant le secteur des services. Ce décalage, qui par le passé pouvait s'expliquer par l'abondance relative d'informations statistiques portant sur l'industrie, est en train de s'estomper. En effet, une place très importante est désormais accordée aux services dans le système statistique français, en raison notamment de leur rôle déterminant dans la compréhension des évolutions de court terme. Bouton et Erkel-Rousse (2003) ont montré que l'enquête de conjoncture dans les services est complémentaire de l'enquête dans l'industrie et permet d'améliorer la prévision du taux de croissance trimestriel du Pib. En plus des soldes d'opinion de l'enquête dans les services, leur étude mobilise un indicateur synthétique trimestriel extrait à l'aide d'une analyse factorielle statique.

Depuis septembre 2004, l'Insee publie chaque mois les résultats de son enquête de conjoncture dans les services ainsi qu'un indicateur synthétique résumant l'information contenue dans cette enquête. Cet indicateur mensuel représente l'information commune aux six principaux soldes d'opinion de l'enquête (cf. graphique I). Ces six soldes d'opinion présentent des fluctuations communes, ce qui suggère d'en extraire un signal – ou facteur – commun. Pour cela, chaque solde est représenté par la somme de deux termes : le premier terme obéissant à une dynamique commune à l'ensemble des séries (le facteur commun), le second, orthogonal au premier, étant une composante spécifique à chaque série.

Ainsi, l'indicateur synthétique est calculé dans le cadre de l'analyse factorielle dynamique, mise en œuvre dans de nombreuses études, en particulier par Geweke (1977), Sargent et Sim (1977), Stock et Watson (1989). Doz et Lenglart (1999) l'ont appliquée à l'enquête de conjoncture de l'Insee dans l'industrie.

Comme dans l'industrie, l'indicateur synthétique dans les services s'interprète comme une mesure du climat des affaires tel qu'il est perçu par les chefs d'entreprise. Il enrichit la panoplie des indicateurs de court terme ; en particulier il contient une information spécifique par rapport à l'indicateur synthétique dans l'industrie manufacturière. Il peut être appliqué aux trois sous-secteurs (services aux

Graphique I
Les six soldes d'opinion pris en compte dans l'analyse factorielle



Lecture : le solde d'opinion correspond à la différence entre le pourcentage d'entrepreneurs ayant répondu positivement à une question donnée et le pourcentage d'entrepreneurs ayant répondu négativement.
Ces soldes apparaissent corrélés entre eux, ce qui justifie la recherche d'une tendance commune. Par exemple, le coefficient de corrélation entre les soldes CAPA et REPA vaut 0,95.
Source : enquête de conjoncture dans les services, soldes d'opinion CVS, Insee.

entreprises, services aux particuliers et activités immobilières) couverts par l'enquête de conjoncture dans les services, ce qui permet d'affiner le diagnostic conjoncturel. Il permet enfin de quantifier le diagnostic conjoncturel, et peut notamment être utilisé pour prévoir le taux de croissance trimestriel de la production de services. Il contribue également à prévoir l'évolution trimestrielle du Pib, en complément de l'indicateur synthétique dans l'industrie manufacturière.

Du fait de sa contribution aux évolutions de la valeur ajoutée, le secteur des services est déterminant pour saisir la conjoncture

Deux raisons principales pouvaient expliquer que jusqu'à la fin des années 1990, les conjoncturistes aient privilégié les résultats des enquêtes de conjoncture dans l'industrie au détriment des informations provenant des autres secteurs :

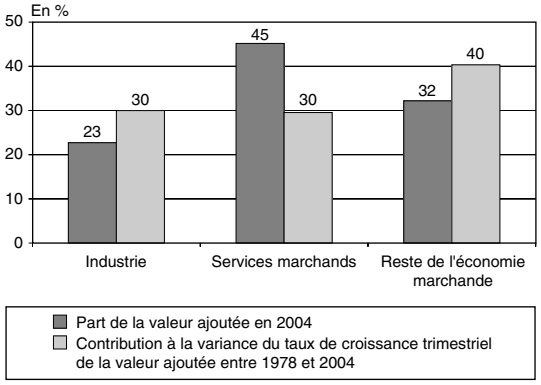
- d'une part, les conjoncturistes se préoccupent surtout des variations du taux de croissance de la valeur ajoutée et à ce titre l'industrie apparaît comme un secteur intéressant : bien qu'elle pèse moins que les services dans la valeur ajoutée marchande, son activité, plus fluctuante que celle des services, explique 30 % de la variance du taux de croissance trimestriel de cette valeur ajoutée (cf. graphique II) ;

- d'autre part, pour des raisons historiques, le suivi de la conjoncture dans l'industrie est assuré par un plus grand nombre d'indicateurs que dans les autres secteurs de l'économie, ces indicateurs étant en outre disponibles sur une période plus longue.

Cependant, en France, comme dans l'ensemble des économies développées, les services marchands occupent une place de plus en plus importante (cf. tableaux 1 et 2) :

- ce secteur représente aujourd'hui 45 % de la valeur ajoutée marchande ;

Graphique II
Répartition de la valeur ajoutée marchande et de la variance de son taux de croissance



Lecture :
23 (industrie) + 45 (services) + 32 (reste) = 100 (valeur ajoutée marchande en 2004)
30 (industrie) + 30 (services) + 40 (reste) = 100 (variance du taux de croissance de la valeur ajoutée entre 1978 et 2004)
Champ : la valeur ajoutée marchande (VA) est calculée sur les postes « EB » à « EP » de la nomenclature économique de synthèse (NES), c'est-à-dire hors : agriculture (EA) ; éducation, santé, action sociale (EQ) ; administration (ER). L'industrie recouvre les postes EB à EG de la NES, les services marchands les postes EM à EP. La partie « reste de l'économie marchande » comprend : la construction (EH), le commerce (EJ), les transports (EK) et les services financiers (EL).
Source : Comptes nationaux trimestriels, données en volume aux prix de 2000, Insee.

Tableau 1
Décomposition de la valeur ajoutée marchande

En %

	Décomposition de la valeur ajoutée…	
	… entre 1978 et 1982	… entre 2000 et 2004
Industrie	25	23
Services marchands	40	45
1. Services aux entreprises	17	21
2. Services aux particuliers	9	7
3. Activités immobilières	14	17
Reste de l'économie marchande	36	32
Ensemble de l'économie marchande	100	100

Source : Comptes nationaux trimestriels, données en volume aux prix de 2000, Insee.

Tableau 2
Décomposition de l'emploi salarié marchand

En %

	Décomposition de l'emploi salarié…	
	… entre 1978 et 1982	… entre 2000 et 2004
Industrie	41	27
Services marchands	20	35
Services aux entreprises	11	21
Services aux particuliers	7	12
Activités immobilières	1	2
Reste de l'économie marchande	40	39
Ensemble de l'économie marchande	100	100

Source : Département de l'emploi et des revenus d'activité, Insee.

- il explique 30 % des variations de cette valeur ajoutée entre 1978 et 2004 (1), soit autant que l'industrie (cf. graphique II) ;

- entre 1980 et 2004, sa part dans les effectifs salariés des secteurs concurrentiels a progressé de 15 points, passant de 20 % à 35 %, en lien notamment avec l'externalisation de certaines fonctions dans l'industrie et le développement de l'intérim.

Il semble donc important de s'intéresser plus précisément à l'activité dans les services. L'enquête de conjoncture dans les services qu'effectue l'Insee depuis 1988 (cf. encadré 1) apparaît alors comme un outil privilégié. Dans cet article, le secteur des services marchands correspondra au champ de cette enquête, aussi dénommée, dans la suite, enquête *Services*. Il comprend les services aux entreprises (postes et télécommunications, conseil et assistance, services opérationnels), les services aux particuliers (hôtels et restaurants, activités récréatives, culturelles et sportives, services personnels) et les activités immobilières (promotion, gestion et location immobilières). Ce secteur pourra être aussi appelé secteur des services, en omettant la référence à son caractère marchand.

L'enquête de conjoncture dans les services constitue une source d'information précieuse pour capter les fluctuations infra-annuelles dans ce secteur

Les questions posées dans l'enquête *Services* portent à la fois sur le passé proche et sur les anticipations des entrepreneurs. Elles permettent de capter les évolutions de court terme, de rendre compte de la situation courante et de prévoir le trimestre à venir. En outre, le questionnaire est suffisamment riche et varié pour donner une vision assez complète de la conjoncture dans les services, les entrepreneurs étant interrogés sur l'activité, l'emploi, les prix, etc.

Les enquêtes de conjoncture apportent des indications précoces sur le passé récent et les perspectives d'évolution à court terme du comportement des acteurs économiques pour chaque grand secteur d'activité. En effet, leur conception favorise la rapidité d'obtention des résultats, ceux-ci étant publiés à la fin du mois de collecte. Ainsi, elles sont des sources d'informations économiques rapidement disponibles, avant les indicateurs quantitatifs infra-annuels (2) (comptes nationaux trimestriels, indices de chiffres d'affaires, effectifs salariés, etc.). Enfin, leurs résultats sont soumis à de très faibles révisions : les résultats bruts sont révisés lors de la publication de l'enquête suivante, en prenant en compte les réponses tardives ; les séries corrigées des variations saisonnières sont légèrement révisées chaque année lorsque les coefficients saisonniers sont réestimés.

L'interprétation des résultats de ces enquêtes peut être rendue difficile pour au moins deux raisons :

- d'une part, un même solde d'opinion (3) peut présenter au mois le mois une évolution volatile, rendant malaisée sa lecture ;

- d'autre part, un mois donné, les soldes d'opinion peuvent afficher des fluctuations opposées.

La réponse proposée consiste à résumer l'information commune aux principaux soldes d'opinion issus de l'enquête de conjoncture dans les services à travers un indicateur synthétique. Cet indicateur, plus lisible que les soldes considérés séparément, facilite l'interprétation des résultats de l'enquête. En outre, l'approche retenue pour le construire permet de tenir compte de la fréquence hétérogène des séries.

L'indicateur synthétique mensuel permet bien d'appréhender la conjoncture dans les services

La recherche d'indicateurs synthétiques a déjà été explorée dans le cadre de l'enquête mensuelle de conjoncture auprès des industriels. La solution apportée est d'extraire un indicateur synthétique qui facilite la lecture des résultats de l'enquête (Doz et Lenglart, 1995 et 1999). Chaque solde d'opinion s'écrit comme la somme d'un terme qui suit une dynamique commune à tous les soldes d'opinion étudiés et d'un terme propre à la série considérée. L'objectif consiste alors à estimer par analyse factorielle cette composante commune, qui correspond à l'indicateur synthétique publié tous les mois dans les *Informations rapides*. Cet indicateur présente

---

1. De 1978 à 1990, l'industrie et les services marchands expliquent respectivement 29 % et 28 % des évolutions de la valeur ajoutée marchande. De 1991 à 2004, leur contribution gagne 3 points et s'élève respectivement à 32 % et 31 %. Parallèlement, la contribution du reste de l'économie marchande aux évolutions de cette valeur ajoutée diminue de 6 points entre ces deux périodes.

2. L'information conjoncturelle produite par l'Insee est accessible sur le site internet de l'Institut : www.insee.fr. Depuis la page d'accueil, cliquer sur « Conjoncture » puis « Indicateurs de conjoncture ».

3. La notion de solde d'opinion est définie dans l'encadré 1.

l'avantage d'être un signal plus aisément interprétable, car unique et moins volatil que les soldes d'opinion (Doz et Lenglart, 1995).

Dans le cadre de l'enquête *Services*, six soldes d'opinion sont pris en compte pour construire un indicateur synthétique (Casaux, Cornec, Deperraz et Lefebvre, 2004) (cf. encadré 1). La méthodologie utilisée pour le construire est analogue à celle mise en œuvre dans l'industrie. Cependant, l'extraction d'un tel indicateur à partir des résultats de l'enquête *Services* se heurte à trois difficultés particulières :

- les séries étudiées n'ont pas toutes la même fréquence (trois séries sont mensuelles, les trois autres sont trimestrielles) ;

- certaines séries changent de fréquence en cours de période (les soldes d'opinion sur le chiffre d'affaires passé et le chiffre d'affaires prévu deviennent mensuelles à partir de juin 2000) ;

Encadré 1

L'ENQUÊTE MENSUELLE DE CONJONCTURE DANS LES SERVICES

Présentation de l'enquête

L'Insee effectue depuis janvier 1988 une enquête d'opinion auprès des entreprises de services marchands. Le champ publié recouvre les services aux entreprises hors activités de courrier, télécommunications et administration d'entreprise ; les services aux particuliers ; les activités immobilières.

Cette enquête était initialement trimestrielle. Depuis juin 2000, la plupart des questions sont posées mensuellement. Leurs résultats sont publiés tous les mois depuis septembre 2004.

Les questions posées

Chaque mois, les entreprises sont interrogées sur l'évolution de leur activité au cours des trois derniers mois ainsi que sur leurs perspectives d'activité pour les trois prochains mois. Les entrepreneurs donnent aussi leur sentiment sur l'évolution générale de leur secteur. L'Insee leur demande également de juger l'évolution récente et future de leurs effectifs et des prix de vente de leurs prestations. Une fois par trimestre, les entrepreneurs des services répondent à des questions complémentaires, portant notamment sur l'évolution récente et future de leur résultat d'exploitation.

Les questions posées sont qualitatives et pour la plupart trimodales (par exemple : chiffre d'affaires « en hausse », « stable » ou « en baisse »). La liste complète des questions est donnée dans la note méthodologique de l'enquête disponible sur le site internet de l'Insee (www.insee.fr, cliquer sur « Conjoncture » puis « Indicateurs de conjoncture » et « Autres indicateurs »).

Présentation des résultats

Dans un premier temps, les données individuelles des entreprises d'une même strate élémentaire sont agrégées en utilisant un système de pondération issu des réponses individuelles aux questions structurelles. La donnée structurelle utilisée comme pondération dépend de la question. Selon le cas, il s'agit des effectifs de l'entreprise, de son chiffre d'affaires total ou de son chiffre d'affaires par type de prestation. Chaque strate élémentaire correspond au croisement entre un secteur fin d'activité de services, exprimé dans la nomenclature d'activité française à 700 postes (NAF 700), et d'une tranche de taille, délimitée par deux seuils de chiffres d'affaires.

Dans un second temps, les résultats ainsi obtenus au niveau des strates élémentaires sont agrégés en utilisant un système de pondérations reflétant l'importance relative de chacune des strates élémentaires dans l'ensemble du champ de l'enquête. Ce système de pondérations est entièrement fondé sur des données extérieures à l'enquête (*Enquête Annuelle d'Entreprise* ou source fiscale). On retrouve ainsi une structure proche de la structure des services en France au sens du champ de l'enquête.

Les résultats sont présentés sous la forme de soldes d'opinion. Un solde d'opinion est la différence entre le pourcentage pondéré de réponses « en hausse » et le pourcentage pondéré de réponses « en baisse ». Les séries publiées sont corrigées des variations saisonnières (CVS). Lorsque la série ne présente pas de caractère saisonnier, la série CVS est identique à la série brute. L'interprétation des résultats est fondée sur l'évolution des séries plutôt que sur leur niveau. Les soldes d'opinion peuvent également être commentés en comparaison à leur moyenne de longue période afin de tenir compte du comportement de réponse usuel des chefs d'entreprise.

Les six soldes d'opinion retenus dans l'analyse factorielle

Il s'agit d'extraire un signal commun aux six principaux soldes d'opinion issus de l'enquête de conjoncture dans les services : chiffre d'affaires passé, chiffre d'affaires prévu, perspectives générales, résultat d'exploitation passé, résultat d'exploitation prévu, demande prévue. Les trois premières séries sont mensuelles depuis juin 2000 tandis que les trois autres sont restées trimestrielles. De plus, la question sur les perspectives générales a été introduite en juin 2000. Ces soldes d'opinion sont corrigés des variations saisonnières (CVS).

- le solde d'opinion sur les perspectives générales d'activité n'existe pas avant juin 2000.

L'analyse factorielle statique n'est donc pas adaptée à ce cas. À l'instar de Doz et Lenglart (1995, 1999), nous traitons ce problème dans le cadre de l'analyse factorielle dynamique. En effet, ce cadre méthodologique est suffisamment général pour prendre en compte des séries de fréquence différente ainsi que des ruptures de fréquence (cf. encadré 2). L'indicateur synthétique ainsi obtenu s'interprète comme une mesure du climat conjoncturel dans les services tel qu'il est perçu par les chefs d'entreprise. Il offre un signal précoce sur l'activité économique et constitue un outil précieux pour le diagnostic conjoncturel.

Cet indicateur synthétique est robuste à différentes spécifications

La question de la robustesse du nouvel indicateur est essentielle. En effet, il est souhaitable qu'il présente des propriétés en partie indépendantes des spécifications ou de la méthode utilisée. Il apparaît également important qu'il soit peu révisé au cours du temps. La robustesse de l'indicateur synthétique dans les services peut d'abord être examinée sous les angles de sa sensibilité à la modélisation *ARMA* et à la période d'estimation.

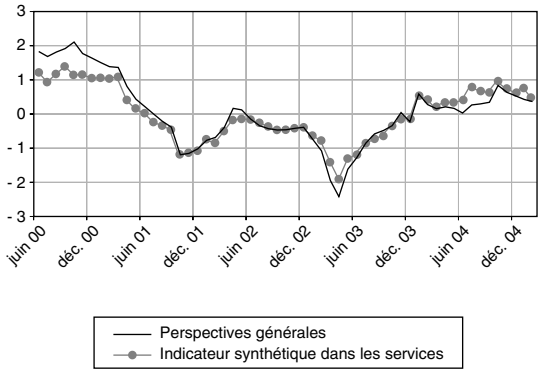
Le facteur commun a été estimé en utilisant différentes modélisations *ARMA* (p, q) avec p et q compris entre 0 et 3. Ces différentes estimations ne présentent pas de différence graphiquement visible. De même, lorsque la période d'estimation commence en 1991 ou en 1997, ou bien lorsque la série des perspectives générales n'est pas utilisée dans l'analyse, les écarts graphiquement visibles sont extrêmement faibles (les données et graphiques correspondant à ces calculs sont disponibles sur demande auprès des auteurs).

Signe de sa robustesse, l'indicateur synthétique dans les services est très peu sensible à l'ordre de la représentation *ARMA* testée, ainsi qu'à la période d'estimation retenue. Par ailleurs, la série sur les perspectives générales n'étant pas disponible avant 2000, on pourrait craindre que ce défaut d'information biaise l'estimation. Il n'en est rien : l'indicateur calculé sans les perspectives générales est très proche de l'indicateur fondé sur six soldes d'opinion. Pour autant, la série des perspectives générales apparaît utile au calcul du facteur commun. En effet, elle est bien corrélée avec les autres soldes et permet de disposer d'un solde d'opinion mensuel supplémentaire : on mobilise ainsi davantage d'information pour actualiser l'indicateur au mois le mois.

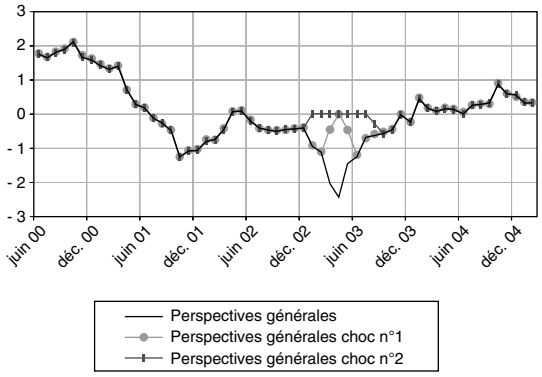
La ressemblance entre l'indicateur synthétique et le solde d'opinion sur les perspectives générales d'activité (cf. graphique III-A) suggère que ce

Graphique III
Réponse de l'indicateur synthétique à un choc sur la série des perspectives générales d'activité

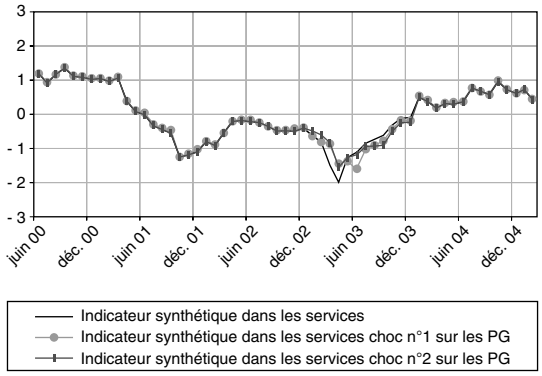
A – Perspectives générales et indicateur synthétique dans les services



B – Solde d'opinion sur les perspectives générales avec et sans choc



C – Indicateur synthétique avec un choc sur les perspectives générales (PG)



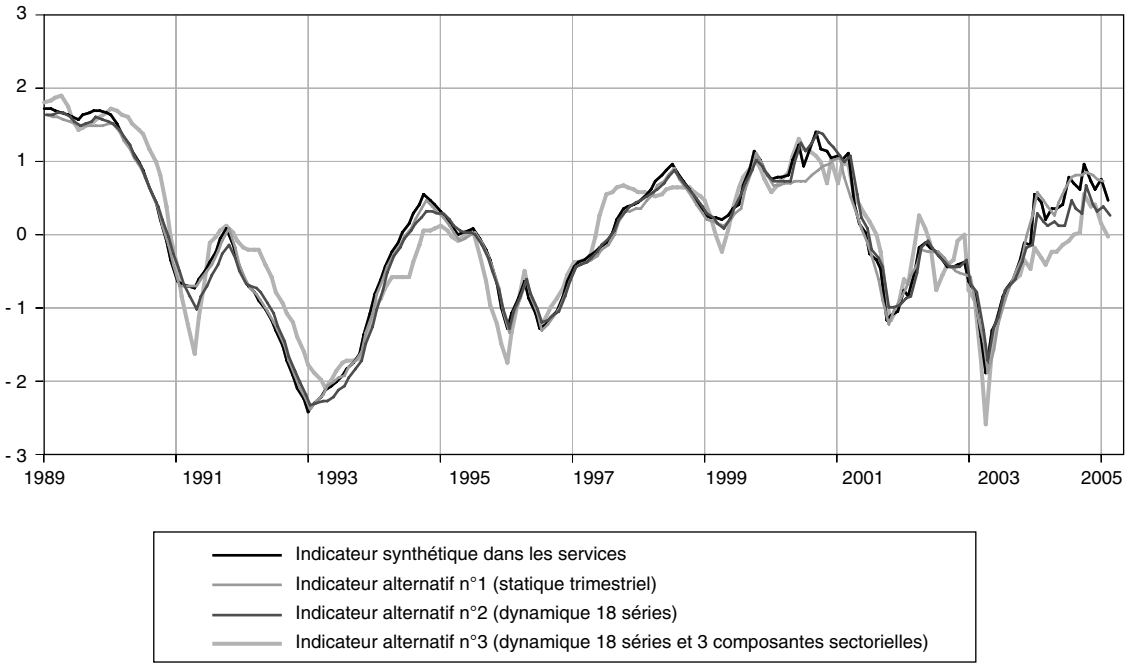
Lecture : les séries sont centrées-réduites c'est-à-dire qu'elles ont subi une transformation affine de telle sorte que leur moyenne soit nulle et leur écart-type égal à 1.
Source : Insee, enquête de conjoncture dans les services et calcul des auteurs.

dernier constitue lui-même un bon résumé des résultats de l'enquête *Services*. Ce constat pourrait faire croire que l'indicateur synthétique ne fait que reproduire ce solde d'opinion. Ce n'est pas le cas puisqu'un choc appliqué aux perspectives générales n'est que partiellement reflété dans l'évolution de l'indicateur synthétique (cf. graphiques III-B et III-C). Cette simulation contribue à justifier l'intérêt de l'indicateur pour la mesure du climat des affaires dans les services.

Par ailleurs, l'indicateur synthétique services (ISS) apparaît très proche d'indicateurs alternatifs extraits à partir de trois modèles différents (cf. tableau 3 et graphique IV). Lorsqu'on modifie le modèle à composantes inobservables sous-jacent, tout en restant dans le cadre de l'analyse factorielle, on obtient des signaux très voisins. Cette comparaison confirme la robustesse de l'indicateur services par rapport aux spécifications du modèle.

Le premier indicateur alternatif (ISS_STA) est le résultat d'une analyse factorielle *statique* sur les séries *trimestrielles*. Il est donc lui-même trimestriel et utilise uniquement cinq soldes d'opinion, car l'analyse statique ne permet pas d'utiliser la question sur les perspectives géné-

Graphique IV
Comparaison de différents indicateurs synthétiques



Lecture : les séries sont centrées-réduites c'est-à-dire qu'elles ont subi une transformation affine de telle sorte que leur moyenne soit nulle et leur écart-type égal à 1.
Source : Insee, enquête de conjoncture dans les services et calculs des auteurs.

Tableau 3
Nombre de séries, nombre de paramètres et dimension du vecteur d'état des quatre modèles d'analyse factorielle estimés

	Indicateur	Nombre de séries	Nombre de paramètres à estimer	Dimension du vecteur d'état
Indicateur synthétique services	ISS (1)	6	21	9
Premier indicateur alternatif	ISS_STA (2)	5	10	Sans objet
Deuxième indicateur alternatif	ISS_DYN_2 (1)	18	57	21
Troisième indicateur alternatif	ISS_DYN_3 (1)	18	78	24

1. Indicateurs dynamiques mensuels.
2. Indicateur statique trimestriel.

Lecture : les paramètres à estimer comprennent les pondérations (λ_i) *, l'écart-type des innovations* (σ_i) *, les paramètres du modèle ARMA(2,1) suivi par le facteur commun* $(\lambda_1, \lambda_2, \theta)$ *, le coefficient autorégressif des résidus* (ρ_i) *. Par exemple, dans le cas du deuxième indicateur alternatif ISS_DYN_2, les 57 paramètres du modèle sont :* $(\lambda_i)_{i=1\ldots18}, (\sigma_i)_{i=1\ldots18}, \lambda_1, \lambda_2, \theta, (\rho_i)_{i=1\ldots18}$ *, soit* $(18.3+3)=57$ *paramètres. Le vecteur d'état représente l'état latent de la conjoncture dans les services. Cet état est a priori inconnu mais il est relié aux variables observées, les soldes d'opinion, grâce auxquelles il peut être estimé (cf. encadré 2).*

Encadré 2

MÉTHODOLOGIE DE LA CONSTRUCTION DE L'INDICATEUR SYNTHÉTIQUE DANS LES SERVICES

Le modèle à composantes inobservées

À chaque mois t, chaque solde d'opinion i s'exprime comme la somme de deux composantes inobservées :

- un terme $\lambda_i F_t$ proportionnel au facteur commun F_t ;
- une composante u_{it} spécifique au solde d'opinion i considéré, également appelée résidu.

La dynamique du facteur commun et des composantes spécifiques est modélisée par un processus *ARMA* (processus autorégressif avec moyenne mobile, cf. Hamilton (1994) ou Gouriéroux et Monfort, 1997). Cette dynamique a été représentée par un modèle *ARMA*(2,1) tandis qu'un modèle *AR*(1) a été retenu pour représenter la dynamique des résidus. Les paramètres du modèle sont estimés au moyen du filtre de Kalman. Cette approche est suffisamment flexible pour prendre en compte les séries de fréquence variable issues de l'enquête *Services*. L'indicateur synthétique est l'espérance du facteur commun conditionnelle à l'information passée. L'application du filtre de Kalman à des séries temporelles est notamment décrite par Hamilton (1994) et par Kim et Nelson (1999), qui l'illustrent à travers plusieurs exemples concrets. D'autres méthodologies (non paramétriques) ont été proposées pour construire des indicateurs synthétiques à partir d'un grand nombre de séries, notamment par Forni *et al.* (1998, 2000) et Stock et Watson (2002).

Ainsi, on obtient le modèle paramétrique suivant :

$$\begin{cases} y_{it} = \lambda_i F_t + u_{it} \\ F_t = \varphi_1 F_{t-1} + \varphi_2 F_{t-2} + \varepsilon_t - \theta \varepsilon_{t-1} \\ u_{it} = \rho_i u_{it-1} + \varepsilon_{it} \end{cases}$$

(ε_t) et (ε_{it}) sont les innovations de (F_t) et de (u_{it}) respectivement ; (ε_t) et (ε_{it}) sont des bruits blancs gaussiens indépendants de variance respective 1 et σ_i^2. En effet, le facteur commun F_t étant défini à une constante multiplicative près, on pose : $V(\varepsilon_t) = 1$ afin de rendre le modèle identifiable. y_{it} représente la valeur du i^e solde d'opinion ($i = 1...6$) au mois t. Les soldes d'opinion sont corrigés des variations saisonnières et centrés réduits. Les paramètres à estimer sont : $\lambda_i, \varphi_1, \varphi_2, \theta, \sigma_i, \rho_i$.

Rappelons que :

- ce modèle est mensuel ;
- toutes les variables ne sont pas observées chaque mois (par exemple, les séries trimestrielles sont observées tous les trois mois) ;
- avant juin 2000 (date de mensualisation de l'enquête), seules les séries trimestrielles sont observées ;
- le solde d'opinion sur les perspectives générales n'est disponible qu'à partir de juin 2000.

Représentation espace-état du modèle latent

Ce modèle à composantes inobservées admet une représentation dite espace-état linéaire (cf. Hamilton (1994) ou Kim et Nelson (1999) pour une présentation générale des modèles espace-état), qui rend le calcul de la vraisemblance plus aisé. À l'aide de cette représentation, il sera possible d'utiliser le filtre de Kalman pour calculer la vraisemblance du modèle.

$$\underset{(n_t,1)}{y_t} = \underset{(n_t,9)(9,1)}{Z_t\ \alpha_t} \quad (\textit{équation de mesure})$$

$$\underset{(9,1)}{\alpha_t} = \underset{(9,9)(9,1)}{A\alpha_{t-1}} + \underset{(9,7)(7,1)}{R\ \eta_t} \quad (\textit{équation d'état})$$

$$\alpha_1 \sim N(0,\Sigma) \quad (\textit{condition initiale})$$

Les dimensions des vecteurs et matrices de cette représentation espace-état sont indiquées entre parenthèses. Dans cette représentation :

y_t est le vecteur colonne des soldes d'opinion observés pour chaque mois t ; la dimension n_t de ce vecteur change au cours du temps car les séries trimestrielles sont observées seulement un mois sur trois. Ainsi, avant juin 2000, la dimension de y_t est égale à 0 ou à 5 selon le mois tandis qu'à partir de juin 2000, elle vaut 3 (soldes d'opinion sur le chiffre d'affaires passé, le chiffre d'affaires prévu et les perspectives générales) ou 6 (soldes d'opinion sur le chiffre d'affaires passé, le chiffre d'affaires prévu, les perspectives générales, le résultat d'exploitation passé, le résultat d'exploitation prévu et la demande prévue) ;

Encadré 2 (suite)

Z_t est la matrice de mesure ; le nombre de lignes de Z_t est égal au nombre d'observations au mois t et varie donc en fonction du temps. Par exemple,

$$Z_t = \begin{pmatrix} \lambda_1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \lambda_2 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \lambda_3 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ \lambda_4 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ \lambda_5 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \lambda_6 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$ lorsque les 6 soldes sont observés

ou

$$Z_t = \begin{pmatrix} \lambda_1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \lambda_2 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \lambda_3 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$ lorsque seuls les 3 soldes mensuels sont observés ;

$\alpha_t = (F_t, F_{t-1}, \varepsilon_t, u'_t)'$ est le vecteur d'état, son contenu dépend étroitement du processus *ARMA* retenu pour le facteur commun et les résidus, α_t est de taille 9 ;

$u_t = (u_{1t}, ..., u_{6t})'$ est le vecteur des résidus ;

$$\underset{(7,1)}{\eta_t} = \begin{pmatrix} \varepsilon_t \\ \varepsilon_{1t} \\ \vdots \\ \varepsilon_{6t} \end{pmatrix}; \underset{(9,9)}{A} = \left(\begin{array}{ccc|ccc} \varphi_1 & \varphi_2 & -\theta & & 0 & \\ 1 & 0 & 0 & & & \\ 0 & 0 & 0 & & & \\ \hline & & & \rho_1 & & 0 \\ & 0 & & & \ddots & \\ & & & 0 & & \rho_6 \end{array}\right); \underset{(9,7)}{R} = \left(\begin{array}{c|cc} 1 & 0 & \\ 0 & 0 & \\ 1 & 0 & \\ \hline 0 & \multicolumn{2}{c}{Id(6)} \end{array}\right); Var(\eta_t) = D = \left(\begin{array}{c|ccc} 1 & \multicolumn{3}{c}{0} \\ \hline & \sigma_1^2 & & 0 \\ 0 & & \ddots & \\ & 0 & & \sigma_6^2 \end{array}\right).$$

Représentation espace-état du modèle observé

La représentation espace-état du modèle mensuel latent ne permet pas d'appliquer le filtre de Kalman directement car la matrice Z_t est nulle deux mois sur trois de janvier 1988 à juin 2000. Cependant, il est possible de changer la fréquence du modèle et d'obtenir des séries temporelles irrégulières. On obtient alors un modèle observé $y_{t'}$ avec :

t' = janvier 1988, avril 1988, juillet 1988, octobre 1988 … janvier 2000, avril 2000, juin 2000, juillet 2000 … janvier 2005, février 2005 (dernière observation utilisée dans cet article).

Les nouvelles dates t' correspondent aux mois où une variable au moins est observée. Soit t_0 le mois à partir duquel on dispose d'observations mensuelles (t_0 = juin 2000). Le modèle observé admet aussi une représentation espace-état :

$y_{t'} = Z_{t'} \alpha_{t'}$
$\alpha_{t'} = A_{t'} \alpha_{t'-1} + \chi_{t'}$

où :

$y_{t'}$ est le vecteur colonne des soldes d'opinion pour chaque mois t', sa dimension n'est jamais nulle ;

$\alpha_{t'}, u_{t'}, Z_{t'}$ sont inchangés ;

$(\chi_{t'})$ est un bruit blanc ;

Pour $t' \leq t_0, A_{t'} = A^3, Var(\chi_{t'}) = RDR' + ARDR'A' + A^2RDR'A'^2$; les matrices $A_{t'}$ et $Var(\chi_{t'})$ se déduisent du modèle latent en exprimant α_t en fonction de α_{t-3} : $\alpha_t = A^3\alpha_{t-3} + R\,\eta_t + AR\,\eta_{t-1} + A^2R\,\eta_{t-2}$;

Pour $t' > t_0, A_{t'} = A, Var(\chi_{t'}) = Var(R\eta_{t'}) = RDR'$.

Ainsi, contrairement à la première représentation, $A_{t'}$ et $Var(\chi_{t'})$ dépendent du temps.

Le filtre de Kalman est un algorithme itératif qui permet de calculer la vraisemblance du modèle, cf. Hamilton (1994). Pour l'initialisation du filtre, nous choisissons simplement $\alpha_1 \sim N(0, Id(9))$ plutôt que la solution stationnaire de l'équation d'état. Ce choix est effectué pour son avantage en temps de calcul et le fait que l'estimateur ainsi construit est asymptotiquement équivalent à l'estimateur du maximum de vraisemblance.

rales introduite en juin 2000. Cet indicateur est notamment utilisé dans Bouton et Erkel-Rousse (2003) en complément de l'indicateur synthétique dans l'industrie afin de prévoir les variations trimestrielles du Pib français. La ressemblance entre ISS et ISS_STA est d'autant plus remarquable que le second n'intègre pas les perspectives générales.

Les principaux avantages de l'analyse factorielle dynamique sont :

- la possibilité d'actualiser l'indicateur chaque mois à partir des nouvelles informations disponibles ;

- sa capacité à combiner des séries de fréquences différentes.

Le *deuxième indicateur alternatif* dynamique (ISS_DYN_2) est fondé sur le même modèle à composantes inobservées que ISS. Cependant, au lieu d'utiliser les six soldes d'opinion relatifs à l'ensemble des services, il utilise les 18 séries portant sur les sous-secteurs des services (six séries pour chacun des trois sous-secteurs : activités immobilières, services aux entreprises, services aux particuliers). Ce deuxième indicateur alternatif est également très proche de l'indicateur synthétique à 6 soldes.

Comme le précédent, le *troisième indicateur alternatif* (ISS_DYN_3) est extrait des 18 séries sectorielles au moyen d'une analyse factorielle dynamique. Cependant, il est fondé sur un modèle à composantes inobservées plus élaboré, dans lequel chaque solde d'opinion est la somme de trois termes : une composante commune à l'ensemble des services, une composante sectorielle et une composante spécifique. Ainsi, ce modèle comporte trois facteurs sectoriels en plus d'un facteur commun à toutes les séries. Chaque solde d'opinion y_{ijt} s'écrit :

$y_{ijt} = \lambda_{ij} F_t + \mu_{ij} F_{jt} + u_{ijt}$, où $i = 1 \ldots 6$ décrit les six soldes d'opinion, $j = 1 \ldots 3$ décrit les trois sous-secteurs des services. Chaque facteur sectoriel est modélisé par un processus $AR(1)$:

$F_{jt} = \gamma_j F_{jt-1} + \varepsilon_{jt}$ où $j = 1 \ldots 3$.

Ce troisième indicateur alternatif s'écarte toutefois parfois des trois autres, notamment en fin de période.

Encadré 2 (suite)

Estimation des paramètres

Les paramètres sont estimés par la méthode du maximum de vraisemblance. On obtient les estimations suivantes pour la dynamique *ARMA*(2,1) suivie par le facteur commun, dans le cas de l'ensemble des services marchands :

$$\underset{(écart-type)}{} \quad F_t = \underset{(0,05)}{1,90}\, F_{t-1} - \underset{(0,05)}{0,91} F_{t-2} + \varepsilon_t - \underset{(0,09)}{0,87}\, \varepsilon_{t-1}$$

L'annexe 1 contient les résultats détaillés des estimations.

Définition de l'indicateur synthétique

Une fois les paramètres estimés, l'indicateur synthétique dans les services (ISS) peut être calculé. Cet indicateur correspond à l'espérance conditionnelle du facteur commun connaissant l'information jusqu'à la date t : $ISS_t = \hat{F}_{t|t} = E(F_t | I_t)$.

Discussion des résultats statistiques

Les coefficients estimés semblent présenter une racine unitaire dans la représentation *ARMA* du facteur commun. Cette observation est déjà soulignée par Doz et Lenglart (cf. Doz et Lenglart, 1999). Bien que les tests usuels (Dickey-Fuller augmenté, Phillips-Perron, Schmidt-Phillips, Elliott-Rothenberg-Stock) acceptent l'hypothèse nulle de non-stationnarité, nous conservons l'a priori de stationnarité des soldes d'opinion dans les services par analogie avec l'enquête de conjoncture dans l'industrie. En effet, sur période longue, l'hypothèse de non-stationnarité des soldes d'opinion de l'enquête *Industrie* est rejetée. Ces séries sont disponibles mensuellement depuis 1976. Par contre, ces tests effectués avec les mêmes séries sur période courte, c'est-à-dire en ne retenant que les points observés depuis 1988, conduisent à accepter l'hypothèse nulle de non-stationnarité. Ainsi, l'approche de type modèle adoptée dans cet article n'est pas tant validée à l'aune de critères statistiques usuels que par ses propriétés de robustesse (cf. texte).

L'indicateur synthétique fournit une grille de lecture de la conjoncture dans les services

On peut lire dans cet indicateur synthétique l'évolution de la conjoncture dans les services au cours des quinze dernières années. En effet, il est notamment bien corrélé avec l'évolution de la production de services (4) (cf. graphique V) (le coefficient de corrélation entre l'indicateur synthétique et le glissement annuel de la production de services s'élève à 0,77).

- De 1990 à 1993, l'activité dans les services marchands connaît un ralentissement après la forte croissance de la fin des années 1980. Ce ralentissement, accentué pendant la période 1992-1993, est bien retracé par l'indicateur synthétique, qui s'inscrit alors en baisse sensible.

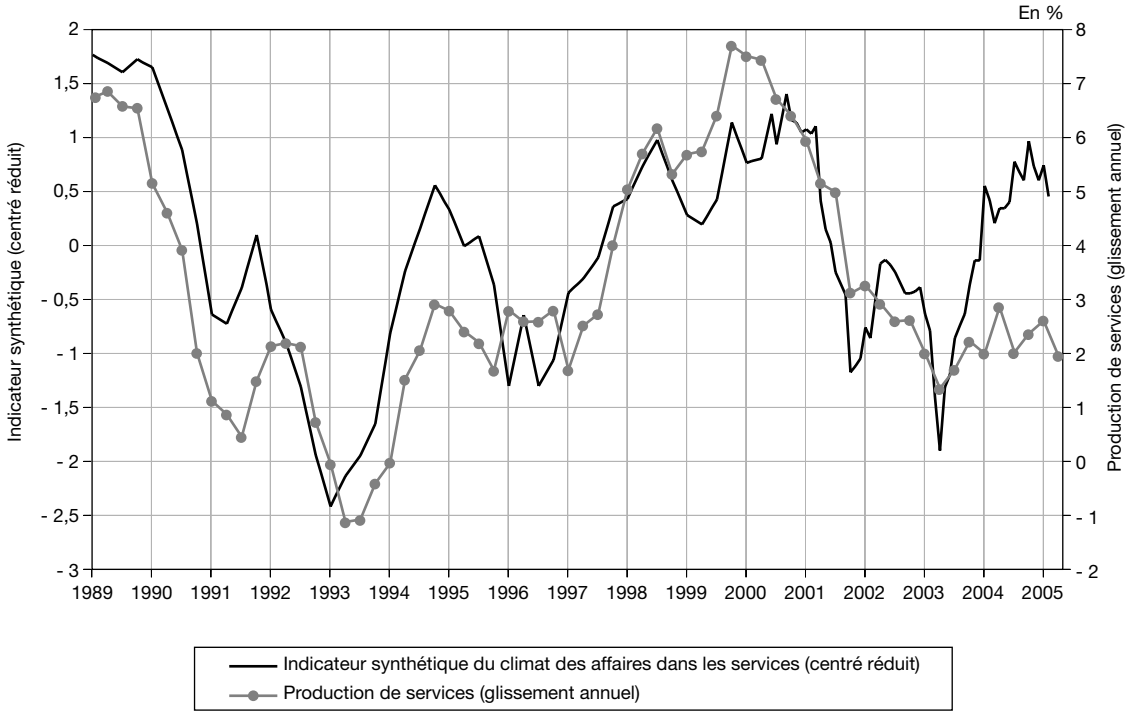
- De 1996 à 2000, les services marchands sont dynamiques, en particulier les services aux entreprises. Une explication réside certainement dans l'engouement suscité par les nouvelles technologies. Cette dynamique est temporairement interrompue fin 1998 – début 1999. La crise financière dans plusieurs pays émergents, en particulier en Asie du Sud-Est, entraîne alors un « trou d'air » dans l'économie.

- Au cours des années 2001-2003, l'activité ralentit fortement. En 2002, l'indicateur synthétique se redresse, contrairement à la production, dont le rythme de croissance se stabilise : il semble que des signaux de reprise ont conduit les entrepreneurs à faire preuve d'un excès d'optimisme. À l'inverse, au printemps 2003, l'indicateur synthétique décroît plus fortement que l'activité : divers événements financiers (notamment l'affaire Enron et ses conséquences sur le cabinet d'audit et de conseil Arthur Andersen) et surtout la guerre en Irak auraient accentué le pessimisme des chefs d'entreprise.

- À partir de la mi-2003, l'indicateur synthétique confirme la reprise de l'activité dans les services. Au second semestre de 2004, l'activité apparaît plus hésitante, puis tend à s'essouffler début 2005, après le point haut atteint à l'enquête de conjoncture de novembre 2004.

4. Production de services en volume aux prix de 2000, corrigée des variations saisonnières et des jours ouvrables (CVS-CJO). Cet agrégat issu des comptes nationaux trimestriels est la somme des productions de services immobiliers, de services aux entreprises et de services aux particuliers. Pour plus d'information sur son calcul, le lecteur peut se reporter à l'Insee méthodes n° 108, « Méthodologie des comptes trimestriels », accessible sur le site web de l'Insee : www.insee.fr → Publications → Insee méthodes.

Graphique V
Indicateur synthétique du climat des affaires et production de services



Lecture : l'indicateur synthétique est centré-réduit c'est-à-dire qu'il a subi une transformation affine de telle sorte que sa moyenne soit nulle et son écart-type égal à 1.
Source : Insee, comptes nationaux trimestriels et enquête de conjoncture dans les services.

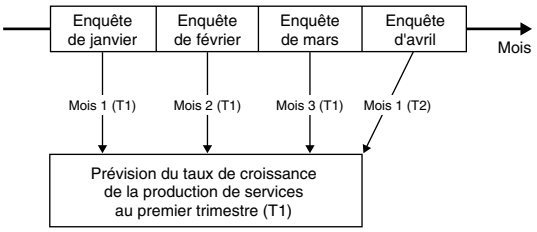
Sur la période récente, un décrochage entre l'évolution de la production de services et l'indicateur synthétique apparaît. Celui-ci traduit un écart entre l'opinion des entrepreneurs et les indicateurs économiques. Cependant, cet écart doit être nuancé dans la mesure où les données des comptes trimestriels sur la période récente sont à ce stade susceptibles de révisions. En effet, au moment de la rédaction de cet article, les comptes trimestriels 2004 et 2005 étaient fondés sur des indices infra-annuels et n'étaient pas encore calés sur les comptes annuels. En mai 2006, les comptes trimestriels 2004 ont été calés sur les comptes annuels et sont devenus « semi-définitifs ». Par ailleurs, l'indicateur synthétique apparaît bien corrélé avec l'évolution du Pib (cf. tableau 4, schéma et graphique VI). Ainsi, cet indicateur permet non seulement de bien capter la conjoncture des services mais aussi la conjoncture macroéconomique. Ce second aspect est développé dans Bouton et Erkel-Rousse (2003), qui démontrent l'apport de l'enquête de conjoncture dans les services dans la prévision à court terme de l'activité. En particulier, ils utilisent les résultats de l'enquête *Services* conjointement à ceux de l'enquête *Industrie* pour prévoir le taux de croissance trimestriel du Pib.

Tableau 4
Matrice des corrélations entre le glissement annuel du Pib et les indicateurs synthétiques dans l'industrie et dans les services

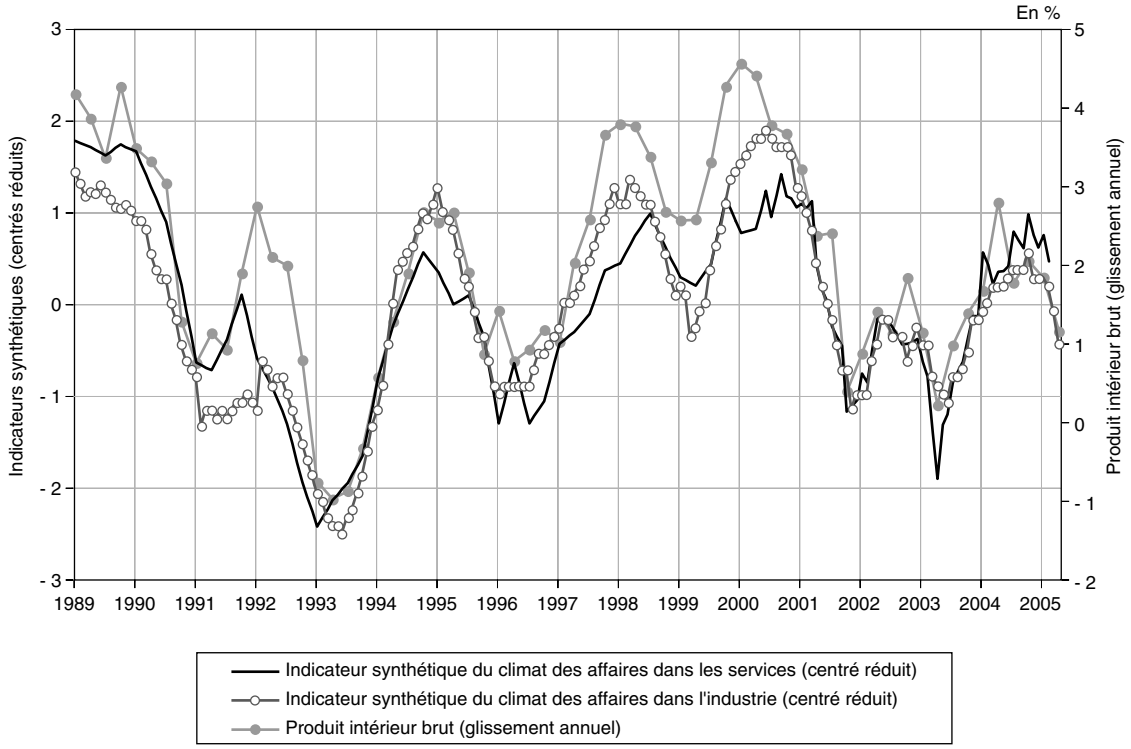
	Pib	Indicateur industrie	Indicateur services
Pib	1,00	0,89	0,86
Indicateur industrie	0,89	1,00	0,88
Indicateur services	0,86	0,88	1,00

Lecture : le glissement annuel est la variation entre la valeur du Pib d'un trimestre donné par rapport à la valeur du Pib au même trimestre de l'année précédente.

Schéma
Actualisation de la prévision du taux de croissance trimestriel de la production de services du premier trimestre en fonction de l'information disponible



Graphique VI
Indicateur synthétique du climat des affaires et produit intérieur brut



Lecture : les indicateurs synthétiques sont centrés-réduits c'est-à-dire qu'ils ont subi une transformation affine de telle sorte que leur moyenne soit nulle et leur écart-type égal à 1.
Source : Insee, comptes nationaux trimestriels et enquêtes de conjoncture dans les services et dans l'industrie.

L'analyse des indicateurs synthétiques sectoriels permet de préciser le diagnostic conjoncturel

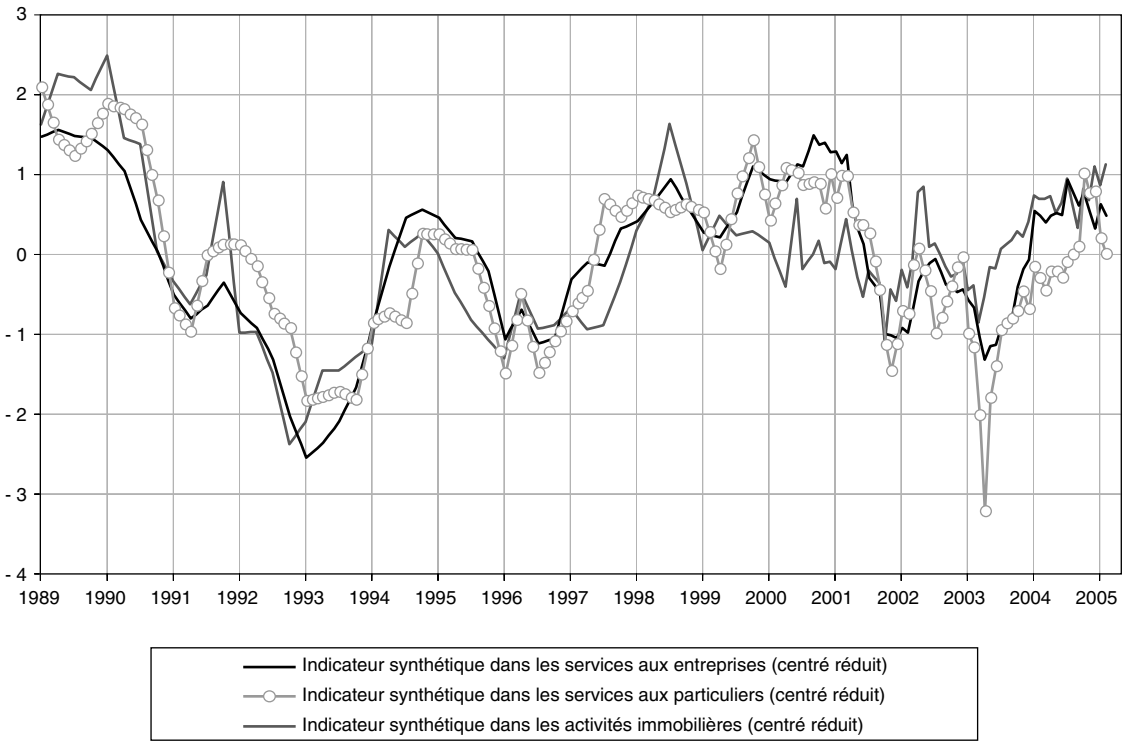
L'indicateur synthétique est appliqué aux trois sous-secteurs (services aux entreprises, services aux particuliers, activités immobilières) couverts par l'enquête (cf. graphique VII). La comparaison entre l'indicateur global et les indicateurs sectoriels permet de préciser le diagnostic conjoncturel.

- Les *services aux entreprises* représentent près de la moitié de la valeur ajoutée des services marchands. L'indicateur synthétique de ce secteur est d'ailleurs très proche de l'indicateur global. De 1994 à 1995, le climat des affaires dans les services aux entreprises se redresse de façon marquée, l'activité bénéficiant du regain de l'investissement dans l'industrie. En 1997, la remontée de l'indicateur synthétique s'explique en partie par le boom du travail temporaire lié à la reprise industrielle.

- Après la récession de 1993, le redressement de l'activité dans les *services aux particuliers* est moins soutenu que dans les autres secteurs. En particulier, en 1994 et 1995, l'appréciation du franc décourage les touristes étrangers et pèse sur l'hôtellerie-restauration. La vague d'attentats de 1995 pénalise encore la fréquentation touristique. En 2001, la chute de l'indicateur dans les services aux particuliers renvoie à la conjonction de plusieurs éléments : attentats du 11 septembre, ralentissement américain, etc. Cette morosité s'accentue encore début 2003, avec notamment le déclenchement de la guerre en Irak, qui tire l'indicateur synthétique à la baisse. Ces facteurs ont notamment pesé sur le tourisme. Dans les enquêtes de conjoncture, cela se traduit par des anticipations très pessimistes de la part des entrepreneurs de services aux particuliers (leurs prévisions de chiffre d'affaires, de résultat d'exploitation et de demande sont orientées à la baisse), et donc par la chute de l'indicateur synthétique. Au second semestre de 2003 et en 2004, le redémarrage des services aux particuliers demeure modéré, contrairement à celui des autres secteurs. C'est ce qu'indique l'indicateur synthétique de ce secteur, qui amorce une remontée plus graduelle que l'indicateur regroupant l'ensemble des services.

- Après une période de forte croissance, les *activités immobilières* ralentissent de façon marquée au début des années 1990. De 1996 à 2001, la croissance des activités immobilières

Graphique VII
Indicateurs synthétiques sectoriels



Lecture : les séries sont centrées-réduites c'est-à-dire qu'elles ont subi une transformation affine de telle sorte que leur moyenne soit nulle et leur écart-type égal à 1.
Source : Insee, enquête de conjoncture dans les services et calcul des auteurs.

apparaît assez irrégulière, comme en attestent les fluctuations de l'indicateur synthétique de ce secteur. Après une année 2000 stable et une année 2001 en baisse, la production du secteur immobilier n'a cessé de progresser sous les effets simultanés de conditions financières favorables et de l'amortissement dit « Besson », ce qui se traduit par une remontée marquée de l'indicateur synthétique.

De plus, le nouvel indicateur synthétique peut être utilisé pour prévoir la production trimestrielle de services…

Comme de nombreux résultats issus des enquêtes de conjoncture, l'indicateur synthétique de climat des affaires dans les services peut contribuer à quantifier le diagnostic conjoncturel. En particulier, il est possible d'estimer une relation économétrique entre cet indicateur et le taux de croissance trimestriel de la production de services (5). Ce taux de croissance constitue la variable endogène de l'équation ; il est expliqué par ses valeurs retardées et les valeurs courante et passées de l'indicateur synthétique. Plusieurs équations peuvent être envisagées selon l'information disponible. En effet, le conjoncturiste souhaite actualiser sa prévision de la production trimestrielle de services au fil des publications mensuelles des enquêtes de conjoncture. À cette fin, une démarche simple est proposée utilisant l'indicateur synthétique dans les services :

- Trimestrialiser les données mensuelles de manière à obtenir trois séries trimestrielles à partir d'une série mensuelle ;

- Estimer des modèles d'étalonnages de la production trimestrielle en fonction des séries trimestrielles obtenues.

L'approche proposée est celle de Dubois et Michaux (2004). Trois séries trimestrielles sont créées en découpant l'indicateur mensuel selon la place du mois dans le trimestre (cf. tableau 5). *In fine*, cette approche permet d'élaborer différents modèles de prévision adaptés à l'information disponible chaque mois.

Au total, *quatre modèles* de prévision de la production trimestrielle de services sont estimés. Ces modèles utilisent successivement les enquêtes des premier, deuxième ou troisième mois du trimestre courant ou du premier mois du trimestre suivant. Ainsi, ils permettent d'affiner la prévision au fil des mois en intégrant l'information nouvelle. Les résultats montrent notamment une amélioration de la qualité des estimations entre le premier et le deuxième mois du trimestre puis une stabilisation de la qualité d'ajustement des équations (cf. tableau 6).

… ou le produit intérieur brut, en complément de l'indicateur synthétique dans l'industrie

L'indicateur synthétique dans les services apporte également un éclairage sur l'activité globale. Ainsi, il peut être utilisé conjointement à l'indicateur synthétique du climat des affaires dans l'industrie pour prévoir le taux de croissance trimestriel du Pib (6). Comme pour la production de services, il est possible de mobiliser l'information disponible mensuellement

5. Production de services en volume aux prix de 2000, corrigée des variations saisonnières et des jours ouvrables (CVS-CJO). Source : comptes nationaux trimestriels. Cet agrégat est la somme des productions de services immobiliers, de services aux entreprises et de services aux particuliers.
6. Produit intérieur brut total en volume aux prix de 2000 (CVS-CJO).

Tableau 5
Découpage de l'indicateur mensuel en trois indicateurs trimestriels selon la place du mois dans le trimestre

Mois	Indicateur mensuel	Indicateur Mois 1	Indicateur Mois 2	Indicateur Mois 3
Janvier	X	X		
Février	X		X	
Mars	X			X
Avril	X	X		
Mai	X		X	
Juin	X			X
Juillet	X	X		
…	…	…	…	…

Lecture : l'indicateur Mois 1 est un indicateur trimestriel dont la valeur au trimestre T est la valeur de l'indicateur mensuel au premier mois du trimestre T.

afin de prévoir ce taux de croissance trimestriel. Avant 2000, la série trimestrielle de l'indicateur synthétique dans les services a été interpolée en utilisant l'indicateur lissé (7). Ensuite, les indicateurs services et industrie mensuels ont été découpés en trois séries trimestrielles, en fonction de la position du mois dans le trimestre (cf. tableau 5).

Le premier modèle consiste à estimer le Pib avec les indicateurs disponibles le premier mois du trimestre. Comme pour la production de services, la qualité d'ajustement de ces étalonnages s'améliore dès le deuxième mois du trimestre (cf. tableau 7) et se stabilise ensuite.

Ne pourrait-on prévoir le Pib en utilisant un seul de ces indicateurs synthétiques ? Pour évaluer l'utilité pour le prévisionniste de l'information apportée par l'indicateur services, on compare les estimations du taux de croissance trimestriel du Pib fondées sur ce seul indicateur avec des estimations utilisant l'indicateur synthétique du climat des affaires dans l'industrie, selon la méthode proposée par Davidson et McKinnon (1981).

Si y_t désigne le taux de croissance trimestriel du Pib, $\hat{y}_t^i$ son estimation obtenue à partir d'une régression utilisant l'enquête *Industrie*, l'enquête *Services* apporte une information complémentaire à l'enquête *Industrie* si, dans la régression $(y_t - \hat{y}_t^i) = \alpha\,(\hat{y}_t^s - \hat{y}_t^i) + v_t$, qui relie l'« erreur de prévision » résultant de la seule utilisation des données concernant l'industrie à la différence entre les deux estimations issues des enquêtes *Services* et *Industrie*, le coefficient α estimé est statistiquement différent de 0. De la même façon, si dans la régression $(y_t - \hat{y}_t^s) = \beta\,(\hat{y}_t^s - \hat{y}_t^i) + w_t$, le coefficient β estimé est statistiquement non nul, l'indicateur synthétique industrie contient une information spécifique par rapport à l'indicateur synthétique services.

Le tableau 8 présente les résultats de ces tests selon la dernière enquête disponible. Le taux de croissance trimestriel du Pib est successivement estimé en utilisant les enquêtes des 1er, 2e et 3e mois du trimestre courant et celles du 1er mois du trimestre suivant (Mois 1 (T + 1)). La période d'estimation s'étend du premier trimestre de 1990 au dernier trimestre de 2003.

7. L'indicateur synthétique lissé est égal à l'espérance conditionnelle du facteur commun sachant toute l'information disponible, soit les enquêtes trimestrielles puis mensuelles depuis 1988. De 1988 à 2000, cet indicateur lissé mensuel est très proche du résultat d'une interpolation linéaire de l'indicateur filtré trimestriel.

Tableau 6
Taux de croissance trimestriel de la production de services
Qualité d'ajustement des modèles selon la dernière enquête disponible

Modèle	Mois 1 (T)	Mois 2 (T)	Mois 3 (T)	Mois 1 (T + 1) (1)
R^2 ajusté	0,56	0,62	0,61	0,61
Écart-type des résidus (2)	0,44	0,41	0,41	0,41

1. Mois 1 (T + 1) : premier mois du trimestre suivant.
2. L'écart-type des résidus a été calculé dans l'échantillon « in sample » (c'est-à-dire que l'estimation et la validation ont été effectuées à l'aide du même échantillon) sur la période 1990 T1 – 2003 T4, en raison de la faiblesse de la taille de l'échantillon initial. L'écart-type de la variable endogène, le taux de croissance trimestriel de la production de services, s'élève à 0,66.

Lecture : le modèle Mois 3 (T) consiste à estimer le taux de croissance trimestriel de la production de services en utilisant l'enquête du troisième mois du trimestre courant. Par exemple, la production de services du premier trimestre est estimée avec l'enquête du mois de mars. Pour les résultats détaillés, cf. annexe 3.

Tableau 7
Taux de croissance trimestriel du Pib
Qualité d'ajustement des modèles selon la dernière enquête disponible

Modèle	Mois 1 (T)	Mois 2 (T)	Mois 3 (T)	Mois 1 (T + 1) (1)
R^2 ajusté	0,49	0,61	0,61	0,58
Écart-type des résidus (2)	0,34	0,30	0,29	0,30

1. Mois 1 (T + 1) : Premier mois du trimestre suivant.
2. L'écart-type des résidus a été calculé dans l'échantillon « in sample » (c'est-à-dire que l'estimation et la validation ont été effectuées à l'aide du même échantillon) sur la période 1990 T1 - 2003 T4, en raison de la faiblesse de la taille de l'échantillon initial. L'écart-type de la variable endogène, le taux de croissance trimestriel du Pib, s'élève à 0,47.

Lecture : le modèle Mois 3 (T) consiste à estimer le taux de croissance trimestriel du Pib en utilisant les enquêtes industrie et services du troisième mois du trimestre courant. Par exemple, le Pib du premier trimestre est estimé avec les enquêtes du mois de mars.

Quelle que soit la dernière enquête disponible, en retenant un seuil de 10 % pour les tests, l'indicateur services apporte une information supplémentaire par rapport à l'indicateur industrie pour prévoir le Pib. L'indicateur industrie contient, quant à lui, une part d'information spécifique par rapport à l'indicateur services sauf pour ce qui concerne le premier mois du trimestre.

Ce premier diagnostic est fondé sur des étalonnages prenant en compte l'évolution au trimestre le trimestre des deux indicateurs. Or, contrairement à l'indicateur services, qui est interpolé entre 1988 et 2000, l'indicateur industrie est mensuel sur l'ensemble de la période d'estimation. Aussi est-il possible d'affiner le diagnostic en tenant compte de la dynamique mensuelle de cet indicateur, c'est-à-dire de ses évolutions au mois le mois.

Par exemple, en début de trimestre, l'estimateur du taux de croissance du Pib utilisant l'enquête *Industrie* :

$$\hat{y}_t^i = 0,47 + 0,20 ISI\_m1_t + 1,06\left(ISI\_m1_t - ISI\_m3_{t-1}\right)$$

se substitue à l'estimateur précédent, qui n'utilisait que les enquêtes du premier mois de chaque trimestre :

$$\hat{y}_t^i = 0,47 + 0,18 ISI\_m1_t + 0,34(ISI\_m1_t - ISI\_m1_{t-1}).$$

Ce nouvel estimateur fait intervenir le terme $\left(ISI\_m1_t - ISI\_m3_{t-1}\right)$, qui représente l'évolution mensuelle de l'indicateur dans l'industrie entre le troisième mois du trimestre précédent et le premier mois du trimestre courant.

Tableau 8
Estimations du taux de croissance trimestriel du Pib
Apport relatif des enquêtes *Industrie* et *Services* en fonction de l'information disponible

A – Qualité d'ajustement de l'estimation du Pib en fonction de l'indicateur synthétique services

Modèle	Mois 1 (T)	Mois 2 (T)	Mois 3 (T)	Mois 1 (*T* + 1)
R^2 ajusté de la régression du taux de croissance trimestriel du Pib en fonction de l'indicateur synthétique services.				
R^2 ajusté	0,46	0,52	0,49	0,51

B – Qualité d'ajustement de l'estimation du Pib en fonction de l'indicateur synthétique industrie

Modèle	Mois 1 (T)	Mois 2 (T)	Mois 3 (T)	Mois 1 (*T* + 1)
R^2 ajusté de la régression du taux de croissance trimestriel du Pib en fonction de l'indicateur synthétique industrie.				
R^2 ajusté	0,37	0,57	0,54	0,49

C – Apport relatif de l'enquête *Services* par rapport à l'enquête *Industrie*

Modèle	Mois 1 (T)	Mois 2 (T)	Mois 3 (T)	Mois 1 (*T* + 1)
P-value de la statistique de Student du coefficient α dans la régression $(y_t - \hat{y}_t^i) = \alpha(\hat{y}_t^s - \hat{y}_t^i) + v_t$.				
P-value	0,002	**0,072**	**0,063**	0,005
Une *P-value* proche de zéro signifie que α est statistiquement non nul et que l'indicateur synthétique services contient une information spécifique par rapport à l'indicateur synthétique industrie. *P-value* (Mois 1) = 0,002 ; au seuil de deux pour mille (et donc, *a fortiori*, au seuil de 5 %), le coefficient α est jugé différent de 0, l'enquête *Services* apporte de l'information supplémentaire par rapport à l'enquête *Industrie*. En gras : *P-value* > 5 %.				

D – Apport relatif de l'enquête *Industrie* par rapport à l'enquête *Services*

Modèle	Mois 1 (T)	Mois 2 (T)	Mois 3 (T)	Mois 1 (*T* + 1)
P-value de la statistique de Student du coefficient β dans la régression $(y_t - \hat{y}_t^s) = \beta(\hat{y}_t^s - \hat{y}_t^i) + w_t$.				
P-value	**0,254**	0,001	0,004	0,025
Une *P-value* proche de zéro signifie que β est statistiquement non nul et que l'indicateur synthétique industrie contient une information spécifique par rapport à l'indicateur synthétique services. *P-value* (Mois 2) = 0,001 ; au seuil de un pour mille (et donc, *a fortiori*, au seuil de 5 %), le coefficient β est jugé différent de 0, l'enquête *Industrie* apporte de l'information supplémentaire par rapport à l'enquête *Services*. En gras souligné : *P-value* > 10 %.				

Lecture : chaque régression du mois m utilise uniquement les enquêtes du même mois de chaque trimestre. P-value : sous l'hypothèse où le coefficient α (ou β) soit nul, la statistique de Student suit une loi normale $N(0,1)$. La P-value représente la probabilité qu'une variable aléatoire suivant une telle loi soit supérieure, en valeur absolue, à cette statistique. Elle est donc comprise entre 0 et 1. Par rapport à la statistique de Student, la P-value présente l'avantage d'être directement comparable au seuil retenu pour effectuer le test. Une P-value proche de zéro signifie que le coefficient est statistiquement non nul.

L'apport de l'indicateur services pour la prévision du Pib est confirmé (cf. tableau 9). En outre, l'indicateur industrie apparaît désormais utile à la prévision dès le premier mois du trimestre.

Au total, ces deux indicateurs de climat des affaires apparaissent complémentaires dans la prévision du Pib.

* *
*

L'approche statistique présentée dans cette étude complète la boîte à outils de l'Insee en matière d'analyse conjoncturelle. Elle permet des prévisions des taux de croissance trimestriels de la production de services et du Pib, avec une actualisation de ces prévisions au mois le mois.

En s'appuyant sur la technique du filtre de Kalman, mise en œuvre pour extraire un indicateur mensuel combinant des séries de fréquences différentes, elle permet d'envisager de nouvelles utilisations des enquêtes de conjoncture.

L'écart récent entre l'opinion des entrepreneurs dans les services, résumée par l'indicateur synthétique, et les indicateurs économiques (cf. graphiques V et VI) mériterait d'être approfondi. Cette question pourrait notamment être explorée à travers la construction d'indicateurs synthétiques mêlant à la fois des informations qualitatives issues des enquêtes de conjoncture et des données quantitatives telles que la production par branche ou le Pib, en s'inspirant de Mariano et Murasawa (2003) et Cornec (2004). La méthodologie présentée dans cette étude ouvre également la voie à la construction d'indicateurs synthétiques « tous secteurs », mêlant des séries mensuelles, bimestrielles et trimestrielles issues de différentes enquêtes de conjoncture. □

Tableau 9
Estimations du taux de croissance trimestriel du Pib
Apport relatif des enquêtes *Industrie* et *Services* en fonction de l'information disponible en tenant compte de la dynamique mensuelle de l'indicateur *Industrie*

A – Qualité d'ajustement de l'estimation du Pib en fonction de l'indicateur synthétique services

Modèle	Mois 1 (T)	Mois 2 (T)	Mois 3 (T)	Mois 1 (T + 1)
R^2 ajusté de la régression du taux de croissance trimestriel du Pib en fonction de l'indicateur synthétique services.				
R^2 ajusté	0,46	0,52	0,49	0,51

B – Qualité d'ajustement de l'estimation du Pib en fonction de l'indicateur synthétique industrie

Modèle	Mois 1 (T)	Mois 2 (T)	Mois 3 (T)	Mois 1 (T + 1)
R^2 ajusté de la régression du taux de croissance trimestriel du Pib en fonction de l'indicateur synthétique industrie.				
R^2 ajusté	0,43	0,56	0,56	0,58

C – Apport relatif de l'enquête *Services*

Modèle	Mois 1 (T)	Mois 2 (T)	Mois 3 (T)	Mois 1 (T + 1)
R^2 ajusté de la régression du taux de croissance trimestriel du Pib en fonction de l'indicateur synthétique industrie.				
R^2 ajusté	0,43	0,56	0,56	0,58
P-value de la statistique de Student du coefficient α dans la régression $(y_t - \hat{y}_t^i) = \alpha\,(\hat{y}_t^s - \hat{y}_t^i) + v_t$.				
P-value	0,010	0,008	0,047	0,026
Une *P-value* proche de zéro signifie que α est statistiquement non nul et que l'indicateur synthétique services contient une information spécifique par rapport à l'indicateur synthétique industrie. *P-value* (Mois 1) = 0,008 ; au seuil de huit pour mille (et donc, *a fortiori*, au seuil de 5 %), le coefficient α est jugé différent de 0, l'enquête *Services* apporte de l'information supplémentaire par rapport à l'enquête *Industrie*.				

D – Apport relatif de enquête *Industrie*

Modèle	Mois 1 (T)	Mois 2 (T)	Mois 3 (T)	Mois 1 (T + 1)
P-value de la statistique de Student du coefficient β dans la régression $(y_t - \hat{y}_t^s) = \beta\,(\hat{y}_t^s - \hat{y}_t^i) + w_t$.				
P-value	0,043	0,001	0,001	0,000
Une *P-value* proche de zéro signifie que β est statistiquement non nul et que l'indicateur synthétique industrie contient une information spécifique par rapport à l'indicateur synthétique services. *P-value* (Mois 2) = 0,001 ; au seuil de un pour mille (et donc, *a fortiori*, au seuil de 5 %), le coefficient β est jugé différent de 0, l'enquête *Industrie* apporte de l'information supplémentaire par rapport à l'enquête *Services*.				

Lecture : cf. tableau 8.

BIBLIOGRAPHIE

Besley D.A., Kuh E. et Welch R.E. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Colinearity,* ed. John Wiley and Sons, Inc. New-York.

Bouton F. et Erkel-Rousse H. (2003), « Conjonctures sectorielles et prévision à court terme de l'activité : l'apport de l'enquête de conjoncture dans les services », *Économie et Statistique*, numéro spécial *Analyse conjoncturelle : entre statistique et économie*, n° 359-360, pp. 35-68.

Brière L., Duclos E., Héricher C., Okham M. et Raton I. (2005), « Les services marchands en 2004 : les services aux entreprises reprennent », *Insee Première*, n° 1030.

Casaux S., Cornec M., Deperraz T. et Lefebvre I. (2004), « Présentation des indicateurs synthétiques résumant le climat de affaires dans les services en France et en zone euro », *Note de conjoncture*, décembre 2004, Insee.

Cornec M. (2004), « Une datation mensuelle de la conjoncture française », *Note de conjoncture*, juin 2004, Insee.

Davidson R. et McKinnon J.G. (1981), « Several Tests for Model Specification in Presence of Alternative Hypothese », *Econometrica*, vol. 49, n° 3, pp. 781-793.

Doornik J.A. et Hansen H. (1994), « A Practical Test for Univariate and Multivariate Normality », *Discussion paper*, Nuffield College.

Doz C. et Lenglart F. (1995), « Une grille de lecture pour l'enquête mensuelle de l'industrie », *Note de conjoncture*, décembre 1995, Insee.

Doz C. et Lenglart F. (1999), « Analyse factorielle dynamique : test du nombre de facteurs, estimation et application à l'enquête de conjoncture dans l'industrie », *Annales d'Économie et de Statistique*, n° 54, pp. 91-127.

Dubois É. (2004), « Grocer 1.0 : an Econometric Toolbox for Scilab », *document de travail*, disponible à l'adresse internet http://dubois.ensae.net/grocer.html.

Dubois É. et Michaux E. (2004), « Étalonnages à l'aide d'enquêtes de conjoncture : de nouveaux résultats », *Économie et Prévision*, à paraître. Version préliminaire : http://dubois.ensae.net/biblio.html.

Fabre J. et Prost C. (2005), « Méthodologie des comptes trimestriels », *Insee Méthodes*, n° 108.

Forni M., Hallin M., Lippi M. et Reichlin L. (2000), « The Generalized Dynamic Factor Model : Identification and Estimation », *The Review of Economics and Statistics*, vol. 82, n° 4, pp. 540-552.

Forni M. et Reichlin L. (1998), « Let's Get Real : a Factor Analytic Approach to Disaggregated Business Cycle Dynamics », *Review of Economic Studies*, vol. 65, pp. 453-473.

Geweke J. (1977), « The Dynamic Factor Analysis of Economic Time Series », *in* D.J. Aigner et A.S. Goldberger (eds.), *Latent Variables in Socio-Economic Models*, pp. 365-383, North-Holland, Amsterdam.

Godfrey L.G. (1978), « Testing for Higher Order Serial Correlation in Regression Equations when the Regressors Include Lagged Dependent Variables », *Econometrica*, vol. 46, n° 6, pp. 1303-1313.

Gouriéroux C. et Monfort A. (1997), *Time Series and Dynamic Models*, Cambridge University Press, Cambridge, New York.

Hamilton J.D. (1994), *Time Series Analysis*, Princeton University Press.

Hendry D.F. (1979), « Predictive Failure and Econometric Modelling in Macro-Economics : The Transactions Demand for Money », *in Economic Modelling*, Ormerod P. (ed.), Heinemann, London, pp. 217–242.

Insee Méthodes (2005), « Méthodologie des comptes trimestriels », n° 108, accessible sur le site web de l'Insee : www.insee.fr → Publications → Insee méthodes.

Kim C. et Nelson C. (1999), *State-Space Models with Regime Switching : Classical and Gibbs-Sampling Approaches with Applications*, MIT Press, Cambridge, MA.

Krolzig H.M. et Hendry D.F. (2001), « Computer Automation of General-to-Specific Model Selection Procedures », *Journal of Economic Dynamics and Control*, vol. 25, n° 6-7, pp. 831-866.

Mariano R. et Murasawa Y. (2003), « A New Coincident Index of Business Cycles Based on Monthly and Quarterly Series », *Journal of Applied Econometrics*, vol. 18, n° 4, pp. 427-443.

Nicholls D.F. et Pagan A.R. (1983), « Heteroscedasticity in Models with Lagged Dependent Variables », *Econometrica*, vol. 51, n° 4, pp. 1233-1242.

Sargent T.J. et Sims C.A. (1977), « Business Cycle Modelling Without Pretending to Have Too Much *a priori* Economic Theory », *in* C.A. Sims (ed.), *New Methods in Business Cycle Research*, pp. 45-109, Federal Reserve Bank of Minneapolis, Minneapolis.

Stock J.H. et Watson M.W. (1989), « New Indexes of Coincident and Leading Economic Indicators », *NBER Macroeconomics Annual 1989*, MIT Press, Cambridge, pp. 351-394.

Stock J.H. et Watson M.W. (2002a), « Macroeconomic Forecasting Using Diffusion Indexes », *Journal of Business and Economic Statistics*, vol. 20, n° 2, pp. 147-162.

Stock J.H. et Watson M.W. (2002b), « Forecasting Using Principal Components From a Large Number of Predictors », *Journal of the American Statistical Association*, vol. 97, pp. 1167-1179.

Les **notes de conjoncture** sont accessibles sur le site web de l'Institut : www.insee.fr → Conjoncture → Analyse de la conjoncture → Archives de la note de conjoncture (Rechercher une note / un dossier).

ANNEXE 1

ESTIMATION DES PARAMÈTRES

Les soldes d'opinion de l'enquête *Services* utilisés dans l'analyse factorielle sont les suivants :

Tableau A
Les séries prises en compte dans l'analyse factorielle

Libellé	Solde d'opinion	Fréquence	Disponible depuis	Notes
CAPA	Chiffre d'affaires passé	Trimestrielle / mensuelle	1988 T1 / juin 2000	Mensuelle depuis juin 2000
CAPRE	Chiffre d'affaires prévu	Trimestrielle / mensuelle	1988 T1 / juin 2000	Mensuelle depuis juin 2000
PG	Perspectives générales	Mensuelle	juin 2000	Non disponible avant juin 2000
REPA	Résultat d'exploitation passé	Trimestrielle	1988 T1	
REPRE	Résultat d'exploitation prévu	Trimestrielle	1988 T1	
DEM (1)	Demande prévue	Trimestrielle	1988 T1	

(1) La question *Demande prévue* est mensuelle depuis septembre 2004. Actuellement, seul le solde d'opinion trimestriel est publié dans l'*Informations rapides* et utilisé dans l'analyse factorielle car la série mensuelle est encore trop courte pour pouvoir être corrigée des variations saisonnières.

Les quatre tableaux $B-1$ à $B-4$ donnent les valeurs estimées des coefficients λ_i (le coefficient, pour chaque solde d'opinion, du facteur commun estimé à partir des informations fournies par les six soldes), ρ_i (le coefficient autorégressif du résidu de l'équation relative à chaque solde), et σ_i (l'écart-type de l'innovation de l'équation portant sur chaque résidu).

Les paramètres φ_1, φ_2 et θ des processus *ARMA* (il s'agit dans tous les cas de processus *ARMA*(2,1)) sont donnés en dessous de chaque tableau.

Ainsi, on obtient le modèle paramétrique suivant, estimé 4 fois (pour les services, les activités immobilières, les services aux entreprises, les services aux particuliers) :

$$\begin{cases} y_{it} = \lambda_i F_t + u_{it} & \textit{L'équation relative à chaque solde} \\ F_t = \varphi_1 F_{t-1} + \varphi_2 F_{t-2} + \varepsilon_t - \theta \varepsilon_{t-1} & \textit{Le facteur commun aux six soldes} \\ u_{it} = \rho_i u_{it-1} + \varepsilon_{it} & \textit{Le résidu de l'équation d'un solde} \end{cases}$$

Tableau B-1
Paramètres estimés pour l'ensemble des services marchands

	λ_i		ρ_i		σ_i	
	Estimation	**Écart-type**	**Estimation**	**Écart-type**	**Estimation**	**Écart-type**
CAPA	0,21	0,02	0,64	0,08	0,30	0,03
CAPRE	0,25	0,03	0,60	0,13	0,20	0,02
PG	0,28	0,03	0,87	0,09	0,17	0,04
REPA	0,21	0,02	0,00	0,01	0,33	0,03
REPRE	0,22	0,02	0,38	0,34	0,25	0,04
DEM	0,23	0,03	0,00	0,01	0,24	0,03

La dynamique du facteur commun (*ARMA*(2,1), entre parenthèses les écarts-types) :

$$F_t = \underset{(0,05)}{1,90} F_{t-1} - \underset{(0,05)}{0,91} F_{t-2} + \varepsilon_t - \underset{(0,09)}{0,87} \varepsilon_{t-1}$$

Lecture : les coefficients en grisé sont non significatifs.

Tableau B-2
Paramètres estimés pour les activités immobilières

	λ_i		ρ_i		σ_i	
	Estimation	**Écart-type**	**Estimation**	**Écart-type**	**Estimation**	**Écart-type**
CAPA	0,26	0,05	0,47	0,11	0,62	0,06
CAPRE	0,28	0,06	0,52	0,12	0,66	0,06
PG	0,56	0,28	0,89	0,21	0,42	0,28
REPA	0,30	0,05	0,30	0,56	0,53	0,11
REPRE	0,31	0,05	- 0,68	0,10	0,40	0,06
DEM	0,23	0,06	0,69	0,09	0,56	0,07
La dynamique du facteur commun (*ARMA*(2,1), entre parenthèses les écarts-types) : $F_t = \underset{(0,22)}{0,19} F_{t-1} + \underset{(0,21)}{0,71} F_{t-2} + \varepsilon_t + \underset{(0,30)}{0,53} \varepsilon_{t-1}$						

Lecture : le coefficient en grisé n'est pas significatif.

Tableau B-3
Paramètres estimés pour les services aux entreprises

	λ_i		ρ_i		σ_i	
	Estimation	**Écart-type**	**Estimation**	**Écart-type**	**Estimation**	**Écart-type**
CAPA	0,14	0,02	0,67	0,07	0,25	0,02
CAPRE	0,17	0,03	0,39	0,14	0,26	0,02
PG	0,19	0,03	0,80	0,11	0,19	0,03
REPA	0,14	0,02	- 0,08	0,06	0,31	0,03
REPRE	0,15	0,02	- 0,01	0,01	0,31	0,03
DEM	0,15	0,02	- 0,59	0,14	0,23	0,04
La dynamique du facteur commun (*ARMA*(2,1), entre parenthèses les écarts-types) : $F_t = \underset{(0,06)}{1,86} F_{t-1} - \underset{(0,06)}{0,88} F_{t-2} + \varepsilon_t - \underset{(0,13)}{0,67} \varepsilon_{t-1}$						

Lecture : les coefficients en grisé sont non significatifs.

Tableau B-4
Paramètres estimés pour les services aux particuliers

	λ_i		ρ_i		σ_i	
	Estimation	**Écart-type**	**Estimation**	**Écart-type**	**Estimation**	**Écart-type**
CAPA	0,32	0,04	0,51	0,12	0,46	0,04
CAPRE	0,37	0,04	0,21	0,19	0,34	0,03
PG	0,46	0,06	0,89	0,10	0,18	0,08
REPA	0,31	0,04	- 0,01	Nd	0,53	0,05
REPRE	0,34	0,04	0,76	0,09	0,22	0,04
DEM	0,34	0,04	0,00	Nd	0,41	0,04
La dynamique du facteur commun (*ARMA*(2,1), entre parenthèses les écarts-types) : $F_t = \underset{(1,26)}{1,18} F_{t-1} - \underset{(1,16)}{0,24} F_{t-2} + \varepsilon_t - \underset{(1,28)}{0,23} \varepsilon_{t-1}$						

Les pondérations λ_i sont toutes significatives, quel que soit le secteur, quelle que soit l'enquête. Ainsi, le facteur commun latent contribue à l'évolution de tous les soldes d'opinion retenus dans l'analyse factorielle. Plus précisément, la dynamique commune explique entre 73 % et 95 % de la variance de ces soldes (cf. tableau C).

La pondération la plus élevée est celle associée à la série des perspectives générales. Les coefficients autorégressifs ρ_i des résidus des variables mensuelles (*CAPA, CAPRE, PG*) sont également significatifs.

En revanche, les coefficients autorégressifs des résidus des variables trimestrielles (*REPA*, *REPRE*, *DEM*) ne le sont pas toujours. Ce résultat laisse à penser que ces résidus sont assimilables à des bruits blancs, traduisant le fait que les soldes d'opinion correspondants fluctuent autour du facteur commun et que l'information propre à chaque série ne dépend pas de son passé.

Tableau C
La part de variance expliquée par le facteur commun (en %)

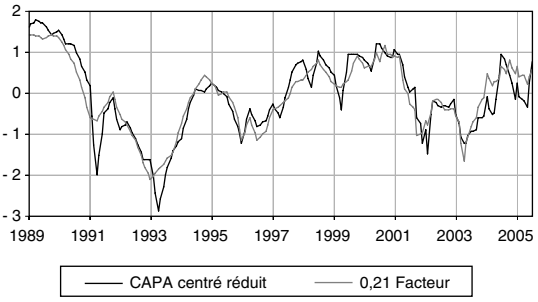
	Ensemble des services	Activités immobilières	Services aux entreprises	Services aux particuliers
CAPA	73	37	73	66
CAPRE	95	44	93	87
PG	91	54	91	97
REPA	81	65	80	69
REPRE	89	69	88	84
DEM	94	39	90	81

ANNEXE 2

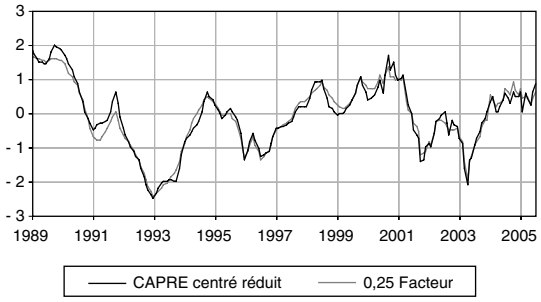
LA PART DE LA DYNAMIQUE COMMUNE DANS LES SOLDES D'OPINION

Graphique A
Les soldes d'opinion et leur composante proportionnelle au facteur commun

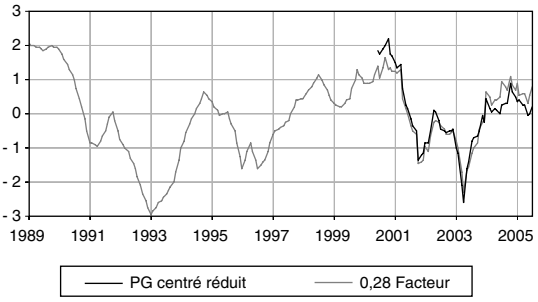
A1 – Chiffre d'affaires passé (CAPA)



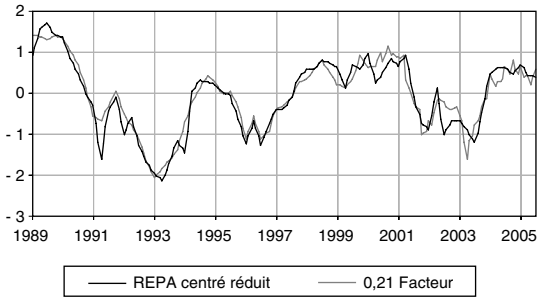
A2 – Chiffre d'affaires prévu (CAPRE)



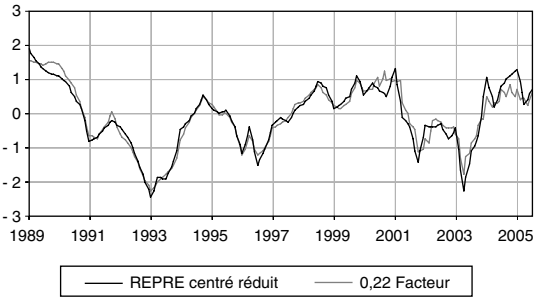
A3 – Perspectives générales (PG)



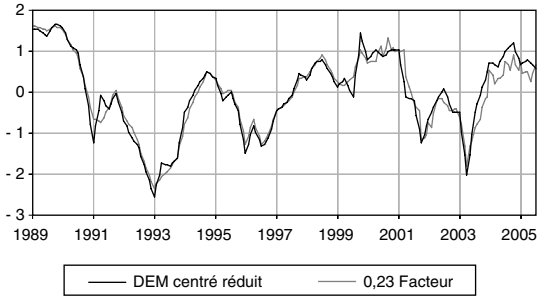
A4 – Résultat d'exploitation passé (REPA)



A5 – Résultat d'exploitation prévu (REPRE)



A6 – Demande prévue (DEM)



Lecture : chaque graphique représente un solde d'opinion centré réduit et sa composante proportionnelle au facteur commun ; par exemple : « REPA centré réduit » et « 0,21 Facteur », ce second terme désigne $\hat{\lambda}_{REPA}\,\hat{F}_t$.
Les soldes sont centrés réduits c'est-à-dire qu'ils ont subi une transformation affine de telle sorte que leur moyenne soit nulle et leur écart-type égal à 1.

ANNEXE 3

MODÈLES DE PRÉVISION DU TAUX DE CROISSANCE TRIMESTRIEL DE LA PRODUCTION DE SERVICES ET DU PIB

A – Estimation du taux de croissance de la production de services selon la dernière enquête *Services* disponible

Variable	Modalité	Mois 1 (T)	Mois 2 (T)	Mois 3 (T)	Mois 1 (*T* + 1) (1)
Constante		0,40 (3,0)	0,38 (3,1)	0,43 (3,5)	0,42 (3,4)
Taux de croissance trimestriel de la production de services	Tctps (- 2)	0,53 (4,1)	0,57 (4,6)	0,51 (4,2)	0,53 (4,3)
Indicateur services obtenu en conservant les enquêtes du premier mois de chaque trimestre (trimestriel)	*ISS _ m*1(+1)				0,44 (5,4)
	*ISS _ m*1	0,59 (4,56)			
	*ISS _ m*1(-1)	- 0,43 (- 3,1)			- 0,20 (- 2,1)
Indicateur services obtenu en conservant les enquêtes du deuxième mois de chaque trimestre (trimestriel)	*ISS _ m*2		0,77 (5,4)		
	*ISS _ m*2(-2)		- 0,59 (- 3,7)		
Indicateur services obtenu en conservant les enquêtes du troisième mois de chaque trimestre (trimestriel)	*ISS _ m*3			0,70 (5,0)	
	*ISS _ m*3(-1)			- 0,45 (- 2,9)	
R^2 ajusté		0,56	0,62	0,61	0,61
Écart-type des résidus (2)		0,44	0,41	0,41	0,41
Durbin-Watson		1,75	1,78	1,89	1,91
Indice de conditionnement		3	5	5	4
Chow (50 %) (3)		0,49	0,56	0,40	0,38
Chow (90 %) (3)		0,89	0,87	0,51	0,62
Normalité (3)		0,07	0,06	0,35	0,33
Autocorrélation (ordre 4) (3)		0,58	0,24	0,53	0,51
Hétéroscédasticité (3)		0,32	0,60	0,78	0,44

1. Mois 1 (*T* + 1) : Premier mois du trimestre suivant.

2. L'écart-type des résidus peut être comparé à celui du taux de croissance trimestriel de la production de services, qui s'élève à 0,66 sur la période d'estimation (1990 T1 – 2003 T4).

3. *P-value* du test.

Lecture : quatre modèles de prévision de la production trimestrielle de services sont estimés. Ces modèles utilisent successivement les enquêtes des premier, deuxième ou troisième mois du trimestre courant ou du premier mois du trimestre suivant. Entre parenthèses sont indiquées les statistiques de Student des paramètres estimés. Période d'estimation : 1990 T1 – 2003 T4.

B – Estimation du taux de croissance du Pib selon les dernières enquêtes *Industrie* et *Services* disponibles

Variable	Modalité	Mois 1 (T)	Mois 2 (T)	Mois 3 (T)	Mois 1 (T + 1) (1)
Constante		0,75 (8,1)	0,76 (9,4)	0,71 (8,3)	0,73 (8,9)
Taux de croissance trimestriel du Pib	Tctpib (- 1)	- 0,35 (- 2,6)	- 0,36 (- 3,0)	- 0,36 (- 3,0)	- 0,29 (- 2,4)
Indicateur industrie obtenu en conservant les enquêtes du premier mois de chaque trimestre (trimestriel)	*ISI _ m*1(+1) – *ISI _ m*1				0,51 (5,8)
	*ISI _ m*1 – *ISI _ m*1(–1)	0,43 (4,3)			
Indicateur services obtenu en conservant les enquêtes du premier mois de chaque trimestre (trimestriel)	*ISS _ m*1(+1)				0,45 (7,1)
	*ISS _ m*1	0,39 (5,7)			
Indicateur industrie obtenu en conservant les enquêtes du deuxième mois de chaque trimestre (trimestriel)	*ISI _ m*2 – *ISI _ m*2(–1)		0,44 (5,5)		
Indicateur services obtenu en conservant les enquêtes du deuxième mois de chaque trimestre (trimestriel)	*ISS _ m*2		0,43 (6,8)		
Indicateur industrie obtenu en conservant les enquêtes du troisième mois de chaque trimestre (trimestriel)	*ISI _ m*3 – *ISI _ m*3(–1)			0,79 (7,1)	
	*ISI _ m*3(–2)			0,17 (2,0)	
Indicateur services obtenu en conservant les enquêtes du troisième mois de chaque trimestre (trimestriel)	*ISS _ m*3			0,28 (2,7)	
R^2 ajusté		0,49	0,61	0,61	0,58
Écart-type des résidus (2)		0,34	0,30	0,29	0,30
Durbin-Watson		1,94	1,83	1,98	1,98
Indice de conditionnement		4	4	6	4
Chow (50 %) (3)		0,51	0,55	0,45	0,59
Chow (90 %) (3)		0,63	0,50	0,37	0,43
Normalité (3)		0,15	0,84	0,33	0,95
Autocorrélation (ordre 4) (3)		0,59	0,88	0,97	0,72
Hétéroscédasticité (3)		0,43	0,29	0,64	0,45

1. Mois 1 (T + 1) : Premier mois du trimestre suivant.
2. L'écart-type des résidus peut être comparé à celui du taux de croissance trimestriel du produit intérieur brut, qui s'élève à 0,47 sur la période d'estimation (1990 T1 – 2003 T4).
3. *P-value* du test.

Lecture : quatre modèles du Pib sont estimés. Ces modèles utilisent successivement les enquêtes des premier, deuxième ou troisième mois du trimestre courant ou du premier mois du trimestre suivant. Entre parenthèses sont indiquées les statistiques de Student des paramètres estimés. Période d'estimation : 1990 T1 – 2003 T4.

C – Lecture des tests de spécifications

Durbin-Watson	La statistique de Durbin et Watson permet de tester l'autocorrélation des résidus à l'ordre 1.
Indice de conditionnement	Cet indice permet de diagnostiquer les problèmes éventuels de multicolinéarité des régresseurs. Les situations pathologiques correspondent à un indice de conditionnement maximal supérieur à 30, cf. Belsley, Kuh et Welsch (1980).
Chow (50 %) Chow (90 %)	Test d'échec prédictif de Chow sur respectivement 50 % et 90 % de la période, cf. Hendry (1979). Il permet de tester la stabilité des paramètres estimés en réestimant le modèle sur un sous-ensemble de l'échantillon (ici, 50 % et 90 % des observations).
Normalité	Test de normalité de Doornik et Hansen (1994).
Autocorrélation (ordre 4)	Test du multiplicateur de Lagrange d'autocorrélation des résidus jusqu'à l'ordre 4, cf. Godfrey (1978).
Hétéroscédasticité	Test d'hétéroscédasticité quadratique entre les régresseurs, cf. Nicholls et Pagan (1983).

Ces tests de spécifications sont notamment préconisés par Krolzig et Hendry (2000).

Chapter 6

Simulating spot electricity prices with regenerative blocks

SIMULATING SPOT ELECTRICITY PRICES WITH REGENERATIVE BLOCKS

Matthieu Cornec
OSIRIS Department, R & D Division
Électricité de France
1, avenue du Général de Gaulle
Clamart, France
email: matthieu.cornec@edf.fr

Hugo Harari-Kermadec
Laboratory of Statistics, CREST
and Ceremade, Université Paris-Dauphine
Paris, France
email: harari@ensae.fr

ABSTRACT
In the early 90's, electricity production has moved from state monopolies to competitive markets, with a new element of risk: the wholesale price uncertainty. Thus, modelling electricity-prices together with exogenous variables is crucially needed for plant scheduling, generation asset management and option pricing. Recent literature on empirical time-series has revealed the difficulty for traditional financial parametric models to catch up with the complex features of electricity prices: mean-reversion, multi-scale seasonality, erratic extreme behaviour with fast reverting spikes. We propose to consider an empirical approach, inspired from the Bootstrap literature. We make use of the regenerative structure of time-series seen as Markov chains to construct almost independent data blocks. Eventually, this approach is applied to the simulation of Powernext electricity prices conditionnaly on temperature.

KEY WORDS
Semiparametric and Nonparametric Methods, Time-Series Models, Simulation Tools, Electric Utilities, Statistical Modelling, Electricity Market Modelling

1 Introduction

Before the 90's, electricity was a monopoly in most countries, government owned or regulated. This changed radically: major countries have deregulated generation and supply activities. One of the important consequences of this reorganization is that prices are now set according to the rule of supply and demand. However, the rich behaviour of electricity prices, caused by its nonstorability, rules out traditional financial models. This remark advocates for a non parametric approach, which is not the common strategy of the literature. Instead of considering parametric model assumptions, we examine assumptions on the dependence structure of the data. We propose to use an algorithm based on Markov chains, introduced by [1, 2]. The model we describe in the subsequent sections will capture the following characteristics observed in electricity markets: seasonal patterns and periodicities, price spikes, mean reversion, price dependent volatilities, long-term nonstationarity.

The paper is organized as follows. In the next section, we summarize the special characteristics of price formation in electricity markets. In section 3, we give a short review of literature on models for electricity prices. A detailed description of our model is given in section 4. Finally, in section 5, some simulation results are described.

2 Stylized facts about spot electricity prices

2.1 The spot market

Spot electricity market is originally a day-ahead market. Every day is divided into 24 hourly spot contracts with physical delivery. Spot hourly prices are day ahead prices determined simultaneously for all 24 hours of the next day. Thus, there are no a priori reasons to model spot price as an hourly series. All interesting statistical features are present in average daily peak price (hours between 8am and 8pm) and off peak price (remaining hours). Principal Component Analysis shows that two daily factors explain about 80% of variance. In this article, we thus concentrate on daily peak and off peak prices.

2.2 Main features

The following features are observed in all electricity markets (cf. figure 1):

Seasonality The daily spot market exhibit a weekly seasonality, that can be explained via the merit order curve.

Price spikes Electricity spot prices show extreme spikes. The price tail distribution is thus heavier than normal or lognormal distribution.

Mean reversion Prices have the tendency to revert rapidly from price spikes to a mean level. Characteristic times of mean reversion have a magnitude of days or at most weeks and can be explained with changes of weather conditions or recovery from power plant outages.

Price dependent volatilities Empirically, there is a strong relationship between price levels and volatility levels.

Long-term nonstationarity Due to the increasing costs of other commodities and uncertainty about factors in the long-term future.

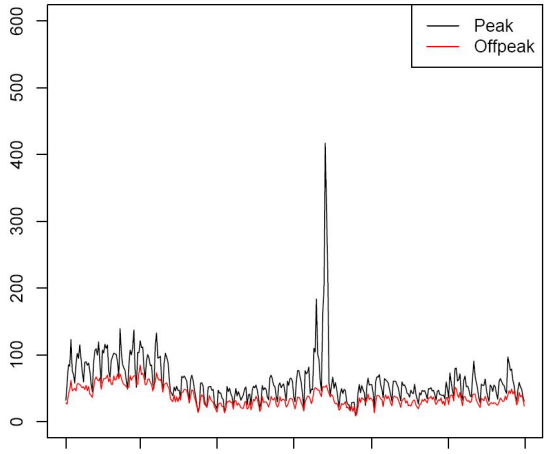


Figure 1. 2006 Peak and Off-peak prices

3 A short review of econometric models for spot electricity prices

In the last few years, there has been a quickly growing literature on stochastic models for electricity prices. Indeed, the traditional models used in financial markets are not appropriate due to the special characteristics of electricity prices. In this section, we will give a short review of some of the parametric models considered so far in the econometric literature. A detailed review of the following summary can be found in [3].

The first modelling step was to adjust some of the traditional models from financial assets to the characteristics of electricity. To address the spiky feature, [13] proposed a random walk jump-diffusion model, adopted from [14]. However, this specification does not address another characteristic of electricity prices: the mean-reversion. This was done in [12]. This "financial asset' models deal with important characteristics of price dynamics, namely mean-reversion and spikes, but it still assumes deterministic price volatility. [8] extended this class of models by adding nonlinearities in the price dynamics, such as regime-switching and stochastic volatility. [9] generalized previous approaches by a seasonal GARCH. To dissociate jumps from large spikes, regime-switching was proposed as an alternative to jump-diffusion. ([11]). In [4], AR-GARCH price process with a seasonal component in volatility is combined with tools of extreme value theory. Eventually, other models [5, 6] aim to forecast electricity prices.

Following [1, 2], we introduce a nonparametric approach, to address the complex features of spot electricity prices described in section 2. Moreover, it is our belief that this can be easily extended to any other electricity market, thanks to the flexibility of the nonparametric modeling.

4 Description of the model

In this section, we describe our model. It is a general model for an electricity spot market that addresses the main electricity prices characteristics. In this paper, we calibrate it to data of the Powernext Energy Exchange PWX, but this can be easily adapted to other market. Recall we describe the spot market price $(P_t)_t$ by a two-dimension discrete time stochastic process with days as time units: $(P_t)_t = \begin{pmatrix} P_t^{\text{peak}} \\ P_t^{\text{offpeak}} \end{pmatrix}_t$.

We introduce three factors: the long term process Z_t, the temperature process θ_t, the short term process ε_t and the following additional quantities (logarithmic) price temperature curves $f(t,.)$ and the variance temperature curve $\sigma(.)$.

The price equation can be written as

$$P_t = \exp\left(Z_t + f(t,\theta_t) + \sigma(\theta_t)\varepsilon_t\right)$$

Z_t can be interpreted as the nonstationary economic trend, in link with the non-decreasing price trend of other commodities such as fuels and gas.

The process θ_t describes the temperature. The functions $f : \mathbb{Z}\times\mathbb{R} \to \mathbb{R}$ and $\sigma : \mathbb{R} \to \mathbb{R}_+$ depend on the actual time t and on the temperature θ_t. It describes the non-linear relation between price and temperature. The dependence in time t is explained by the calendar seasonality of load.

Eventually, ε_t describes the residual short term market fluctuations.

This section is divided in two parts. The first one is dedicated to pre-processing. The pursued goal here is to extract the residuals ε_t. In the second part, we describe our Model and apply it to the extracted residuals.

4.1 Pre-processing

We describe here briefly the methodology of the pre-processing. With the previous notations, the pre-processing algorithm is the following:

- Take the logarithm to preserve the nonnegativity of prices, and also the order between peak and off peak prices.

$$p_t = \begin{pmatrix} \log\left(P_t^{\text{peak}} - P_t^{\text{offpeak}}\right) \\ \log P_t^{\text{offpeak}} \end{pmatrix}_t$$

- evaluate $Z_{t,i}$ with a smooth spline of $p_{t,i}$ on time t.

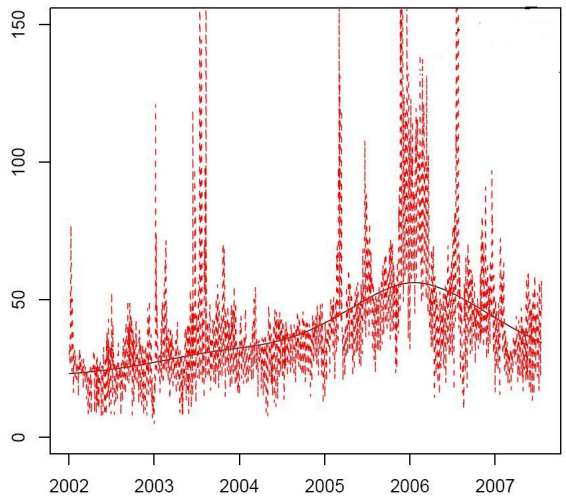


Figure 2. Peak prices (red) and their trend (black)

- evaluate $f(t,\theta_t)$. First, compute the weekly multiplicative seasonality by averaging the stationarized return over the weekdays: $S_d^w \simeq \frac{7}{T}\sum_{t:d(t)=d}\Delta(p_t - Z_t)$ where $d(t)$ is the weekday of time t and Δ is the finite difference operator. To complete the calendar effects, it is sufficient to regress $\Delta(p_t - Z_t) - S^w_{d(t)}$ on indicators of public days and holidays. We thus obtain the calendar effect at time t denoted S_t.

  Secondly, regress the seasonally adjusted and stationarized price $p_t - Z_t - S_t$ on temperature-based explicative variables: $\{h(\theta_t), \Delta h(\theta_t)\}$ where h is a smooth spline of $p_t - Z_t - S_t$ on temperature θ_t. At the end, an estimate of $f(t,\theta_t)$ is $S_t + \widehat{\alpha}_1 h(\theta_t) + \widehat{\alpha}_2 \Delta h(\theta_t)$.

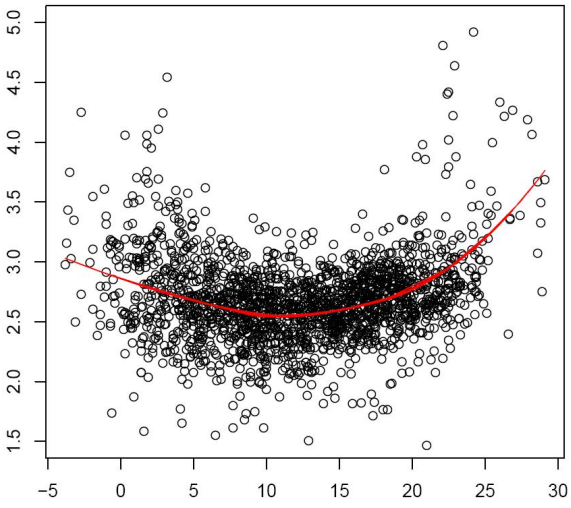


Figure 3. logarithm of the peak prices versus temperature, in red the smooth spline

- evaluate $\sigma^2(\theta_t)$ with a smooth spline of $(p_t - Z_t - f(t,\theta_t))^2$ on temperature θ_t (cf. figure 3).

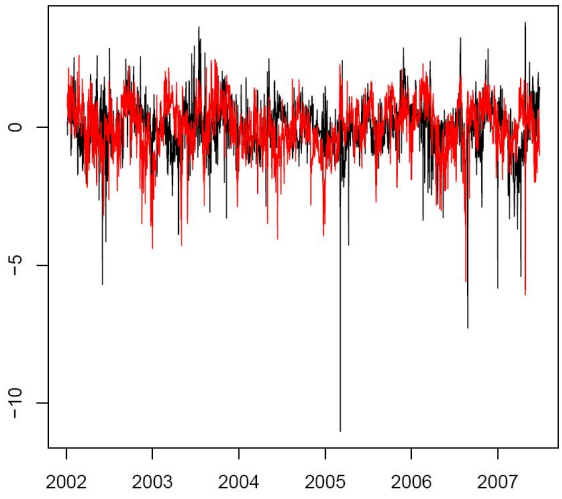


Figure 4. Extracted residuals, offpeak (red) and peak (black)

- eventually, we obtain the extracted residuals (cf. figure 4) $\varepsilon_t = \frac{p_t - Z_t - f(t,\theta_t)}{\sigma(t,\theta_t)}$. Recall that these residuals are stationary, seasonally adjusted and independent of temperature.

4.2 Pseudo-Regenerative Blocks

We recall here an algorithm introduced by [1, 2] that takes into account the dependence in the residual time-series ε_t. This method uses the regeneration properties of ε_t to cut the time-series into almost independent blocks. Simulation are then obtained by Bootstrap on the blocks.

The main assumption is to consider the time-series as a Markov chain of order 1. This means that the future states depend on the past states only through the present state:

$$(\varepsilon_{t+1}|\varepsilon_t, \ldots, \varepsilon_1) = (\varepsilon_{t+1}|\varepsilon_t) = \Pi(\varepsilon_t, \varepsilon_{t+1}).$$

This assumption differs from the AR(1) model in that we do not suppose that the dependence is linear. The values of ε_t being not discrete but real, the Markov model is in fact very general. The restriction to the order 1 is based on the spot market functioning, where prices are defined for the next day and on the non storability of electricity. Therefore, all the available information from the past is incorporated in these prices.

To take into account the Markovian structure of the residuals, we cut the chain in blocks as independent as possible. This can be done by the splitting technique introduced in [16] under very weak assumptions. The idea is to extend the original chain to a "virtual" chain with an atom: when the chain hits the atom, present and past are totally independent and the chain can be cut. Such a hitting time

is called a regeneration time. Before we recall the splitting technique, the notion of *small set* must be introduced. Let's consider a Markov chain ε_t with transition density p. A set S is a small set if there exist $\delta > 0$ and a density ϕ supported by S such that, for all $(x, a) \in S^2$,

$$p(x, a) \geq \delta\phi(a). \tag{1}$$

The idea to construct the split chain (ε, W) is the following:

- if $\varepsilon_t \notin S$, generate W_t as a Bernoulli random value $Ber(\delta)$.
- if $\varepsilon_t \in S$, generate W_t as a Bernoulli random value $Ber\left(\delta\phi(\varepsilon_{t+1})/p(\varepsilon_t, \varepsilon_{t+1})\right)$.

The construction principle is that, under the minimization condition (1), $p(x, a)$ can be seen on S as a mixture: $p(x, a) = (1-\delta)(p(x, a) - \delta\phi(a))/(1-\delta) + \delta\phi(a)$, which is constant (independent of x) when the second component is picked (see [15] and [1] for details).

This construction ensures that the split chain (ε, W) is atomic, and preserves the marginal distribution of ε (see [15]). The atom is then $A = S \times \{1\}$.

An assumption on the regenerative properties of the chain is needed: the expectation of the return time to S is finite.

H0 : $\sup_{x \in S} \mathbb{E}_x[\tau_S] < \infty$

Unfortunately, the Nummelin technique involves the transition density p of the chain, which must therefore be estimated. An estimator p_n can be given by standard kernel methods. Some conditions, satisfied by the usual kernel estimators, must be fulfilled.

H1 For a sequence $(\alpha_n)_{n \in \mathbb{N}}$ decreasing to 0 as $n \to \infty$, p is estimated by p_n at the rate α_n for the mean square error when error is measured by the L^∞ loss over $S \times S$:

$$\mathbb{E}\left[\sup_{(x,y) \in S \times S} |p_n(x, y) - p(x, y)|^2\right] = \mathcal{O}(\alpha_n).$$

H2 The minorizing density ϕ is such that $\inf_{x \in S} \phi(x) > 0$.

H3 The densities p and p_n are bounded over S^2 and $\inf_{x,y \in S} p_n(x, y)/\phi(y) > 0$.

Since ϕ is set by the statistician, we can use for instance the uniform distribution over S. This ensures **H2**. Similarly, it is not difficult to construct an estimator p_n satisfying the constraints of **H3**.

Now we detail the simulation algorithm.

Pseudo-Regenerative Blocks Construction

1. Find an estimator p_n of the transition density (for instance a Nadaraya-Watson estimator).
2. Choose a small set S and a density ϕ on S and evaluate $\delta = \min_{x,y \in S} \left\{ \frac{p_n(x,y)}{\phi(y)} \right\}$.
3. When ε hits S, generate W_t as a Bernoulli with probability $\delta\phi(\varepsilon_{t+1})/p_n(\varepsilon_t, \varepsilon_{t+1})$. If $W_t = 1$, the approximate split chain (ε_t, W_t) hits the atom $A = S \times \{1\}$ and i is an approximate regenerative time. These times define the approximate regenerative times $\hat{\tau}(j)$.
4. Count the number of visits to A up to time n: $l_n + 1 = \sum_{i=1}^{n} \mathbb{1}_{(\varepsilon_i, W_i) \in A}$.
5. Divide the observed trajectory $(\varepsilon_1, ..., \varepsilon_n)$ into $l_n + 2$ blocks corresponding to the pieces of the sample path between approximate return times to the atom A,

$$\begin{aligned} B_0 &= (\varepsilon_1, ..., \ \varepsilon_{\hat{\tau}(1)}), \\ B_1 &= (\varepsilon_{\hat{\tau}(1)+1}, ..., \ \varepsilon_{\hat{\tau}(2)}), \\ &..., \\ B_{l_n} &= (\varepsilon_{\hat{\tau}(l_n)+1}, ..., \ \varepsilon_{\hat{\tau}(l_n+1)}), \\ B^{(n)}_{l_n+1} &= (\varepsilon_{\hat{\tau}(l_n+1)+1}, ..., \ \varepsilon_n), \end{aligned}$$

   with the convention $B^{(n)}_{l_n+1} = \emptyset$ when $\hat{\tau}(l_n + 1) = n$.
6. Drop the first block B_0, and the last one $B^{(n)}_{l_n+1}$ (possibly empty when $\hat{\tau}(l_n + 1) = n$).
7. Repeat steps 3 to 6 in order to obtain more pseudo-regenerative blocks: the stochastic generation of the Y_t's will lead to different $\hat{\tau}(j)$ and then to different blocks.
8. To simulate a time-series of length L, choose with equal probability among the blocks, with replacement and abut the chosen blocks. Once the total length is larger than L, cut the resulting time-series at L and drop the remaining part of the last block.

The choice of the small set S is crucial to obtain a number of block l_n large enough. Hopefully, it is a much easier task to choose S than a bandwidth in kernel methods. The dynamic of l_n as S varies is the following; l_n increases when S is translated to a region often visited; the size of S has a balanced effect on l_n: if S increases, the number of hitting times increases, but δ decreases and therefore the rate of approximate regenerative times decreases also. S must therefore be centered on the most visited region and it size can be established a posteriori by maximizing l_n.

Figure 5 gives an example of a residual time-series and the corresponding pseudo regenerative times.

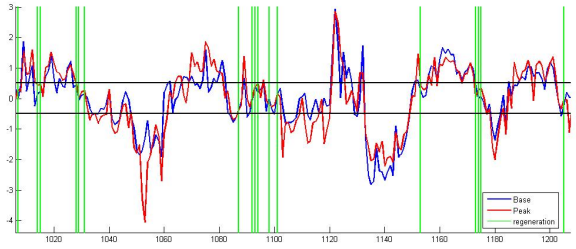


Figure 5. A trajectory of Peak and Base (Off-peak) prices and the pseudo regenerative times (vertical lines).

Figure 6 shows a simulation based on blocks obtain at Figure 5.

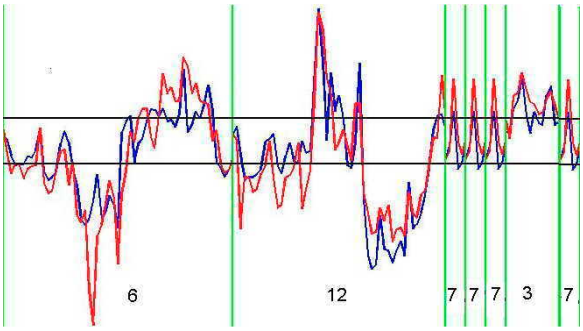


Figure 6. A simulation based on the previous blocks.

5 Simulations

In this part, we illustrate how the Model described previously can be used to simulate spot prices.

5.1 Simulation methodology

Suppose we are given scenarios of temperature $(\widehat{\theta}_t)_t$. The pursued goal here is to generate spot price scenarios consistent with the temperature simulations. For this, we apply the following simulation strategy:

1. choose an economic trend $\widehat{Z}_t$. This can be done either by considering economic projections or by using the prices of futures.

2. simulate the influence of temperature and calendar effects by $f(t, \widehat{\theta}_t)$ and $\sigma(t, \widehat{\theta}_t)$.

3. simulate new time-series of residuals by Bootstrap $(\widehat{\varepsilon}_t)_t$ according to the methodology described at the end of section 4.

4. at the end, put all together to simulate the spot price: $\widehat{P}_t = \exp(\widehat{Z}_t + f(t, \widehat{\theta}_t) + \sigma(t, \widehat{\theta}_t)\widehat{\varepsilon}_t)$.

5.2 Simulation examples

We illustrate our approach with the following example: we consider the year 2006 with the same economic trend but with the 2005 temperature. The figures (7, 8) give two different simulations of spot prices for the year 2006, all having the typical spot price features.

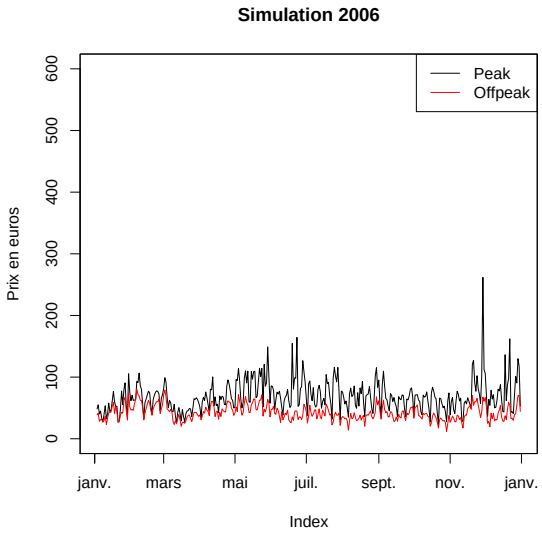


Figure 7. A simulation of year 2006 with 2005 temperature.

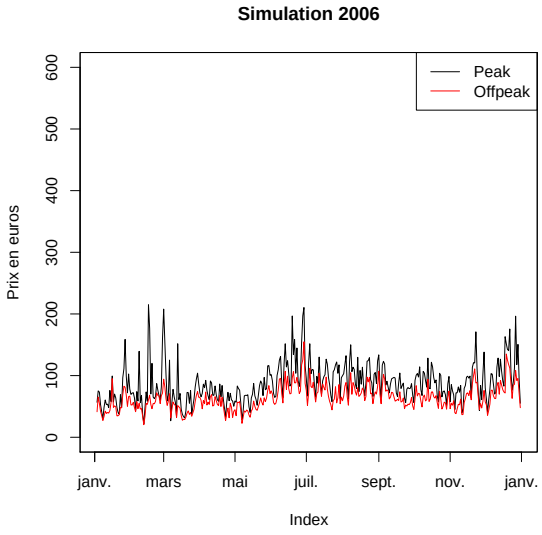


Figure 8. Another simulation of year 2006 with 2005 temperature.

Furthermore, if we compare the peak prices empirical density with the empirical density of the simulations, we can see the tails are quite similar (figures 9, 10)

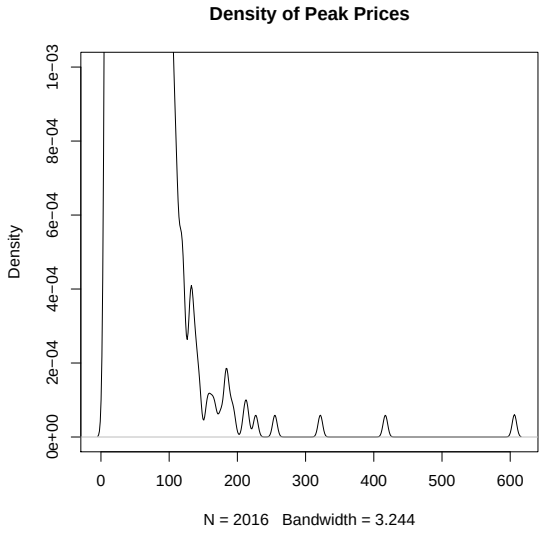


Figure 9. Empirical density of peak electricity prices.

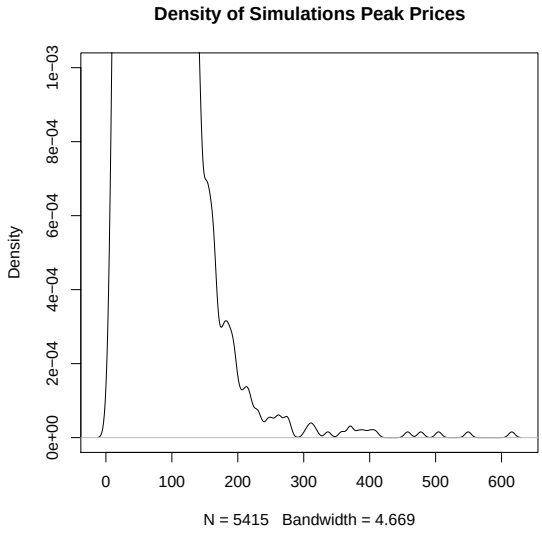


Figure 10. Empirical density of Reba simulations for peak prices.

6 Conclusion

In this article, we exhibit a non parametric approach for time series inspired by the bootstrap literature. Combined with pre-processing, this model addresses the main complex features of Powernext spot electricity prices. Moreover, we argue it could be applied easily to other electricity markets.

References

[1] P. Bertail and S. Clémençon. Regeneration-based statistics for Harris recurrent Markov chains. In P. Bertail, P. Doukhan, and P. Soulier, editors, *Dependence in Probability and Statistics*, volume 187 of *Lecture Notes in Statistics*. Springer, 2006.

[2] P. Bertail and S. Clémençon. Second-order properties of regeneration-based bootstrap for atomic markov chains. *TEST*, 16(1):109–122, 2007.

[3] D. Bunn and N. Karakatsani. Forecasting electricity prices. 2003. Working paper.

[4] H. Bystrom. Extreme value theory and extremely large electricity price changes. 2002. Working paper.

[5] A. J. Conejo, M. A. Plazas, R. Espinola, and A. B. Molina. Day-ahead electricity price forecasting using the wavelet transform and arima models. *IEEE Transactions on Power Systems*, 202:1035–1042, 2005.

[6] J. C. Cuaresma, J. Hlouskova, A. Kossmeier, and M. Obersteiner. Forecasting electricity spot-prices using linear univariate time-series models. *Applied Energy*, 77:87–106, 2004.

[7] S. Deng. Stochastic models of energy commodity prices and their applications: Mean reversion with jumps and spikes. *The U.S. Power Market*, 2000. Working paper, PWP-073.

[8] A. Escribano, J. L. Peña, and P. Villaplana. Modelling electricity prices: International evidence. *The U.S. Power Market*, 2002. Working paper.

[9] R. Huisman and R. Mahieu. Regime jumps in electricity prices. *Energy Economics*, 25(5):425–434, 2003.

[10] B. Johnson and G. Barz. Selecting stochastic processes for modelling electricity prices. *Energy Modelling and the Management of Uncertainty, Risk Publications*, 1999.

[11] V. Kaminski. The challenge of pricing and risk managing electricity derivatives. *The U.S. Power Market*, 3:149–71, 1997.

[12] R. Merton. Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 3:125–144, 1976.

[13] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer, 1996.

[14] E. Nummelin. A splitting technique for Harris recurrent chains. *Z. Wahrsch. Verw. Gebiete*, 43:309–318, 1978.

Bibliography

[AL68] D. M. Allen, The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16, 125-127,1968.

[ARL08] S. Arlot. *Rééchantillonnage et sélection de modèles.* Thèse de doctorat, Université Paris-Sud XI, Décembre 2007.

[ARL07] S. Arlot, Model selection by resampling penalization. En révision pour *Electronic Journal of Statistics* 2008.

[ATE92] M. Atteia. Hilbertian kernels and spline functions. North-Holland, 1992.

[BB95] P. Barbe and P. Bertail. The weighted bootstrap, volume 98 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1995.

[BEN04] Y. Bengio and Y. Grandvalet. No Unbiased Estimator of the Variance of K-Fold Cross-Validation. *Journal of Machine Learning Research* 5, 1089-1105, 2004.

[BC04a] P. Bertail and S. Clémençon, Edgeworth expansions for suitably normalized sample mean statistics of atomic Markov chains. *Prob. Th. Rel. Fields* 130, 388-414, 2004.

[BC05] P. Bertail and S. Clémençon,Regeneration-based statistics for Harris recurrent Markov chains. Dependence in Probability and Statistics. Lecture Notes in Statistics. Bertail, P. and Doukhan, P. and Soulier, P. Springer, 2005.

[BC07] P. Bertail and S. Clémençon, Second-order properties of regeneration-based bootstrap for atomic Markov chains. *TEST* 16, 1,109-122, 2007.

[BE03] Bouton F. and Erkel-Rousse H. , Conjonctures sectorielles et prévision à court terme de l'activité : l'apport de l'enquête de conjoncture dans les services . *Economie et Statistique*, numéro spécial " Analyse conjoncturelle : entre statistique et économie ", nř359-360 - 2002, April 2003.

[BIA08] G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers, *Journal of Machine Learning Research*, Vol. 9, pp. 2015–2033,2008.

[BIS05] M. Markatou, H. Tian, S. Biswas, G. Hripcsak. Analysis of Variance of Cross-Validation Estimators of the Generalization Error. *Journal of Machine Learning Research* 1127-1168, 2005.

[BREI84] L. Breiman, J. H. Friedman, R. Olshen, and C. J. Stone. Classification and regression trees. *The Wadsworth statistics probability series*. Wadsworth International Group, 1984.

[BREI92] L. Breiman, and Spector, P., Submodel selection and evaluation in regression: The X-random case *International Statistical Review*, 60, 291-319,1992.

[BRE96] L. Breiman. Bagging predictors. *Machine Learning*, 24:2, 123–140,1996.

[BRE98] L. Breiman. Arcing classifiers. *The Annals of Statistics*, 26:3, 801–849,1998.

[BKL99] A. Blum, A., Kalai, A., and Langford, J.. Beating the hold-out: Bounds for k-fold and progressive cross-validation. *Proceedings of the International Conference on Computational Learning Theory*,1999.

[BE01] O. Bousquet and A. Elisseef. Algorithmic stability and generalization performance *In Advances in Neural Information Processing Systems* 13: Proc. NIPS'2000, 2001.

[BE02] O. Bousquet and A. Elisseef. Stability and generalization. *Journal of Machine Learning Research*, 2002.

[Bierbrauer2005] M. Bierbrauer and S. Trueck and R. Weron. Modeling electricity prices with regime switching models. http://ideas.repec.org/p/wpa/wuwpem/0502005.html. Feb, 2005.

[BUH00] P. Buhlmann, and B. Yu. Explaining Bagging. , Vol. 30, No. 4 (Aug., 2002), pp. 927-961 *The Annals of Statistics*,2002.

[BUJ00] A. Buja and W. Stuetzle. The effect of bagging on variance, bias and mean squared error. *Technical Report, AT&T Labs-Research*,2000.

[BUJ02] A. Buja and W. Stuetzle, "Observations on Bagging", University of Pennsylvania and University of Washington, Seattle,2002.

[BTW07] F. Bunea, A.B. Tsybakov and M.H. Wegkamp, M. H. Sparsity oracle inequalities for the Lasso. *Electron. J. Statist.*, 1 169?194, 2007.

[Bunn2003] D.W. Bunn and N. Karakatsani, Forecasting Electricity Prices. Working Paper, London Business School, 2003.

[Burger2004] M. Burger, B. Klar, A. Müller and G. Schindlmayr. A spot market model for pricing derivatives in electricity markets. *Quantitative Finance*. 4, 109-122, 2004.

[BUR89] P. Burman. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76:503– 514, 1989.

[Bystrom2001] H. Bystrom. Extreme value theory and extremely large electricity price changes. Working paper, Lund University, 2001.

[COD06] M. Cornec and T.Deperraz. Un nouvel indicateur synthétique mensuel résumant le climat des affaires dans les services en France. *Economie et Statistique*, 395-396, 13-38, 2006.

[Conejo2005] A.J. Conejo, M.A. Plazas, and R. Espinola and A.B. Molina. Day-ahead electricity price forecasting using the wavelet transform and ARIMA models. *IEEE Transactions on Power Systems*,202, 1035-1042, 2005.

[COR06] M. Cornec, Analyse factorielle dynamique multi-fréquence appliquée à la datation de la conjoncture française *Economie et Prévision*, nN172, pp. 29-43, 2006.

[COR09A] M.Cornec. Concentration inequalities of the cross-validation estimator for Empirical Risk Minimiser. Technical Report, 2009.

[CHK08] M.Cornec and H. Harari Kermadec. Simulating spot electricity prices with regenerative blocks. *Proceedings of IASTED ASM* , 2008.

[COR09B] M.Cornec. Concentration inequalities of the cross-validation estimate for stable predictors. Technical Report, 2009.

[Cuaresma2004] J.C. Cuaresma and J.Hlouskova and A. Kossmeier and M. Obersteiner, Forecasting electricity spot-prices using linear univariate time-series models. *Applied Energy*,77, 87-106, 2004.

[deJong2002] C. De Jong and R. Huisman. Option Formulas for Mean-Reverting Power Prices With Spikes. Working paper, Erasmus University, 2002.

[Deng2000] S. Deng. Stochastic Models of Energy Commodity Prices and their applications: Mean Reversion with Jumps and Spikes. *The U.S. Power Market*. Working paper, PWP-073. University of California Energy Institute, 2000.

[DGL96] L. Devroye, L. Gyorfi, and G. Lugosi. A Probabilistic Theory of Pattern Recognition. Number 31 in *Applications of Mathematics*. Springer, 1996.

[DW79] L. Devroye and T. Wagner. Distribution-free performance bounds for potential function rules. *IEEE Trans. Inform. Theory*, 25(5):601 604, 1979.

[DEWA79] L. P. Devroye and T. J. Wagner. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Transactions on Information Theory*, IT-25(2):202?207, 1979.

[DE96] T. J. DiCiccio and B. Efron. Bootstrap confidence intervals. *Statist. Sci.,* 11(3):189:228,1996.

[DL95] C. Doz et F. Lenglart.Une grille de lecture pour l'enquête mensuelle de l'industrie. *Note de conjoncture de décembre* , Insee, 1995.

[D0M97] P.Domingos. Why does bagging work? A Bayesian account and its implications. *In Proceedings of the Third International Conference on Knowledge Discovery and Data Mining* (pp. 155–158), AAAI Press, 1997.

[DUD03] S. Dudoit and M. J. van der Laan. Asymptotics of cross-validated risk estimation in model selection and performance assessment. *Statistical Methodology* , 2(2), 131Ű154., 2005.

[DUD04] S. Dudoit, M. J. van der Laan, S. Keles, A. M. Molinaro, S. E. Sinisi, and S. L. Teng. Loss-based estimation with cross-validation: Applications to microarray data analysis. *SIGKDD Explorations, Microarray Data Mining Special Issue*, 2004.

[DUD04BIS] M.J. van der Laan, S. Dudoit, A. van der Vaart ,The cross-validated adaptive epsilon-net estimator in *Statist. Decisions* ,24 373–395,2006.

[ELI05] A. Elisseeff, T. Evgeniou, & M. Pontil. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6: 55-79,2005.

[ELI04] T. Evgeniou, M. Pontil, and A. Elisseeff, "Leave One Out Error, Stability, and Generalization of Voting Combinations of Classifiers", *Machine Learning*, 55(1), 71-97,2004.

[EFR79] B.Efron. Bootstrap methods : another look at the jackknife. *Ann. Statist.,* 7(1):1-26,1979.

[EFR82] B.Efron. The jackknife, the bootstrap and other resampling plans, volume 38 of *CBMS-NSF Regional Conference Series in Applied Mathematics.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa.,1982.

[EFR93] B. Efron, R.J. & Tibshirani. An introduction to the bootstrap, Vol. 57 of *monographs on statistics and applied probability.* Chapman & Hall,1993.

[Escribano2002] A. Escribano and J.L. Peña and P. Villaplana, Modelling Electricity Prices: International Evidence. *The U.S. Power Market* Working paper, 2002.

[Ethier1998] R. Ethier and T. Mount. Estimating The Volatility of Spot Prices In Restructured Electricity Markets And The Implications For Option Values. *Working paper* Cornell University, 1998.

[FRI00] J.H. Friedman, & P. Hall. On bagging and non-linear estimation. *TJ. Statist. Plann. Inference* ,137 669Ű683,2007.

[FRE95] Y. Freund and R. Schapire. A decision-theoretic generalization of on-linelearning and an application to boosting. In *Proc. of the Second European Conference on Computational Learning Theory*. LNCS, March 1995.

[FRO03] M. Fromont. *Quelques problèmes de sélection de modèles : construction de tests adaptifs, ajustement de pénalités par des méthodes de bootstrap.* Thèse de doctorat, Université Paris-Sud XI, December 2003.

[GE77] J. Geweke. The dynamic factor analysis of economic time series in D.J. Aigner and A.S. Goldberger (eds.), "Latent Variables in Socio-Economic Models", pp. 365-383, Amsterdam: North-Holland, 1977.

[GEI75] S. Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70:320–328, 1975.

[GRA04] Y. Grandvalet. Bagging Equalizes Influence. *Machine Learning*. 55,3: 251 – 270.,2004.

[GYO02] L. Gyorfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer-Verlag, New York, 2002.

[HA91] J.D. Hamilton. "Time series Analysis", Princeton University Press, 1991.

[HAL92] P.Hall. The boostrap and Edgeworth expansion. Springer Series in Statistics. Springer-Verlag, New York, 1992.

[HTF01] T. Hastie, R. Tibshirani, and J. H. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, 2001.

[HOEF63] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58, 13-30,1963.

[HOL96] S. B. Holden. Cross-validation and the PAC learning model. *Research Note* RN/96/64, Dept. of CS, Univ. College, London, 1996.

[HOL96bis] S. B. Holden. PAC-like upper bounds for the sample complexity of leave-one-out cross validation. *In Proceedings of the Ninth Annual ACM Workshop on Computational Learning Theory*, pages 41 50, 1996.

[Huisman2001] R. Huisman and R. Mahieu. Regime jumps in electricity prices. *Energy Economics* 25,5, 425-434, 2001.

[Johnson1999] Johnson and Barz, Selecting Stochastic Processes For Modelling Electricity Prices. *Energy Modelling and the Management of Uncertainty, Risk Publications*, 1999.

[Kaminski1997] V., Kaminski. The Challenge of Pricing And Risk Managing Electricity Derivatives. *The U.S. Power Market*. 3, 149-71, 1997.

[KN89] C. Kim and C.Nelson,State-Space Models with Regime Switching: Classical and Gibbs-Sampling Approaches with Applications,MIT Press, Cambridge, MA, 1989.

[KR99] M. Kearns and D. Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11:1427 1453, 1999.

[KEA95] M. Kearns. A bound on the error of cross validation, with consequences for the training-test split. *In Advances in Neural Information Processing Systems 8*. The MIT Press, 1995

[KMNR95] M. J. Kearns, Y. Mansour, A. Ng,, and D. Ron. An experimental and theoretical comparison of model selection methods. *In Proceedings of the Eighth Annual ACM Workshop on Computational Learning Theory*, 1995.

[KUT02] S. Kutin. Extensions to McDiarmid's inequality when differences are bounded with high probability. *Technical report*, Department of Computer Science, The University of Chicago, 2002. In preparation.

[KUNIY02] S. Kutin and P. Niyogi. Almost-everywhere algorithmic stability and generalization error, 2002. *Technical report* TR-2002-03, University of Chicago.

[KUNIY01] S. Kutin and P. Niyogi. The interaction of stability and weakness in AdaBoost. *Technical Report* TR-2001-30, Department of Computer Science, The University of Chicago, 2001.

[LM68] P. A. Lachenbruch,; M. Mickey, Estimation of error rates in discriminant analysis. *TechnometricsLM68* Estimation of error rates in discriminant analysis. *Technometrics* 10, 1-11, 1968.

[Li87] K-C Li. Asymptotic optimality for cp, cl, cross-validation and generalized cross-validation: Discrete index set. *Annals of Statistics*, 15:958–975, 1987.

[Lucia2002] Escribano, Lucia, J. J. and Schwartz, E., Electricity Prices And Power Derivatives: Evidence From The Nordic Power Exchange. *Review of Derivatives Research* 5, 5-50, 2002.

[Lug03] G Lugosi. Concentration-of-measure inequalities presented at *the Machine Learning Summer School 2003*, Australian National University, Canberra, 2003.

[MAS03] P. Massart. Concentration Inequalities and Model Selection. Ecole d'été de Probabilités de Saint-Flour 2003, Lecture Notes in Mathematics, Springer, 2007.

[McC76] P. J. McCarthy. The use of balanced half-sample replication in crossvalidation studies. *Journal of the American Statistical Association*, 71: 596–604, 1976.

[McD89] C. McDiarmid. On the method of bounded differences. *In Surveys in combinatorics*, 1989 (Norwich, 1989), pages 148 188. Cambridge Univ. Press, Cambridge, 1989.

[McD98] C. McDiarmid. Concentration. In Probabilistic Methods for Algorithmic Discrete Mathematics, pages 195 248. Springer, Berlin, 1998.

[MM03] M. Mariano and Y. Murasawa, A New Coincident Index of Business Cycles Based on Monthly and Quarterly Series. *Journal of Applied Econometrics* , vol. 18, pp. 427-443, 2003.

[MN92] D. M. Mason and M. A. Newton. A rank statistics approach to the consistency of a general bootstrap. *Ann. Statist.,* 20(3):1611-1624,1992.

[Merton1976] R.C. Merton. Option Pricing When Underlying Stock Returns are Discontinuous. *Journal of Financial Economics.* 3, 125-144, 1978.

[MT96] S. P. Meyn and R. L. Tweedie. Markov Chains and Stochastic Stability. Springer, 1996.

[NUM78] E. Nummelin. A splitting technique for Harris recurrent chains. *Z. Wahrsch. Verw. Gebiete*, 43:309Ű318, 1978.

[PET07] M.L. Petersen, A. Molinaro, S.E. Sinisi, M.J. van der Laan (2007) Cross-validated Bagged Learning. *J. Multiv. Analysis*: 98 (9): 1693-1704,2007.

[PIC84] R. R. Picard and R. D. Cook. Cross-validation of regression models. *Journal of the American Statistical Association*, 79:575–583, 1984.

[PW93] J. Praestgaard and J A. Wellner. Exchangeably weigthed boostraps of the general empirical process. *Ann. Prob.,* 21(4):2053:2086,1993.

[QUE49] M H. Quenouille. Approximate tests of correlation in time-series. *J.Roy. Statist. Soc. Ser. B.,* 11:68-84,1949.

[RIP96] B. D. Ripley. Pattern recognition and neural networks. *Cambridge University Press*, Cambridge, New York, 1996.

[ROG63] C. Rogers. Covering a sphere with spheres. *Mathematika*, vol. 10,pp. 157-164, 1963.

[SCH98] R. Schapire, Y. Freund, P.Bartlett, & W.S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26:5, 1651–1686?1998.

[SHAO93] J. Shao. Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88:486–494, 1993.

[SS77] T.J. Sargent et C.A. Sims, Business cycle modelling without pretending to have too much a priori economic theory. *New Methods in Business Cycle Research"*, pp. 45-109, Minneapolis: Federal Reserve Bank of Minneapolis. 1977.

[ST95] J. Shao and D. S. Tu. The jackknife and boostrap. Springer Series in Statistics. Springer-Verlag, New York, 1992.

[STO74] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B*, 36, 111,147,1974.

[STO77] M. Stone.Asymptotics for and against cross-validation. *Biometrika*, 64, 29-35,1977.

[SW89] J.H. Stock et M.W. Watson , New indexes of coincident and leading economic indicators. *NBER Macroeconomics Annual*,MIT Press, Cambridge, 351- 394. 1989.

[TAN97] Taniguchi, M., & Tresp, V.. Averaging regularized estimators. *Neural Computation*, 9:7, 1163–1178,1997.

[TUK58] J. Tuley. Bias and confidence in not-quite large samples. *Ann. Math. Statist.,* 29:614,1958.

[VAL84)] L.G. Valiant. A theory of learnable. *Proc. of the 1984, STOC*, pages 436-445,1984.

[Vaart96] A. W. van der Vaart and J. Wellner. Weak Convergence and Empirical *Processes*. Springer-Verlag, New York, 1996.

[VA71] V. Vapnik, and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16, 264-280,1971.

[VC71] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280,1971.

[VA82] V. Vapnik. Estimation of Dependences Based on Empirical Data. Springer-Verlag,1982.

[Vap95] V. Vapnik. The Nature of Statistical Learning Theory. Springer, 1995.

[Vap98] V. Vapnik. Statistical learning theory. John Wiley and Sons Inc., New York. A Wiley-Interscience Publication,1998.

[WU86] JC-F J. Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.,* 14(4):1261:1350,1986.

[YAN07] Y. Yang, Consistency of Cross Validation for Comparing Regression Procedures. *Annals of Statistics*, Volume 35, Number 6, 2450-2473,2007.

[ZHA93] P. Zhang. Model selection via multifold cross-validation. *Annals of Statistics*, 21:299–313, 1993.

[ZHA00] T Zhang . A leave-one-out cross validation bound for kernel methods with applications in learning. *14th Annual Conference on Computational Learning Theory*, 2001 - Springer.