Scattering Representations for Recognition

Joan Bruna

Scattering Representations for Recognition

Joan Bruna Estrach
Ecole Polytechnique
Palaiseau, France

PhD thesis

Jury
Stéphanie Allassonnière (Examiner)
Emmanuel Bacry (Examiner)
Stéphane Jaffard (Reviewer)
Stéphane Mallat (Adviser)
Yves Meyer (President)
Alain Trouvé (Reviewer)

Submitted November 2012
Defended February 2013

Abstract

This thesis addresses the problem of pattern and texture recognition from a mathematical perspective. These high level tasks require signal representations enjoying specific invariance, stability and consistency properties, which are not satisfied by linear representations.

Scattering operators cascade wavelet decompositions and complex modulus, followed by a lowpass filtering. They define a non-linear representation which is locally translation invariant and Lipschitz continuous to the action of diffeomorphisms. They also define a texture representation capturing high order moments and which can be consistently estimated from few realizations.

The thesis derives new mathematical properties of scattering representations and demonstrates its efficiency on pattern and texture recognition tasks. Thanks to its Lipschitz continuity to the action of diffeomorphisms, small deformations of the signal are linearized, which can be exploited in applications with a generative affine classifier yielding state-of-the-art results on handwritten digit classification. Expected scattering representations are applied on image and auditory texture datasets, showing their capacity to capture high order moments information with consistent estimators. Scattering representations are particularly efficient for the estimation and characterization of fractal parameters. A renormalization of scattering coefficients is introduced, giving a new insight on fractal description, with the ability in particular to characterize multifractal intermittency using consistent estimators.

Acknowledgements

Stéphane, j'imagine que tu as déjà lu quelques fois cette page auparavant, et j'espère pour le bien de la science que plein d'autres vont suivre. Mais pour moi c'est l'opportunité de témoigner à quel point j'ai été ravi de pouvoir travailler avec toi pendant toutes ces années. Ton talent hors norme m'a impressionné en tant que mathématicien et scientifique, mais je me souviendrai toujours aussi de cette vitalité, passion et envie de partager, qui m'ont inspiré bien au déla de mon travail. Merci infiniment.

This thesis would not have been possible without the agreement of Zoran Inc. to grant me part-time status while being an employee, in particular thank you Dave Auld for having accepted such a generous deal. Merci à toi Christophe, d'abord pour avoir accepté de me libérer, et aussi pour ton amitié que je garde très précieusement.

Thank you Mike Glinsky, for having invited me to Perth and having tried (and failed) to teach me some Quantum Field Theory. Merci à mes collègues de thèse, Joakim, Laurent et Irène, pour avoir accepté de manger aussi souvent au libanais. Merci en particuler à Irène pour toutes les corrections et suggestions qui ont abouti à ce document.

Merci Jean-François et Emmanuel pour m'avoir introduit au monde fascinant des multifractales, je m'estime très heureux d'avoir pu profiter de votre talent et de votre gentillesse extrême. Merci aussi aux membres du jury, en particulier aux professeurs A. Trouvé et S. Jaffard pour avoir accepté d'être rapporteurs.

Thank you Mark and Jean-Baptiste for all your support and advice during the difficult periods, I didn't go to Kentucky hopefully.

Emy, tu sais à quelle point cette thèse a été importante pour moi. Mais sache qu'elle l'est vraiment peu par rapport à ce que tu m'as donné.

Per últim, gràcies pare per haver-me sempre guiat en la bona direcció i per haver-me inculcat el gust per les matemàtiques; aquesta tesi és en gran part gràcies a tu.

Als meus pares Anna Maria i Joaquim,
a les meves germanes Laia i Maria,
i a la meva estimada esposa Emy.

Contents

List of Notations

$x(u)$: Function defined on a continuous domain $u \in \mathbb{R}^d$.

$x[n]$: Discrete signal defined for $n \in \mathbb{Z}^d$.

$\delta(u)$: Dirac distribution.

$\mathbf{L}^2(\mathbb{R}^d)$: Finite energy functions $x(u)$ such that $\int |x(u)|^2 du < \infty$.

$\mathbf{L}^1(\mathbb{R}^d)$: Integrable functions $x(u)$ such that $\int |x(u)| du < \infty$.

$\mathbf{l}^2$: Finite energy discrete signals $x[n]$ such that $\sum_n |x[n]|^2 < \infty$.

$\mathbf{l}^1$: Summable discrete signals $x[n]$ such that $\sum_n |x[n]| < \infty$.

$\|x\|_p$: $\mathbf{L}^p(\mathbb{R}^d)$ norm of the function $x(u)$: $\left(\int |x(u)|^p du\right)^{1/p}$.

$\|A\|$:

- If A is an $\mathbf{L}^2(\mathbb{R}^d)$ linear operator, $\|A\| = \sup_{\|x\|=1} \|Ax\|$.
- If A is an element of $\mathbf{L}^2(\mathbb{R}^d)$, $\|A\| = \|A\|_2$.
- If $A = \{A_i\,,\, i \in \mathcal{I},\, A_i \in \mathbf{L}^2(\mathbb{R}^d)\}$, then

$$\|A\|^2 = \sum_{i \in \mathcal{I}} \|A_i\|^2 \ .$$

$|x|$: Euclidean norm of a finite-dimensional $x \in \mathbb{R}^d$.

$\hat{x}$: The Fourier transform of $x \in \mathbf{L}^1(\mathbb{R}^d) \cup \mathbf{L}^2(\mathbb{R}^d)$: $\hat{x}(\omega) = \int x(u) e^{-i\omega u} du$.

$x \star g(u)$: Convolution operator: $x \star g(u) = \int x(v) g(u-v) dv$.

$X(t)$: Stochastic process defined for $t \geq 0$.

$E(X)$: Expected Value of the random variable X.

$R_X(\tau)$: Auto-covariance of the stationary process $X(t)$.

$X \stackrel{l}{=} Y$: The random variables X and Y follow the same probability distribution.

$\{X(t)\}_t \stackrel{l}{=} \{Y(t)\}_t$: Equality in distribution between the processes $X(t)$ and $Y(t)$.

$X_n \xrightarrow{P} X$: The sequence of random variables X_n converges in probability to X: $\forall \delta > 0\,,\ \lim_{n\to\infty} \mathrm{Prob}(|X_n - X| > \delta) = 0$.

$X_n \xrightarrow{d} X$: The sequence of random variables X_n converges in distribution to X: $\lim_{n\to\infty} F_n(t) = F(t)$ at all t where $F(t)$ is continuous, where F_n and F are the cdf of X_n and X respectively.

$F(n) \simeq G(n)$ $(n \to a)$: There exists C_1, C_2 such that

$$C_1 \leq \liminf_{n\to a} \frac{F(n)}{G(n)} \leq \limsup_{n\to a} \frac{F(n)}{G(n)} \leq C_2 \ .$$

Chapter 1

Introduction

A fundamental topic in image and audio processing is to find appropriate metrics to compare images and sounds. One may construct a metric as an Euclidean distance on a signal representation Φ, applied on signals x, x':

$$d(x, x') = \|\Phi(x) - \Phi(x')\| \ .$$

This puts all the structure of the problem in the construction of a signal representation, whose role is to encode the relevant signal information and to capture the right notion of similarity for a given task.

In this thesis, we are interested in the recognition and discrimination of a variety of different objects, including sound and image patterns such as handwritten digits, and stationary processes, which model a variety of image and auditory textures, as well as multifractal measures. Linear representations, such as orthogonal wavelet decompositions, define a metric which is equivalent to the Euclidean norm. While in some problems, such as denoising or compression, this metric is well adapted to assess the similarity between an original signal and its corrupted or reconstructed version, it does not capture the correct notion of similarity in high level tasks such as pattern or texture recognition.

Objects and sounds are perceived and recognized in the human brain in a fraction of a second, under a variety of physical transformations, including translations, rotations, illumination changes or transpositions. Similarly, we are able to identify non-gaussian textures, modeled as random stationary processes, from very few realizations. This motivates the study of signal and texture representations which incorporate these symmetries.

Signals may be translated, rotated, but they can also be deformed, warped, occluded, without affecting recognition. An efficient invariant representation also needs to be stable with respect to the amount of deformation applied to a signal. As we shall see, this is a fundamental property which explains the success of popular image and audio descriptors such as SIFT [Low04] or MFCC [Mer76]. On the other hand, the texture representation is another outstanding problem, since most textures appearing in nature are realizations of non-gaussian processes, and hence are not fully characterized from their spectral densities. It is thus necessary to incorporate information from higher order statistics into the representation, but their direct estimation has large variance, which limits their efficiency to discriminate non-gaussian textures. An efficient texture representation should thus capture information from high order moments, in such a way that it can be estimated consistently from few realizations.

Scattering operators, introduced by S. Mallat in [Mal12], build locally invariant, stable and informative signal representations by cascading wavelet modulus decompositions followed by a lowpass averaging filter. They also define a representation for processes with stationary increments, capturing high moments, and which can be estimated consistently for a wide class of ergodic processes. Scattering representations have the structure of a convolutional network, a subclass of neural networks introduced by LeCun [FKL10] cascading filter banks with pooling and nonlinearities, which are learnt for each specific task using a gradient descent strategy. Rather than being learnt, the scattering network is obtained from the invariance, stability and informative requirements, which lead to wavelet filter banks and to point-wise non-linearities.

This thesis provides further insight on the properties of scattering representations, with special focus on its applications in pattern and texture recognition. In particular, we show that the first two layers of the scattering network already constitute a powerful representation, capturing stable geometric information and with the ability to discriminate non-gaussian textures.

The first part of the thesis delves into the mathematical properties of scattering operators, deriving new results characterizing their non-linearities, and relating the signal decay with the regularity of its scattering transform. Next we develop the necessary tools for image processing applications, and then concentrate on the pattern recognition problem, and show how the stability of the scattering metric can be exploited to build robust, efficient linear generative classifiers, achieving state-of-the-art results on handwritten digit classification.

The second half of the thesis is devoted to the texture representation problem. We first demonstrate that scattering texture representations are highly discriminative descriptors of non-gaussian textures with multifractal behavior, and that they can be consistently estimated from few realizations. We study the discrimination of image and auditory textures, obtaining state-of-the-art results on a dataset of material textures, as well as the reconstruction from scattering representations.

We finally focus on the study of multifractals. Fractals are fundamental mathematical objects, characterized by their self-similarity scaling laws. Their identification and discrimination requires access to high order moments, which are difficult to estimate or even might fail to exist. We introduce a renormalization of scattering coefficients which captures a new characteristic scaling law of the multifractal, and which allows stable identification of several fractal parameters.

1.2 The Scattering Representation

The construction of signal representations for recognition starts with the invariance and stability properties. These properties can be stated in mathematical terms as follows. If $x \in \mathbf{L}^2(\mathbb{R}^d)$ and G denotes a given group of transformations of $\mathbb{R}^d$ and $L_\varphi x(u) = x(\varphi(u))$ denotes the action of an element $\varphi \in G$ in $\mathbf{L}^2(\mathbb{R}^d)$, the invariance to the action of G is obtained by requiring

$$\forall \varphi \in G\,,\ \forall x \in \mathbf{L}^2(\mathbb{R}^d),\ \Phi(L_\varphi x) = \Phi(x)\ .$$

On the other hand, if now we consider a diffeomorphism $\varphi \in Diff(\mathbb{R}^d)$, the stability to deformations is expressed as a Lipschitz continuity condition with respect to a metric $\|\varphi\|$ on the space of diffeomorphisms measuring the amount of deformation:

$$\forall \varphi \in Diff(\mathbb{R}^d)\,, \forall x \in \mathbf{L}^2(\mathbb{R}^d)\ \ \|\Phi(L_\varphi x) - \Phi(x)\| \leq C\|x\|\|\varphi\|\ .$$

Chapter 2 reviews the scattering transform for deterministic functions and processes, together with its mathematical properties. It also studies these properties on image processing applications, and obtains two new mathematical results: the first one characterizing its non-linearities from stability constraints, and the second one giving a partial answer to a conjecture stated in [Mal12], relating the signal sparsity with the regularity of its scattering representation in the transformed domain.

Scattering representations [Mal12] construct invariant, stable and informative signal representations by cascading wavelet modulus decompositions followed by a lowpass filter. A wavelet decomposition operator at scale J is defined as

$$\mathcal{W}_J x = \{x \star \psi_\lambda\}_{\lambda \in \Lambda_J}\ ,$$

where $\psi_\lambda(u) = 2^{-dj}\psi(2^{-j}r^{-1}u)$ and $\lambda = 2^j r$, with $j < J$ and $r \in G$ belongs to a finite rotation group G of $\mathbb{R}^d$. Each rotated and dilated wavelet thus extracts the energy of x located at a given scale and orientation given by λ. Wavelet coefficients are not translation invariant, and their average does not produce any information since wavelets have zero mean. A translation invariant measure can be extracted out of each wavelet sub-band λ by introducing a non-linearity which restores a non-zero, informative average value. This is for instance achieved by computing the complex modulus and averaging the result

$$\int |x \star \psi_\lambda|(u)du\ .$$

The information lost by this averaging is recovered by a new wavelet decomposition $\{|x \star \psi_\lambda| \star \psi_{\lambda'}\}_{\lambda' \in \Lambda_J}$ of $|x \star \psi_\lambda|$, which produces new invariants by iterating the same procedure. Let $U[\lambda]x = |x \star \psi_\lambda|$ denote the wavelet modulus operator corresponding to the subband λ. Any sequence $p = (\lambda_1, \lambda_2, ..., \lambda_m)$ defines a *path*, i.e, the ordered product of non-linear and non-commuting operators

$$U[p]x = U[\lambda_m] \ldots U[\lambda_2]\,U[\lambda_1]x = |\ ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \ldots | \star \psi_{\lambda_m}|\ ,$$

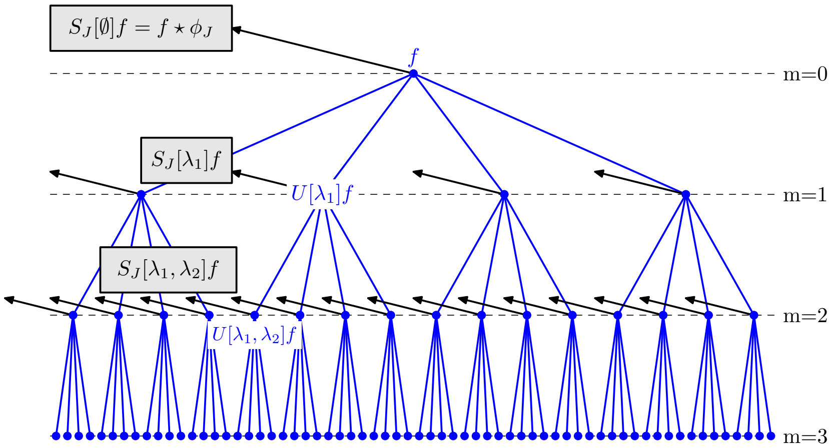


Figure 1.1: Convolutional structure of the windowed scattering transform. Each layer is computed from the previous by applying a wavelet modulus decomposition U on each envelope $U[p]f$. The outputs of each layer are obtained via a lowpass filter ϕ_J.

with $U[\emptyset]x = x$.

Many applications in image and audio recognition require locally translation invariant representations, but which keep spatial or temporal information beyond a certain scale 2^J. A windowed scattering transform computes a locally translation invariant representation by computing a lowpass average at scale 2^J with a lowpass filter $\phi_{2^J}(u) = 2^{-2J}\phi(2^{-J}u)$. For each path $p = (\lambda_1, \dots, \lambda_m)$ with $\lambda_i \in \Lambda_J$ we define the windowed scattering transform as

$$S_J[p]x(u) = U[p]x \star \phi_{2^J}(u) = \int U[p]x(v)\phi_{2^J}(u-v)\,dv \ ,$$

A Scattering transform has the structure of a convolutional network, but its filters are given by wavelets instead of being learnt. Thanks to this structure, the resulting transform is locally translation invariant and stable to deformations. The scattering representation enjoys several appealing properties. In particular, with the scattering norm, defined as

$$\|S_J x\|^2 = \sum_{m\geq 0} \sum_{p \in \Lambda_J^m} \|S_J[p]x\|^2 \ ,$$

and for an appropriate choice of the wavelets, the scattering transform is a non-expansive $\mathbf{L}^2(\mathbb{R}^d)$ operator, $\|S_J x - S_J x'\| \leq \|x - x'\|$, which moreover preserves the $\mathbf{L}^2$ norm: $\|S_J x\| = \|x\|$. In particular, this implies that in practice the first m_{max} layers of the transform capture most of the signal energy and thus the network depth can be limited. We shall see that for most practical applications, the first two layers carry most of the signal energy and also provide enough discriminative information for recognition.

Scattering transforms also define a representation for stationary processes $X(t)$, which will be the appropriate tool to study image and auditory textures as well as

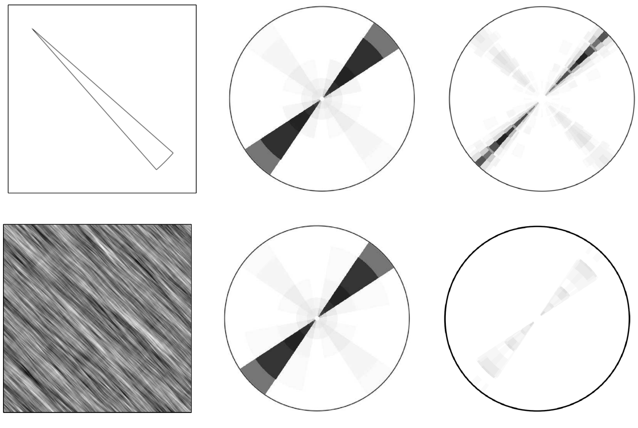


Figure 1.2: (a) Two images $x(u)$, $x'(u)$. (b) First order scattering coefficients $S_J[\lambda_1]x$ displayed with the mapping described in Section 2.6. They are the same for both images. (c) Second order scattering coefficients $S_J[\lambda_1, \lambda_2]x$. They are different for each image, thus showing that second order scattering can efficiently measure the sparsity of the signal.

multifractal processes. For each p, the expected scattering is defined as

$$\overline{S}X(p) = E(U[p]X(t)) \ .$$

As opposed to the power spectrum, scattering coefficients depend upon high moments of the process.

Scattering networks, similarly as other convolutional networks, require a nonlinear step at the output of its filter banks in order to create new invariants. If one imposes such nonlinearities to be non-expansive and that they commute with any diffeomorphism, then we show in Section 2.4. that they are necessarily pointwise. The non-expansive property is important to ensure that the overall representation is stable with respect to additive noise. The commutation with diffeomorphisms allows these nonlinearities to be intertwined in a cascade of filter bank decompositions without affecting the stability to deformations. Moreover, they enable a systematic procedure to obtain invariant coefficients. The modulus is preferred since it also provides energy conservation: $\||x|\|_2 = \|x\|_2$.

If one considers the limit $J \to \infty$, then the windowed scattering transform converges to a translation invariant representation of $\mathbf{L}^2(\mathbb{R}^d)$ defined on an uncountable path set $\overline{\mathcal{P}}_\infty$ which contains infinite length paths of arbitrarily large scales. One can define a measure and a metric on this set, and show [Mal12] that with the appropriate renormalisation, given by the response of a Dirac, the limit scattering maps functions from $\mathbf{L}^2(\mathbb{R}^d)$ to $L^2(\overline{\mathcal{P}}_\infty, d\mu)$. $\overline{S} : L^2(\mathbb{R}^d) \to L^2(\mathcal{P}_\infty, d\mu)$,

$$\overline{S}x(p) = \mu_p^{-1} \int U[p]x(u)\, du \quad \text{with} \quad \mu_p = \int U[p]\delta(u)\, du \ .$$

This renormalized limit scattering has striking ressemblances with the continuous Fourier transform. In particular, we show in Section 2.5 that for functions $x(u)$ belonging to $\mathbf{L}^1(\mathbb{R}^d)$, then the scattering representation $\overline{S}x(p)$ is continuous on a weaker topology of $\overline{\mathcal{P}}_\infty$, given by paths with bounded slope and finite order.

1.3 Image and Pattern Classification

The properties of scattering operators are exploited in the context of signal classification in Chapter 3. Given K signal classes, we observe L samples for each class, $x_{k,l}\,, l = 1..L, k = 1..K$, which are used to estimate a classifier $\hat{k}(x)$ assigning a class amongst K to each new signal x.

Complex object recognition problems require more forms of invariance other than those modeled by physical transformations. For instance, image datasets such as Caltech or Pascal exhibit large variability in shape, appearance, clutter, as shown in Figure 1.3. Similarly, challenging problems such as speech recognition have to take into account variability of speaker. However, even these datasets are exposed to variability coming from physical transformations, and hence most object recognition architectures require a feature extraction step which eliminates such variability, while building stability to deformations and keeping enough discriminative information. The efficiency of scattering representations for such a role is first tested in environments where physical transformations and deformations account for most of the variability.

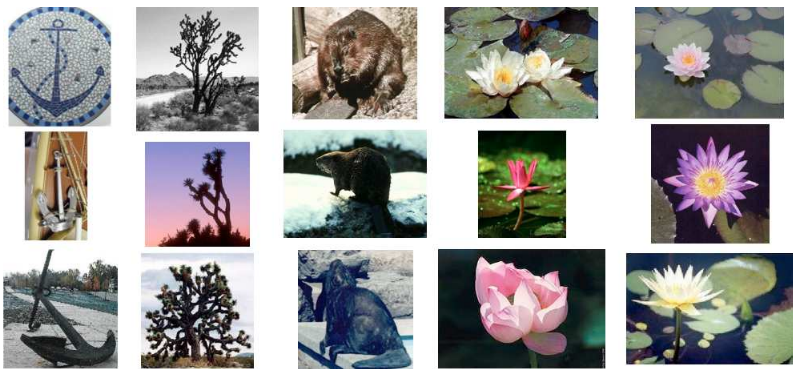


Figure 1.3: Examples from the Caltech dataset. Intra-class variability includes changes in shape, appearance, clutter.

The Lipschitz continuity property implies that small deformations are linearized by the scattering transform. We construct a generative classification architecture which exploits this linearization property as follows. For each signal class, we learn a collection of embedded affine spaces $\mathbf{A}_{d,k} = \mu_k + V_{d,k}$ in the scattering domain which best approximate the observed samples. The best affine approximation spaces are easily computed with a Principal Component Analysis (PCA) which diagonalises the class-conditional empirical covariance matrix

$$\Sigma_k = \frac{1}{L} \sum_l (S_J x_{k,l} - \mu_k)(S_J x_{k,l} - \mu_k)^T \ ,$$

where μ_k is the scattering class empirical average $\mu_k = \frac{1}{L} \sum_l S_J x_{k,l}$.

Signal classes in the scattering domain are thus approximated with affine spaces of varying dimension. The resulting classifier associates a signal x to the class $\hat{k}$ yielding the best approximation space:

$$\hat{k}(x) = \underset{k \leq K}{\operatorname{argmin}} \|S_J x - P_{\mathbf{A}_{d,k}}(S_J x)\| \ . \tag{1.1}$$

As the dimension increases, the approximation error for each class is reduced, but not necessarily the discrimination between affine spaces of different classes. The classifier thus learns the dimension d yielding the best trade-off on a validation subset of the available samples. This generative architecture is particularly efficient with small training samples due to the fact that learning is limited to class-conditional auto-correlation matrices.

The discriminability of scattering coefficients can be improved by renormalising its coefficients. Renormalisation is an important aspect of many classifiers [FKL10; Bur98], which often requires significant expertise to properly adjust. We explore a robust equalization strategy which replaces the Dirac renormalisation by the observed maximum response, and which improves classification rates. We also compare this strategy with the scattering transfer

$$T_J[\lambda_1, \lambda_2]x(u) = \frac{S_J[\lambda_1, \lambda_2]x(u)}{S_J[\lambda_1]x(u)} \ ,$$

a renormalization which will be fundamental in the study of multifractal processes.

The linear generative architecture is tested on handwritten recognition, in the MNIST and USPS datasets. The variability in such problems is well modeled by translations and elastic deformations. We show that second order scattering coefficients capture stable, discriminative information leading to state-of-the-art classification results. The generative linear classifier is compared with a Gaussian kernel SVM, as a function of the training set size. For small training sizes, the generative classifier outperforms state-of-the-art methods obtained with convolutional networks. As the training size grows, the richer Gaussian SVM classifier overtakes all previously reported classification results.

1.4 Texture Discrimination and Reconstruction from Scattering

Chapter 4 studies the efficiency of scattering representations to discriminate and reconstruct image and auditory textures. These problems require a statistical treatment of the observed variability as realizations of stationary processes.

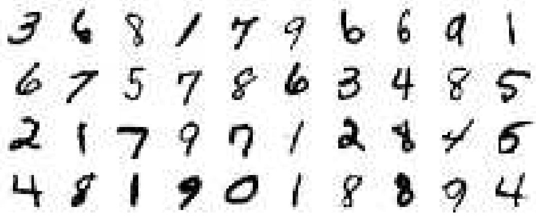


Figure 1.4: Examples from the MNIST dataset. Intra-class variability is well modeled by non-rigid geometric transformations.

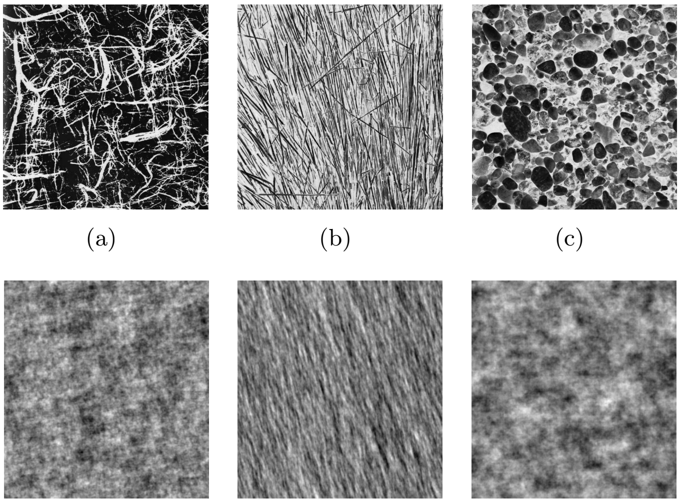
Stationary processes admit a spectral representation. Its spectral density is computed from second moments, and completely characterizes Gaussian processes. However, second moments are not enough to discriminate between most real-world textures.

Texture classification and reconstruction requires a representation for stationary processes capturing high order statistics in order to discriminate non-Gaussian processes. One can include high order moments $E(|X(t)|^n)$, $n \geq 2$, in the texture representation, but their estimation has large variance due to the presence of large, rare events produced by the expansive nature of x^n for $n \geq 1$. In addition, discrimination requires a representation which is stable to changes in viewpoint or illumination.

Julesz [Jul62] conjectured that the perceptual information of a stationary texture $X(t)$ was contained in a collection of statistics $\{E(g_k(X(t)))\, , \, k \in \mathcal{K}\}$. He originally stated his hypothesis in terms of second-order statistics, measuring pairwise interactions, but he later reformulated it in terms of *textons*, local texture descriptors capturing interactions across different scales and orientations. Textons can be implemented with filter-banks, such as oriented Gabor wavelets [LM01], which form the basis for several texture descriptors.

The expected scattering representation is defined for processes with stationary increments. First order scattering coefficients average wavelet amplitudes, and yield similar information as the average spectral density within each wavelet subband. Second order coefficients depend upon higher order moments and are able to discriminate between non-Gaussian processes. The expected scattering is estimated consistently from windowed scattering coefficients, thanks to the fact that it is computed with non-expansive operators. Besides, thanks to the stability of wavelets to deformations, the resulting texture descriptor is robust to changes in viewpoint which produce non-rigid small deformations.

The CUREt dataset [DVGNK99] contains a collection of different materials taken under different lighting and pose conditions. On top of the stochastic variability within different realizations of the stationary textured material, we thus observe a variability modeled by a low-dimensional manifold. In this context, the scattering generative classifier of Chapter 3 outperforms previous state-of-the-art methods. The expected scattering representation is consistently estimated with a delocalized scattering, which reduces the intra-class variance while keeping discriminative information. The residual intra-class variability, due to pose and illumination changes, is reduced thanks to the



(a) (b) (c)

(d) (e) (f)

Figure 1.5: First row: Three different examples x_i from Brodatz dataset, $i = 1 \ldots 3$. Bottom row: Reconstruction obtained by equalizing white gaussian noise with each spectral density $\hat{R}_{x_i}$, $i = 1 \ldots 3$, so that the textures on each column have the same second order moments.

low-dimensional affine spaces.

Auditory textures are highly non-gaussian, and also require a descriptor which captures high order moments and which can be estimated with reduced variance. First we consider the discrimination between different sounds, and we verify the fact that second order scattering coefficients are necessary to discriminate non-gaussian textures. This fact is confirmed by studying the texture reconstruction problem. By following the same statistical framework as [MS11], we reconstruct audio realizations from a given expected scattering descriptor using a gradient descent scattering reconstruction algorithm, which iteratively modifies a sample by modulating its wavelet coefficients with an envelope, whose spatial variations encoded by new wavelet coefficients, which become the descent variables. While first order coefficients produce sounds with a characteristic gaussian signature, the reconstructions obtained by adding second order scattering coefficient produce reconstructions perceptually similar to the originals.

1.5 Multifractal Scattering

Many image textures exhibit an intermittent behavior at different scales, and are thus well modeled as multifractals. Fractals are singular almost everywhere, and its identification and discrimination requires an analysis of their singularities. Chapter 5 is dedicated to the the estimation and characterization of fractal parameters from scattering representations, which explains its efficiency on the task of classifying multifractal textures. Fractals are fundamental mathematical objects, defined from by scaling laws relating how a process or measure $X(t)$ relates to $D_s X(t) = X(s^{-1}t)$, a dilated version of $X(t)$. Examples include stochastic processes such as Fractional Brownian Motions, and also deterministic measures, such as Cantor measures.

A fundamental signature of a multifractal is given by its spectrum of singularity, which measures the Hausdorff dimension $\mathcal{D}(h)$ of the sets Ω_h where the fractal has Hölder singularity given by an exponent h. Under appropriate self-similarity conditions, it can be shown [Jaf97; BKM08a] that this spectrum of singularity can be recovered from the scaling exponents $\zeta(q)$, measuring the power law of the increments of the process as the scale converges to 0. For fractal processes $X(t)$, it is defined as

$$\forall q\ ,\ \lim_{l \to 0^+} E(|X(t) - X(t-l)|^q) l^{-\zeta(q)} = C_q\ .$$

These scaling exponents can be recovered from a wavelet decomposition [BKM08a], thanks to their vanishing moment: $E(|X \star \psi_j|^q) \simeq 2^{j\zeta(q)}$. While monofractal processes are characterized by a linear exponent, $\zeta(q) = qH$, multifractal processes have a strictly concave $\zeta(q)$. The scaling exponents thus contain crucial information on the fractal. However, they are difficult to estimate, since high order moments are dominated by large, rare events which increase the variance of their estimators.

The expected scattering representation is computed from first moments of wavelet modulus decompositions, and is well defined for processes with stationary increments having finite first moments. Since the self-similarity of a fractal is expressed through dilations, it is translated in the scattering domain by relationships between different scattering paths. These relationships can be exploited to construct a new scale independent fractal descriptor. In this chapter we introduce the scattering transfer, constructed from first and second order scattering coefficients:

$$TX(j_1, j_2) = \frac{\overline{S}X(j_1, j_2)}{\overline{S}X(j_1)}\ .$$

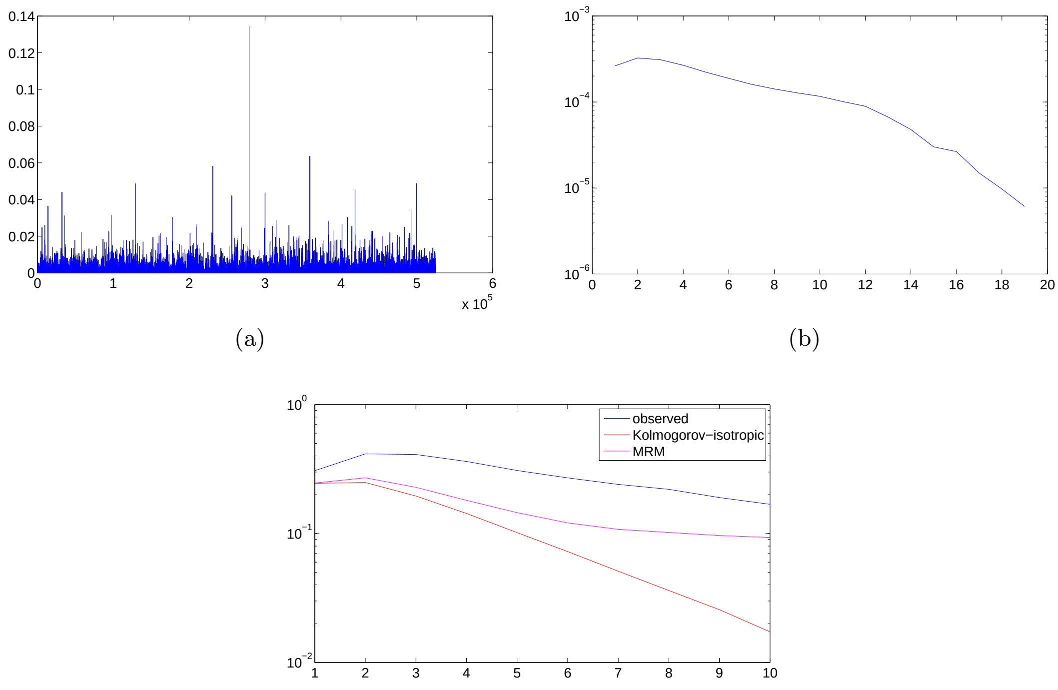
It is defined as a function of two path variables, but the self-similarity of X implies that the scattering transfer is only a function of path increments: $TX(j_1, j_1+l) = \overline{T}X(l)$. This normalized scattering measure, together with the expected first order scattering, defines a new tool to study fractal phenomena, capturing information which allows identification and discrimination of several self-similar fractal families. It is consistently estimated from windowed scattering coefficients by combining information from different scales. The scattering transfer is a fractal geometric descriptor, which is shown to be nearly invariant to the action of the derivative operator. As a result, it provides a relative measure of how the singularities of the fractal are distributed in space.

This function is computed from first and second order coefficients, but it also has the capacity to control the behavior of higher order scattering coefficients. For a wide class of fractals, scattering coefficients can be asymptotically predicted from the scattering transfer and its first order coefficients. This asymptotic property corresponds to a Markov propagation across scattering paths.

Multifractal processes are constructed with an integral scale, which restricts the self-similarity to a scale range of the form $(-\infty, J)$. The curvature of its scaling exponents $\zeta(q)$, referred as the intermittency, gives important information on the degree of multifractality of the process. Its analysis requires access to information contained in high order moments. We show that this information can be extracted consistently from the scattering transfer. Thanks to another asymptotic property, denoted Markov energy property, the intermittency $2\zeta(1) - \zeta(2)$, measuring the curvature of $\zeta(q)$, is obtained from the smallest ρ satisfying the equation

$$\sum_{l \geq 1} \overline{T}X(l)^2 \rho^l = 1$$

by $2\zeta(1) - \zeta(2) = -\log_2(\rho)$. The scattering transfer is thus an efficient measure to discriminate monofractal processes -having scaling exponents with no curvature, from multifractal processes.



(a)

(b)

(c)

Figure 1.6: (a) Measured energy dissipation $F(t)$ in a turbulent flow, as a function of time. (b) estimated first order scattering coefficients $\overline{S}F(j)$. The decay observed across scales is consistent with the isotropic dissipation model of Kolmogorov and the '$k^{-5/3}$' law. (c) Comparison of the scattering transfer $\overline{T}F(l)$ estimated from the observed turbulent flows with two energy dissipation models. The scattering transfer provides insight on the adequacy of multifractal models.

The scattering transfer is computed on several fractal families, including Fractional Brownian Motions, Lévy processes and Multifractal Random Measures, as well as empirical data from turbulent flows, revealing important properties of the fractal. We also compare the estimation of the intermittency $2\zeta(1) - \zeta(2)$ using the scattering transfer with a direct estimation of the moments, and also with an estimation based on the logarithm of the increments, showing a consistency in pair with the state-of-the-art.

Chapter 2

Invariant Scattering Representations

2.1 Introduction

Image and audio recognition require a metric to compare signals which is stable with respect to deformations and additive noise and which respects the invariants given by transformation groups, while being able to discriminate between different objects or textures. A metric of the form

$$d(x, x') = \|\Phi(x) - \Phi(x')\|$$

for signals in $\mathbf{L}^2(\mathbb{R}^d)$ translates these invariance, stability and discriminability properties into the signal representation Φ of $\mathbf{L}^2(\mathbb{R}^d)$ functions.

Linear signal representations are complete and stable to additive noise, which explains their use on tasks such as compression or denoising. However, they define a metric equivalent to the Euclidean metric, which is not continuous with respect to geometric deformations. Therefore, they are unable to achieve stability to deformations or local invariance while keeping high frequency information.

This opens a Pandora box of non-linear signal representations with the prescribed stability, invariance and discriminability properties. Translation invariance can be obtained with the Fourier modulus, but high frequencies are unstable to small dilations. Stability to small deformations can be recovered by grouping frequencies into dyadic intervals with wavelets, leading to the scattering representation introduced by S. Mallat in [Mal12].

Scattering operators construct invariant, stable and informative signal representations by cascading wavelet modulus decompositions followed by a lowpass filter. A windowed scattering transform has the structure of a convolutional network [FKL10], but its filters are given by wavelets instead of being learnt. Scattering operators also define a representation for processes with stationary increments, which will be the appropriate tool to study image and auditory textures, in Chapter 4, as well as multifractals, in Chapter 5. It is computed from first moments of cascaded wavelet modulus, which are contractive operators. As a result, it defines a representation that is consistently estimated from windowed scattering coefficients for a wide class of ergodic processes, and which includes information from high order moments.

The choice of the non-linearities is an important factor determining the complexity and the properties of the representation. As in other convolutional neural networks, the non-linearities in the scattering operator are given by point-wise operators. We shall see that this choice can be justified from the stability conditions imposed on the representation.

Besides its stability to additive noise and geometric deformations, the scattering representation enjoys other remarkable properties. In particular, as the scale of the lowpass window grows, and after an appropriate renormalization, the scattering transform converges into a translation invariant integral transform, defined on an uncountable path space. The resulting integral transform shares some properties with the Fourier transform; in particular, signal decay can be related with the regularity on the integral scattering domain.

The rest of the chapter is structured as follows. Section 2.2 considers signal representations for recognition, and formulates the stability properties as Lipschitz continuity conditions on the representation. Section 2.3 reviews the windowed scattering representation for $\mathbf{L}^2(\mathbb{R}^d)$ functions and its stability and energy conservation properties, as well as the expected scattering for processes with stationary increments, and the integral scattering transform. Section 2.4 gives the characterization of point-wise non-linear operators from stability properties. Then, in Section 2.5 we give a partial positive answer to a conjecture stated in [Mal12], predicting that the integral scattering transform of a $\mathbf{L}^1(\mathbb{R}^d)$ function is continuous. Finally, in Section 2.6, we study scattering representations on image processing applications; we give examples of scattering wavelets and we explore the properties of scattering representations as local image and texture descriptors.

2.2 Signal Representations and Metrics for Recognition

This section reviews some of the existing tools for signal representation and discrimination in recognition tasks. Invariance and stability are formulated in terms of Lipschitz continuity conditions, and are then studied on a variety of representations.

2.2.1 Local translation invariance, Deformation and Additive Stability

In recognition tasks, the action of small geometric deformations and small additive perturbations produce small changes in the appearance of objects and textures. This motivates the study of signal representations defining an Euclidean metric stable to those perturbations.

These stability properties can be expressed mathematically as Lipschitz continuity properties. Stability to additive noise is guaranteed by imposing Φ to be non-expansive:

$$\forall\, x, \tilde{x} \in \mathbf{L}^2(\mathbb{R}^d)\ ,\ \|\Phi(x) - \Phi(\tilde{x})\| \leq \|x - \tilde{x}\|\ . \tag{2.1}$$

Indeed, it results that the metric defined by Φ is Lipschitz continuous with respect to the Euclidean norm of an additive perturbation:

$$d(x+h,x) = \|\Phi(x+h) - \Phi(x)\| \leq \|h\| \ .$$

On the other hand, stability to deformations is achieved by controlling the behavior of Φ under the action of diffeomorphisms $u \mapsto u - \tau(u)$, where $\tau : \mathbb{R}^d \to \mathbb{R}^d$ is an invertible displacement field. The amount of deformation can be measured with a metric on the space of diffeomorphisms. If $|\tau(u)|$ denotes the Euclidean norm in $\mathbb{R}^d$, $|\nabla\tau(u)|$ denotes the operator norm of $\nabla\tau(u)$ and $|H\tau(u)|$ is the sup norm of the Hessian tensor, then the norm of the space of $\mathbf{C}^2$ diffeomorphisms measures the amount of deformation over any compact subset $\Omega \subset \mathbb{R}^d$ as

$$\|\tau\| = \sup_{u\in\Omega} |\tau(u)| + \sup_{u\in\Omega} |\nabla\tau(u)| + \sup_{u\in\Omega} |H\tau(u)| \ .$$

This deformation metric penalizes displacement fields by its maximum amplitude $\sup_{u\in\Omega}|\tau(u)|$ and maximum elasticity $\sup_{u\in\Omega}|\nabla\tau(u)|$. In most contexts, however, rigid displacement fields, corresponding to translations, do not affect recognition to the same extent as non-rigid, elastic deformations. This motivates the notion of locally translation invariant representations. We say that Φ is Lipschitz continuous to the action of $\mathbf{C}^2$ diffeomorphisms and locally translation invariant at scale 2^J if for any compact $\Omega \subset \mathbb{R}^d$ there exists C such that, for all $x \in \mathbf{L}^2(\mathbb{R}^d)$ supported in Ω and all $\tau \in \mathbf{C}^2$,

$$d(L[\tau]x,x) = \|\Phi(L[\tau]x) - \Phi(x)\| \leq C\|x\| \left(2^{-J} \sup_{u\in\Omega}|\tau(u)| + \sup_{u\in\Omega}|\nabla\tau(u)| + \sup_{u\in\Omega}|H\tau(u)| \right) \ . \tag{2.2}$$

The reference scale 2^J controls the amount of translation invariance required on the representation, by diminishing the influence of the amplitude of τ in the deformation metric. If τ is a displacement field with maximum amplitude $\sup_{u\in\Omega}|\tau(u)| \ll 2^J$, then (2.2) shows that the representation stability is controlled by the amount of elastic deformation applied to x. On the other hand, the scale of local invariance also controls the amount of delocalization of the representation. Pattern recognition tasks often require signal representations which keep spatial information up to a certain resolution, whereas we will ask stationary texture representations to be fully translation invariant, by letting the local invariance scale J go to infinity.

2.2.2 Kernel Methods

Kernel methods refer to a general theory in the machine learning framework, whose main purpose consists in embedding data in a high dimensional space, in order to express complex relationships in the data in terms of linear scalar products.

For a generic input space $\mathcal{X}$, a *feature map* $\Phi : \mathcal{X} \longrightarrow \mathcal{H}$ maps data into a Hilbert space $\mathcal{H}$. Linear classification methods access the transformed data $\Phi(x)$ only through scalar products of the form [STC04]

$$\langle \Phi(x), \Phi(x') \rangle \ .$$

Rather than building the mapping explicitly, the popular "Kernel Trick" exploits Mercer's theorem. It states that a continuous, symmetric and positive definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defines an integral operator of $L^2(\mathcal{X})$, which diagonalizes in an orthonormal basis [MNY06] $\{\phi_n\}_n$ of $L^2(\mathcal{X})$, with non-negative eigenvalues. As a result, $K(x, x')$ admits a representation

$$K(x, x') = \sum_{n \geq 1} \lambda_n \phi_n(x) \phi_n(x') \ ,$$

which yields

$$K(x, x') = \langle \Phi(x), \Phi(x') \rangle \ ,$$

with $\Phi(x) = (\lambda_n^{1/2} \phi_n(x))_n$. In Kernel methods it is thus sufficient to construct positive definite kernels K on $\mathcal{X}^2$ in order to extend linear classification tools to more complex relationships.

Support Vector Machines (SVMs) are particular instances of kernel methods, which construct separating hyperplanes in supervised learning tasks. We shall discuss these methods in further detail on Chapter 3.

Despite their success and effectiveness in a number of machine learning tasks, the high dimensional embeddings induced by kernel methods do not automatically enjoy the stability properties to additive noise or deformations. The kernel needs to be chosen accordingly.

2.2.3 Deformable Templates

The theory of deformable templates, pioneered by Grenader in [Gre93], introduced the notion of group action to construct metrics in a generic object space. A deformable template is defined as an element $x \in \mathcal{X}$ on which a group action $G \times \mathcal{X}$, $(g, x) \mapsto g.x \in \mathcal{X}$ is defined. This action defines through the orbits $\{g.x\,,\ g \in G\}$ a family of "deformed" objects.

This structure allows us to incorporate the group action into the notion of similarity between elements of $\mathcal{X}$. A metric d on $\mathcal{X}$ can be constructed from a metric $\tilde{d}$ on the lifted product space $G \times \mathcal{X}$ [MY01]. If $\tilde{d}$ is *left-invariant* [1] , then

$$d(x, x') = \inf\{\tilde{d}((id, x), (g, g.x'))\,,\ g \in G\} \tag{2.3}$$

defines a metric from the set distance between the orbits of x and x' under the action of G.

If G denotes the group of $\mathbb{R}^d$ diffeomorphisms, acting on $\mathcal{X} = \mathbf{L}^2(\mathbb{R}^d)$ by composition, then the construction (2.3) has the capacity to express the similarity between x and its deformed version $g.x$ in terms of the amount of deformation g.

[1] A distance $\tilde{d}$ on $G \times \mathcal{X}$ is left-invariant [MY01] if, for all $h, g, g' \in G$ and all x, $x' \in \mathcal{X}$,

$$\tilde{d}(h.(g, x), h.(g', x')) = \tilde{d}((g, x), (g', x'))$$

The computation of the distance $d(x,x')$ in (2.3) requires to optimize a deformation cost function, which in general is an ill-posed inverse problem. In some applications, however, such as medical imaging, it uncovers essential information about the underlying deformation process.

In general, there is no simple way to construct meaningful left-invariant metrics on the product space $G\times\mathcal{X}$. In [TY05], Trouvé and Younes construct a differential structure on this space, based on infinitesimal deformations and amplitude variations. This structure then enables the definition and computation of geodesics. More specifically, the authors consider infinitesimal perturbations of a template $I\in\mathcal{X}$ of the form

$$\begin{array}{rcl} G\times\mathcal{X} & \longrightarrow & \mathcal{X}\ , \\ (\tau,\sigma) & \longmapsto & \tilde{I}^{\epsilon}_{(v,z)}(u) = I(u-\epsilon\tau(u)) + \beta\epsilon\sigma(u) + o(\epsilon)\ , \end{array}$$

for small $\epsilon > 0$. The vector field $\tau(u)$ thus carries the geometrical transformation, whereas the scalar function $\sigma(u)$ represents infinitesimal amplitude variations. One then constructs the tangent space T_I from these infinitesimal variations:

$$T_I = \left\{ \lim_{\epsilon\to 0} \frac{\tilde{I}^{\epsilon}_{(\tau,\sigma)} - I}{\epsilon}\ ,\ (\tau,\sigma)\in G\times\mathcal{X} \right\}\ .$$

By considering a norm $|\cdot|_W$ on $G\times\mathcal{X}$, one can define a metric $|\cdot|_{W'}$ on the tangent space T_I, which leads to the geodesic distance

$$d_W(I_0,I_1) = \inf\left\{ \int_0^1 |\dot{\gamma}(t)|_{W'}dt\ ,\gamma(0)=I_0\ ,\gamma(1)=I_1 \right\}\ ,$$

where the infimum is taken over all paths γ joining the two images I_0, I_1.

The computation of such geoedesics thus requires to solve a variational problem, which in particular estimates deformations minimizing a trade-off between the geometric and photometric components of the perturbation of the form

$$\hat{\varphi} = \text{argmin}\, \|I_1 - I_0\circ\varphi\|_2^2 + D_G(Id,\varphi)^2\ ,$$

where D_G is a metric on the space of diffeomorphisms.

The estimation of diffeomorphisms is a difficult inverse problem. Several methods have been proposed, for instance in [Tro95; VMYT04], with applications in medical imaging [RACW10] and classification [AAT07].

A particular instance of such problem is the optical flow estimation. In this case, the goal is to estimate the motion field between consecutive frames of a video sequence $(I_t)_t$. Under the hypothesis that illumination is constant along object trajectories, $I(x+v_t(x))_t = C(x)$, and that I is differentiable, we obtain the differential equation

$$\langle\nabla I, v_t\rangle + \frac{\partial I_t}{\partial t} = 0\ .$$

Estimating the motion field v_t is thus an ill-posed inverse problem since there are more unknowns than equations. There exists a vast literature on how to regularize this inverse problem, from Horn and Shunk [HS81] to wavelet based approaches [Ber99].

2.2.4 Fourier Modulus, Autocorrelation and Registration Invariants

Translation invariant representations can be obtained from registration, auto-correlation or Fourier modulus operators. However, the resulting representations are not Lipschitz continuous to deformations.

A representation $\Phi(x)$ is translation invariant if it maps global translations $x_c(u) = x(u-c)$ by $c \in \mathbb{R}^d$ of any function $x \in \mathbf{L}^2(\mathbb{R}^d)$ to the same image:

$$\forall\, x \in \mathbf{L}^2(\mathbb{R}^d)\,,\, \forall\, c \in \mathbb{R}^d\ ,\ \Phi(x_c) = \Phi(x)\ . \tag{2.4}$$

The Fourier transform modulus is an example of a translation invariant representation. Let $\hat{x}(\omega)$ be the Fourier transform of $x(u) \in \mathbf{L}^2(\mathbb{R}^d)$. Since $\widehat{x_c}(\omega) = e^{-ic.\omega}\, \hat{x}(\omega)$, it results that $|\widehat{x_c}| = |\hat{x}|$ does not depend upon c.

A Fourier modulus is translation invariant and stable to additive noise, but unstable to small deformations at high frequencies [Mal12], as illustrated with the following dilation example. Let $\tau(x) = sx$ denote a linear displacement field where $|s|$ is small, and let $x(u) = e^{i\xi u}\theta(u)$ be a modulated version of a lowpass window $\theta(u)$. Then the dilation $x_\tau(u) = L[\tau]x(u) = x((1+s)u)$ moves the central frequency of $\hat{x}$ from ξ to $(1+s)\xi$. If $\sigma_\theta^2 = \int |\omega|^2 |\hat{\theta}(\omega)|^2 d\omega$ measures the frequency spread of θ, then

$$\sigma_x^2 = \int |\omega - \xi|^2 |\hat{x}(\omega)|^2 d\omega = \sigma_\theta^2\ ,$$

and

$$\begin{aligned}
\sigma_{x_\tau}^2 &= (1+s)^{-d} \int (\omega - (1+s)\xi)^2 |\hat{x}((1+s)^{-1}\omega)|^2 d\omega \\
&= \int |(1+s)(\omega-\xi)|^2 |\hat{x}(\omega)|^2 d\omega = (1+s)^2 \sigma_x^2\ .
\end{aligned}$$

It follows that if the distance between the central frequencies of x and x_τ, $s\xi$, is large compared to their frequency spreads, $(2+s)\sigma_\theta$, then the frequency supports of x and x_τ are nearly disjoint and hence

$$\| |\hat{x}_\tau| - |\hat{x}| \| \sim \|x\|\ ,$$

which shows that $\Phi(x) = |\hat{x}|$ is not Lipschitz continuous to deformations, since ξ can be arbitrarily large.

The autocorrelation of x

$$R_x(v) = \int x(u) x^*(u-v) du$$

is also translation invariant: $R_x = R_{x_c}$. Since $R_x(v) = x \star \overline{x}(v)$, with $\overline{x}(u) = x^*(-u)$, it follows that the autocorrelation representation $\Phi(x) = R_x$ satisfies

$$\widehat{R_x}(\omega) = |\hat{x}(\omega)|^2\ .$$

The Plancherel formula thus proves that it has the same instabilities as a Fourier transform:

$$\|R_x - R_{x_\tau}\| = (2\pi)^{-1} \| |\hat{x}|^2 - |\hat{x}_\tau|^2 \|\ .$$

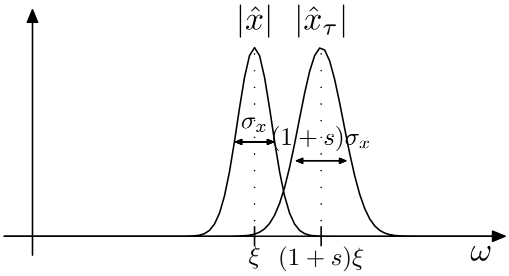


Figure 2.1: Dilation of a complex bandpass window. If $\xi \gg \sigma_x s^{-1}$, then the supports are nearly disjoint.

Besides deformation instabilities, the Fourier modulus and the autocorrelation lose too much information. For example, a Dirac $\delta(u)$ and a linear chirp e^{iu^2} are two signals having Fourier transforms whose moduli are equal and constant. Very different signals may not be discriminated from their Fourier modulus.

A canonical invariant [KDGH07; Soa09] $\Phi(x) = x(u - a(x))$ registers $x \in \mathbf{L}^2(\mathbb{R}^d)$ with an anchor point $a(x)$, which is translated when x is translated:

$$a(x_c) = a(x) + c\ .$$

It thus defines a translation invariant representation: $\Phi x_c = \Phi x$. For example, the anchor point may be a filtered maximum $a(x) = \arg\max_u |x \star h(u)|$, for some filter $h(u)$. A canonical invariant $\Phi x(u) = x(u - a(x))$ carries more information than a Fourier modulus, and characterizes x up to a global absolute position information [Soa09]. However, it has the same high-frequency instability as a Fourier modulus transform. Indeed, for any choice of anchor point $a(x)$, applying the Plancherel formula proves that

$$\|x(u - a(x)) - x'(u - a(x'))\| \geq (2\pi)^{-1}\, \||\hat{x}(\omega)| - |\hat{x}'(\omega)|\|\ .$$

If $x' = x_\tau$, the Fourier transform instability at high frequencies implies that $\Phi x = x(u - a(x))$ is also unstable with respect to deformations.

2.2.5 SIFT and HoG

SIFT (Scale Invariant Feature Transform) is a local image descriptor introduced by Lowe in [Low04], which achieved huge popularity thanks to its invariance and discriminability properties.

The SIFT method originally consists in a keypoint detection phase, using a Differences of Gaussians pyramid, followed by a local description around each detected keypoint. The keypoint detection computes local maxima on a scale space generated by isotropic gaussian differences, which induces invariance to translations, rotations and partially to scaling. The descriptor then computes histograms of image gradient amplitudes, using 8 orientation bins on a 4×4 grid around each keypoint, as shown in Figure 2.2.

Dense SIFT [BZM07] bypasses the detection phase by computing the SIFT descriptors over a uniformly subsampled grid, which improves discriminability on recognition tasks [BZM07]. The resulting signal representation is locally translation invariant, thanks to the averaging created by the orientation histrograms. Moreover, it is stable to deformations and robust to changes in illumination thanks to a renormalisation of its coefficients.

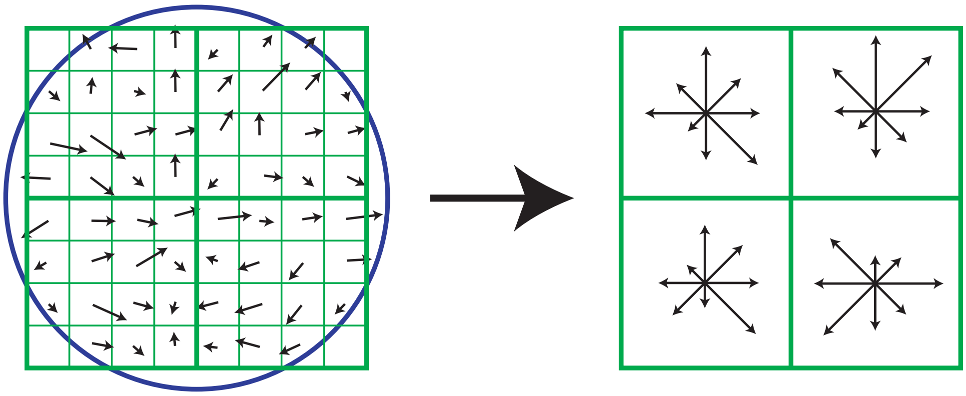


Figure 2.2: Local SIFT descriptor diagram from [Low04]. Around each detected keypoint, image gradients are computed along a grid of 16×16 pixels. The gradient amplitudes are pooled into 4 histograms, each of them consisting in 8 orientation bins. The resulting descriptor is locally stable to deformations.

SIFT has been extended and refined into a variety of similar descriptors [BETVG08; KS04]. In particular, the DAISY descriptor [TLF10] showed that SIFT coefficients can be approximated by local averages of wavelet coefficient amplitudes.

SIFT coefficients provide local image descriptors which are locally invariant to translations, and stable to additive noise and geometric and amplitude deformations. However, they operate at a relatively small scale of 2^2 pixels, which limits its invariance properties.

Histogram of Oriented Gradients (HoG) [DT05] is a similar image descriptor, which computes gradients at several image scales over a dense, overlapping grid, and pools amplitudes with similar orientations. The pooled vector is then normalized to have unit L_p norm, with $p = 1$ or $p = 2$ [DT05].

2.2.6 Convolutional Networks

Convolutional Networks [LBBH98] are a specific class of neural networks which obtain invariant signal representations by cascading trainable filter banks with non-linearities and subsampling and pooling operators. They have been successfully applied to a variety of image and audio recognition tasks [FKL10]. Each layer of the network is connected to the previous one by establishing "connections" or filters, whose output is processed with a non-linearity such as sigmoids or rectifications. The spatial localization is progressively lost by successively "pooling", or subsampling, the resulting feature maps with local averages or general L^p norms. Figure 2.3, from [LBBH98], displays a particular convolutional network.

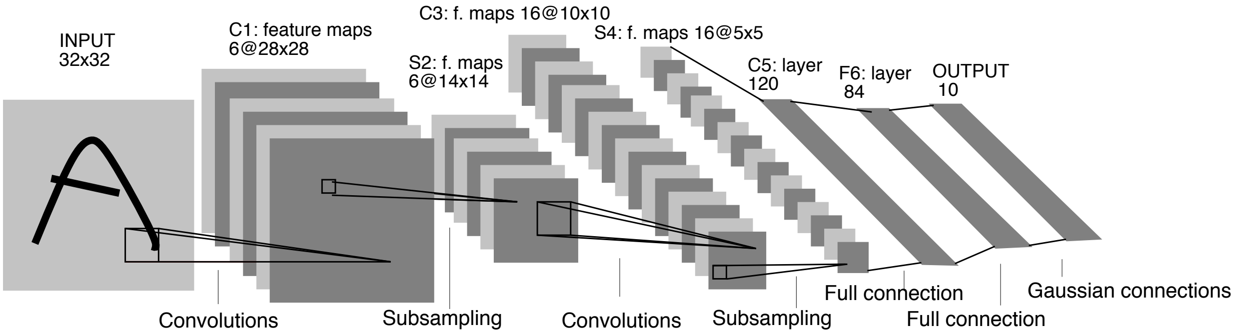


Figure 2.3: Architecture of a convolutional network from [LBBH98]. Each layer is obtained from the previous one by a filter bank convolution followed by a non-linear pooling operator.

Convolutional network architectures were originally learnt over a collection of labeled examples, using a backpropagation algorithm [LBBH98], which optimizes a classification error loss function using a gradient descent across the network. As opposed to general neural networks, convolutional networks incorporate a translation invariance prior, which greatly reduces the number of parameters to learn and hence its efficiency in a number of recognition tasks. Recently, convolutional network architectures have been trained with unsupervised data [RBL07] with *sparse autoencoders*, which learn a filter bank with the ability to encode its inputs with a sparse representation.

2.3 Scattering Review

This section reviews the Scattering transform introduced in [Mal12] and its mathematical properties. Section 2.3.1 reviews windowed scattering transforms and its construction from Littlewood-Paley wavelet decompositions. Section 2.3.2 introduces the scattering metric and reviews the scattering energy conservation property, and Section 2.3.3 reviews the Lipschitz continuity property of scattering transforms with respect to deformations. Section 2.3.4 briefly describes the integral scattering transform, and finally Section 2.3.5 presents the expected scattering transform for processes with stationary increments.

2.3.1 Windowed Scattering transform

In order to achieve stability to deformations, scattering operators are constructed from a Littlewood-Paley wavelet decomposition.

A wavelet transform is defined by dilating a mother wavelet $\psi \in \mathbf{L}^2(\mathbb{R}^d)$ with scale factors $\{a^j\}_{j\in\mathbb{Z}}$ for $a > 1$. In image processing applications one usually sets $a = 2$, whereas audio applications need smaller dilation factors, typically $a \leq 2^{1/8}$. Wavelets are not only dilated but also rotated along a discrete rotation group G of $\mathbb{R}^d$. As a result, a dilation by a^j and a rotation by $r \in G$ of ψ produce

$$\psi_{a^j r}(u) = a^{-dj}\psi(a^{-j}r^{-1}u) \ . \tag{2.5}$$

The wavelets are thus normalized in $\mathbf{L}^1(\mathbb{R}^d)$, such that $\|\psi_{a^j r}\|_1 = \|\psi\|_1$, which means that their Fourier transforms satisfy $\hat{\psi_{a^j r}}(\omega) = \hat{\psi}(a^j r\omega)$. In order to simplify notations, we denote $\lambda = a^j r \in a^{\mathbb{Z}} \times G$ and $|\lambda| = a^j$, and define $\psi_\lambda(u) = a^{-dj}\psi(\lambda^{-1}u)$. This notation will be used throughout the rest of the thesis.

Scattering operators can be defined for general mother wavelets, but of particular interest are the complex wavelets that can be written as

$$\psi(u) = e^{i\eta u}\theta(u) \ ,$$

where θ is a lowpass window whose Fourier transform is real and has a bandwidth of the order of π. As a result, after a dilation and a rotation, $\hat{\psi_\lambda}(\omega) = \hat{\theta}(\lambda\omega - \eta)$ is centered at $\lambda^{-1}\eta$ and has a support size proportional to $|\lambda|^{-1}$. In Section 2.6.1 we shall specify the wavelet families used along all numerical experiments.

A Littlewood-Paley wavelet transform is a redundant representation which computes the following filter bank, without subsampling:

$$\forall u \in \mathbb{R}^d, \forall \lambda \in a^{\mathbb{Z}} \times G \ , W_\lambda x(u) = x \star \psi_\lambda(u) = \int x(v)\psi_\lambda(u-v)dv \ . \tag{2.6}$$

If x is real and the wavelet is chosen such that $\hat{\psi}$ is also real, then $W_{-\lambda}x = W_\lambda x^*$, which implies that in that case one can assimilate a rotation r with its negative version $-r$ into an equivalence class of positive rotations $G^+ = G/\{\pm 1\}$.

A wavelet transform with a finite scale 2^J only considers the subbands λ satisfying $|\lambda| \leq 2^J$. The low frequencies which are not captured by these wavelets are recovered by a lowpass filter ϕ_J whose spatial support is proportional to 2^J: $\phi_J(u) = 2^{-dJ}\phi(2^{-J}u)$. The wavelet transform at scale 2^J thus consists in the filter bank

$$\mathcal{W}_J x = \{x \star \phi_J, (W_\lambda x)_{\lambda \in \Lambda_J}\} \ ,$$

where $\Lambda_J = \{a^j r : r \in G^+, |\lambda| \leq 2^J\}$. Its norm is defined as

$$\|\mathcal{W}_J x\|^2 = \|x \star \phi_J\|^2 + \sum_{\lambda \in \Lambda_J} \|W_\lambda x\|^2 \ .$$

$\mathcal{W}_J$ is thus a linear operator from $\mathbf{L}^2(\mathbb{R}^d)$ to a product space generated by copies of $\mathbf{L}^2(\mathbb{R}^d)$. It defines a frame of $\mathbf{L}^2(\mathbb{R}^d)$, whose bounds are characterized by the following Littlewood-Paley condition:

Proposition 2.3.1 *If there exists $\epsilon > 0$ such that for almost all $\omega \in \mathbb{R}^d$ and all $J \in \mathbb{Z}$*

$$1 - \epsilon \leq |\hat{\phi}(2^J\omega)|^2 + \frac{1}{2}\sum_{j \leq J}\sum_{r \in G} |\hat{\psi}(2^j r\omega)|^2 \leq 1 \ ,$$

then $\mathcal{W}_J$ *is a frame with bounds given by* $1-\epsilon$ *and* 1*:*

$$(1-\epsilon)\|x\|^2 \leq \|\mathcal{W}_J x\|^2 \leq \|x\|^2 \quad , x \in \mathbf{L}^2(\mathbb{R}^d) \ . \tag{2.7}$$

In particular, this Littlewood-Paley condition implies that $\hat{\psi}(0) = 0$ and hence that the wavelet must have at least a vanishing moment. When $\epsilon = 0$, the wavelet decomposition preserves the Euclidean norm and we say that it is unitary.

Wavelet coefficients are not translation invariant but translate as the input is translated, and their average $\int W_\lambda x(u)du$ does not produce any information since wavelets have zero mean. A translation invariant measure which is also stable to the action of diffeomorphisms can be extracted out of each wavelet sub-band λ, by introducing a non-linearity which restores a non-zero, informative average value. This is for instance achieved by computing the complex modulus and averaging the result

$$\int |x \star \psi_\lambda|(u)du \ .$$

The information lost by this averaging is recovered by a new wavelet decomposition $\{|x\star\psi_\lambda| \star \psi_{\lambda'}\}_{\lambda' \in \Lambda_J}$ of $|x \star \psi_\lambda|$, which produces new invariants by iterating the same procedure. Let $U[\lambda]x = |x \star \psi_\lambda|$ denote the wavelet modulus operator corresponding to the subband λ. Any sequence $p = (\lambda_1, \lambda_2, ..., \lambda_m)$ defines a *path*, i.e, the ordered product of non-linear and non-commuting operators

$$U[p]x = U[\lambda_m] \ldots U[\lambda_2]\, U[\lambda_1]x = |\ ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \ldots | \star \psi_{\lambda_m}|\ ,$$

with $U[\emptyset]x = x$.

Similarly as with frequency variables, one can manipulate path variables $p = (\lambda_1, \ldots, \lambda_m)$ in a number of ways. The scaling and rotation by $a^l g \in a^{\mathbb{Z}} \times G^+$ of a path p is denoted $a^l gp = (a^l g\lambda_1, \ldots, a^l g\lambda_m)$, and the concatenation of two paths is written $p + p' = (\lambda_1, \ldots, \lambda_m, \lambda'_1, \ldots, \lambda'_{m'})$.

Many applications in image and audio recognition require locally translation invariant representations, but which keep spatial or temporal information beyond a certain scale 2^J. A windowed scattering transform computes a locally translation invariant representation by applying a lowpass filter at scale 2^J with $\phi_{2^J}(u) = 2^{-2J}\phi(2^{-J}u)$.

Definition 2.3.2 *For each path* $p = (\lambda_1, \ldots, \lambda_m)$ *with* $\lambda_i \in \Lambda_J$ *and* $x \in \mathbf{L}^1(\mathbb{R}^d)$ *we define the windowed scattering transform as*

$$S_J[p]x(u) = U[p]x \star \phi_{2^J}(u) = \int U[p]x(v)\phi_{2^J}(u-v)\, dv\ ,$$

A Scattering transform has the structure of a convolutional network, but its filters are given by wavelets instead of being learnt. Thanks to this structure, the resulting transform is locally translation invariant and stable to deformations. The scattering representation enjoys several appealing properties described in sections 2.3.2 and 2.3.3.

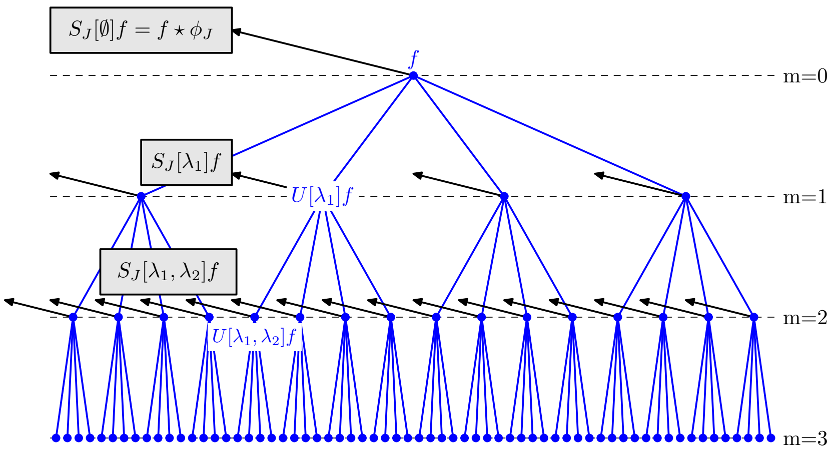


Figure 2.4: Convolutional structure of the windowed scattering transform. Each layer is computed from the previous by applying a wavelet modulus decomposition U on each envelope $U[p]f$. The outputs of each layer are obtained via a lowpass filter ϕ_J.

2.3.2 Scattering metric and Energy Conservation

The windowed scattering representation is obtained by cascading a basic propagator operator,

$$\mathcal{U}_J x = \{x \star \phi_J, (U[\lambda]x)_{\lambda \in \Lambda_J}\} . \tag{2.8}$$

The first layer of the representation applies $\mathcal{U}_J$ to the input function, whereas successive layers are obtained by applying $\mathcal{U}_J$ to each output $U[p]x$. Since $U[\lambda]U[p] = U[p+\lambda]$ and $U[p]x \star \phi_J = S_J[p]x$, it follows that

$$\mathcal{U}_J U[p]x = \{S_J[p]x, (U[p+\lambda]x)_{\lambda \in \Lambda_J}\} . \tag{2.9}$$

If Λ_J^m denotes the set of paths of length or *order* m, it follows from (2.9) that the $(m+1)$-th layer given by Λ_J^{m+1} is obtained from the previous layer via the propagator $\mathcal{U}_J$. We denote $\mathcal{P}_J$ the set of paths of any order up to scale 2^J, $\mathcal{P}_J = \cup_m \Lambda_J^m$.

The propagator $\mathcal{U}_J$ is non-expansive, since the wavelet decomposition $\mathcal{W}_J$ is non-expansive from (2.7) and the modulus is also non-expansive. As a result,

$$\|\mathcal{U}_J x - \mathcal{U}_J x'\|^2 = \|x \star \phi_J - x' \star \phi_J\|^2 + \sum_{\lambda \in \Lambda_J} \| |W_\lambda x| - |W_\lambda x'| \|^2 \leq \|x - x'\|^2 .$$

Moreover, if the wavelet decomposition is unitary, then the propagator $\mathcal{U}_J$ is also unitary.

For any path set Ω, the Euclidean norm defined by the scattering coefficients $S_J[p]$, $p \in \Omega$ is

$$\|S_J[\Omega]x\|^2 = \sum_{p \in \Omega} \|S_J[p]x\|^2 .$$

Since $S_J[\mathcal{P}_J]$ is constructed by cascading the non-expansive operator $\mathcal{U}_J$, it results that $S_J[\mathcal{P}_J]$ is also non-expansive:

Proposition 2.3.3 *The windowed scattering transform is non-expansive:*

$$\forall x, x' \in \mathbf{L}^2(\mathbb{R}^d) \ , \quad \|S_J[\mathcal{P}_J]x - S_J[\mathcal{P}_J]x'\| \leq \|x - x'\| \ . \tag{2.10}$$

The windowed scattering thus defines a metric which is continuous with respect to the $\mathbf{L}^2(\mathbb{R}^d)$ euclidean metric, and thus it is stable to additive noise. But in fact the scattering metric also preserves the signal energy, thus showing that all high-frequency information is encoded in scattering coefficients.

Theorem 2.3.4 *A scattering wavelet ψ is said to be admissible if there exists $\eta \in \mathbb{R}^d$ and $\rho \in \mathbf{L}^2(\mathbb{R}^d) \geq 0$, with $|\hat{\rho}(\omega)| \leq |\hat{\phi}(2\omega)|$, $\hat{\rho}(0) = 1$, such that the function*

$$\hat{\Psi}(\omega) = |\hat{\rho}(\omega - \eta)|^2 - \sum_{k=1}^{\infty} k \left(1 - |\hat{\rho}(2^{-k}(\omega - \eta))|^2\right)$$

satisfies

$$\inf_{1 \leq |\omega| \leq 2} \sum_{j=-\infty}^{\infty} \sum_{r \in G} \hat{\Psi}(2^{-j} r^{-1} \omega) |\hat{\psi}(2^{-j} r^{-1} \omega)|^2 > 0 \ . \tag{2.11}$$

If ψ satisfies the Littlewood-Paley condition (2.7) with $\epsilon = 0$ and is admissible, then

$$\forall x \in \mathbf{L}^2(\mathbb{R}^d) \ , \quad \lim_{m \to \infty} \|U[\Lambda_J^m]x\|^2 = \lim_{m \to \infty} \sum_{n \geq m} \|S_J[\Lambda_J^n]x\|^2 = 0 \ , \tag{2.12}$$

and

$$\|S_J[\mathcal{P}_J]x\| = \|x\| \ . \tag{2.13}$$

The proof of this theorem shows that the scattering energy propagates progressively towards the low frequencies, thanks to the demodulation effect of the complex modulus. The energy of $U[\Lambda_J^m]x$ is concentrated along scale increasing paths $p = (\lambda_1, \ldots, \lambda_m)$ with $|\lambda_i| < |\lambda_{i+1}|$, which greatly reduces the computational needs for numeric applications. The decay of $\|U[\Lambda_J^m]x\|$ also means that in practice only the first m_0 layers of the transform carry significant energy. Section 2.6 will show that most applications require at most 2 or 3 layers.

2.3.3 Local Translation Invariance and Lipschitz Continuity to Deformations

The windowed scattering metric defined in the previous section is non-expansive, which gives stability to additive perturbations. Moreover, it is also stable to the action of diffeomorphisms, and becomes translation invariant as the localization scale 2^J increases.

The limit of $\|S_J[\mathcal{P}_J]x - S_J[\mathcal{P}_J]x'\|$ as $J \to \infty$ is well defined thanks to the following non-expansive property:

Proposition 2.3.5 *For all $x, x' \in \mathbf{L}^2(\mathbb{R}^d)$ and $J \in \mathbb{Z}$,*

$$\|S_{J+1}[\mathcal{P}_{J+1}]x - S_{J+1}[\mathcal{P}_{J+1}]x'\| \leq \|S_J[\mathcal{P}_J]x - S_J[\mathcal{P}_J]x'\| \ .$$

As a result, the sequence $d_{x,x'}(J) = \|S_J[\mathcal{P}_J]x - S_J[\mathcal{P}_J]x'\|$ is positive and non-increasing as J increases, and hence it converges. This limit metric is translation invariant:

Theorem 2.3.6 *Let $x_c(u) = x(u-c)$. Then for admissible scattering wavelets satisfying (2.11) we have*

$$\forall x \in \mathbf{L}^2(\mathbb{R}^d) \ , \forall c \in \mathbb{R}^d \ , \ \lim_{J\to\infty} \|S_J[\mathcal{P}_J]x - S_J[\mathcal{P}_J]x_c\| = 0 \ . \tag{2.14}$$

But, most importantly, the windowed scattering transform defines a stable metric with respect to the action of diffeomorphisms, which can model non-rigid deformations. A diffeomorphism maps a point $u \in \mathbb{R}^d$ to $u - \tau(u)$, where $\tau(u)$ is a vector displacement field satisfying $\|\nabla\tau\|_\infty < 1$. It acts on functions $x \in \mathbf{L}^2(\mathbb{R}^d)$ by composition: $L[\tau]x(u) = x(u-\tau(u))$. The following central theorem computes an upper bound of $\|S_J[\mathcal{P}_J]L[\tau]x - S_J[\mathcal{P}_J]x\|$. For that purpose, we assume an admissible scattering wavelet, and we define the auxiliary norm

$$\|U[\mathcal{P}_J]x\|_1 = \sum_{m\geq 0} \|U[\Lambda_J^m]x\| \ .$$

Theorem 2.3.7 *There exists C such that every $x \in \mathbf{L}^2(\mathbb{R}^d)$ with $\|U[\mathcal{P}_J]x\|_1 < \infty$ and $\tau \in C^2(\mathbb{R}^d)$ with $\|\nabla\tau\|_\infty \leq 1/2$ satisfy*

$$\|S_J[\mathcal{P}_J]L[\tau]x - S_J[\mathcal{P}_J]x\| \leq C\|U[\mathcal{P}_J]x\|_1 K(\tau) \ , \tag{2.15}$$

with

$$K(\tau) = 2^{-J}\|\tau\|_\infty + \|\nabla\tau\|_\infty \max(1, \log\frac{\sup_{u,u'}|\tau(u)-\tau(u')|}{\|\nabla\tau\|_\infty}) + \|H\tau\|_\infty \ ,$$

and for all $m \geq 0$, if $\mathcal{P}_{J,m} = \cup_{n<m}\Lambda_J^n$, then

$$\|S_J[\mathcal{P}_{J,m}]L[\tau]x - S_J[\mathcal{P}_{J,m}]x\| \leq Cm\|x\|K(\tau) \ . \tag{2.16}$$

This theorem shows that the effect of a diffeomorphism produces in the scattering domain an error bounded by a term proportional to $2^{-J}\|\tau\|_\infty$, which corresponds to the local translation invariance, plus a deformation error proportional to $\|\nabla\tau\|_\infty$. When x has compact support, the following corollary shows that the windowed scattering metric is Lipschitz continuous to the action of diffeomorphisms:

Corollary 2.3.8 *For any compact set $\Omega \subset \mathbb{R}^d$ there exists C such that for all $x \in \mathbf{L}^2(\mathbb{R}^d)$ supported in Ω with $\|U[\mathcal{P}_J]x\|_1 < \infty$ and for all $\tau \in C^2(\mathbb{R}^d)$ with $\|\nabla\tau\|_\infty \leq 1/2$, then*

$$\|S_J[\mathcal{P}_{J,m}]L[\tau]x - S_J[\mathcal{P}_{J,m}]x\| \leq C\|U[\mathcal{P}_J]x\|_1 \left(2^{-J}\|\tau\|_\infty + \|\nabla\tau\|_\infty + \|H\tau\|_\infty\right) \ . \tag{2.17}$$

The translation error term, proportional to $2^{-J}\|\tau\|_\infty$, can be reduced to a second-order error term, $2^{-2J}\|\tau\|_\infty^2$, by considering a first order Taylor approximation of each $S_J[p]x$ [Mal12].

2.3.4 Integral Scattering transform

The metric defined by the windowed scattering transform converges towards a translation invariant metric as the scale of the lowpass window 2^J increases. However, the transform itself does not converge without a proper renormalisation.

In order to define the convergence of $S_J[\mathcal{P}_J]$, it is first necessary to embed the countable set of paths $\mathcal{P}_J$ into a non-countable set $\overline{\mathcal{P}}_\infty$, and to equip it with a measure and a metric. In the limit when $J \to \infty$, a path $p = (\lambda_1, \ldots, \lambda_m)$ of length m belongs to the finite product set Λ_∞^m, with $\Lambda_\infty = a^{\mathbb{Z}} \times G^+$. A path of infinite length is a sequence $(\lambda_n)_n \in \Lambda_\infty^\infty$. One can construct a product topology in Λ_∞^∞ by defining cylinders as the open sets: $C_m(\lambda) = \{(q_n)_n \in \Lambda_\infty^\infty \,:\, q_{m+1} = \lambda\}$. These sets define a sigma algebra, on which one can define a measure. If $\mathcal{P}_\infty$ denotes the set of all finite paths $\mathcal{P}_\infty = \cup_{m\geq 0}\Lambda_\infty^m$, then this sigma algebra is also generated by the cylinder sets

$$p = (\lambda_1, \ldots, \lambda_m) \in \mathcal{P}_\infty,\ C(p) = \{(q_n)_n \in \Lambda_\infty^\infty \,:\, q_1 = \lambda_1, \ldots, q_m = \lambda_m\} \ . \tag{2.18}$$

A measure on Λ_∞^∞ is constructed from the scattering of the Dirac distribution $U[p]\delta$, by defining $\mu(C(p)) = \|U[p]\delta\|^2$ for all $p \in \mathcal{P}_\infty$. A neighborhood can be defined using cylinder sets of scale 2^J:

$$C_J(p) = \bigcup_{\substack{\lambda\in\Lambda_\infty \\ |\lambda|^{-1}\leq a^{-J}}} C(p+\lambda) \subset C(p) \ . \tag{2.19}$$

These neighborhoods define a distance in Λ_∞^∞:

$$d(q,\tilde{q}) = \inf_{q,\tilde{q}\in C_J(p)} \mu(C_J(p)) \ ,$$

but Λ_∞^∞ is not complete with this metric. Its completion is achieved by embedding the set $\mathcal{P}_\infty$ of finite paths, $\overline{\mathcal{P}}_\infty = \Lambda_\infty^\infty \cup \mathcal{P}_\infty$, defined by adding each $p \in \mathcal{P}_\infty$ to the cylinders $C(p)$ and $C_J(p)$ without modifying their measure.

An integral scattering transform can now be constructed to be an element of $\mathbf{L}^2(\overline{\mathcal{P}}_\infty, d\mu)$. For that purpose, we define

$$S_J x(q,u) = \sum_{p\in\mathcal{P}_J} \frac{S_J[p]x(u)}{\|S_J[p]\delta\|} \mathbf{1}_{C_J(p)}(q) \ , q \in \overline{\mathcal{P}}_\infty \ ,$$

where $\mathbf{1}_{C_J(p)}(q)$ is the indicator function of $C_J(p)$ in $\overline{\mathcal{P}}_\infty$. This extension can be seen as a scattering energy density on $\overline{\mathcal{P}}_\infty \times \mathbb{R}^d$. It has a spatial resolution of 2^{-J} along its spatial coordinate, and a frequency resolution 2^J along its path variable q. As $J \to \infty$, we seek to define a scattering transform which depends only upon $q \in \overline{\mathcal{P}}_\infty$. This is achieved by first defining the marginals

$$\forall q \in \overline{\mathcal{P}}_\infty \ , \overline{S}_J x(q) = \left(\int |S_J x(q,u)|^2 dx \right)^{1/2} = \sum_{p\in\mathcal{P}_J} \frac{\|S_J[p]x\|}{\|S_J[p]\delta\|} \mathbf{1}_{C_J(p)}(q) \ .$$

$\overline{S}_J$ is a non-expansive operator of $\mathbf{L}^2(\mathbb{R}^d)$ which preserves the norm, and with the property that $\|\overline{S}_J x - \overline{S}_J x'\|_{\overline{\mathcal{P}}}$ is non-decreasing and bounded as $J \to \infty$. For $x \in \mathbf{L}^1(\mathbb{R}^d)$, $\overline{S}_J$ converges pointwise in $\overline{\mathcal{P}}_\infty$ to an integral scattering transform:

Proposition 2.3.9 *If $f \in \mathbf{L}^1(\mathbb{R}^d)$ then*

$$\forall p \in \overline{\mathcal{P}}_\infty \ , \ \lim_{J\to\infty} \overline{S}_J x(p) = \frac{1}{\mu_p} \int U[p]x(u)du \ ,$$

with $\mu_p = \int U[p]\delta(u)du$.

Finally, one can extend the integral scattering to $\overline{\mathcal{P}}_\infty$ for functions in $\mathbf{L}^2(\mathbb{R}^d)$, as the limit of windowed normalized scattering:

$$\forall q \in \overline{\mathcal{P}}_\infty \ , \ \overline{S}x(q) = \liminf_{J\to\infty} \overline{S}_J x(q) \ . \tag{2.20}$$

In [Mal12], one can find sufficient conditions for which $\overline{S}_J x$ converges strongly to $\overline{S}x$, which then preserves the $\mathbf{L}^2(\mathbb{R}^d)$ norm of x.

2.3.5 Expected Scattering for Processes with stationary increments

This section reviews the definitions and basic properties of the expected scattering of random processes [Mal12]. The role of the $\mathbf{L}^2(\mathbb{R}^d)$ norm in the deterministic setting is replaced by the mean square norm $E(|X|^2)^{1/2}$.

If $X(t)$ is a stationary process or has stationary increments, meaning that $\delta_s X(t) = X(t) - X(t-s)$ is stationary for all s, then $X \star \psi_\lambda$ is also stationary, and taking the modulus preserves stationarity. It results that for any path $p = (\lambda_1, ..., \lambda_m) \in \overline{\mathcal{P}}_\infty$, the process

$$U[p]X = |...|X \star \psi_{\lambda_1}| \star ...| \star \psi_{\lambda_m}|$$

is stationary, hence its expected value does not depend upon the spatial position t.

Definition 2.3.10 *The expected scattering of X is defined for all $p \in \overline{\mathcal{P}}_\infty$ by*

$$\overline{S}X(p) = E(U[p]X) = E(|...|X \star \psi_{\lambda_1}| \star ...| \star \psi_{\lambda_m}|) \ .$$

The expected scattering defines a representation for the process $X(t)$ which carries information on high order moments of $X(t)$, as we shall see in later sections. It also defines a metric between stationary processes, given by

$$\|\overline{S}X - \overline{S}Y\|^2 = \sum_{p \in \overline{\mathcal{P}}_\infty} |\overline{S}X(p) - \overline{S}Y(p)|^2 \ .$$

The scattering representation of $X(t)$ is estimated by computing a windowed scattering transform of a realization x of $X(t)$. If $\Lambda_J = \{\lambda = 2^j \ ; \ 2^{-j} > 2^{-J}\}$ denotes the set of scales smaller than J, and $\mathcal{P}_J$ is the set of finite paths $p = (\lambda_1, .., \lambda_m)$ with $\lambda_k \in \Lambda_J\ \forall k$, then the windowed scattering at scale J of a realization $x(t)$ is

$$S_J[\mathcal{P}_J]x = \{U[p]x \star \phi_J\, ,\, p \in \mathcal{P}_J\}\ .$$

Since $\int \phi_J(u)du = 1$, we have $E(S_J[\mathcal{P}_J]X) = E(U[p]X) = \overline{S}X(p)$, so S_J is an unbiased estimator of the scattering coefficients contained in $\mathcal{P}_J$.

When the wavelet ψ satisfies the Littlewood-Paley condition (2.7), the non-expansive nature of the operators defining the scattering transform implies that $\overline{S}$ and $S_J[\mathcal{P}_J]$ are also non-expansive:

Proposition 2.3.11 *If X and Y are finite second order processes with stationary increments, then*

$$E(\|S_J[\mathcal{P}_J]X - S_J[\mathcal{P}_J]Y\|^2) \;\leq\; E(|X - Y|^2)\ , \tag{2.21}$$

$$\|\overline{S}X - \overline{S}Y\|^2 \;\leq\; E(|X - Y|^2)\ , \tag{2.22}$$

in particular

$$\|\overline{S}X\|^2 \leq E(|X|^2)\ . \tag{2.23}$$

The $\mathbf{L}^2(\mathbb{R}^d)$ energy conservation theorem (2.3.4) yields an equivalent energy conservation property for the mean squared power:

Theorem 2.3.12 *If the wavelet ψ satisfies an admissibility condition (2.11), and if X is stationary, then*

$$E(\|S_J[\mathcal{P}_J]X\|^2) = E(|X|^2)\ . \tag{2.24}$$

Expected scattering coefficients are estimated with the windowed scattering $S_J[p]X = U[p]X \star \psi_J$ for each $p \in \mathcal{P}_J$. If $U[p]X$ is ergodic, $S_J[p]X$ converges in probability to $\overline{S}X(p) = E(U[p]X)$ when $J \to \infty$. A process $X(t)$ with stationary increments is said to have a mean squared consistent scattering if the total variance of $S_J[\mathcal{P}_J]X$ converges to zero as J increases:

$$\lim_{J\to\infty} E(\|S_J[\mathcal{P}_J]X - \overline{S}X\|^2) = \sum_{p\in\mathcal{P}_J} E(|S_J[p]X - \overline{S}X(p)|^2) = 0\ . \tag{2.25}$$

This condition implies that $S_J[\mathcal{P}_J]X$ converges to $\overline{S}X$ with probability 1. Mean square consistent scattering is observed numerically on a variety of processes, including gaussian and non-gaussian fractal processes. It is conjectured in [Mal12] that Gaussian stationary processes X whose autocorrelation R_X is in $\mathbf{L}^1$ have a mean squared consistent scattering. As a consequence of Theorem 2.3.12, mean squared consistency implies an expected scattering energy conservation:

Corollary 2.3.13 *For admissible wavelets as in Theorem 2.3.12, $S_J[\mathcal{P}_J]X$ is mean squared consistent if and only if*

$$\|\overline{S}X\|^2 = E(|X|^2)\ .$$

Expected scattering coefficients depend upon normalized high order moments of X. If one expresses $|U[p]X|^2$ as

$$|U[p]X(t)|^2 = E(|U[p]X|^2)(1 + \epsilon(t)) \ ,$$

then, assuming $|\epsilon| \ll 1$, a first order approximation of

$$U[p]X(t) = \sqrt{|U[p]X(t)|^2} \approx E(|U[p]X|^2)^{1/2}(1 + \epsilon/2)$$

yields

$$U[p+\lambda]X = |U[p]X \star \psi_\lambda| \approx \frac{||U[p]X|^2 \star \psi_\lambda|}{2E(|U[p]X|^2)^{1/2}} \ ,$$

thus showing that $\overline{S}X(p) = E(U[p]X)$ for $p = (\lambda_1, \ldots, \lambda_m)$ depends upon normalized moments of X of order 2^m, determined by the cascade of wavelet sub-bands λ_k. As opposed to a direct estimation of high moments, scattering coefficients are computed with a non-expansive operator which allows consistent estimation with few realizations. In Chapter 4 we shall see that this is a fundamental property which enables texture recognition and classification from scattering representations.

The scattering representation is related to the sparsity of the process through the decay of its coefficients $\overline{S}X(p)$ as the order $|p|$ increases. Indeed, the ratio of the first two moments of X

$$\rho_X = \frac{E(|X|)}{E(|X|^2)^{1/2}}$$

gives a rough measure of the fatness of the tails of X.

For each p, the Littlewood-Paley unitarity condition satisfied by ψ gives

$$E(|U[p]X|^2) = E(U[p]X)^2 + \sum_\lambda E(|U[p+\lambda]X|^2) \ ,$$

which yields

$$1 = \rho_{U[p]X} + \frac{1}{E(|U[p]X|^2)} \sum_\lambda E(|U[p+\lambda]X|^2) \ . \tag{2.26}$$

Thus, the fraction of energy that is trapped at a given path p is given by the relative sparsity $\rho_{U[p]X}$.

This relationship between sparsity and scattering decay across the orders is of particular importance for the study of point processes, which are sparse in the original spatial domain, and for regular image textures, which are sparse when decomposed in the first level UX of the transform. In particular, the scattering transform can easily discriminate between white noises of different sparsity, such as Bernouilli and Gaussian.

The autocovariance of a real stationary process X is denoted

$$RX(\tau) = E\Big((X(x) - E(X))\,(X(x-\tau) - E(X))\Big) \ .$$

Its Fourier transform $\widehat{R}X(\omega)$ is the power spectrum of X. Replacing X by $X \star \psi_\lambda$ in the conservation energy formula (2.3.12) implies that

$$\sum_{p \in \mathcal{P}_J} E(|S_J[p+\lambda]X|^2) = E(|X \star \psi_\lambda|^2) \ . \tag{2.27}$$

These expected squared wavelet coefficients can also be written as a filtered integration of the Fourier power spectrum $\widehat{R}X(\omega)$

$$E(|X \star \psi_\lambda|^2) = \int \widehat{R}X(\omega)\, |\hat{\psi}(\lambda^{-1}\omega)|^2 \, d\omega \ . \tag{2.28}$$

These two equations prove that summing scattering coefficients recovers the power spectrum integral over each wavelet frequency support, which only depends upon second-order moments of X. However, scattering coefficients $\overline{S}X(p)$ depend upon moments of X up to the order 2^m if p has a length m. Scattering coefficients can thus discriminate textures having same second-order moments but different higher-order moments.

2.4 Characterization of Non-linearities

This section characterizes from a stability point of view the nonlinearities necessary in any invariant signal representation in order to produce its locally invariant coefficients; and in particular in scattering representations.

Every stable, locally invariant signal representation incorporates a non-linear operator in order to produce its coefficients. Neural networks, and in particular convolutional networks, introduce rectifications and sigmoids at the outputs of its "hidden units", whereas SIFT descriptors compute the norm of the filtered image gradient prior to its pooling into the local histograms.

Filter bank outputs are by definition translation covariant, not invariant. Indeed, if $y(u) = x \star h(u)$, then a translation of the input $x_c(u) = x(u-c)$ produces a translation in the output by the same amount, $x_c \star h(u) = y(u-c) = y_c$.

Translation invariant measures require non-linear operators because the only linear measurement which is translation invariant is the signal average. Indeed, the following proposition shows that a bounded, translation invariant linear operator in $\mathbf{L}^2(\mathbb{R}^d) \cap \mathbf{L}^1(\mathbb{R}^d)$ is a multiple of the signal average operator $\int x(u)du$. We denote $T_c x(u) = x(u-c)$, $c \in \mathbb{R}^d$ the translation operator.

Proposition 2.4.1 *Let Q be a linear functional of $\mathbf{L}^2(\mathbb{R}^d) \cap \mathbf{L}^1(\mathbb{R}^d)$ such that*

$$\forall c \in \mathbb{R}^d \ , \forall x \in \mathbf{L}^2(\mathbb{R}^d) \cap \mathbf{L}^1(\mathbb{R}^d) \ , Qx = QT_c x$$

and bounded in $\mathbf{L}^2(\mathbb{R}^d)$: $\|Qx\|_2 \leq C\|x\|_2$ for all $x \in \mathbf{L}^2(\mathbb{R}^d) \cap \mathbf{L}^1(\mathbb{R}^d)$. Then $Qx = C(\int x(u)du)$.

Proof: Let $(\mathbf{V}_k)_{k\in\mathbb{Z}}$ be a multiresolution analysis generated by a scaling function $\varphi \in \mathbf{L}^2(\mathbb{R}^d) \cap \mathbf{L}^1(\mathbb{R}^d)$, and fix $j \in \mathbb{Z}$. Let us first prove the result for $x_j = P_{\mathbf{V}_j} x$. x_j can be written using the orthonormal basis $\{\varphi_j(u - k2^j)\}_{k\in\mathbb{Z}}$, with $\varphi_j(u) = 2^{-jd/2}\varphi(2^{-j}u)$:

$$x_j(u) = \sum_k c_k \varphi_j(u - 2^j k) \ ,$$

with $c_k = \langle x(u), \varphi_j(u - 2^j k)\rangle$. But

$$Qx_j(u) = \sum_k c_k Q\varphi_j(u - 2^j k) = Q\varphi_j(u) \sum_k c_k \ , \tag{2.29}$$

thanks to the fact that Q is linear and $Q\varphi_j(u - 2^j k) = QT_{2^j k}\varphi_j = Q\varphi_j$. Moreover,

$$\begin{aligned} \int x_j(u)du &= \int \sum_k c_k \varphi_j(u - 2^j k)du \\ &= \sum_k c_k \left(\int \varphi_j(u - 2^j k)du \right) = \left(\sum_k c_k \right) \int \varphi_j(u)du \ , \end{aligned}$$

which implies, by substituting in (2.29), that

$$Qx_j = Q(\varphi_j) \left(\int \varphi_j(u)du \right)^{-1} \left(\int x_j(u)du \right) = C \left(\int x_j(u)du \right) \ ,$$

where C only depends upon the scaling function and the resolution. We finally extend the result to $\mathbf{L}^2(\mathbb{R}^d) \cap \mathbf{L}^1(\mathbb{R}^d)$ with a density argument. Given $x \in \mathbf{L}^2(\mathbb{R}^d) \cap \mathbf{L}^1(\mathbb{R}^d)$ and $\epsilon > 0$, there exists a resolution j such that $\|x - P_{V_j}x\| \leq \epsilon$. Let $x_j = P_{V_j}x$. Since $\int x(u)du = \int x_j(u)du$, it follows that

$$\begin{aligned} \|Qx - Qx_j\| &= \|Qx - C\int x(u)du\| \\ &= \|Q(x - x_j)\| \leq \|Q\|\|x - x_j\| \leq \|Q\|\epsilon \ , \end{aligned}$$

which concludes the proof since Q is a bounded operator $\Box$.

The local version of this result characterizes linear operators $\tilde{Q}$ which keep spatial localization at scale 2^J and which are locally translation invariant for displacements $c \in \mathbb{R}^d$ with $|c| \ll 2^J$:

$$\|\tilde{Q}x - \tilde{Q}T_c x\| \leq C\|x\|2^{-J}|c| \ . \tag{2.30}$$

If $\tilde{Q}$ is implemented as a convolution with a filter q (possibly followed by a downsampling), then (2.30) implies that the operator

$$x \mapsto q \star h_c \star x \ ,$$

with $h_c(u) = \delta(u) - \delta(u - c)$, must satisfy

$$\forall c \in \mathbb{R}^d \ , \sup_\omega |\hat{q}(\omega)||\hat{h_c}(\omega)| \leq 2^{-J}|c| \ . \tag{2.31}$$

Since $|\hat{h_c}(\omega)|^2 = 2|1-\cos(\omega.c)|^2$, (2.31) forces $\hat{q}$ to have its energy concentrated at frequencies $\omega \leq C2^{-J}$, which characterizes local smoothing kernels.

Translation invariant measures can be obtained by integrating any operator M of $\mathbf{L}^2(\mathbb{R}^d)$ along the orbit generated by the translation group:

$$\overline{M}x = \int MT_c x d\mu(c) \ .$$

Here, $T_c x(u) = x(u-c)$ is the translation operator by $c \in \mathbb{R}^d$ and $d\mu$ is a left-invariant Haar measure of the translation group. If the operator M commutes with translations, then $\overline{M}x$ becomes the average of the function Mx:

$$\overline{M}x = \int MT_c x d\mu(c) = \int Mx(u)du \ .$$

Non-linear operators commuting with translations thus give a systematic procedure to obtain translation invariant measures in a convolutional network architecture. As mentioned in Section 2.2.1, besides local translation invariance it is fundamental to enforce stability to additive noise and to the action of diffeomorphisms. Additive stability is guaranteed by imposing non-linear operators which are non-expansive.

In order to preserve the overall stability of the network to the action of diffeomorphisms, one can ask the non-linearities to not only commute with translations, but with any diffeomorphism. This property puts all the geometric stability requirements into the design of the filter bank. Indeed, if $\mathcal{W}$ is a filter bank stable to the action of diffeomorphisms, and M is a non-expansive operator commuting with such diffeomorphisms, then $\mathcal{W}M$ is also stable:

$$\|\mathcal{W}ML[\tau]x - \mathcal{W}Mx\| = \|\mathcal{W}L[\tau]Mx - \mathcal{W}Mx\| \leq C\|Mx\|\|\tau\| \leq C\|x\|\|\tau\| \ .$$

Moreover, the commutation property of M preserves the geometrical information encoded by the filters. The following theorem proves that non-linear operators of $\mathbf{L}^2(\mathbb{R}^d)$ which are non-expansive and which commute with the action of diffeomorphisms are necessarily point-wise.

Theorem 2.4.2 *If M is an operator of $L^2(\mathbb{R}^d)$ which is non-expansive, ie $\|Mf - Mg\| \leq \|f-g\|$, and commutes with the action of diffeomorphisms, then M is a pointwise operator: $Mf(u) = \rho(f(u))$ almost everywhere.*

Proof: Let $\mathbf{1}_\Omega$ be the indicator of a compact ball $\Omega \subset \mathbb{R}^d$. Let us first show that $M\mathbf{1}_\Omega = \rho\mathbf{1}_\Omega$. Let $\phi \in Diff(\mathbb{R}^d)$ be a diffeomorphism of $\mathbb{R}^d$. For $f \in \mathbf{L}^2(\mathbb{R}^d)$, we denote $L_\phi f = f \circ \phi$. Given $f \in \mathbf{L}^2(\mathbb{R}^d)$, let

$$G(f) = \{\phi \in Diff(\mathbb{R}^d)\ , L_\phi f = f\}$$

denote the isotropy group of f, ie the subgroup of diffeomorphisms leaving f unchanged up to a set of zero measure. If $\phi \in G(f)$, then

$$\|Mf - L_\phi Mf\| = \|Mf - ML_\phi f\| \leq \|f - L_\phi f\| = 0 \ ,$$

which means that $\phi \in G(M(f))$ too.

If $f = c\mathbf{1}_\Omega$, then its isotropy group contains any diffeomorphism ϕ satisfying

$$\phi(\Omega) = \Omega\,, \quad \phi(\overline{\Omega}) = \overline{\Omega}\ ,$$

where $\overline{\Omega} = \mathbb{R}^d - \Omega$. Thus, Mf is also invariant to the action of any ϕ satisfying the above conditions. It results that Mf must also be constant within both Ω and $\overline{\Omega}$ up to a set of zero measure. Indeed, otherwise we could find two subsets $I_1, I_2 \subset \Omega$ of strictly positive measure $\mu(I_1) = \mu(I_2) > 0$, such that

$$\int_{I_1} Mf(x)d\mu(x) \neq \int_{I_2} Mf(x)d\mu(x)\ ,$$

but then a diffeomorphism ϕ such that $\phi \in G(\mathbf{1}_\Omega)$ and mapping I_1 to I_2, does not satisfy $\|Mf - L_\phi Mf\|_2 = 0$, which is a contradiction.

Since Mf belongs to $\mathbf{L}^2(\mathbb{R}^d)$ and $\overline{\Omega}$ has infinite measure, it results that $Mf(x) = 0\ \forall x \in \overline{\Omega}$, and hence

$$M(c\mathbf{1}_\Omega) = \rho(c,\Omega)\mathbf{1}_\Omega\ ,$$

with $\rho(c,\Omega) = (Mc\mathbf{1}_\Omega)(x_0)$ for any $x_0 \in \Omega$. Since the hypercube Ω can be obtained from the unit ball Ω_0 of $\mathbb{R}^d$ with a similarity transform T_Ω , $\Omega = T_\Omega\Omega_0$, we have $M(c\mathbf{1}_\Omega) = M(T_\Omega c\mathbf{1}_{\Omega_0}) = T_\Omega M(c\mathbf{1}_{\Omega_0})$, which shows that $\rho(c,\Omega)$ does not depend upon Ω, and we shall write it $\rho(c)$.

Let us now consider $f \in C^\infty$ with compact support Ω. Fix a point $x_0 \in \Omega$. We consider a sequence of diffeomorphisms $(\phi_n)_{n\in\mathbb{N}}$ which progressively warp f towards $f(x_0)\mathbf{1}_\Omega$:

$$\lim_{n\to\infty} \|L_{\phi_n} f - f(x_0)\mathbf{1}_\Omega\| = 0\ , \tag{2.32}$$

For that purpose, we construct ϕ_n such that $\phi_n(x) = x$ for $x \in \overline{\Omega}$ for all n, and such that it maps a neighborhood of radius 2^{-n} of x_0 to the set $\Omega_n \subset \Omega$ defined as

$$\Omega_n = \{x \in \Omega\ , \mathrm{dist}(x, \overline{\Omega}) \geq 2^{-n}\}\ .$$

Thanks to the fact that the domain Ω is regular, such diffeormorphisms can be constructed for instance by expanding the rays departing from x_0 at the neighborhood of x_0 and contracting them as they approach the border $\partial\Omega$. Since f is C^∞ and it is compactly supported, it is bounded, and hence

$$\begin{aligned}\|L_{\phi_n}(Mf) - M(f(x_0)\mathbf{1}_\Omega)\| &= \|M(L_{\phi_n}f) - M(f(x_0)\mathbf{1}_\Omega)\| \\ &\leq \|L_{\phi_n}f - f(x_0)\mathbf{1}_\Omega\|\ ,\end{aligned}$$

and it results from (2.32) that $\lim_{n\to\infty} L_{\phi_n}(Mf) = M(f(x_0)\mathbf{1}_\Omega)$ in $\mathbf{L}^2(\mathbb{R}^d)$. Since the diffeomorphisms ϕ_n expand the neighborhood of x_0 and $M(f(x_0)\mathbf{1}_\Omega) = \rho(f(x_0))\mathbf{1}_\Omega$, then necessarily $Mf(x_0) = M(f(x_0)\mathbf{1}_\Omega)(x_0)$, and hence $Mf(x_0) = \rho(f(x_0))$, which only depends upon the value of f at x_0.

Since C^∞, compact support functions are dense in $\mathbf{L}^2(\mathbb{R}^d)$ and M is Lipschitz continuous, for any $f \in \mathbf{L}^2(\mathbb{R}^d)$ and $\epsilon > 0$ we can find $f_0 \in C^\infty$ such that

$$\|Mf - Mf_0\| = \|f - f_0\| < \epsilon ,$$

and hence Mf can be approximated by a pointwise operator with arbitrary precision, and as a result $Mf(x) = \rho(f(x))$ almost everywhere for all $f \in \mathbf{L}^2(\mathbb{R}^d)$. □

Point-wise non-linearities are thus necessary to preserve the stability to additive noise and to the action of diffeomorphisms, while keeping all the geometrical information. If moreover one wishes a unitary signal representation, $\|Mx\| = \|x\|$, then $|\rho(y)| = |y|\,,\ \forall y$, which is obtained for instance by choosing $Mx = |x|$, and corresponds to the choice in the scattering decomposition.

The point-wise characterization requires both the non-expansive property and the commutation with respect to the action of diffeomorphisms. Indeed, one can find counterexamples whenever each of these conditions is dropped. An operator of the form $Mf = \sup_u \rho(f(u))$, where ρ is pointwise, commutes with diffeomorphisms, but fails to be non-expansive. On the other hand, the commutation property on the whole group of diffeomorphisms seems to be necessary in order to characterize point-wise operators. For instance, if one relaxes the commutation condition to the subgroup of diffeomorphisms given by the affine group, then a counter-example by I. Waldspurger constructs an operator M which is non-expansive and commutes with affine transformations, but is not point-wise. It dilutes an element $f \in \mathbf{L}^2(\mathbb{R}^d)$ by expanding its support progressively.

Max-pooling is a very popular non-linear pooling operator used in several object recognition architectures [BBLP10]. It computes local maxima values within a neighborhood of a given size:

$$M_P x(u) = \max_{|u'-u| \leq R} |x(u')| \ .$$

If max-pooling is followed by a downsampling using a critical downsampling step R, then the resulting non-linear operator is non-expansive, but fails to commute with the action of diffeomorphisms.

2.5 On the L_1 continuity of Integral Scattering

Section 2.3.4 reviewed the extension of the windowed scattering to an integral, translation invariant transform, defined on an uncountable path set $\overline{\mathcal{P}}_\infty$. In [Mal12], it was observed that this integral scattering transform shares some striking resemblance with the Fourier transform. In particular, it was conjectured that when $x \in \mathbf{L}^1(\mathbb{R}^d)$, then $\overline{S}x$ is continuous with respect to the metric defined from the cylinder topology.

In this section, we prove a partial affirmative result of this conjecture. The cylinder topology from (2.18) defines neighborhoods in $\overline{\mathcal{P}}_\infty$, formed by finite and infinite sequences $(q_k)_k$ of subband indicies $q_k \in \Lambda_\infty$, satisfying constraints on its first terms.

The $\mathbf{L}_1$ continuity of the integral scattering states that when $x \in \mathbf{L}^1(\mathbb{R}^d)$, then for each $p \in \overline{\mathcal{P}}_\infty$ and each $\epsilon > 0$, one can find $J > 0$ such that the neighborhood $C_J(p)$

satisfies

$$\forall q \in C_J(p) \ , \ |\overline{S}x(q) - \overline{S}x(p)| \leq \epsilon \ . \tag{2.33}$$

We shall prove a weaker version of this result, which considers a subset of $\overline{\mathcal{P}}_\infty$ formed by finite paths of bounded slope. Let $p = (\lambda_1, \ldots, \lambda_m) \in \mathcal{P}_\infty$ be a path of finite order. Its slope is defined as

$$\Delta(p) = \sup_k \sup_{k'>k} \frac{|\lambda_k|}{|\lambda_{k'}|} \ .$$

Scale increasing paths, which concentrate most of the scattering energy, satisfy $\Delta(p) < 0$ since $|\lambda_{k'}|^{-1} < |\lambda_k|^{-1}$, $\forall k' > k$. The following theorem proves that when one approaches a path $p \in \mathcal{P}_\infty$ with paths $q \in C_J(p)$ of bounded slope and finite order, then $\overline{S}x(q)$ converges towards $\overline{S}x(p)$ for $x \in \mathbf{L}^1(\mathbb{R}^d)$. For simplicity, we write the result using dyadic wavelets ψ_λ obtained with $a = 2$: $\lambda = r2^j$.

Theorem 2.5.1 *Let $x \in \mathbf{L}^1(\mathbb{R}^d)$, and let $\overline{S}x$ be the integral scattering transform*

$$\overline{S}x(p) = \frac{1}{\mu_p} \int U[p]x(u)du \ ,$$

where $\mu_p = \int U[p]\delta(u)du$. Then, for any $p \in \mathcal{P}_\infty$, B , $m \in \mathbb{N}$, and $\epsilon > 0$, there exists $J > 0$ such that for any $q \in \mathcal{P}_\infty$ satisfying $q \in C_J(p)$, $|q| \leq m$ and $\Delta(q) \leq B$, then

$$|\overline{S}x(p) - \overline{S}x(q)| \leq \epsilon \ . \tag{2.34}$$

Proof: Fix $p \in \mathcal{P}_\infty$, and let $q \in C_J(p) \cap \mathcal{P}_\infty$ be a path in the neighborhood of p. We can thus write $q = p + \tilde{q}$, with $\tilde{q} = \tilde{\lambda} + \check{q} \in \mathcal{P}_\infty$ satisfying $|\tilde{\lambda}|^{-1} \leq 2^{-J}$. We have

$$\begin{aligned} \overline{S}x(q) &= \frac{\int U[q]x(u)du}{\int U[q]\delta(u)du} = \frac{\int U[\tilde{q}]U[p]x(u)du}{\int U[\tilde{q}]U[p]\delta(u)du} \\ &= \frac{\int U[\tilde{q}]U[p]x(u)du}{\int U[\tilde{q}]\delta(u)du} \cdot \frac{\int U[\tilde{q}]\delta(u)du}{\int U[\tilde{q}]U[p]\delta(u)du} \\ &= \overline{S}(U[p]x)(\tilde{q}) \cdot \big(\overline{S}(U[p]\delta)(\tilde{q})\big)^{-1} \ . \end{aligned} \tag{2.35}$$

The following lemma proves that if x is in $\mathbf{L}^1(\mathbb{R}^d)$ and is positive, then $\overline{S}x$ has a particularly simple form on "small" paths $\tilde{q} \in \mathcal{P}_\infty$ with finite order and finite excursion:

Lemma 2.5.2 *Let m, $B \in \mathbb{N}$, and let*

$$\mathcal{A}_{J,m} = \{q \in \overline{\mathcal{P}}_\infty \, ; \, q = (\lambda_1, \ldots, \lambda_m), \, |q| = m, \, |\lambda_1| = 2^J, \, \Delta(q) \leq M\} \ . \tag{2.36}$$

If $x \in \mathbf{L}^1(\mathbb{R}^d)$, $x \geq 0$, then

$$\lim_{J\to\infty} \sup_{q \in \mathcal{A}_{J,m}} \left| \overline{S}x(q) - \int x(u)du \right| = 0 \ . \tag{2.37}$$

If we apply Lemma 2.5.2 to $f_1 = U[p]x$ and $f_2 = U[p]\delta$, then the identity (2.35) implies that for any $\epsilon > 0$ there exists $J > 0$ such that

$$\forall q \text{ s.t. } q \in C_J(p),\, |q| \leq m,\, \Delta(q) \leq B \ , \left| \overline{S}x(q) - \frac{\int U[p]x(u)du}{\int U[p]\delta(u)du} \right| \leq \epsilon \ ,$$

which implies (2.34) since $\overline{S}x(p) = (\int U[p]\delta(u)du)^{-1} \int U[p]f(u)du$.

We shall then prove (2.37). Fix $J > 0$, and let $q \in \mathcal{A}_{J,m}$. By definition (2.36), we can write $q = r2^J + \tilde{q}$, and without loss of generality, we can assume that $r = 1$. let $D_j x(u) = 2^{-jd} x(2^{-j}u)$ be a dilation operator normalized in $\mathbf{L}^1(\mathbb{R}^d)$. A change of variables shows that

$$\begin{aligned} D_j x \star \psi_\lambda(u) &= 2^{-jd} \int x(2^{-j}v)\psi_\lambda(u-v)dv \\ &= \int x(v)\psi_\lambda(u - 2^j v)dv = \int x(v)\psi_\lambda(2^j(2^{-j}u - v))dv \\ &= 2^{2^{-jd}} \int x(v)\psi_{2^{-j}\lambda}(2^{-j}u - v)dv \\ &= D_j(x \star \psi_{2^{-j}\lambda})(u) \ , \end{aligned} \tag{2.38}$$

and by cascading this property we obtain that

$$U[p]D_j x = D_j U[2^{-j}p]x \ ,$$

or equivalently $U[p]x = D_j U[2^{-j}p]D_{-j}x$. By setting $j = J$, we obtain

$$\overline{S}x(2^J + \tilde{q}) = \overline{S}D_{-J}x(1 + \tilde{q}2^{-J}) = \frac{\int U[1+\tilde{q}2^{-J}]D_{-J}x(u)du}{\int U[1+\tilde{q}2^{-J}]\delta(u)du} \ , \tag{2.39}$$

since $D_j\delta = \delta \ \forall j$ with the $\mathbf{L}^1(\mathbb{R}^d)$ normalization. Now, if $\overline{x} = \int x(u)du$, (2.39) can be decomposed as

$$\begin{aligned} \overline{S}x(2^J + \tilde{q}) &= \\ &= \frac{\int \overline{x}U[1+\tilde{q}2^{-J}]\delta(u)du}{\int U[1+\tilde{q}2^{-J}]\delta(u)du} + \frac{\int \left(U[1+\tilde{q}2^{-J}]D_{-J}x(u) - \overline{x}U[1+\tilde{q}2^{-J}]\delta(u)\right) du}{\int U[1+\tilde{q}2^{-J}]\delta(u)du} \\ &= \overline{x} + \frac{\int \left(U[1+\tilde{q}2^{-J}]D_{-J}x(u) - U[1+\tilde{q}2^{-J}]\overline{x}\delta(u)\right) du}{\int U[1+\tilde{q}2^{-J}]\delta(u)du} \ , \end{aligned} \tag{2.40}$$

The path $2^{-J}q = 1+\tilde{q}2^{-J}$ is obtained by a translation in scale of q, and hence it satisfies $|2^{-J}q| = |q|$ and $\Delta(2^{-J}q) = \Delta(q)$. We will prove (2.37) by showing that

$$\inf_{q \in \mathcal{A}_{1,m}} \int U[q]\delta(u)du > 0 \ , \tag{2.41}$$

and

$$\lim_{J\to\infty} \sup_{q\in\mathcal{A}_{1,m}} \left| \int \left(U[q]D_{-J}x(u) - U[q]a\delta(u)\right) du \right| = 0 \ . \tag{2.42}$$

Let us first prove (2.41), by induction on the maximum path order m.

Let $m = 2$. In that case, the set $\mathcal{A}_{1,2}$ contains paths $q = (1, \lambda)$, where the scale of λ is lower bounded by $|\lambda|^{-1} \leq M$. We need to see that

$$\inf_{|\lambda|^{-1} \leq M} \int ||\psi| \star \psi_\lambda|(u)du = \||\psi| \star \psi_\lambda\|_1 > 0 \ .$$

From (2.38) we deduce that if $j = |\lambda|$, then

$$\||\psi| \star \psi_\lambda\|_1 = \|D_j(D_{-j}|\psi| \star \psi)\|_1 = \|D_{-j}|\psi| \star \psi\|_1 \ .$$

Since $|\psi| \in \mathbf{L}^1(\mathbb{R}^d)$ and $|\psi| \geq 0$, it follows that $D_{-j}|\psi|$ is an approximation of the identity in $\mathbf{L}^1(\mathbb{R}^d)$ as $j \to \infty$, with

$$\forall j \ , \int D_{-j}|\psi|(u)du = \|\psi\|_1 \ ,$$

and hence

$$\lim_{j \to \infty} \|D_{-j}|\psi| \star \psi - \|\psi\|_1 \psi\|_1 = 0 \ . \tag{2.43}$$

But

$$\begin{aligned} \left| \int ||\psi| \star \psi_\lambda|(u)du - \|\psi\|_1 \int |\psi|(u)du \right| &= \left| \int |D_{-j}|\psi| \star \psi|(u)du - \|\psi\|_1 \int |\psi|(u)du \right| \\ &\leq \int \left| |D_{-j}|\psi| \star \psi|(u) - \|\psi\|_1 |\psi|(u) \right| du \\ &\leq \|D_{-j}|\psi| \star \psi - \|\psi\|_1 \psi\|_1 \ . \end{aligned}$$

As a result, $\forall \epsilon > 0$ there exists J such that if $|\lambda| > J$, then

$$\left| \int ||\psi| \star \psi_\lambda|(u)du - \|\psi\|_1^2 \right| \leq \epsilon \ .$$

If ϵ is chosen such that $\epsilon < \|\psi\|_1^2/2$, and J_ϵ is the corresponding J, then the paths $q \in \mathcal{A}_{1,2}$, $q = (1, \lambda)$ with $|\lambda| > J_\epsilon$ satisfy

$$\forall q \in \mathcal{A} \ , q = (1, \lambda) \ , |\lambda| > J_\epsilon \ , \int U[q]\delta(u)du > \|\psi\|_1^2 - \epsilon = \frac{\|\psi\|_1^2}{2} > 0 \ . \tag{2.44}$$

On the other hand, there are only a finite number of paths $q \in \mathcal{A}_{1,2}$ with $|\lambda| \leq J_\epsilon$, since by definition $|\lambda| \geq M^{-1}$. As a result,

$$\inf_{\substack{q \in \mathcal{A} \\ q=(1,\lambda), |\lambda| \leq J_\epsilon}} \int U[q]\delta(u)du = \alpha_0 > 0 \ . \tag{2.45}$$

By combining (2.44) and (2.45) we obtain that

$$\inf_{q \in \mathcal{A}} \int U[q]\delta(u)du \geq \min(\alpha_0, \frac{\|\psi\|_1^2}{2}) = \alpha > 0 \ . \tag{2.46}$$

Let us now suppose the result true for $m = m_0 - 1$. We shall prove that it is also true for $m = m_0$. Let

$$\inf_{q \in \mathcal{A}_{1,m_0-1}} \int U[q]\delta(u)du = \alpha > 0 \ .$$

For each $l > 0$, we shall decompose the set $\mathcal{A}_{1,m_0}$ in terms of the maximum jump of the path:

$$\mathcal{A}_{1,m_0} = \mathcal{B}_l \cup (\mathcal{A}_{1,m_0} \setminus \mathcal{B}_l) \ ,$$

with

$$\mathcal{B}_l = \left\{ q \in \mathcal{A}_{1,m_0} \, , q = (\lambda_1, \dots, \lambda_{m_0}); \, \chi(q) = \max_k \left(\frac{|\lambda_k|}{\sum_{k'<k} |\lambda_{k'}|} \right) \geq 2^l \right\} \ .$$

The maximum jump $\chi(q)$ of a path thus measures the largest decrease on the scale, with respect to the current cumulated support of $U[\lambda_1, \dots, \lambda_k]$. Since the set $\mathcal{A}_{1,m}$ contains paths of finite order and finite slope, the maximum jump is lower bounded by a constant M_0 depending on M and the order m_0.

Let $q \in \mathcal{B}_l$. We can write $q = q_0 + \lambda + q_1$, where $q_0 = (\lambda'_1, \dots, \lambda'_{k'})$ satisfies

$$|\lambda| \geq (\sum_{i \leq k'} |\lambda'_i|) 2^l \ . \tag{2.47}$$

If $\lambda = 2^j r$, we have

$$\begin{aligned} U[q_0 + \lambda]\delta &= U[\lambda]U[q_0]\delta = |U[q_0]\delta \star \psi_\lambda| \\ &= D_j(D_{-j}U[q_0]\delta \star \psi_{2^0 r}) \end{aligned} \tag{2.48}$$

We will now exploit again the fact that $f_j(u) = D_{-j}U[q_0]\delta(u)$ is an approximation of the identity in $\mathbf{L}^1(\mathbb{R}^d)$. Let $\gamma = \int f_j(u)du$, which does not depend upon j. We have

$$\begin{aligned} \|D_j(D_{-j}U[q_0]\delta \star \psi_{2^0 r}) - \gamma D_j \psi_{2^0 r}\|_1 &= \|(f_j \star \psi_{2^0 r}) - \gamma \psi_{2^0 r}\|_1 \\ &= \int \left| \int (\psi_{2^0 r}(u - t) - \psi_{2^0 r}(u)) f_j(t) dt \right| du \\ &\leq \int \|T_t \psi_{2^0 r} - \psi_{2^0 r}\|_1 f_j(t) dt \ , \end{aligned} \tag{2.49}$$

where $T_t h(u) = h(u - t)$ is the translation operator. Since the translation operator $t \mapsto T_t h$ is continuous in $\mathbf{L}^1(\mathbb{R}^d)$ for any $h \in \mathbf{L}^1(\mathbb{R}^d)$, then for each $\epsilon > 0$, we can find $\eta > 0$ which only depends upon ψ such that

$$\forall |t| < \eta \ , \ \|T_t \psi_{2^0 r} - \psi_{2^0 r}\|_1 < \epsilon/2 \ . \tag{2.50}$$

On the other hand,

$$\begin{aligned} \int_{|t|>\eta} \|T_t \psi_{2^0 r} - \psi_{2^0 r}\|_1 f_j(t) dt &\leq 2\|\psi\|_1 \int_{|t|>\eta} f_j(t) dt \\ &= 2\|\psi\|_1 \int_{|t|>\eta} D_{-j}U[q_0]\delta(t) dt \\ &= 2\|\psi\|_1 \int_{|t|>2^j \eta} U[q_0]\delta(t) dt \ . \end{aligned} \tag{2.51}$$

By construction, the scale 2^j is such that

$$2^j \geq (\sum_{i \leq k'} |\lambda'_i|) \cdot 2^l ,$$

from (2.47). Since the wavelet ψ has fast decay, $U[q_0]\delta(t)$ satisfies

$$|U[q_0]\delta(t)| \leq C_1/(C_2 + (|t|/K))^n ,$$

where C_i and n only depend upon ψ and $K = \sum_{i \leq k'} |\lambda'_i|$ is proportional to the effective support of the cascade of convolutions given by $U[q_0]h = |||h \star \psi_{\lambda'_1}| \star \ldots | \star \psi_{\lambda'_{k'}}|$. As a result, the error in (2.51) can be bounded by

$$\begin{aligned} \int_{|t|>\eta} \|T_t \psi_{2^0 r} - \psi_{2^0 r}\|_1 f_j(t) dt &\leq C \|\psi\|_1 \epsilon(l) \int U[q_0]\delta(t) dt \\ &\leq C\|\psi\|_1 \gamma \epsilon(l) , \end{aligned} \tag{2.52}$$

where $\epsilon(l) \to 0$ as $l \to \infty$. By using (2.50) and (2.52) we can now bound (2.49) with

$$\begin{aligned} \|(f_j \star \psi_{2^0 r}) - \gamma \psi_{2^0 r}\|_1 &\leq \int \|T_t \psi_{2^0 r} - \psi_{2^0 r}\|_1 f_j(t) dt \\ &= \int_{|t|<\eta} \|T_t \psi_{2^0 r} - \psi_{2^0 r}\|_1 f_j(t) dt + \int_{|t|>\eta} \|T_t \psi_{2^0 r} - \psi_{2^0 r}\|_1 f_j(t) dt \\ &\leq \epsilon/2\gamma + C\|\psi\|_1 \gamma \epsilon(l) \\ &\leq \|\psi\|_1^{|q_0|} (\epsilon/2 + C\|\psi\|_1 \epsilon(l)) , \end{aligned} \tag{2.53}$$

since $\gamma = \int U[q_0]\delta(u) du \leq \|\psi\|_1^{m_0}$ using the Young inequality $\|f \star g\|_1 \leq \|f\|_1 \|g\|_1$.

Since

$$\begin{aligned} \|U[\lambda]f - U[\lambda]g\|_1 &= \||f \star \psi_\lambda| - |g \star \psi_\lambda|\|_1 \\ &\leq \|f \star \psi_\lambda - g \star \psi_\lambda\|_1 = \|(f-g) \star \psi_\lambda\|_1 \\ &\leq \|f - g\|_1 \|\psi\|_1 , \end{aligned}$$

it follows that

$$\| U[p]f - U[p]g\|_1 \leq \|f - g\|_1 \|\psi\|_1^{|p|} . \tag{2.54}$$

As a result of (2.53), any path $q \in \mathcal{B}_l$, which was decomposed as $q = q_0 + \lambda + q_1$, satisfies

$$\begin{aligned} \left| \int U[q]\delta(u) du - \gamma \int U[\lambda + q_1]\delta(u) du \right| &\leq \\ \|U[q]\delta - \gamma U[q_1]U[\lambda]\delta\|_1 &= \|U[q_1]U[\lambda]U[q_0]\delta - \gamma U[q_1]U[\lambda]\delta\|_1 \\ &\leq \|\psi\|_1^{|q_1|} \|U[q_0]\delta \star \psi_\lambda - \gamma \psi_\lambda\|_1 \\ &\leq \|\psi\|_1^{m_0} (\epsilon/2 + C\|\psi\|_1 \epsilon(l)) , \end{aligned} \tag{2.55}$$

by applying (2.54) on $U[q_1]$. (2.55) implies that for any $\epsilon > 0$ one can find sufficiently large l such that $\int U[q]\delta(u)du$ is at distance at most ϵ from $\gamma \int U[\tilde{q}]\delta(u)du$, where $|\tilde{q}| < |q|$ and $\alpha \leq \gamma \leq \|\psi\|_1^{m_0}$. By applying the induction hypothesis with $\epsilon = \alpha/2$, we conclude that

$$\forall q \in \mathcal{B}_l \quad , \quad \int U[q]\delta(u)du \geq \alpha^2/2 > 0 \ . \tag{2.56}$$

On the other hand, the set $\mathcal{A}_{1,m_0} \setminus \mathcal{B}_l$ contains only a finite number of paths, since their slope is bounded by $\Delta(q) \leq B$, and thus

$$\min_{q \in \mathcal{A}_{1,m_0} \setminus \mathcal{B}_l} \int U[q]\delta(u)du = \alpha_0 > 0 \ .$$

We conclude that

$$\forall q \in \mathcal{A}_{1,m_0} \quad , \quad \int U[q]\delta(u)du \geq \min(\alpha^2/2, \alpha_0) > 0 \ , \tag{2.57}$$

which proves (2.41).

Let us finally prove (2.42). Since $x \in \mathbf{L}^1(\mathbb{R}^d)$ and $x \geq 0$, $D_{-J}x$ is also an approximation of the identity, which, with $\overline{x} = \int x(u)du$, satisfies

$$\forall h \in \mathbf{L}^1(\mathbb{R}^d) \ , \quad \lim_{J\to\infty} \|D_{-J}x \star h - \overline{x}h\|_1 = 0 \ . \tag{2.58}$$

If $q \in \mathcal{A}$, $q = \lambda_1 + \tilde{q}$ with $\lambda_1 = 2^0 r$, and hence $U[q]D_{-J}x = U[\tilde{q}]|D_{-J}x \star \psi_{\lambda_1}|$. Then, by using again (2.54), it results that

$$\begin{aligned}
\left|\int U[q]D_{-J}x(u)du - \overline{x}\int U[q]\delta(u)du\right| &= \left|\int (U[q]D_{-J}x(u) - \overline{x}U[q]\delta(u))du\right| \\
&\leq \int |U[q]D_{-J}x(u) - \overline{x}U[q]\delta(u)|\, du \\
&= \|U[q]D_{-J}x - \overline{x}U[q]\delta\|_1 \\
&= \|U[\tilde{q}]|D_{-J}x \star \psi_{\lambda_1}| - \overline{x}U[\tilde{q}]|\psi_{\lambda_1}|\|_1 \\
&\leq \|\psi\|_1^{|\tilde{q}|}\||D_{-J}x \star \psi_{\lambda_1}| - \overline{x}|\psi_{\lambda_1}|\|_1 \\
&\leq \|\psi\|_1^{|\tilde{q}|}\|D_{-J}x \star \psi_{\lambda_1} - \overline{x}\psi_{\lambda_1}\|_1 \ , \qquad (2.59)
\end{aligned}$$

which can be made arbitrarily small thanks to (2.58). This proves (2.42), which concludes the proof of Lemma 2.5.2, and hence of (2.34) $\square$.

Theorem 2.5.1 thus shows that integrable functions have a normalized scattering transform which enjoys some form of regularity. The regularity is measured on paths with finite order and finite slope. In order to prove the full version of the conjecture, Theorem 2.5.1 needs to be extended to handle two new asymptotic regimes: paths with arbitrarily large order and arbitrarily large slope.

The strategy used to prove (2.37), which finds a lower bound for the denominator and an upper bound for the numerator, is not powerful enough to study the regularity on paths with infinite order or slope; indeed, in that case

$$\inf_q \int U[q]\delta(u)du = 0 \ .$$

The general case thus requires to show that for those paths, the numerator $\int U[q]x(u)du$ has the same decay law as the denominator, with a proportionally factor given by $\int x(u)du$.

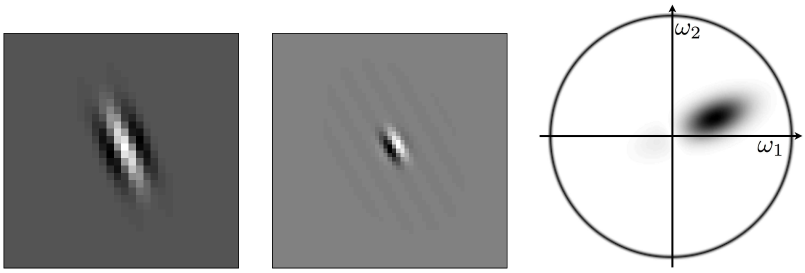
2.6 Scattering Networks for Image Processing

This section concentrates on image processing applications of scattering representations. It introduces several scattering wavelet families and studies the properties of its associated scattering operators for object and texture representations.

2.6.1 Scattering Wavelets

This section describes several wavelet families used to implement scattering representations.

The Littlewood-Paley wavelet transform of x, $\{x \star \psi_\lambda(u)\}_\lambda$, defined in (2.6), is a redundant transform with no orthogonality property. Section 2.3.1 explained that it is stable and invertible if the wavelet filters $\hat{\psi}_\lambda(\omega)$ cover the whole frequency plane. On discrete images, to avoid aliasing, we only capture frequencies in the circle $|\omega| \leq \pi$ inscribed in the image frequency square. Most camera images have negligible energy outside this frequency circle.



(a) (b) (c)

Figure 2.5: Complex Morlet wavelet. (a): Real part of $\psi(u)$. (b): Imaginary part of $\psi(u)$. (c): Fourier modulus $|\hat{\psi}(\omega)|$.

Let $u.u'$ and $|u|$ denote the inner product and norm in $\mathbb{R}^2$. A Morlet wavelet ψ is an example of complex wavelet given by

$$\psi(u) = \alpha\,(e^{iu.\xi} - \beta)\,e^{-|u|^2/(2\sigma^2)}\ ,$$

where $\beta \ll 1$ is adjusted so that $\int \psi(u)\,du = 0$. Its real and imaginary parts are nearly quadrature phase filters. Figure 2.5 shows the Morlet wavelet with $\sigma = 0.85$ and $\xi = 3\pi/4$, used in all classification experiments. The Morlet wavelet ψ shown in Figure 2.5 together with $\phi(u) = \exp(-|u|^2/(2\sigma^2))/(2\pi\sigma^2)$ for $\sigma = 0.7$ satisfy (2.7) with $\epsilon = 0.25$.

Cubic spline wavelets are an important family of unitary wavelets satisfying the Littlewood-Paley condition (2.7) with $\epsilon = 0$. They are obtained from a cubic-spline orthogonal Battle-Lemairé wavelet, defined from the conjugate mirror filter [Mal08]

$$\hat{h}(\omega) = \sqrt{\frac{S_8(\omega)}{2^8 S_8(2\omega)}} ,$$

with

$$S_n(\omega) = \sum_{k=-\infty}^{\infty} \frac{1}{(\omega + 2k\pi)^n} ,$$

which in the case $n = 8$ simplifies to the expression

$$S_8(2\omega) = \frac{5 + 30\cos^2(\omega) + 30\sin^2(\omega)\cos^2(\omega)}{1052^8 \sin^8(\omega)} + \frac{70\cos^4(\omega) + 2\sin^4(\omega)\cos^2(\omega) + 2/3\sin^6(\omega)}{1052^8 \sin^8(\omega)} .$$

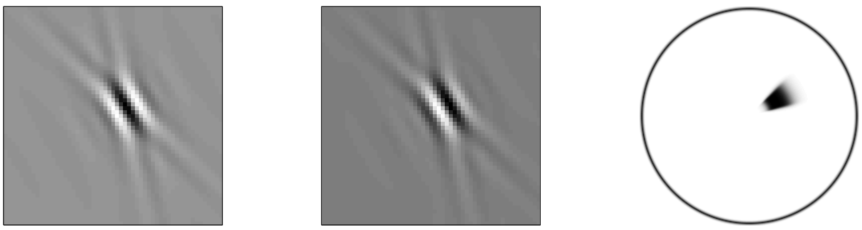
In two dimensions, $\hat{\psi}$ is defined as a separable product in frequency polar coordinates $\omega = |\omega|\eta$, where η is a unit vector:

$$\forall |\omega|, \eta \in \mathbb{R}^+ \times \mathbf{S}^1 \ , \ \hat{\psi}(\omega) = \hat{\psi}_1(|\omega|)\gamma(\eta) \ ,$$

with γ designed such that

$$\forall \eta \ , \ \sum_{r \in G+} |\gamma(r^{-1}\eta)|^2 = 1 \ .$$

Figure 2.6 shows the corresponding two-dimensional filters obtained with spline wavelets, by setting both $\hat{\psi}_1$ and γ to be cubic splines.



(a) (b) (c)

Figure 2.6: Complex cubic Spline wavelet. (a): Real part of $\psi(u)$. (b): Imaginary part of $\psi(u)$. (c): Fourier modulus $|\hat{\psi}(\omega)|$.

2.6.2 Scattering Convolution Network

If $p = (\lambda_1, ..., \lambda_m)$ is a path of length m then the windowed scattering coefficients $S_J[p]x(u)$ of order m are computed at the layer m of a convolution network which is specified. For large scale invariants, several layers are necessary to avoid losing crucial information.

For appropriate wavelets, first order coefficients $S_J[\lambda_1]x$ are equivalent to SIFT coefficients [Low04]. Indeed, SIFT computes the local sum of image gradient amplitudes among image gradients having nearly the same direction, in a histogram having 8 different direction bins. The DAISY approximation [TLF10] shows that these coefficients are well approximated by $S_J[\lambda_1]x = |x \star \psi_{\lambda_1}| \star \phi_{2^J}(u)$ where ψ_{λ_1} are partial derivatives of a Gaussian computed at the finest image scale, along 8 different rotations. The averaging filter ϕ_{2^J} is a scaled Gaussian.

Partial derivative wavelets are well adapted to detect edges or sharp transitions but do not have enough frequency and directional resolution to discriminate complex directional structures. For texture analysis, many researchers [MP90; PS99] have been using averaged wavelet coefficient amplitudes $|x \star \psi_\lambda| \star \phi_{2^J}(u)$, calculated with a complex wavelet ψ having a better frequency and directional resolution.

The translation invariance of $S_J[p]x$ is due to the averaging of $U[p]x$ by ϕ_{2^J}. It has been argued [BBLP10] that an average pooling loses information, which has motivated the use of other operators such as hierarchical maxima [BRP09]. A scattering avoids this information loss by recovering wavelet coefficients at the next layer, which explains the importance of a multilayer network structure.

A scattering is implemented by a deep convolution network [FKL10], having a very specific architecture. As opposed to standard convolution networks, output scattering coefficients are produced by each layer as opposed to the last layer.Filters are not learned from data but are predefined wavelets. Indeed, they build invariants relatively to the action of the translation group which does not need to be learned. Building invariants to other known groups such as rotations or scaling is similarly obtained with predefined wavelets, which perform convolutions along rotation or scale variables [Mal12; SM12].

Different complex quadrature phase wavelets may be chosen but separating signal variations at different scales is fundamental for deformation stability [Mal12]. Using a modulus to pull together quadrature phase filters is also important to remove the high frequency oscillations of wavelet coefficients.

For a fixed position u, windowed scattering coefficients $S_J[p]x(u)$ of order $m = 1, 2$ are displayed as piecewise constant images over a disk representing the Fourier support of the image x. This frequency disk is partitioned into sectors $\{\Omega[p]\}_{p \in \mathcal{P}^m}$ indexed by the path p. The image value is $S_J[p]x(u)$ on the frequency sectors $\Omega[p]$, shown in Figure 2.7.

For $m = 1$, a scattering coefficient $S_J[\lambda_1]x(u)$ depends upon the local Fourier transform energy of x over the support of $\hat{\psi}_{\lambda_1}$. Its value is displayed over a sector $\Omega[\lambda_1]$ which approximates the frequency support of $\hat{\psi}_{\lambda_1}$. For $\lambda_1 = 2^{-j_1} r_1$, there are K rotated sectors located in an annulus of scale 2^{-j_1}, corresponding to each $r_1 \in G$, as shown by Figure 2.7(a). Their area are proportional to $\|\psi_{\lambda_1}\|^2 \sim K^{-1} \, 2^{-j_1}$.

Second order scattering coefficients $S_J[\lambda_1, \lambda_2]x(u)$ are computed with a second wavelet transform which performs a second frequency subdivision. These coefficients are displayed over frequency sectors $\Omega[\lambda_1, \lambda_2]$ which subdivide the sectors $\Omega[\lambda_1]$ of the first

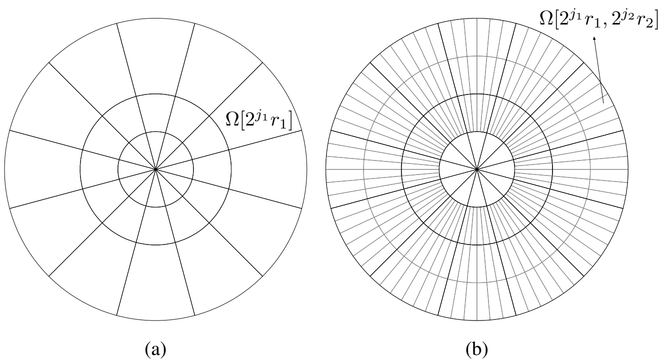


Figure 2.7: To display scattering coefficients, the disk covering the image frequency support is partitioned into sectors $\Omega[p]$, which depend upon the path p. (a): For $m = 1$, each $\Omega[\lambda_1]$ is a sector rotated by r_1 which approximates the frequency support of $\hat{\psi}_{\lambda_1}$. (b): For $m = 2$, all $\Omega[\lambda_1, \lambda_2]$ are obtained by subdividing each $\Omega[\lambda_1]$.

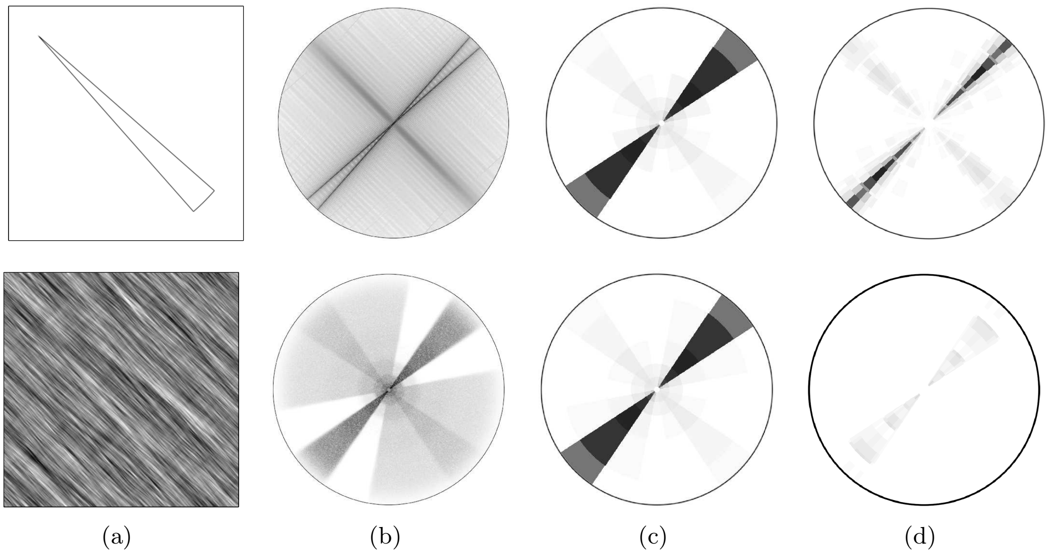


Figure 2.8: (a) Two images $x(u)$. (b) Fourier modulus $|\hat{x}(\omega)|$. (c) First order scattering coefficients $S_J x[\lambda_1]$ displayed over the frequency sectors of Figure 2.7(a). They are the same for both images. (d) Second order scattering coefficients $S_J x[\lambda_1, \lambda_2]$ over the frequency sectors of Figure 2.7(b). They are different for each image.

wavelets $\hat{\psi}_{\lambda_1}$, as illustrated in Figure 2.7(b). For $\lambda_2 = 2^{-j_2} r_2$, the scale 2^{j_2} divides the radial axis and the resulting sectors are subdivided into K angular sectors corresponding to the different r_2. The scale and angular subdivisions are adjusted so that the area of each $\Omega[\lambda_1, \lambda_2]$ is proportional to $\| |\psi_{\lambda_1}| \star \psi_{\lambda_2} \|^2$.

Figure 2.8 shows the Fourier transform of two images, and the amplitude of their scattering coefficients. In this case the 2^J is equal to the image size. The top and bottom images are very different but they have the same first order scattering coefficients. The second order coefficients clearly discriminate these images. Section 2.6.3 shows that the second-order scattering coefficients of the top image have a larger amplitude because the image wavelet coefficients are more sparse. Higher-order coefficients are not displayed because they have a negligible energy, as explained also in Section 2.6.3.

2.6.3 Analysis of Scattering Properties

A convolution network is highly non-linear, which makes it difficult to understand how the coefficient values relate to the signal properties. Scattering convolutional networks, thanks to their construction from stable, Littlewood-Paley wavelet filter banks and the point-wise non-linearities, enjoy energy conservation and stability to additive noise and to the action of diffeomorphisms.

A windowed scattering S_J is computed with a cascade of wavelet modulus operators $\mathcal{U}$ defined in (2.8), and its properties thus depend upon the wavelet transform properties. Sections 2.3.1 and 2.3.2 gave conditions on wavelets to define a scattering transform which is non-expansive and preserves the signal norm. The scattering energy conservation shows that $\|S_J[p]x\|$ decreases quickly as the length of p increases, and is non-negligible only over a particular subset of frequency-decreasing paths. Reducing computations to these paths defines a convolution network with much fewer internal and output coefficients.

Theorem 2.3.4 proves that the energy captured by the m-th layer of the scattering convolutional network, $\sum_{|p|=m} \|S_J[p]x\|^2$, converges to 0 as $m \to \infty$.

The scattering energy conservation also proves that the more sparse the wavelet coefficients, the more energy propagates to deeper layers. Indeed, when 2^J increases, one can verify that at the first layer, $S_J[\lambda_1]x = |x \star \psi_{\lambda_1}| \star \phi_{2^J}$ converges to $\|\phi\|^2 \, \|x \star \psi_\lambda\|_1^2$. The more sparse $x \star \psi_\lambda$, the smaller $\|x \star \psi_\lambda\|_1$ and hence the more energy is propagated to deeper layers to satisfy the global energy conservation (2.13).

Figure 2.8 shows two images having same first order scattering coefficients, but the top image is piecewise regular and hence has wavelet coefficients which are much more sparse than the uniform texture at the bottom. As a result the top image has second order scattering coefficients of larger amplitude than at the bottom. For typical images, as in the CalTech101 dataset [FFFP04], Table 2.1 shows that the scattering energy has an exponential decay as a function of the path length m. Scattering coefficients are computed with cubic spline wavelets, which define a unitary wavelet transform and satisfy the scattering admissibility condition (2.11) for energy conservation. As expected, the energy of scattering coefficients converges to 0 as m increases, and it is already below 1% for $m \geq 3$.

Table 2.1: Percentage of energy $\sum_{p \in \mathcal{P}_\downarrow^m} \|S_J[p]x\|^2/\|x\|^2$ of scattering coefficients on frequency-decreasing paths of length m, depending upon J. These average values are computed on the Caltech-101 database, with zero mean and unit variance images.

J	$m = 0$	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m \leq 3$
1	95.1	4.86	-	-	-	99.96
2	87.56	11.97	0.35	-	-	99.89
3	76.29	21.92	1.54	0.02	-	99.78
4	61.52	33.87	4.05	0.16	0	99.61
5	44.6	45.26	8.9	0.61	0.01	99.37
6	26.15	57.02	14.4	1.54	0.07	99.1
7	0	73.37	21.98	3.56	0.25	98.91

The propagated energy $\|U[p]x\|^2$ decays because $U[p]x$ is a progressively lower frequency signal as the path length increases. Indeed, each modulus computes a regular envelop of oscillating wavelet coefficients. The modulus can thus be interpreted as a non-linear "demodulator" which pushes the wavelet coefficient energy towards lower frequencies. As a result, an important portion of the energy of $U[p]x$ is then captured by the low pass filter ϕ_{2^J} which outputs $S_J[p]x = U[p]x \star \phi_{2^J}$. Hence fewer energy is propagated to the next layer.

Another consequence is that the scattering energy propagates only along a subset of frequency decreasing paths. Since the envelope $|x \star \psi_\lambda|$ is more regular than $x \star \psi_\lambda$, it results that $|x \star \psi_\lambda(u)| \star \psi_{\lambda'}$ is non-negligible only if $\psi_{\lambda'}$ is located at lower frequencies than ψ_λ, and hence if $|\lambda'| < |\lambda|$. Iterating on wavelet modulus operators thus propagates the scattering energy along frequency-decreasing paths $p = (\lambda_1, ..., \lambda_m)$ where $|\lambda_k| < |\lambda_{k-1}|$ for $1 \leq k < m$. We denote by $\mathcal{P}_\downarrow^m$ the set of frequency decreasing (or equivalently scale increasing) paths of length m. Scattering coefficients along other paths have a negligible energy. This is verified by Table 2.1 which shows not only that the scattering energy is concentrated on low-order paths, but also that more than 99% of the energy is absorbed by frequency-decreasing paths of length $m \leq 3$. Numerically, it is therefore sufficient to compute the scattering transform along frequency-decreasing paths. It defines a much smaller convolution network. Section 2.6.4 shows that the resulting coefficients are computed with $O(N \log N)$ operations.

Preserving energy does not imply that the signal information is preserved. Since a scattering transform is calculated by iteratively applying $\mathcal{U}$, inverting S_J requires to invert $\mathcal{U}$. The wavelet transform $\mathcal{W}$ is a linear invertible operator, so inverting $\mathcal{U}z = \{z \star \phi_{2^J}, |z \star \psi_\lambda|\}_{\lambda \in \mathcal{P}}$ amounts to recover the complex phases of wavelet coefficients removed by the modulus. The phase of Fourier coefficients can not be recovered from their modulus, but wavelet coefficients are redundant, as opposed to Fourier coefficients. For particular wavelets, it has been proved that the phase of wavelet coefficients can be recovered from their modulus, and that $\mathcal{U}$ has a continuous inverse [Wal12].

Still, one can not exactly invert S_J because we discard information when computing the scattering coefficients $S_J[p]x = U[p] \star \phi_{2^J}$ of the last layer $\mathcal{P}^{\overline{m}}$. Indeed, the propagated coefficients $|U[p]x \star \psi_\lambda|$ of the next layer are eliminated, because they are not invariant and have a negligible total energy. The number of such coefficients is larger than the total number of scattering coefficients kept at previous layers. Initializing the inversion by considering that these small coefficients are zero produces an error. This error is further amplified as the inversion of $\mathcal{U}$ progresses across layers from $\overline{m}$ to 0. Numerical experiments conducted over one-dimensional audio signals, [AM12a] indicate that reconstructed signals have a good audio quality with $\overline{m} = 2$, as long as the number of scattering coefficients is comparable to the number of signal samples. Audio examples in *www.cmap.polytechnique.fr/scattering* show that reconstructions from first order scattering coefficients are typically of much lower quality because there are much fewer first order than second order coefficients. When the invariant scale 2^J becomes too large, the number of second order coefficients also becomes too small for accurate reconstructions. Although individual signals can be not be recovered, reconstructions of equivalent stationary textures are possible with arbitrarily large scale scattering invariants, as it will be shown in Chapter 4 in auditory texture synthesis.

2.6.4 Fast Scattering Computations

We describe a fast scattering implementation over frequency decreasing paths, where most of the scattering energy is concentrated. A frequency decreasing path $p = (2^{-j_1}r_1, ..., 2^{-j_m}r_m)$ satisfies $0 < j_k \leq j_{k+1} \leq J$. If the wavelet transform is computed over K rotation angles then the total number of frequency-decreasing paths of length m is $K^m\binom{J}{m}$. Let N be the number of pixels of the image x. Since ϕ_{2^J} is a low-pass filter scaled by 2^J, $S_J[p]x(u) = U[p]x \star \phi_{2^J}(u)$ is uniformly sampled at intervals $\alpha 2^J$, with $\alpha = 1$ or $\alpha = 1/2$. Each $S_J[p]x$ is an image with $\alpha^{-2}2^{-2J}N$ coefficients. The total number of coefficients in a scattering network of maximum depth $\overline{m}$ is thus

$$P = N\,\alpha^{-2}\,2^{-2J}\sum_{m=0}^{\overline{m}} K^m\binom{J}{m}\,. \tag{2.60}$$

If $\overline{m} = 2$ then $P \simeq \alpha^{-2}\,N 2^{-2J}K^2J^2/2$. It decreases exponentially when the scale 2^J increases.

Algorithm 1 describes the computations of scattering coefficients on sets $\mathcal{P}_\downarrow^m$ of frequency decreasing paths of length $m \leq \overline{m}$. The initial set $\mathcal{P}_\downarrow^0 = \{\emptyset\}$ corresponds to the original image $U[\emptyset]x = x$. Let $p + \lambda$ be the path which begins by p and ends with $\lambda \in \mathcal{P}$. If $\lambda = 2^{-j}r$ then $U[p+\lambda]x(u) = |U[p]x \star \psi_\lambda(u)|$ has energy at frequencies mostly below $2^{-j}\pi$. To reduce computations we can thus subsample this convolution at intervals $\alpha 2^j$, with $\alpha = 1$ or $\alpha = 1/2$ to avoid aliasing.

At the layer m there are $K^m\binom{J}{m}$ propagated signals $U[p]x$ with $p \in \mathcal{P}_\downarrow^m$. They are sampled at intervals $\alpha 2^{j_m}$ which depend on p. One can verify by induction on m that the layer m has a total number of samples equal to $\alpha^{-2}\,(K/3)^m\,N$. There are also $K^m\binom{J}{m}$ scattering signals $S[p]x$ but they are subsampled by 2^J and thus have

Algorithm 1 Fast Scattering Transform

for $m = 1$ to $\overline{m}$ **do**
  for all $p \in \mathcal{P}_{\downarrow}^{m-1}$ **do**
    Output $S_J[p]x(\alpha 2^J n) = U[p]x \star \phi_{2^J}(\alpha 2^J n)$
  end for
  for all $p + \lambda_m \in \mathcal{P}_{\downarrow}^{m}$ with $\lambda_m = 2^{-j_m} r_m$ **do**
    Compute

$$U[p + \lambda_m]x(\alpha 2^{j_m} n) = |U[p]x \star \psi_{\lambda_m}(\alpha 2^{j_m} n)|$$

  end for
end for
for all $p \in \mathcal{P}_{\downarrow}^{\max}$ **do**
  Output $S_J[p]x(\alpha 2^J n) = U[p]x \star \phi_{2^J}(\alpha 2^J n)$
end for

much less coefficients. The number of operation to compute each layer is therefore driven by the $O((K/3)^m N \log N)$ operations needed to compute the internal propagated coefficients with FFT's. For $K > 3$, the overall computational complexity is thus $O((K/3)^{\overline{m}} N \log N)$.

2.6.5 Analysis of stationary textures with scattering

Section 2.3.5 showed that the scattering representation can be used to describe stationary processes, in such a way that high order moments information is captured and estimated consistently with few realizations.

Image textures can be modeled as realizations of stationary processes $X(u)$. We denote the expected value of X by $E(X)$, which does not depend upon u. The Fourier spectrum $\widehat{R}_X(\omega)$ is the Fourier transform of the autocorrelation

$$R_X(\tau) = E\Big([X(u) - E(X)][X(u-\tau) - E(X)]\Big) .$$

Despite the importance of spectral methods, the Fourier spectrum is often not sufficient to discriminate image textures because it does not take into account higher-order moments.

The discriminative power of scattering representations is illustrated using the two textures in Figure 2.9, which have the same power spectrum and hence same second order moments. Scattering coefficients $S_J[p]X$ are shown for $m = 1$ and $m = 2$ with the frequency tiling illustrated in Figure 2.7. The ability to discriminate the top process X_1 from the bottom process X_2 is measured by a scattering distance normalized by the variance:

$$\rho(m) = \frac{\|S_J X_1[\Lambda_J^m] - E(S_J X_2[\Lambda_J^m])\|^2}{E(\|S_J X_2[\Lambda_J^m] - E(S_J X_2[\Lambda_J^m])\|^2)} .$$

For $m = 1$, scattering coefficients mostly depend upon second-order moments and are thus nearly equal for both textures. One can indeed verify numerically that $\rho(1) = 1$

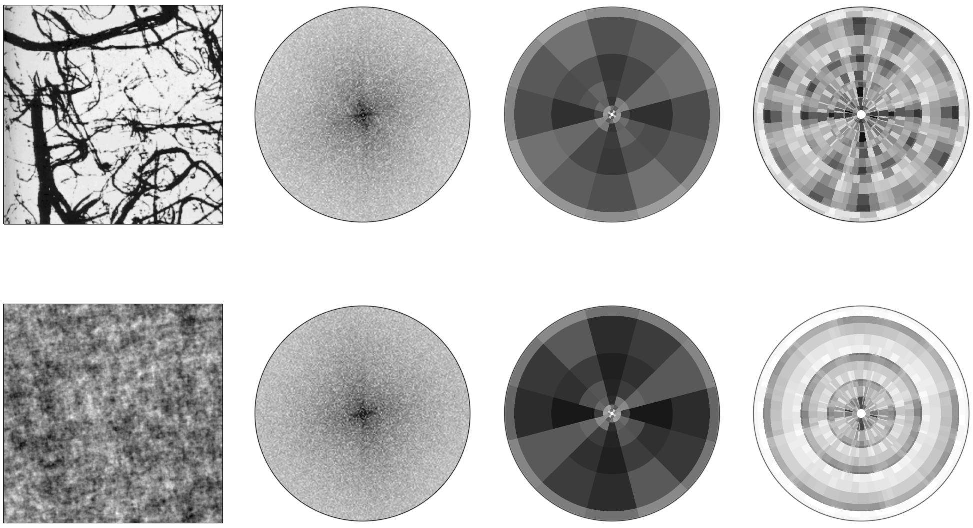


Figure 2.9: Two different textures having the same Fourier power spectrum. (a) Textures $X(u)$. Top: Brodatz texture. Bottom: Gaussian process. (b) Same estimated power spectrum $\widehat{R}X(\omega)$. (c) Nearly same scattering coefficients $S_J[p]X$ for $m = 1$ and 2^J equal to the image width. (d) Different scattering coefficients $S_J[p]X$ for $m = 2$.

so both textures can not be distinguished using first order scattering coefficients. On the contrary, scattering coefficients of order 2 are highly dissimilar because they depend on moments up to order 4, and $\rho(2) = 5$. A scattering representation of stationary processes includes second order and higher-order moment descriptors of stationary processes, which discriminates between such textures. Chapter 4 exploits the consistency and the discriminability of expected scattering representations to the tasks of texture classification and synthesis.

The windowed scattering $S_J[\mathcal{P}_J]X$ estimates scattering coefficients by averaging wavelet modulus over a support of size proportional to 2^J. If X is a stationary process, Section 2.3.5 showed that the expected scattering transform $\overline{S}X$ is estimated with the windowed scattering

$$S_J[\mathcal{P}_J]X = \{U[p]X \star \phi_J \ , p \in \mathcal{P}_J\} \ .$$

This estimate is called mean-square consistent if its total variance over all paths converges:

$$\lim_{J\to\infty} \sum_{p\in\mathcal{P}_J} E(|S_J[p]X - \overline{S}X(p)|^2) = 0 \ .$$

Corollary 2.3.13 showed that mean-square consistency is equivalent to

$$E(|X|^2) = \sum_{p\in\mathcal{P}_\infty} |\overline{S}X(p)|^2 \ ,$$

which in turn is equivalent to

$$\lim_{m\to\infty} \sum_{p\in\mathcal{P}_\infty\,,\,|p|=m} E(|U[p]X|^2) = 0\ . \tag{2.61}$$

If a process $X(t)$ has a mean square consistent scattering, then one can recover the scaling law of its second moments with scattering coefficients:

Proposition 2.6.1 *Suppose that $X(t)$ is a process with stationary increments such that $S_J X$ is mean square consistent. Then*

$$E(|X\star\psi_j|^2) = \sum_{p\in\mathcal{P}_\infty} |\overline{S}X(j+p)|^2\ . \tag{2.62}$$

Proof: If X is such that $S_J X$ is mean square consistent, then the process $X_j = |X\star\psi_j|$ also yields a mean square consistent scattering representation, since for each J

$$\begin{aligned}\sum_{p\in\mathcal{P}_J} E(|S_J[p]X_j - \overline{S}X_j(p)|^2) &= \sum_{p\in\mathcal{P}_J} E(|S_J[j+p]X - \overline{S}X(j+p)|^2)\\ &\leq \sum_{p\in\mathcal{P}_J} E(|S_J[p]X - \overline{S}X(p)|^2)\ ,\end{aligned}$$

which implies that $\lim_{J\to\infty} E(\|S_J[\mathcal{P}_J]X_j - \overline{S}X_j\|^2) = 0$. As a result,

$$E(|X\star\psi_j|^2) = \sum_{p\in\mathcal{P}_\infty} |\overline{S}X_j(p)|^2 = \sum_{p\in\mathcal{P}_\infty} |\overline{S}X(j+p)|^2\ . \tag{2.63}$$

□.

For a large class of ergodic processes including most image textures, it is observed numerically that the total scattering variance $\sum_{p\in\mathcal{P}_J} E(|S_J[p]X - \overline{S}X(p)|^2)$ decreases to zero when 2^J increases. Table 2.2 shows the decay of the total scattering variance, computed on average over the Brodatz texture dataset.

Corollary 2.3.13 showed that this variance decay then implies that

$$\|\overline{S}X\|^2 = \sum_{m=0}^{\infty}\sum_{p\in\Lambda_\infty^m} |\overline{S}X(p)|^2 = E(|X|^2)\ .$$

Table 2.3 gives the percentage of expected scattering energy $\sum_{p\in\Lambda_\infty^m} |\overline{S}X(p)|^2$ carried by paths of length m, for textures in the Brodatz database. Most of the energy is concentrated in paths of length $m\leq 3$.

Table 2.2: Decay of the total scattering variance $\sum_{p \in \mathcal{P}_J} E(|S_J[p]X - \overline{S}X(p)|^2)/E(|X|^2)$ in percentage, as a function of J, averaged over the Brodatz dataset. Results obtained using cubic spline wavelets.

$J=1$	$J=2$	$J=3$	$J=4$	$J=5$	$J=6$	$J=7$
85	65	45	26	14	7	2.5

Table 2.3: Percentage of expected scattering energy $\sum_{p \in \Lambda_\infty^m} |\overline{S}X(p)|^2$, as a function of the scattering order m, computed with cubic spline wavelets, over the Brodatz dataset.

$m=0$	$m=1$	$m=2$	$m=3$	$m=4$
0	74	19	3	0.3

Chapter 3

Image and Pattern Classification with Scattering

3.1 Introduction

This chapter applies the properties of the scattering metric to pattern recognition. In this task, signal classes are affected by several sources of variability, such as geometrical or photometric transformations, non-rigid deformations, shape and texture variability or clutter, as shown in figure 1.3. The effectiveness of a signal representation for recognition thus depends upon its capacity to reduce the intra-class variability while keeping enough signal information to discriminate between different object classes.

The first source of variability, which affects the large majority of pattern recognition tasks, is given by geometric and photometric transformations, and non-rigid deformations. Chapter 2 showed that the metric defined by scattering representations has the capacity to reduce such variability while keeping high frequency information, and hence that scattering descriptors are an effective, universal preprocessing step in complex object recognition tasks, which reduces geometric variability and facilitates the learning of more complex structures.

This claim can be verified by first considering classification problems where signal classes are well modeled as templates $x \in \mathbf{L}^2(\mathbb{R}^d)$, with an associated deformation structure including geometrical transformations and deformations.

Thanks to its Lipschitz continuity property to the action of diffeomorphisms, small deformations are linearized in the scattering domain. This property is exploited in this chapter with a generative linear classifier, which learns for each signal class a low-dimensional affine approximation model. This classifier is estimated with a class-conditional PCA, and can be interpreted as a supervised invariance learning step.

Prior to the supervised learning of the affine spaces, we increase the efficiency of scattering representations by reducing the correlation across paths observed in natural images. Section 3.3 shows that a discrete cosine transform across scale and orientation variables approximates the Karhunen-Loève basis on a dataset of natural images.

We apply this linear generative classifier to the MNIST and USPS handwritten datasets. We show that it outperforms discriminative classifiers such as SVM for small training sizes, achieving state-of-the-art results. For larger training sizes, a discriminative classifier, implemented with a Gaussian Kernel SVM, is shown to outperform previously published state-of-the-art results. When training samples are scarce, linear generative classifiers can take advantage of the Lipschitz regularity of the transformed data more effectively than discriminative SVMs, as shown with a toy model in Section 3.4.3.

Similarly as in other convolutional networks and SVMs, the performance of the PCA classifier is improved by renormalizing scattering coefficients. The renormalization given by the scattering transfer, which will also be used in Chapter 5 for the study of multifractals, yields an improvement with respect to the original scattering metric.

3.2 Support Vector Machines

This Section reviews Support Vector Machines, a popular kernel method to perform discriminative classification.

Support Vector Machines were introduced in the late 70s by Vapnik [Vap79], and initially formulated as a supervised method for classification or regression in high dimensional spaces. In its simplest formulation, given a binary classification task $\mathcal{X} \longrightarrow \{\pm 1\}$, where $\mathcal{X}$ is a Hilbert space, and where training examples $\mathcal{T} = \{(x_i, y_i)\,,\, x_i \in \mathcal{X}\,,\, y_i = \pm 1\,,\, i = 1 \dots I\}$ are observed, a support vector machine [Bur98] constructs a hyperplane in $\mathcal{X}$ which best separates the two classes $\mathcal{T}_+, \mathcal{T}_-$, corresponding to points x_i such that $y_i = +1$ and $y_i = -1$ respectively.

The criteria for best separation is based on the notion of margin. If one first supposes that the training data is linearly separable in $\mathcal{X}$, i.e., that there exists an hyperplane $(\omega, b)\,, \omega \in \mathcal{X}, b \in \mathbb{R}$, such that

$$\forall x_i \in \mathcal{T}_+,\ x_i.\omega - b > 0 \quad , \quad \forall x_i \in \mathcal{T}_-,\ x_i.\omega - b < 0\ , \tag{3.1}$$

then the classifier with smallest generalization error is given by the hyperplane which maximizes the distance to the nearest training examples. Such hyperplane is characterized by

$$\left\{ \begin{array}{l} \min_{(\omega,b)} \|\omega\|^2 \\ \text{s.t. } \forall i\,,\ y_i(\omega.x_i - b) \geq 1\ . \end{array} \right. \tag{3.2}$$

This program is solved by introducing Lagrange multipliers. It results from the Karush-Kuhn-Tucker quadratic programming conditions that the solution ω is a linear combination of the training samples, $\omega = \sum_i \alpha_i y_i x_i$, with $\alpha_i \geq 0$. The indices where $\alpha_i > 0$ are called the support vectors, since they determine the position of the best-separating hyperplane.

In general, however, the training data is not linearly separable, meaning that (3.1) is not satisfied. Cortes and Vapnik [CV95] modified the maximum-margin linear program (3.2) by introducing slack variables, which allow mis-classified examples but set a penalization for such misclassifications. The resulting dual program is

$$\left\{ \begin{array}{l} \max_{\alpha_i} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i . x_j \\ \text{s.t. } \forall i \, , \ 0 \leq \alpha_i \leq C \ , \sum_i \alpha_i y_i = 0 \end{array} \right. \tag{3.3}$$

Support Vector Machines can be generalized to create non-linear decision functions by applying the kernel trick [BGV92]. Indeed, the training phase, given by program (3.3) only uses data through the scalar products $x_i.x_j$. By replacing this linear kernel by a positive-definite kernel $K(x_i, x_j)$, Section 2.2.2 explained that this is equivalent to estimating a separating hyperplane on a higher-dimensional feature space, characterized by Mercer's theroem. Popular kernels include the Gaussian radial basis function $K(x, y) = \exp(-\|x - y\|^2/2\sigma^2)$, as well as the polynomial kernels $K(x, y) = (1 + x.y)^d$.

Support Vector Machines are a very popular discriminative classifier thanks to their good generalization properties and the flexibility given by different kernels. However, we shall see in Section 3.4.3 that on small training sets they can be outperformed by generative classifiers.

3.3 Compression with Cosine Scattering

In many pattern recognition tasks, it is important from a computational point of view to compute local descriptors as small as possible. A discrete cosine transform along scale and orientation variables removes the correlation in scattering coefficients, yielding a local image descriptor with reduced dimensionality.

Natural images have scattering coefficients $S_J[p]X(u)$ which are correlated across paths $p = (2^{j_1} r_1, ..., 2^{j_m} r_m)$, at any given position u. The strongest correlation is between paths of the same length. For each m, scattering coefficients are decorrelated in a Karhunen-Loève basis which diagonalizes their covariance matrix. Figure 3.1 compares the decay of the sorted variances $E(|S_J[p]X - E(S_J[p]X)|^2)$ and the variance decay in the Karhunen-Loève basis computed on paths of length $m = 1$, and on paths of length $m = 2$, over the Caltech image dataset with a Morlet wavelet. The variance decay is much faster in the Karhunen-Loève basis, which shows that there is a strong correlation between scattering coefficients of same path length.

A change of variables proves that a rotation and scaling $X_{2^l r}(u) = X(2^{-l} r u)$ produces a rotation and inverse scaling on the path variable $p = (2^{j_1} r_1, ..., 2^{j_m} r_m)$:

$$\overline{S} X_{2^l r}(p) = \overline{S} X(2^l r p) \ \text{ where } \ 2^l r p = (2^{l+j_1} r r_1, ..., 2^{l+j_m} r r_m) \ .$$

If images are randomly rotated and scaled by $2^l r^{-1}$ then the path p is randomly rotated and scaled [Per10]. In this case, the scattering transform has stationary variations along the scale and rotation variables. This suggests approximating the Karhunen-Loève basis by a cosine basis along these variables. Let us parameterize each rotation r by its angle $\theta \in [0, 2\pi]$. A path p is then parameterized by $([j_1, \theta_1], ..., [j_m, \theta_m])$.

Since scattering coefficients are computed along frequency decreasing paths for which $-J \leq j_k < j_{k-1}$, to reduce boundary effects, a separable cosine transform is computed along the variables $\tilde{j}_1 = j_1$, $\tilde{j}_2 = j_2 - j_1$, ..., $\tilde{j}_m = j_m - j_{m-1}$, and along each angle variable θ_1, θ_2, ..., θ_m. We define the cosine scattering transform as the coefficients obtained by applying this separable discrete cosine transform along the scale and angle variables of $S_J[p]X(u)$, for each u and each path length m. Figure 3.1 shows that the cosine scattering coefficients have variances for $m = 1$ and $m = 2$ which decay nearly as fast as the variances in the Karhunen-Loeve basis. It shows that a DCT across scales and orientations is nearly optimal to decorrelate scattering coefficients. Lower-frequency DCT coefficients absorb most of the scattering energy. On natural images, more than 99% of the scattering energy is absorbed by the 1/3 lowest frequency cosine scattering coefficients.

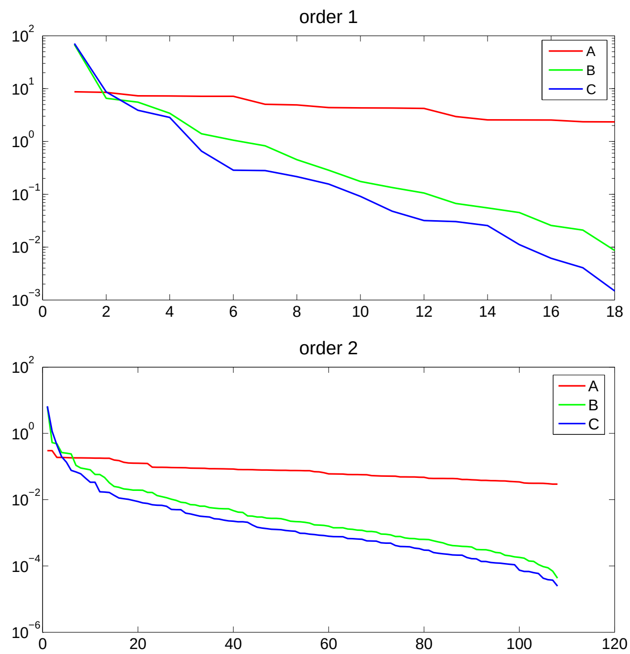


Figure 3.1: (A): Sorted variances of scattering coefficients for $m = 1$ (left) and $m = 2$ (right). (B): Sorted variances of DCT scattering coefficients. (C): Variances in the scattering Karhunen-Loeve basis.

Another source of correlation between scattering paths comes from the self-similarity present in natural textures and geometric structures. If x_α is an image patch containing an edge along an angle α, then for a given sub-band $\lambda = 2^j \theta$,

$$|x_\alpha \star \psi_\lambda| \approx C_\theta (x_\alpha \star |\psi_\lambda|) ,$$

where the envelope $|\psi_\lambda|$ depends mostly on the scale 2^j. This induces a correlation on second order paths along its orientation parameters. Similarly, isolated singularities such as a Dirac impulse produce scale correlations. Moreover, in Chapter 5 we shall see that self-similar processes have redundant scattering coefficients, and that these redundancy can in fact define an invariant quantity.

3.4 Generative Classification with Affine models

Although discriminant classifiers such as SVM have better asymptotic properties than generative classifiers, the situation can be inverted for small training sets [NJ02]. We will consider a simple robust generative classifier based on affine space models computed with a PCA, resulting in a simplified version of the discriminative k q-metrics algorithm [SS09]. Applying a Discrete Cosine Transform on scattering coefficients has no effect on any linear classifier because it is a linear orthogonal transform. However, keeping the 50% lower frequency cosine scattering coefficients reduces computations and has a negligible effect on classification results. The classification algorithm is described directly on scattering coefficients to simplify explanations. Each signal class is represented by a random vector X_k, whose realizations are images of N pixels in the class. Section 3.4.1 presents a linear generative classifier which takes advantage of the Lipschitz regularity properties of scattering operators. Section 3.4.2 introduces a renormalization strategy which improves classification results, similarly as in other classifiers. Finally, in Section 3.4.3 we compare the generative classification strategy with SVM classification, and show that it may be a more effective tool to exploit local regularity of data on small training sets.

3.4.1 Linear Generative Classifier

Chapter 2 showed that the scattering operator S_J is Lipschitz continuous with respect to additive and geometric perturbations of the signal. Section 2.3.2 showed that since both the wavelet modulus and the averaging kernel are contractive $\mathbf{L}^2(\mathbb{R}^d)$ operators, the scattering metric satisfies

$$\|S_J(x+h) - S_Jx\| \leq \|h\| \ .$$

It follows that the application $h \mapsto \varphi_x(h) = x + h$, performing an additive perturbation on x, is Lipschitz continuous at 0 when composed with the scattering operator:

$$\|S_J\varphi_x(h) - S_J\varphi_x(0)\| \leq \|\varphi_x(h) - \varphi_x(0)\| = \|h\| \ . \tag{3.4}$$

In that case, thanks to the Radon-Nikodỳm property of Hilbert spaces, S_J admits a Gâteaux differential almost everywhere [LPT12]. This differential, which is a bounded linear operator, encodes how infinitesimal additive perturbations of $x \in \mathbf{L}^2(\mathbb{R}^d)$ are mapped into the scattering domain:

$$S_J\varphi_x(h) = S_Jx + DS_{Jx}(h) + o(\|h\|) \ , \|h\| \to 0 \ . \tag{3.5}$$

On the other hand, Section 2.3.3 showed that the scattering metric is also upper bounded by an elastic deformation metric. Indeed, if $x \in \mathbf{L}^2(\mathbb{R}^d)$ has compact support and $L[\tau]x(u) = x(u - \tau(u))$, then theorem 2.3.7 showed that the perturbation $\tau \mapsto \varphi_x(\tau) = L[\tau]x$ satisfies

$$\|S_J\varphi_x(\tau) - S_J\varphi_x(0)\| = \|S_JL[\tau]x - S_Jx\| \leq C\|x\|\|\tau\|_G \ , \tag{3.6}$$

where $\|\tau\|_G = 2^{-J}|\tau|_\infty + \|\nabla\tau\|_\infty + \|H\tau\|_\infty$ is the deformation metric defined in Section 2.3.3. As a result, the map $\tilde{S}_{Jx} = S_J \circ \varphi_x$ is also differentiable in the sense of Gâteaux almost everywhere. Its differential $D\tilde{S}_{Jx}$ maps small deformations into the scattering domain:

$$S_J L[\tau]x = S_J x + D\tilde{S}_{Jx}(\tau) + o(\|\tau\|_G) \ , \|\tau\|_G \to 0 \ . \tag{3.7}$$

As a result, both small additive perturbations and small geometric deformations are linearized by the scattering transform. This source of regularity can be exploited in a supervised classification setting using a class-conditional PCA.

We shall represent each signal class by a random process X_k, whose realizations are observed images or audio samples. Let $E(S_J X) = \{E(S_J[p]X(u))\}_{p,u}$ be the family of N_J expected scattering values, computed along all frequency-decreasing paths of length $m \leq m_{\max}$ and all subsampled positions $u = \alpha 2^J n$. The difference $S_J X_k - E(S_J X_k)$ is approximated by its projection in a linear space of low dimension $d \ll N_J$. The covariance matrix of $S_J X_k$ is a matrix of size N_J^2. Let $\mathbf{V}_{d,k}$ be the linear space generated by the d PCA eigenvectors of this covariance matrix having the largest eigenvalues. Among all linear spaces of dimension d, this is the space which approximates $S_J X_k - E(S_J X_k)$ with the smallest expected quadratic error. This is equivalent to approximating $S_J X_k$ by its projection on an affine approximation space:

$$\mathbf{A}_{d,k} = E\{S_J X_k\} + \mathbf{V}_{d,k}.$$

The resulting classifier associates a signal X to the class $\hat{k}$ which yields the best approximation space:

$$\hat{k}(X) = \underset{k \leq K}{\operatorname{argmin}} \|S_J X - P_{\mathbf{A}_{d,k}}(S_J X)\| \ . \tag{3.8}$$

This algorithm is a simple instance of the supervised $k-q$-flats algorithm and its discriminative variants [SS09], where each class is assigned a single affine subspace, rather than learning a collection of k prototypes for each class. The minimization of (3.8) also has similarities with the minimization of a tangential distance [HK02], in the sense that we remove the principal directions of variability to evaluate the distance. However, it is much simpler since it does not evaluate a tangential space which depends upon $S_J x$. Let $\mathbf{V}_{d,k}^\perp$ be the orthogonal complement of $\mathbf{V}_{d,k}$, corresponding to directions of lower variability. This distance is also equal to the norm of the difference between $S_J x$ and the average class "template" $E(S_J X_k)$, projected in $\mathbf{V}_{d,k}^\perp$:

$$\|S_J x - P_{\mathbf{A}_{d,k}}(S_J x)\| = \left\| P_{\mathbf{V}_{d,k}^\perp}\Big(S_J x - E(S_J X_k)\Big) \right\| \ . \tag{3.9}$$

Minimizing the affine space approximation error is thus equivalent to finding the class centroid $E(S_J X_k)$ which is the closest to $S_J x$, without taking into account the first d principal variability directions. The d principal directions of the space $\mathbf{V}_{d,k}$ result from deformations and from structural variability. These d principal directions of variability can also be interpreted in terms of the Gâteaux differentials from (3.5, 3.7). Indeed, the spaces $\mathbf{V}_{d,k}$ encode a subspace of the scattering tangent space generated by the perturbations h with largest deviations $DS_{J E(X)}(h)$. The variability in the scattering domain is thus approximated from an input distribution of deformations or additive perturbations through the corresponding bounded linear operators from (3.5) and (3.7).

The affine space selection is effective if $S_J X_k - E(S_J X_k)$ is well approximated by a projection in a low-dimensional space. This is the case if realizations of X_k are translations and limited deformations of a single template. Indeed, the scattering differential DS_{JX} controls how each geometric or additive perturbation is perceived in the scattering domain. If a given template is deformed along displacement fields which span a low-dimensional space, then the resulting transformed coefficients will also be well approximated by a low-dimensional space. Moreover, if the deformation is enlarged with random translations, this won't affect the approximation power of low-dimensional affine spaces, since translations are geometric perturbations which are atentuated by DS_{JX} thanks to the local translation invariance. Hand-written digit recognition is an example. This is also valid for stationary textures where $S_J X_k$ has a small variance, which can be interpreted as structural variability.

The dimension d must be adjusted so that $S_J X_k$ has a better approximation in the affine space $\mathbf{A}_{d,k}$ than in affine spaces $\mathbf{A}_{d,k'}$ of other classes $k' \neq k$. This is a model selection problem, which requires to optimize the dimension d in order to avoid overfitting [BM97].

The invariance scale 2^J must also be optimized. When the scale 2^J increases, translation invariance increases but it comes with a partial loss of information which brings the representations of different signals closer. One can prove [Mal12] that for any x and x'

$$\|S_{J+1}x - S_{J+1}x'\| \leq \|S_J x - S_J x'\| \ .$$

When 2^J goes to infinity, this scattering distance converges to a non-zero value. To classify deformed templates such as hand-written digits, the optimal 2^J is of the order of the maximum pixel displacements due to translations and deformations. In a stochastic framework where x and x' are realizations of stationary processes, $S_J x$ and $S_J x'$ converge to the expected scattering transforms $\overline{S}x$ and $\overline{S}x'$. In order to classify stationary processes such as textures, the optimal scale is the maximum scale equal to the image width, because it minimizes the variance of the windowed scattering estimator.

A cross-validation procedure is used to find the dimension d and the scale 2^J which yield the smallest classification error. This error is computed on a subset of the training images, which is not used to estimate the covariance matrix for the PCA calculations.

The class-conditional PCA is closely related to the decorrelation step induced by the Cosine Scattering transform of Section 3.3. However, while in the compression phase we kept the coordinates yielding largest variance, the PCA classifier does the opposite and keeps the coefficients corresponding to eigenvectors with smallest variability. Indeed, the cosine scattering transform projects the scattering representation into a linear subspace containing most of the signal energy, independently of the signal class. This space is generated by linear combinations of scattering paths keeping the low-frequency variations across paths along orientation and scale directions. The class-conditional PCA further decorrelates the scattering coefficients within a class, by exploiting the spatial dependencies and class-specific path correlations. But since it is computed for a specific class in a supervised setting, the variance is non-informative and thus we discard the leading principal directions, as opposed to the unsupervised compression step carried out by the Cosine Scattering transform.

Affine space scattering models can be interpreted as generative models computed independently for each class. As opposed to discriminative classifiers such as SVM, they do not estimate cross-terms between classes, besides from the choice of the model dimensionality d. Such estimators are particularly effective for small number of training samples per class. Indeed, if there are few training samples per class then variance terms dominate bias errors when estimating off-diagonal covariance coefficients between classes [BL08].

An affine space approximation classifier can also be interpreted as a robust quadratic discriminant classifier obtained by coarsely quantizing the eigenvalues of the inverse covariance matrix. For each class, the eigenvalues of the inverse covariance are set to 0 in $\mathbf{V}_{d,k}$ and to 1 in $\mathbf{V}_{d,k}^{\perp}$, where d is adjusted by cross-validation. This coarse quantization is justified by the poor estimation of covariance eigenvalues from few training samples. These affine space models will typically be applied to distributions of scattering vectors having non-Gaussian distributions, where a Gaussian Fisher discriminant can lead to important errors.

This affine space classifier can also be related to the "A Contrario" models for detection of geometric structures [DMM01]. This framework is based in the so-called Helmhotz principle, which states that relevant geometric features have a very small probability of occurring in a random context [DMM01]. If one considers a classification task with $E(\|X\|^2) = 1$, and one assumes that the scattering is computed with unitary, admissible wavelets, then random test examples are located in the scattering unit ball. If one assumes a uniform distribution along this ball, then given a subspace $\mathbf{V}_d$ of dimension d, then $\|S_J x - P_{\mathbf{V}_d} x\| \approx (1 - \frac{d}{N}_J)\|S_J x\|$ with high probability. The event that $\|S_J x - P_{\mathbf{V}_{d,k}} S_J x\|$ is significantly smaller than $\|S_J x\|$ thus has a very small probability under the hypothesis that x is drawn from a uniform distribution in the unit ball, and hence it results in an "a contrario" test for belonging to class k.

3.4.2 Renormalization

As in the case of Support Vector Machines, the performance of the affine PCA classifier can be improved by equalizing the descriptor space.

Table 2.1 from Chapter 2 showed that scattering vectors have unequal energy distribution along its path variables, in particular as the order varies.

A first robust equalization method which is widely used on Support Vector Machines is obtained by re-normalizing each $S_J[p]x(u)$ by the maximum $\|S_J[p]x_i\| = \left(\sum_u |S_J[p]x_i(u)|^2 \right)^{1/2}$ over all training signals X_i:

$$\frac{S_J[p]x(u)}{\sup_{x_i} \|S_J[p]x_i\|} . \tag{3.10}$$

This supervised scheme ensures that all paths of the renormalized scattering descriptor have the same maximum energy across the training set. The supremum is preferred to the average in (3.10) since the underlying distribution of $S_J[p]x_i$ generally contains outliers.

The *scattering transfer*, which will be introduced and studied in Chapter 5, brings another renormalization scheme. It is defined for scattering paths $p = p_0 + \lambda$ of order $|p| > 1$ as

$$T_J X[p](u) = \frac{S_J[p]X(u)}{S_J[p_0]X(u)} , \tag{3.11}$$

and for first order coefficients as $T_J X[\lambda](u) = \frac{S_J[\lambda]X(u)}{|X|\star\phi_J(u)}$. Sections 5.3 and 4.7 will prove that this renormalization produces scattering coefficients which are nearly invariant to smooth modulations, which model illumination changes; and to the derivation operator $DX(u) = \frac{dX}{du}(u)$. As we shall see, this strong invariance property can be interpreted as a form of geometric invariant.

3.4.3 Comparison with Discriminative Classification

The classification of observations x into labels y can be stated in probabilistic terms as maximizing the conditional probability $\max_i p(y = c_i|x)$. Generative classifiers exploit the Bayes rule

$$p(y|x) = \frac{p(y, x)}{p(x)} = \frac{p(x|y)p(y)}{p(x)}$$

and estimate the class-conditional probabilities $p(x|y)$ with a specified model, whereas discriminative classifiers directly estimate the posterior probabilities $p(y|x)$, or, equivalently, learn a mapping from observations x to the labels y.

Several authors [Vap98; NJ02] showed that discriminative classifiers have better asymptotic properties than generative classifiers, since they directly estimate the final classification objective function. However, it has also been observed [NJ02; UB05] that under certain circumstances, generative classifiers can outperform discriminative ones, thanks to a better trade-off between the bias and variance of the estimation.

Section 3.4 argued that thanks to the Lipschitz regularity of scattering operators with respect to additive perturbations and geometric deformations, signal classes with a deformable template structure are well approximated by low-dimensional affine spaces in the scattering domain. The linear generative classifier from Section 3.4.1 is able to exploit such regularity by diagonalizing the empirical intra-class covariance of the observed data.

We study the pertinence of such classification strategy by modeling a binary classification problem, where observations $x \in \mathbb{R}^N$ are drawn from Gaussian distributions

$$(x|y = 0) \sim \mathcal{N}(\mu_0, \Sigma_0) , \quad (x|y = 1) \sim \mathcal{N}(\mu_1, \Sigma_1) .$$

The covariance matrices Σ_0 and Σ_1 are constructed as

$$\Sigma_i = U_i D_r U_i^T , \ i = 0, 1 , \tag{3.12}$$

where U_i is a random orthogonal matrix of dimension N and D_r is a diagonal matrix such that

$$D_r(n,n) = \sigma_0^2 n^{-r} \ , \ n = 1 \dots N \ .$$

The parameter r thus controls the decay of the eigenvectors of Σ_i. We also set the signal-to-noise ratio ρ of the model by normalizing the total variance of each class:

$$\rho = \frac{\|\mu_0 - \mu_1\|^2}{\sigma_0^2 \sum_{n=1}^{N} n^{-r}} \ .$$

We compare the linear generative classifier described in Section 3.4.1 with an SVM using a Gaussian Kernel. In both methods, the classifier estimates the hyper-parameters using a validation set, taken off the training set. In the PCA classifier, it is limited to the dimension d of the affine spaces, whereas in the SVM case we estimate the parameters of the Gaussian kernel together with the margin cost [CL11].

Figure 3.2 shows the classification results as a function of the training size T and the decay r of the covariance spectrum. We generated data using $N = 40$, $\rho = 0.1$, with training set sizes of $T = 20, 80, 320, 1280, 2560$ and test size of 1000 samples. We averaged the classification results over 4 runs, and we used a validation set obtained with 20% of the training samples selected at random. When r is small, the covariance spectrum decays slowly, which implies that each class spans a relatively large subspace. In these conditions, the PCA classifier does not outperform the SVM at any regime. As the training set grows, the bias-variance trade-off turns in favor of the richer SVM model. However, as r increases, the Gaussian distributions produce data which is better approximated by low-dimensional linear subspaces; the PCA classifier takes advantage of this regularity more efficiently than the SVM classifier, and as a result it reaches its asymptotic regime faster than the SVM, as shown in figure 3.2. This phenomena is consistent with the theoretical analysis performed on binary data in [NJ02]. We shall see in the next section that real-world data such as handwritten digits exhibit a similar behavior, where for small training sets the PCA classifier outperforms its discriminative counterpart.

3.5 Handwritten Digit Classification

The MNIST database of hand-written digits is an example of structured pattern classification, where most of the intra-class variability is due to local translations and deformations. It comprises at most 60000 training samples and 10000 test samples. If the training dataset is not augmented with deformations, the state of the art was achieved by deep-learning convolutional networks [RHBL07], deformation models [KDGH07; AT07], and dictionary learning [MBP10]. These results are improved by a scattering classifier.

All computations are performed on the reduced cosine scattering representation described in Section 3.3, which keeps the lower-frequency half of the coefficients. Table 3.1 computes classification errors on a fixed set of test images, depending upon the size of the training set, for different representations and classifiers. The affine space selection of

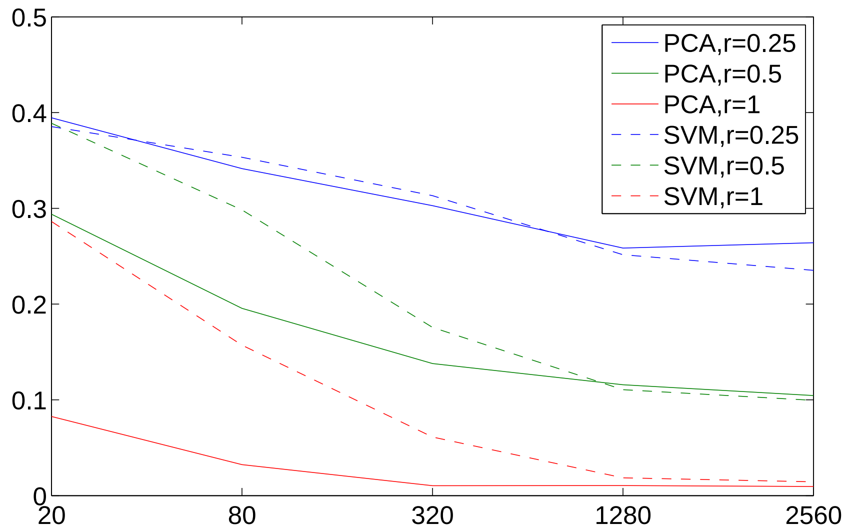
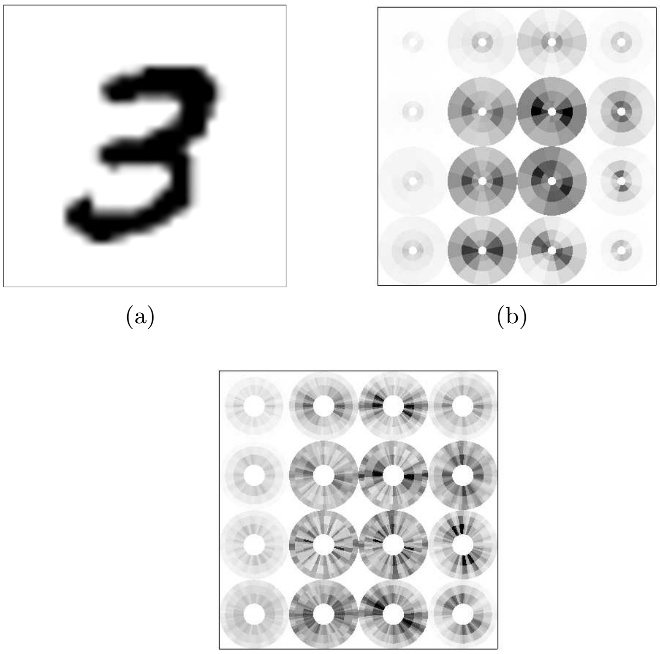


Figure 3.2: Comparison between PCA and SVM classifiers for the simulated data of model (3.12), for different training sizes and different decay of the covariance spectra.

Section 3.4.1 is compared with an SVM classifier using RBF kernels, which are computed using Libsvm [CL11], and whose variance is adjusted using standard cross-validation over a subset of the training set. The SVM classifier is trained with a renormalization which maps all coefficients to $[-1, 1]$. The PCA classifier is trained with the renormalisation (3.10), and similar results were obtained with the scattering transfer renormalization (3.11). The first two columns of Table 3.1 show that classification errors are much smaller with an SVM than with the PCA algorithm if applied directly on the image. The 3rd and 4th columns give the classification error obtained with a PCA or an SVM classification applied to the modulus of a windowed Fourier transform. The spatial size 2^J of the window is optimized with a cross-validation which yields a minimum error for $2^J = 8$. It corresponds to the largest pixel displacements due to translations or deformations in each class. Removing the complex phase of the windowed Fourier transform yields a locally invariant representation but whose high frequencies are unstable to deformations, as explained in Chapter 2. Suppressing this local translation variability improves the classification rate by a factor 3 for a PCA and by almost 2 for an SVM. The comparison between PCA and SVM confirms the fact that generative classifiers can outperform discriminative classifiers when training samples are scarce [NJ02]. As the training set size increases, the bias-variance trade-off turns in favor of the richer SVM classifiers, independently of the descriptor.

Columns 6 and 8 give the PCA classification result applied to a windowed scattering representation for $m_{\max} = 1$ and $m_{\max} = 2$. The cross validation also chooses $2^J = 8$. For the digit '3', Figure 3.3 displays the 4-by-4 array of normalized scattering vectors.

For each $u = 2^J(n_1, n_2)$ with $1 \leq n_i \leq 4$, the scattering vector $S_J[p]X(u)$ is displayed for paths of length $m = 1$ and $m = 2$, as circular frequency energy distributions following Section 2.6.2. Figure 3.4 displays the centroid $E(S_J X_k)$ estimated for the class '1', together with the principal direction of variability, corresponding to the eigenvector of largest eigenvalue of the empirical covariance. The panels (e) and (f) of the figure show that the principal source of variability remaining in the scattering domain is a rotation of the digit, since the principal direction oscillates along the circular arcs $r(p) = \{(r\lambda_1, r\lambda_2)\,;\, p = (\lambda_1, \lambda_2)\,,\, r \in G^+\}$.



(a) (b)

(c)

Figure 3.3: (a): Image $X(u)$ of a digit '3'. (b): Array of scattering vectors $S_J[p]X(u)$, for $m = 1$ and u sampled at intervals $2^J = 8$. (c): Scattering vectors $S_J[p]X(u)$, for $m = 2$.

Increasing the scattering order from $m_{\max} = 1$ to $m_{\max} = 2$ reduces errors by about 30%, which shows that second order coefficients carry important information even at a relatively small scale $2^J = 8$. However, third order coefficients have a negligible energy and including them brings marginal classification improvements, while increasing computations by an important factor. As the learning set increases in size, the classification improvement of a scattering transform increases relatively to windowed Fourier transform because the classification is able to incorporate more high frequency structures, which have deformation instabilities in the Fourier domain as opposed to the scattering domain.

Table 3.1 also shows that below $5 \cdot 10^3$ training samples, the scattering PCA clas-

Table 3.1: MNIST classification results.

Training size	x PCA	x SVM	Wind. PCA	Four. SVM	Scat. $m_{\max}=1$ PCA	Scat. $m_{\max}=1$ SVM	Scat. $m_{\max}=2$ PCA	Scat. $m_{\max}=2$ SVM	Conv. Net.	POP [AT07]
300	14.5	15.4	7.35	7.4	5.7	8	4.7	5.6	7.18	**3**
1000	7.2	8.2	3.74	3.74	2.35	4	2.3	2.6	3.21	**1.75**
2000	5.8	6.5	2.99	2.9	1.7	2.6	**1.3**	1.8	2.53	-
5000	4.9	4	2.34	2.2	1.6	1.6	**1.03**	1.4	1.52	1.11
10000	4.55	3.11	2.24	1.65	1.5	1.23	0.88	1	0.85	**0.8**
20000	4.25	2.2	1.92	1.15	1.4	0.96	0.79	**0.58**	0.76	-
40000	4.1	1.7	1.85	0.9	1.36	0.75	0.74	**0.53**	0.65	-
60000	4.3	1.4	1.80	0.8	1.34	0.62	0.7	**0.43**	0.53	0.68

sifier improves results of a deep-learning convolutional networks, which learns all filter coefficients with a back-propagation algorithm [FKL10], and obtains comparable results as the Patchwork Of Parts model from [AT07], which estimates a mixture of deformable parts. As more training samples are available, the flexibility of the SVM classifier brings an improvement over the more rigid affine classifier, yielding a 0.43% error rate on the original dataset, thus improving upon previous state of the art methods.

To evaluate the precision of the affine space model, we compute the relative affine approximation error, averaged over all classes:

$$\sigma_d^2 = K^{-1} \sum_{k=1}^{K} \frac{E(\|S_J X_k - P_{\mathbf{A_{d,k}}}(S_J X_k)\|^2)}{E(\|S_J X_k\|^2)} .$$

For any $S_J X_k$, we also calculate the minimum approximation error produced by another affine model $A_{d,k'}$ with $k' \neq k$:

$$\lambda_d = \frac{E(\min_{k' \neq k} \|S_J X_k - P_{\mathbf{A_{k',d}}}(S_J X_k)\|^2)}{E(\|S_J X_k - P_{\mathbf{A_{d,k}}}(S_J X_k)\|^2)} .$$

For a scattering representation with $m_{\max} = 2$, Table 3.2 gives the dimension d of affine approximation spaces optimized with a cross validation, with the corresponding values of σ_d^2 and λ_d. When the training set size increases, the model dimension d increases because there are more samples to estimate each intra-class covariance matrix. The approximation model becomes more precise so σ_d^2 decreases and the relative approximation error λ_d produced by wrong classes increases. This explains the reduction of the classification error rate observed in Table 3.1 as the training size increases.

The US-Postal Service is another handwritten digit dataset, with 7291 training samples and 2007 test images 16×16 pixels. The state of the art is obtained with tangent distance kernels [HK02]. Table 3.3 gives results obtained with a scattering transform

Table 3.2: Values of the dimension d of affine approximation models on MNIST classification, of the intra class normalized approximation error σ_d^2, and of the ratio λ_d between inter class and intra class approximation errors, as a function of the training size.

Training	d	σ_d^2	λ_d
300	5	$3 \cdot 10^{-1}$	2
5000	100	$4 \cdot 10^{-2}$	3
40000	140	$2 \cdot 10^{-2}$	4

with the PCA classifier for $m_{\max} = 1, 2$. The cross-validation sets the scattering scale to $2^J = 8$. As in the MNIST case, the error is reduced when going from $m_{\max} = 1$ to $m_{\max} = 2$ but remains stable for $m_{\max} = 3$. Different renormalization strategies can bring marginal improvements on this dataset. If the renormalization is performed by equalizing using the standard deviation of each component, the classification error is 2.3% whereas it is 2.6% if the supremum is normalized.

Table 3.3: Percentage of errors for the whole USPS database.

Tang. Kern.	Scat. $m_{\max} = 2$ SVM	Scat. $m_{\max} = 1$ PCA	Scat. $m_{\max} = 2$ PCA
2.4	2.7	3.24	2.6 / **2.3**

The scattering transform is stable but not invariant to rotations. Stability to rotations is demonstrated over the MNIST database in the setting defined in [LBLL09]. A database with 12000 training samples and 50000 test images is constructed with random rotations of MNIST digits. The PCA affine space selection takes into account the rotation variability by increasing the dimension d of the affine approximation space. This is equivalent to projecting the distance to the class centroid on a smaller orthogonal space, by removing more principal components. The error rate in Table 3.4 is much smaller with a scattering PCA than with a convolution network [LBLL09]. Much better results are obtained for a scattering with $m_{\max} = 2$ than with $m_{\max} = 1$ because second order coefficients maintain enough discriminability despite the removal of a larger number d of principal directions. In this case, $m_{\max} = 3$ marginally reduces the error.

Table 3.4: Percentage of errors on an MNIST rotated dataset [LBLL09].

Scat. $m_{\max} = 1$ PCA	Scat. $m_{\max} = 2$ PCA	Conv. Net.
8	**4.4**	8.8

Table 3.5: Percentage of errors on scaled and/or rotated MNIST digits

Transformed Images	Scat. $m_{\max} = 1$ PCA	Scat. $m_{\max} = 2$ PCA
Without	1.6	0.8
Rotation	6.7	3.3
Scaling	2	1
Rot. + Scal.	12	5.5

Scaling invariance is studied by introducing a random scaling factor uniformly distributed between $1/\sqrt{2}$ and $\sqrt{2}$. In this case, the digit '9' is removed from the database as to avoid any indetermination with the digit '6' when rotated. The training set has 9000 samples (1000 samples per class). Table 3.5 gives the error rate on the original MNIST database and including either rotation, scaling, or both in the training and testing samples. Scaling has a smaller impact on the error rate than rotating digits because scaled scattering vectors span an invariant linear space of lower dimension. Second-order scattering outperforms first-order scattering, and the difference becomes more significant when rotation and scaling are combined, because it provides interaction coefficients which are discriminative even in presence of scaling and rotation variability.

3.6 Towards an Object Recognition Architecture

The previous section showed the efficiency of scattering representations to describe signal classes with a deformable structure. General object recognition datasets, such as Caltech-101 [FFFP04], Pascal [EVGW+] or Imagenet [DDS+09], contain object classes with far more complex sources of variability, including occlusions, clutter or complex changes of shape and/or texture, as shown in figure 1.3.

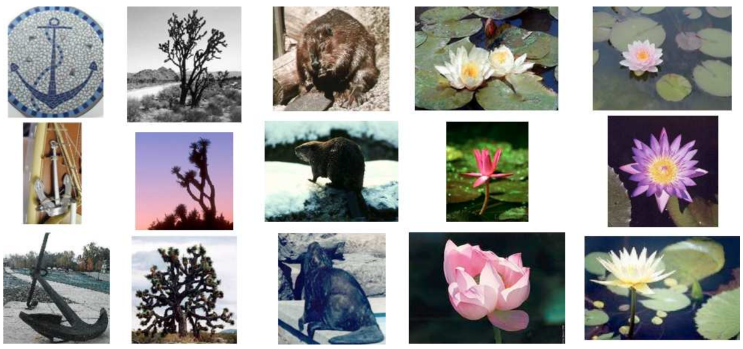


Figure 3.5: Examples from the Caltech dataset. Intra-class variability includes changes in shape, appearance, clutter.

In these cases, the windowed scattering representation brings local image descriptors which capture high frequency information while being stable to non-rigid deformations and locally translation invariant. The residual variability after the scattering network cannot in general be absorbed with linear models, as it was the case in the handwritten digit task. Variability coming from other physical transformation groups, such as rotations or scaling, can be reduced with a scattering operator defined on a roto-translation group [SM12]. However, complex datasets exhibit other sources of variability, which do not result from the action of any transformation group: for instance, the clutter in the background of objects, or the variability in the shapes of, say, chairs.

Many object recognition architectures [LSP06; YYGH09; BBLP10] use SIFT or HoG as their first layer of processing. Once image patches are transformed using these visual descriptors, most object architectures encode the descriptors using supervised or non-supervised methods. Sparse coding strategies [BBLP10] learn a dictionary of visual features using a sparse inducing criteria to encode each transformed path into a sparse code [BBLP10]. These transformed codes are successively delocalized by pooling neighboring codes using max-pooling or average pooling [LSP06]. Sparse coding can be replaced by other encoding strategies; for instance, vector quantization gives rise to bags-of-words architectures [LSP06], and it is learnt with a clustering of visual features. Local coordinate coding [YGZ09] is an encoding strategy half-way between vector quantization and sparse coding, which encodes a visual feature with a linear combination of its closest prototypes.

Finally, deep neural networks, and in particular convolutional networks, have recently achieved state-of-the-art results on several complex object recognition tasks [KSH12]. They learn a huge network of filter banks and non-linearities on large datasets, using both supervised and non-supervised methods. Whereas the first layers of the network learn local structures such as oriented edges and corners, higher layers are able to learn more complex relationships of input images, spanning larger spatial neighborhoods.

These non-linear encoding strategies all require learning from data to some extent. Whereas variability to physical transformations such as translation or rotation is universal and does not need to be learnt, learning becomes important in order to address more complex sources of variability. In this context, wavelet scattering networks may provide the first layers of these general object recognition architectures. The invariance and stability properties of scattering operators have the capacity to simplify the learning task of subsequent layers, since they map image patches into a regular manifold thanks to their Lipschitz continuity properties.

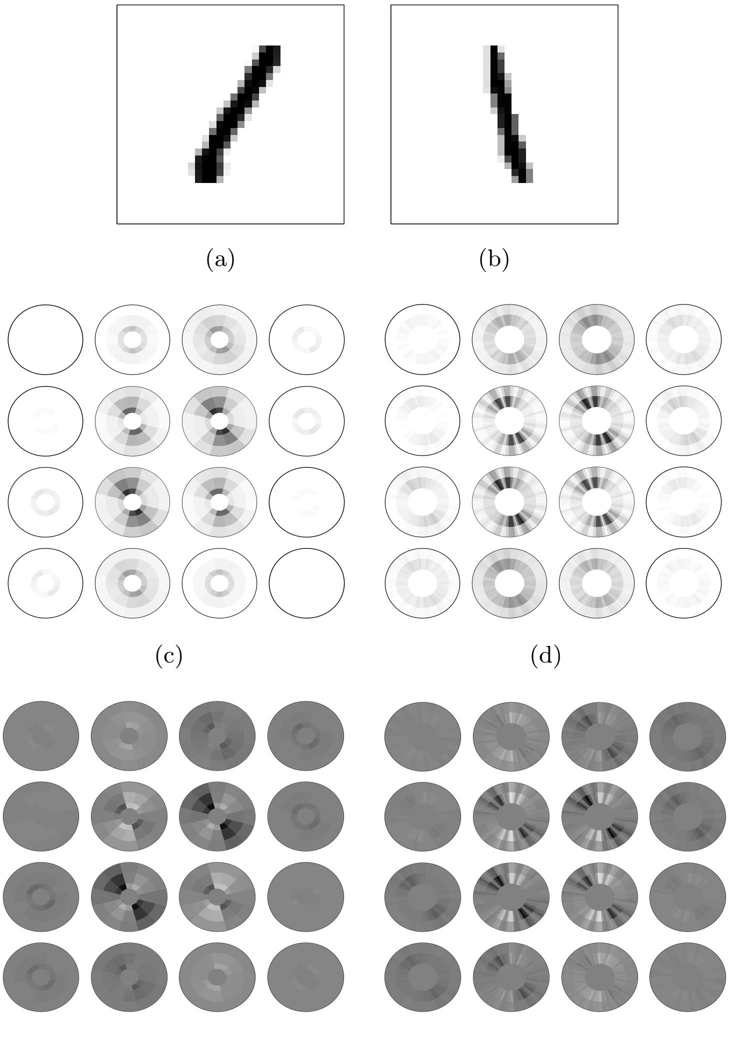

(a) (b)
(c) (d)
(e) (f)

Figure 3.4: Normalized scattering coefficients for the '1' class. (a),(b): Two examples $X_1(u)$, (c) $E(S_J[p]X_1(u))$, for $|p| = 1$, (d) $E(S_J[p]X_1(u))$, for $|p| = 2$. (e) first order coefficients of the first principal direction, (f) second order coefficients of the first principal direction. Observe that this principal direction is capturing a rotation of the digit.

Chapter 4

Texture Discrimination and Synthesis with Scattering

4.1 Introduction

In the previous chapter we modeled object classes as deterministic functions in $\mathbf{L}^2(\mathbb{R}^d)$ with an associated structure of deformations, which might include geometrical and photometric transformations, shape variability, occlusion, etc. While this description is useful for the study of many pattern recognition problems, it is not adequate to represent the class of, say, images of grass or sand, or applause sound clips. These examples are naturally modeled as realizations of stochastic processes, referred as image or auditory textures in our context. Thus, the observed variability requires to be treated from a statistical point of view. Stationary processes are particularly relevant since they express the property that most textures do not depend upon the spatial or temporal reference.

In this chapter we study the problems of texture discrimination and reconstruction. In the first case, we consider a family of textures, modeled as stationary processes $X_1, \dots, X_K$. We observe one or several realizations of each process, which are used to learn a texture model for each class, and then a classifier assigning a label $\{1 \dots K\}$ to each new incoming texture realization. In the texture reconstruction problem, we are given a realization x of an unknown stationary process X, and the goal is to reproduce independent and perceptually similar realizations of x.

50 years ago, Julesz formulated the hypothesis [Jul62] that the perceptual information of a given texture X was contained in a finite collection of statistics $\{E(g_i(X)), i = 1..I\}$, in the sense that if $E(g_i(X)) = E(g_i(Y))$ for all i then X and Y are perceptually equivalent. Julesz originally stated the hypothesis in terms of second-order statistics, which measure pairwise interactions of the form $g_i(X) = g_i(X(u)X(u + \tau_k))$. He later gave counterexamples showing that third or higher order statistics were required, finally leading to a formulation in terms of *textons* [Jul81], local texture descriptors capturing interactions across different scales and orientations.

Texture classification and recognition both require a texture representation capturing enough discriminative information about the underlying process. However, they impose different constraints on the representation. Whereas in texture synthesis is imperative to keep all the statistics which are perceptible in the sense of Julesz, texture classification requires a representation which can be estimated consistently with few realizations, and which is stable to geometric and photometric deformations.

One can then ask for a texture representation which can be used effectively in both classification and reconstruction contexts. The expected Scattering representation incorporates high order moments information, and is consistently estimated for a wide class of ergodic processes thanks to the fact that it is computed from non-expansive operators. Moreover, Section 2.3.3 showed that it provides stability with respect to deformations. This chapter shows that the expected scattering can be used as a texture representation for both classification and reconstruction, thanks to its stability, consistency and highly informative content.

Gaussian processes are a fundamental class of stochastic processes. Thanks to their particularly rich structure, we are able to obtain analytical properties of their first and second order scattering coefficients, which will be of particular interest in Chapter 5. Besides Gaussian processes, we also consider stochastic modulation processes, which are of special interest on auditory texture modeling.

The rest of the chapter is structured as follows. Section 4.3 concentrates on image texture classification, with the Curet material texture dataset [DVGNK99]. Each class is modeled as a process of the form $L_\theta X_i$, where X_i is a stationary process, representing the material under some canonic lighting and viewing conditions, and L_θ is a random geometric and photometric transformation. We show that the expected scattering representation, followed by the generative PCA classifier from Chapter 3, builds a highly discriminative and consistent estimator for each class, yielding a significative improvement over state-of-the-art results. Section 4.4 studies auditory texture classification from a dataset compiled by McDermott&Simoncelli [McD]. We confirm numerically that scattering representations discriminate non-gaussian stationary processes, by adding to the dataset Gaussian processes with the same spectral density as each process in the dataset.

Section 4.5 focuses on texture synthesis from expected scattering representations. For that purpose, we introduce a gradient descent algorithm in Section 4.5.1, which adjusts first and second order scattering coefficients, and we place ourselves in the same statistical framework of [MS11], which approximates the sampling of maximum entropy distributions with samples from the Julesz ensemble. We apply the scattering synthesis to the auditory textures from Section 4.4, confirming the fact that second order scattering representations capture enough information to produce realistic auditory texture reconstruction. Finally, Section 4.6 focuses on first and second order scattering of Gaussian processes, whereas Section 4.7 studies stochastic modulation models, showing that by properly renormalizing second order scattering coefficients, one can separate the influence of the carrier from that of the envelope.

4.2 Texture Representations for Recognition

Texture representation has long been a major research topic in computer vision and signal processing. We give a short overview of the literature, with special attention to the problems of texture discrimination and synthesis.

The law of a stationary process X is an object living in an extremely high dimensional space. For instance, if we assume a real, discrete process $X[n]$ whose samples become statistically independent beyond a window of size S, then its distribution is specified by a density $f_X \in \mathbf{L}^1(\mathbb{R}^S)$, an object of huge complexity which in general cannot be estimated with a limited number of samples. Texture discrimination and reconstruction thus require a concise statistical representation of the process, capturing enough discriminative information for the task at hand and such that it can be consistently estimated from the available observations.

In the discrimination problem, moreover, the representation should be estimated consistently with few realizations, and should also be stable to changes in the acquisition process which are not perceptually relevant, as in Chapter 3. For instance, the representation of a textured material should be stable to changes in pose, illumination or, more generally, geometrical and photometric deformations.

4.2.1 Spectral Representation of Stationary Processes

A stationary process $X(t)$ admits a spectral representation, via the Cramer decomposition

$$\forall\, t \;\; X(t) = \int e^{i\omega t} dZ(\omega) \ ,$$

where $dZ(\omega)$ is a zero-mean random spectral measure, which is decorrelated, $E(Z(I)Z(J)) = 0$ if $I \cap J = \emptyset$, and whose energy, called the spectral density of X, can be obtained from the auto-correlation of $X(t)$:

$$E(|Z(I)|^2) = \int_I \hat{R}_X(\omega) d\omega \ ,$$

where $R_X(\tau) = E(X(t)X * (t+\tau))$. If $X(t)$ is Gaussian, then the spectral density completely characterizes the process from its second order moments, which explains their popularity in many domains of signal processing. However, most textures found in nature are not well modeled as Gaussian processes, as shown in figure 4.1. In particular, it results that high order moments contain essential discriminative information. A useful texture representation thus should depend upon high order moments.

4.2.2 High Order Spectral Analysis

The spectral density $\hat{R}_X(\omega)$ of a stationary process $X(t)$ is obtained from the second order auto-covariance function $R_X(\tau) = E((X(t) - E(X))(X(t+\tau) - E(X))^*)$. Higher order Spectral analysis [Pet99] generalizes the spectral density by considering third and higher order cumulants and their multidimensional Fourier transforms. In particular,

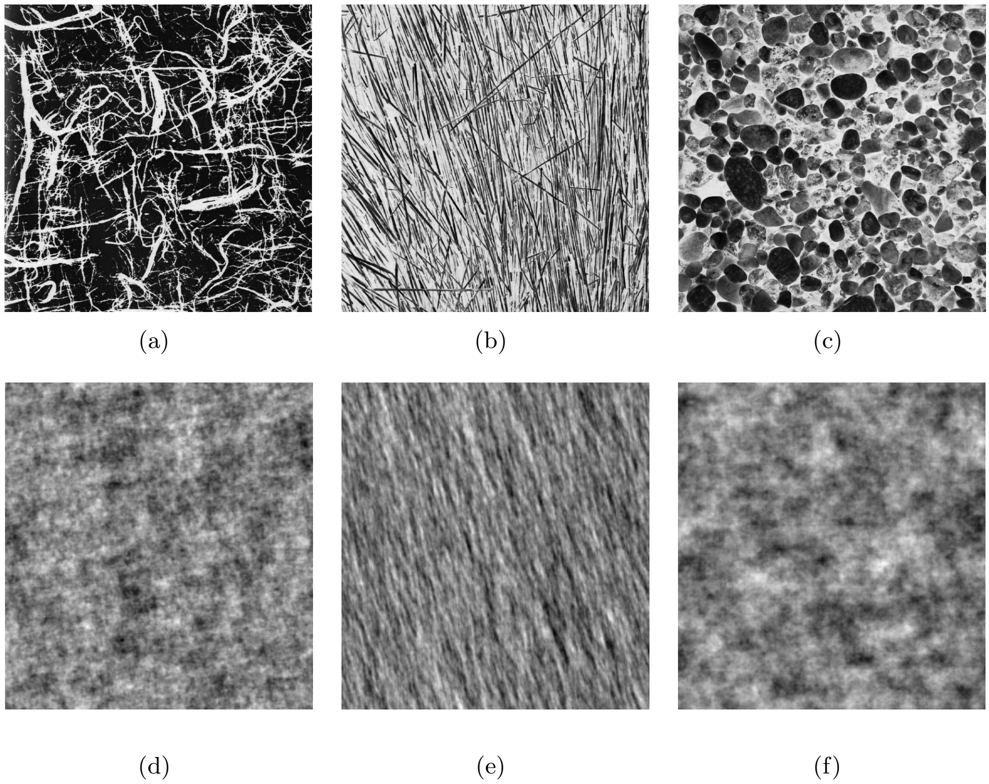


Figure 4.1: First row: Three different examples x_i from Brodatz dataset, $i = 1 \ldots 3$. Bottom row: Reconstruction obtained by equalizing white gaussian noise with each spectral density $\hat{R}_{x_i}$, $i = 1 \ldots 3$, so that the textures on each column have the same second order moments.

the bi-spectrum is defined for a unidimensional stationary process $X(t)$ as the two-dimensional Fourier transform of the third-order cumulant

$$C_3X(\tau_1, \tau_2) = E(X(t)X(t+\tau_1)X(t+\tau_2)) - E(X)(R_X(\tau_1) + R_X(\tau_2) + R_X(\tau_1 - \tau_2)) + 2E(X)^3 .$$

Higher order spectral analysis has been applied to texture discrimination in [HK76] with relative success. However, an accurate estimation of high order cumulants requires large amounts of samples, as shown in [ANVV97], due to the high variance associated to their estimates. Indeed, the estimation of high moments amounts to estimating the expected values $E(g(X(t), \dots, X(t+\tau_{k-1})))$, with

$$g(x_1, \dots x_k) = \prod_{i \leq k} |x_i|^{\alpha_i} \ , \sum_i \alpha_i = m > 1 \ .$$

The function $g(x)$ effectively expands large values of its arguments, making the expected value more and more dominated by rare events which increase the variance of the estimator.

4.2.3 Markov Random Fields

Markov Random Fields are an important family of stationary processes which have been used for both texture discrimination and synthesis. A discrete stationary process $X[n]$ is called a Markov Random Field (MRF) if the conditional probability distribution $f(X[n] = x | X[m] = x_m \, , \, m \neq n)$, which in general depends upon all values of $(x_m)_{m \neq n}$, can be written as a function only of the neighbors $(x_m)_{m \in \mathcal{N}_n}$, for a given neighborhood $\mathcal{N}$, which is independent of n since $X[n]$ is stationary. If $X[n]$ is a Markov random field then its probability distribution is a Gibbs distribution [GG84]. The size of $\mathcal{N}$ governs the complexity of the model, which captures high order statistics up to order $|\mathcal{N}|$.

MRFs have been used in both texture synthesis and discrimination. In [PL98] the authors study image texture synthesis by using non-parametric, multiscale MRFs. In [ZFCV02] the authors extract only first and second-order statistics, along a family of cliques, thus capturing pixel co-occurrences. They use these statistics to define a Gibbs distribution, from which they sample using a Metropolis algorithm. The joint pixel distributions, while capturing high-order statistics, have a complexity which grows exponentially with the size of the neighborhoods, making their estimation soon unfeasible using a small realization of the process.

Moreover, such distributions are not stable to a number of input transformations, such as deformations, perspective effects or pose changes. In [VZ03], MRFs are used in texture classification. Local distributions are learned by doing a vector quantization of the joint distributions. This clustering step generates partial invariance, but it also loses information. Recovering the lost information from the a nonlinear clustering step is difficult, as opposed to linear averaging.

4.2.4 Wavelet based texture analysis

Wavelet filter banks have been used in both texture synthesis and discrimination [HB95; MP90; LM01]. Marginal statistics for each filter output capture important geometrical features of the texture, but [PS99] showed that one should not ignore the co-occurrence or joint statistics across wavelet coefficients in order to achieve good synthesis. Simoncelli and Portilla build a steerable pyramid representation, which is a filter bank of oriented, multiscale complex wavelets, consisting in 4 orientations and 4 scales, and which is designed to satisfy the unitary Littlewood-Paley condition. The authors consider second moments of $X \star \phi_J$ for $J = 1..4$, as well as skewness and kurtosis, together with auto-correlation samples of $|X \star \psi_\lambda|$, $\lambda = r2^j$, where ψ is a generator of the steerable pyramid and λ spans 4 octaves and and 4 different orientations. From the Parseval identity, we know that these autocorrelation samples have the same stability as the power spectrum $E(||\widehat{X \star \psi_\lambda}||^2)$, which is not stable to non-rigid deformations, as seen in 2.2.4.

More recently, synthesis of auditory textures from envelope statistics has been successfully applied in [MS11]. Here, the authors start by decomposing the signal with a filter bank of 30 cochlea filters satisfying the Littlewood-Paley unitary condition. They are bandpass filters whose bandwidth increases with the central frequency, and such that they remain approximately constant on a logarithmic scale. Then they extract the envelope by taking the modulus of the analytic signal for each of the outputs of the filter bank, which corresponds to using complex bandpass filters such as those considered in the scattering decomposition. The envelopes at the output of each filter are next compressed to reproduce the nonlinear response of the auditory system, and are re-decomposed with a new family of 20 bandpass filters, the modulation filters, which have similar design as the cochlea filters but operate at a lower frequency range (similarly as the outputs of second order progressive scattering). The authors then compute statistics of each cochlea envelope and their corresponding modulation bands. The statistics consist in marginal moments (mean, variance, skew and kurtosis) and pairwise correlations for the cochlea envelopes, as well as variance and pairwise cross-correlation for the modulation bands, which form a vector of total dimension approximately 1400.

4.2.5 Maximum Entropy Distributions

Texture synthesis from a given statistical representation has been studied since Julesz. In absence of any other information, the distribution characterized by the expected values $\{E(g_k(X)) \, , k = 1..K\}$ is the maximum entropy distribution. It ensures that the amount of prior information outsize the set of constraints is minimized. The maximum entropy distribution is characterized by the Boltzmann theorem, which assesses that under some conditions on the probability space P, the probability density $p(x)$, $x \in P$, of maximum entropy subject to the constraints

$$E(g_k(X)) = h_k \, , \, k = 1 \ldots K \tag{4.1}$$

has the form

$$p(x) = \frac{1}{Z} \exp\left(\sum_{k=1 \ldots K} -\lambda_k g_k(x) \right) \, , \; x \in P \, . \tag{4.2}$$

Here, Z is the partition function and λ_k are the Lagrange multipliers associated to the constrained entropy optimization, which need to be adjusted such that the resulting density satisfies the expected value constraints (4.1). One can sample from such distribution using Gibbs sampling and MCMC methods, but given the high dimensional nature of the probability space, these methods are computationally intensive and require long iterations. Moreover, the conditional distributions of (4.2) do not have a simple expressions for non-linear constraints g_k. In [ZWM97], the authors proposed a simpler framework for obtaining new samples, consisting in projecting a realization of a high entropy distribution, such as gaussian white noise, to the set of realizations satisfying the expected value constraints, referred as *Julesz Ensemble*:

$$\{x \in \mathbf{L}^2(\mathbb{R}^d)\,,\, g_k(x) = h_k\,,\, k = 1 \dots K\}\ .$$

4.2.6 Exemplar based texture synthesis

Exemplar-based texture synthesis is a another family of texture synthesis methods with successful numerical results [EF01; KEBK05; LH05], which have made them popular in the computer graphics community. The general principle is to consider an original realization of a stationary texture X, and to enlarge it progressively, such that it remains visually similar to X. These methods assume a Markov property valid for a certain neighborhood $\mathcal{N}$. Then, rather than estimating the conditional probability distribution $f(X = x|\mathcal{N}_x = \mathbf{p})$, the method searches for patches in the input which are similar to $\mathbf{p}$, and then retrieves the central pixel value by choosing one of those matches at random. The quality of the synthesized textures thus depends upon the ability to find similar patches within the initial texture realization. Periodic and highly structured textures are examples where the auto-correlation $R_X(\tau) = E(X(t)X(t+\tau))$ has large values for nonzero τ, but for white noise $R_X(\tau) = \sigma^2\delta(\tau)$, and hence the method cannot find similar patches. Exemplar-based methods indirectly estimate high order statistics with a single realization, which has large variance. For synthesis purposes this is not a problem, but discrimination requires consistent estimators, which in particular impose a representation with smaller dimensionality.

Dictionary learning has been successfully applied to the texture synthesis problem in [Pey09]. A dictionary of patches is learnt by maximizing the amount of sparsity on a set of exemplar patches. Then, new realizations are obtained by solving a convex optimization program which enforces the synthesized image to have a sparse representation on the learnt dictionary.

4.2.7 Modulation models for Audio

Stationary processes admit a spectral decomposition, given by the Crámer decomposition theorem. The spectral density completely characterizes gaussian processes. However, it does not explain many stationary phenomena observed in nature.

Modulations correspond to multiplicative phenomena and are an essential step in the generation of sound. Multiplicative processes appear also in the construction of fractal processes, studied in Chapter 5. In its simplest setting, a modulation model is defined as the process

$$X(t) = X_s(t)X_f(t)\ , \tag{4.3}$$

where X_s is a smooth (slow) , positive stationary process and X_f is also stationary (and fast) and independent from X_s. The demodulation step consists in obtaining such decomposition from X. This is an ill-posed inverse problem, without a unique solution, which hence requires regularization. Similarly as in the additive spectral decomposition, it is necessary to specify a scale or frequency separation between the terms. Alternatively, if one assumes that $X_f \neq 0$ with probability 1, then an additive decomposition of $Y = \log |X| = Y_1 + Y_2$ leads to a possible modulation model for $X = |X| e^{i\theta(X)}$:

$$X = e^{Y_1+Y_2} e^{i\theta(X)} = (e^{Y_1})(e^{Y_2+i\theta(X)}) \ .$$

In audio processing, modulations play a central role in the generation of sounds. This fact motivates the study of the associated inverse problem, the demodulation, in order to obtain a signal representation in terms of the modulation components. Generally speaking, it consists in decomposing a signal $x(t)$ as

$$x(t) = a(t)c(t) \ , \tag{4.4}$$

where a is a positive, slow envelope, and c is an oscillatory, wide-band carrier. Demodulation is thus an ill-posed inverse problem which requires prior knowledge on the components to be recovered. Several methods exist in the literature, ranging from the Hilbert transform demodulation of Gabor [Gab46] to probabilistic demodulation methods, introduced in [Tur10], where a generative model is estimated.

The basic modulation model (4.4) is often not rich enough to express natural sounds. In [Sch07; Tur10] the model is generalized to incorporate modulated components at different temporal scales,

$$x(t) = \sum_{j=1}^{J} a_j(t) c_j(t) \ .$$

The components at each scale are estimated using sub-band demodulation algorithms [Sch07]. Turner [Tur10] also studied another generalisation where a signal is modulated multiple times, to form a modulation cascade.

4.3 Image texture discrimination with Scattering representations

Visual texture discrimination remains an outstanding image processing problem because textures are realizations of non-Gaussian stationary processes, which cannot be discriminated using the power spectrum. Depending on the imaging conditions, textures undergo transformations due to illumination, rotation, scaling or more complex deformations when mapped on three-dimensional surfaces.

If $X(t)$ is a stationary process, then Section 2.3.5 showed that under ergodicity conditions that we referred as mean squared scattering consistency, the expected scattering $\overline{S}X$ is estimated from a realization x of X using the windowed scattering $S_J x$, with a total variance which tends to zero as $J \to \infty$:

$$\lim_{J\to\infty} \sum_{p\in\mathcal{P}_J} E(|S_J X[p] - \overline{S}X|^2) = 0 \ . \tag{4.5}$$

For a wide range of ergodic processes, this property is observed numerically, with a variance decaying exponentially with J.

The expected scattering can thus be estimated from a single realization under appropriate ergodicity. However, visual textures are not always well modeled by ergodic processes. In particular, inspired by the CUReT texture database [LM01; VZ09], we shall model each texture class as follows. Let $\tilde{X}_i$ represent each material of the class, with i ranging through the different texture classes, taken under canonic lighting an pose, and let us assume it is a stationary, ergodic process satisfying (4.5). The observed process for each class is modeled as

$$X_i = L_\theta(\tilde{X}_i) \ ,$$

where L_θ is a random photometric and geometric transformation, modeling the change in viewing and lighting conditions. Here θ is a random variable encoding a global illumination and a similarity transform. Since global illumination and rigid affine transformations define both low-dimensional manifolds [BJ03], θ can be thought as a low-dimensional random vector. Although $\tilde{X}_i$ is ergodic, X_i is not ergodic in general, since each realization contains only one instance of the random global deformation vector θ.

For each class, we observe realizations x_k, $k = 1 \dots K$. The windowed scattering transform of each realization $S_J x_k$ converges to $\overline{S}L_{\theta_k}\tilde{X}$ as J increases. For sufficiently large J, the residual variability in $(S_J x_k)_k$ is thus dominated by the low-dimensional transformations L_{θ_k}, expressed in the scattering domain. This suggests using the same affine PCA space classifier from Chapter 3, applied on the transformed realizations $\{S_J x_k\}_k$. Indeed, low-dimensional affine spaces are able to absorb the principal directions of variance, and hence reduce the influence of global lighting and pose during classification.

Texture classification is tested on the CUReT texture database, which includes 61 classes of image textures of $N = 200^2$ pixels. Each texture class gives images of the same material with different pose and illumination conditions. Specularities, shadowing and surface normal variations make classification challenging. Figure 4.2 illustrates the large intra-class variability, after a normalization of the mean and variance of each textured image.

Table 4.1 compares error rates obtained with different classifiers. The database is randomly split into a training and a testing set, with 46 training images for each class as in [VZ09]. Results are averaged over 10 different splits. A PCA affine space classifier applied directly on the image yields a large classification error of 17%. The lowest published classification errors obtained on this dataset are 2% for Markov Random Fields [VZ09], 1.53% for a dictionary of textons [HCFE04], 1.4% for Basic Image Features [CG10] and 1% for histograms of image variations [Bro05]. To estimate the Fourier spectrum, windowed Fourier transforms are computed over half-overlapping windows of

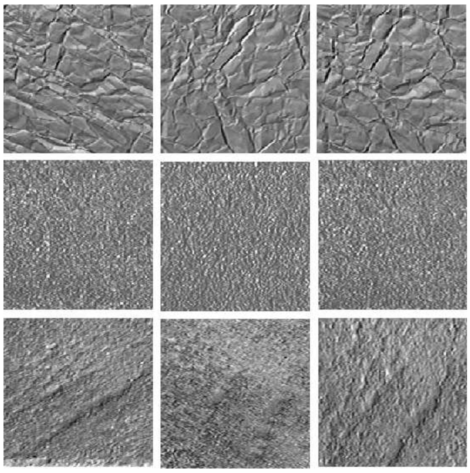


Figure 4.2: Examples of textures from the CUReT database with normalized mean and variance. Each row corresponds to a different class, showing intra-class variability in the form of stochastic variability and changes in pose and illumination.

Table 4.1: Percentage of classification errors of different algorithms on CUReT.

Training size	X PCA	MRF [VZ09]	Textons [HCFE04]	BIF [CG10]	Histo. [Bro05]	$E(\lvert\hat{X}\rvert^2)$ PCA	$\overline{S}$ $m_{max}=1$ PCA	$\overline{S}$ $m_{max}=2$ PCA
46	17	2	1.5	1.4	1	1	0.5	**0.2**

size 2^J, and their squared modulus is averaged over the whole image. This averaging is necessary to reduce the spectrum estimator variance, which does not decrease when the window size 2^J increases. The cross-validation sets the optimal window scale to $2^J = 32$, whereas images have a width of 200 pixels. The error drops to 1%.

For the scattering PCA classifier, the cross validation chooses an optimal scale 2^J equal to the image width to reduce the scattering estimation variance. Indeed, contrarily to a power spectrum estimation, the variance of the scattering vector decreases when 2^J increases. Figure 4.3 displays the scattering coefficients $S_J[p]x$ of order $m = 1$ and $m = 2$ of an example x from the Curet dataset.

A PCA classification with only first order coefficients ($m_{max} = 1$) yields a classification error of 0.5%. Although first-order scattering coefficients are strongly correlated with second order moments, whose values depend on the Fourier spectrum, the classification error is improved relatively to a power spectrum estimator because $S_J[\lambda_1]X = |X \star \psi_{\lambda_1}| \star \phi_{2^J}$ is an estimator of a first order moment $\overline{S}X(\lambda_1) = E(|X \star \psi_{\lambda_1}|)$ and thus has a lower variance than second order moment estimators. A PCA classification with first and second order scattering coefficients ($m_{max} = 2$) reduces the error to 0.2%. Indeed, scattering coefficients of order $m = 2$ depend upon moments of order

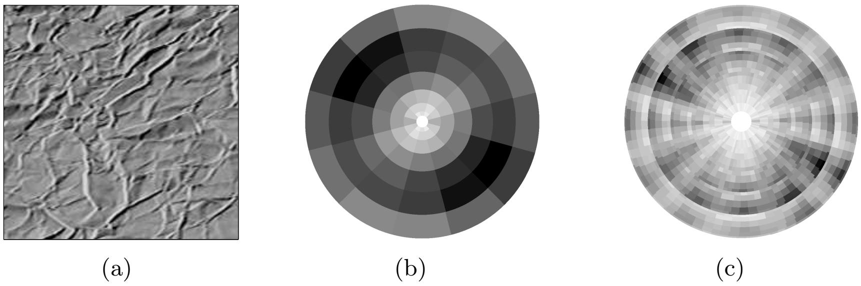


Figure 4.3: (a): Example of CureT texture $X(u)$. (b): Scattering coefficients $S_J[p]X$, for $m = 1$ and 2^J equal to the image width. (c): Scattering coefficients $S_J[p]X(u)$, for $m = 2$.

4, which are necessary to differentiate textures having same second order moments as in Figure 4.1. Moreover, the estimation of $\overline{S}X(\lambda_1, \lambda_2) = E(||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|)$ has a low variance because X is transformed by a non-expansive operator, as opposed to X^q for high order moments $q \geq 2$.

For $m_{max} = 2$, the cross validation chooses affine space models of small dimension $d = 16$. However, they still produce a small average linear approximation error. By recalling the approximation error and separation ratio measures of affine PCA spaces introduced in previous chapter,

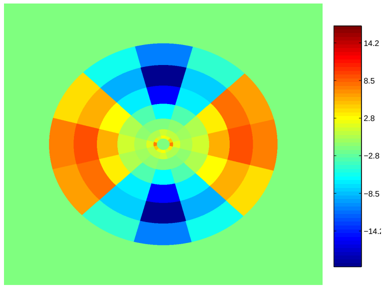
$$\sigma_d^2 = C^{-1} \sum_{k=1}^{C} \frac{E(\|S_J X_k - P_{\mathbf{A_k}}(S_J X_k)\|^2)}{E(\|S_J X_k\|^2)} ,$$

and

$$\rho_d^2 = C^{-1} \sum_{k=1}^{C} \frac{E(\min_{l \neq k} \|S_J X_k - P_{\mathbf{A}_l}(S_J X_k)\|^2)}{E(\|S_J X_k - P_{\mathbf{A_k}}(S_J X_k)\|^2)} ,$$

we obtain an average approximation error $\sigma_d^2 = 2.5 \cdot 10^{-1}$ and a separation ratio of $\rho_d^2 = 3$.

The PCA classifier provides partial rotation invariance by removing principal components. Figure 4.4 shows the first principal direction for a given material in the class, seen in the scattering domain. We can clearly distinguish differences along coefficients of the form p, $R_\alpha p$, where $R_\alpha p = (R_\alpha \lambda_1, \dots, R_\alpha \lambda_m)$ is the path obtained by rotating all its subbands by an angle α. As a result, by removing such principal directions, the algorithm is effectively averaging scattering coefficients along path rotation parameters, which comes with a loss of discriminability. A more efficient rotation invariant texture classification is obtained by cascading the translation invariant scattering $\overline{S}$ with a second rotation invariant scattering, as studied in [SM12]. This cascade transforms each layer of the translation invariant scattering network with new wavelet convolutions along rotation parameters, followed by modulus and average pooling operators, which are cascaded. A combined translation and rotation scattering yields a translation and



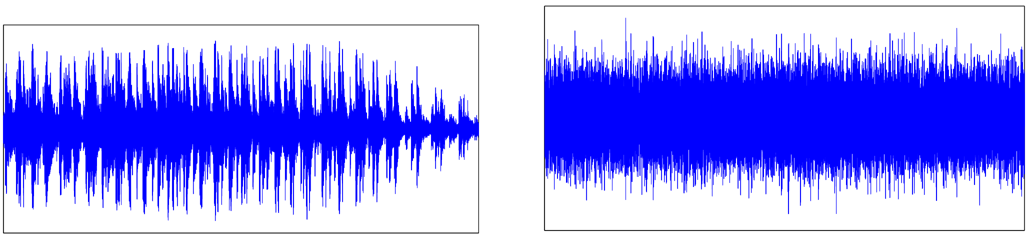
(a) (b)

Figure 4.4: First principal direction of a texture class #4 in the scattering domain. (a) scattering coefficients for $m = 1$. (b) scattering coefficients for $m = 2$. Reddish colors indicate positive values, blueish colors indicate negative values.

rotation invariant representation, with the same stability properties to deformations as those shown in Chapter 2 [Mal12].

4.4 Auditory texture discrimination

Auditory textures such as rain, insect noises or applauses are well modeled as stationary processes. However, and similarly as in the visual case, they are poorly characterized by their power spectra, which imposes audio representations to capture high order moments. Figure 4.5 shows an example of an applause realization taken from [McD] and a white gaussian noise realization, with its spectral power adjusted to match the applause. The characteristic clapping produces an impulsive behavior in the original clip, which is not captured by second order statistics.



(a) (b)

Figure 4.5: (a) example of an applause. The clip has a length of 8 seconds, sampled at a rate of 20 KHz. (b) gaussian process with the same spectral density as those form example (a). The spectral density is estimated with a window of length 1 second.

In order to illustrate the capacity of scattering representations to capture high order moments information, we considered the dataset created by McDermott and Simoncelli,

Table 4.2: Percentage of correct neighborhood predictions for several audio representations.

Φ	ρ
$\lvert\hat{x}\rvert$	83%
$S_J\,,\, m_{max}=1\,,\, Q=1$	91%
$S_J\,,\, m_{max}=1\ Q=16$	91%
$\mathbf{S_J\,,\, m_{max}=2\,,\, Q=1}$	**97%**
$S_J\,,\, m_{max}=2\ Q=16$	91%
$S_J\,,\, m_{max}=3\,,\, Q=1$	96%
$S_J\,,\, m_{max}=3\,,\, Q=16$	91%

available in [McD]. It contains a collection of 30 auditory textures, such as wind, rain, or helicopter. The authors also offer synthesis clips using several reconstruction strategies, which include the synthesis from the spectral density. We then consider the dataset of 60 classes, formed by the 30 original plus the 30 gaussian equivalents. Each clip is split into 6 smaller signals, which correspond to approximately 1 seconds of sound. For each example x_i we compute its audio representation $\Phi(x_i)$, and measure the percentage ρ of examples for which the nearest neighbor of $\Phi(x_i)$, in the sense of the distance $d(x,x') = \|\Phi(x) - \Phi(x')\|$, comes from the same class as x_i.

Table 4.2 shows the classification rates for several choices of Φ. As expected, the Fourier modulus representation performs poorly, since by construction there are always two classes whose realizations have the same power spectrum. The rate is better than pure chance probably because the estimation of the power spectrum in the construction of the samples differs from the one used to construct Φ. We observe, as predicted, that second order scattering coefficients bring a significant improvement over first order coefficients, thanks to the fact that they depend upon moments of order higher than 2. Third order coefficients do not improve the results, mostly due to the fact that they have little influence on the metric $\|\Phi(x) - \Phi(x')\|$. The relative influence of high order scattering coefficients can be modified by different renormalisation strategies, such as those studied in Section subsection 3.4.2. A remarkable point is the influence of the choice of the wavelet. The best results are obtained by selecting dyadic splines with $Q = 1$ voice per octave, as opposed to the standard filter banks used in audio recognition, which are more selective in frequency with $Q = 16$ voices per octave. Scattering with narrowband wavelets becomes more discriminative to variations in frequency, which are important for the representation of pitch, at the expense of being less stable to frequency fluctuations. Auditory textures have features which are not well localized in frequency, which explains the interest in working with wavelets with larger bandwidths.

4.5 Texture synthesis with Scattering

This section studies the texture reconstruction from the expected scattering representation. We follow the same statistical framework as in [PS99; MS11], by using as the set of statistical measures the expected scattering representation.

We start by introducing a scattering reconstruction algorithm in Section 4.5.1, which adjusts expected scattering coefficients with a gradient descent using a family of wavelet modulation operators. We then focus on the auditory texture synthesis, where we apply the scattering reconstruction algorithm on a family of sound textures.

4.5.1 Scattering Reconstruction Algorithm

Section 4.2.5 showed that, given a texture representation $\Phi(X) = \{E(g_k(X)\,, k = 1 \dots K\}$, sampling from the maximum entropy distribution determined by the constraints $\{\Phi(X)_k = y_k\,, k \leq K\}$ can be approximated by a uniform sampling on the Julesz ensemble:

$$\Phi^{-1}(y) = \{x\,;\, \overline{g_k(x)} = y_k\,,\, k \leq K\}\ .$$

Under this framework, the texture synthesis from scattering representations requires an algorithm which can adjust the averaged scattering coefficients of a given realization to a specified vector.

Given an element $y \in \mathrm{Im}\, S_J$, the goal is then to solve for $x \in \mathbf{L}^2(\mathbb{R})$

$$S_J x = y\ . \tag{4.6}$$

We place ourselves in the discrete case where x is a signal of size N. The scattering image is also a finite dimensional space of dimension given by $|\Gamma| = (J + \binom{J}{2})N2^{-J}$. In particular, when $J = \log_2 N$, $|\Gamma| = (J + \binom{J}{2})$.

Problem (4.6) can be solved with a gradient descent algorithm, which minimizes

$$\mathcal{C}(x) = \|S_J x - y\|^2\ . \tag{4.7}$$

$\mathcal{C}(x)$ is convex in the variable $S_J x$. Although S_J is not differentiable with respect to x, Section 2.3.3 showed that thanks to the Lipschitz property, its first variations [LPT12] are approximated by a bounded linear operator DS_{Jx}. As a result, at any given point $S_J x$, the direction of steepest descent in the scattering domain, $S_J x - y$, can be recovered with a perturbation h satisfying

$$DS_{Jx}(h) = \epsilon(S_J x - y)\ . \tag{4.8}$$

It follows that the efficiency of the gradient descent depends upon choosing appropriate descent variables, with the capacity to modify scattering coefficients along any given direction of the scattering domain.

Scattering coefficients are computed with a non-linear operator which cascades wavelet decompositions with complex modulus. A good family of perturbation operators thus must be able to factorize the influence of wavelet decompositions and the modulus. If

$\sigma : \mathbb{R}^d \to \mathbb{C}$ denotes a smooth complex valued function with $\|\sigma\|_\infty < \infty$, we define the linear modulation operator $M[\sigma]$ as

$$\forall x \in \mathbf{L}^2(\mathbb{R}^d),\ \forall u\,,\ M[\sigma]x(u) = \sigma(u)\,\cdot\,x(u)\ .$$

By properly specifying the regularity of the envelopes, modulation operators have a particularly simple interaction with wavelet decompositions and also with complex modulus. The regularity of the envelope σ can be controlled with a multiresolution analysis [Mal08] $\mathbf{V}_k^\infty$, $k \in \mathbb{Z}$, defined in Appendix A.1 as a collection of embedded subspaces generated by translated versions of a dilated scaling function ϕ. Proposition A.1.1 shows that if σ is an envelope in $\mathbf{V}_k^\infty$, that is, carrying details up to a scale 2^k, and ψ_j is a wavelet localized at scale 2^j, then

$$\forall x \in \mathbf{L}^2(\mathbb{R}^d)\ ,\ \|\sigma\,\cdot\,(x \star \psi_j) - (\sigma\,\cdot\,x) \star \psi_j\| \leq C 2^{j-k} |\sigma|_\infty \|x\|\ .$$

Therefore, when $k \gg j$, this proposition shows that the modulation with σ nearly commutes with the wavelet decomposition at scale 2^j.

This near commutation property of modulations is exploited to construct a family of wavelet modulation operators, which modulate each wavelet sub-band with a slowly varying complex envelope.

Definition 4.5.1 *A wavelet modulation operator of $\mathbf{L}^2(\mathbb{R}^d)$ is defined for any complex multiscale envelope $\sigma(\lambda, u)$, $\lambda \in \Lambda_J$, $u \in \mathbb{R}^d$, with $\sup_{\lambda,u} |\sigma(\lambda, u)| < 1$ and $\sigma(\lambda, \cdot) \in \mathbf{V}_j^\infty$, $\lambda = 2^j r$, as*

$$\overline{M}[\overline{\sigma}]x = x + Re \sum_\lambda \tilde{W}_\lambda M[\sigma(\lambda, \cdot)] W_\lambda x\ ,\ x \in \mathbf{L}^2(\mathbb{R}^d). \tag{4.9}$$

Each envelope $\sigma(\lambda, u)$ defines a perturbation of a signal x with the capacity to modulate each sub-band $x \star \psi_j$ along a specified envelope $\sigma(\lambda, \cdot)$. Appendix A.3 shows that, thanks to the near commutation property, the influence of wavelet modulation perturbations on scattering coefficients is nearly decorrelated when using the wavelet decomposition coefficients of the envelopes $\sigma(\lambda, \cdot)$ as descent variables. Whereas first order scattering coefficients are influenced by amplifying or attenuating all wavelet coefficients in a subband, second order scattering coefficients are influenced by modulations with envelopes having their own spatial or temporal variations. These variations are carried by wavelet decomposition coefficients of each $\sigma(\lambda, \cdot)$.

Let Γ be the space of first and second order progressive scattering paths

$$\Gamma = \mathcal{P}_\downarrow^1 \cup \mathcal{P}_\downarrow^2 = \{j_1 \in \mathbb{Z}\,, j_1 \leq J\} \cup \{(j_1, j_2) \in \mathbb{Z}^2\,;\, j_1 < j_2 \leq J\}\ .$$

For convenience, we can identify first order paths $j_1 \in \mathcal{P}^1$ with $(j_1, J{+}1)$ and parametrize $\Gamma = \{(j_1, j_2) \in \mathbb{Z}^2\,;\, j_1 < j_2 \leq J+1\}$.

We approximate the direction of steepest descent characterized in (4.8) by projecting it into the subspace $\nabla_{\overline{M}} S_{Jx,\epsilon}$ generated by a family of wavelet modulation operators.

Appendix A.3 also shows that wavelet modulations are Lipschitz $\mathbf{L}^2(\mathbb{R}^d)$ operators, and hence, by the same argument exposed in Section 2.3.3, the first variations of $S_J\overline{M}[\sigma]x$ are approximated for sufficiently small ϵ by

$$S_J\overline{M}[\sigma]x = S_J x + \nabla_{\overline{M}} S_{J x,\epsilon}(\sigma) + \epsilon o(|\sigma|_\infty) \ , \epsilon \to 0 \ .$$

As a result, the perturbations of steepest descent can be obtained by resolving a linear system.

The gradient descent algorithm proceeds as follows. We initialize x_0 with a sample from white gaussian noise, and iteratively update x_m by searching for the wavelet modulation $\overline{M}[\gamma_m\sigma_m]$ with step $\gamma_m > 0$ yielding the steepest descent in the cost function of (4.7):

$$x_{m+1} = \overline{M}[\gamma_m\sigma_m]x_m \ . \tag{4.10}$$

Let us now characterize the multiscale envelope $\sigma(j,u)$ yielding the steepest descent, in terms of its complex wavelet coefficients $\Theta = (\theta_p[j,n])_{p,j,n}$:

$$\sigma(j,u) = \sum_{|j'-j_2|\leq\Delta_2, j'>j} \sum_n \theta[j',n]\Psi_{j'}(u-2^{j'}n) \quad , \tag{4.11}$$

where $\{\Psi_j(u-2^jn)\}_n$ is a wavelet basis for the space of details $\mathbf{W}_j$ at scale 2^j determined by the multiresolution analysis.

Let $z_0 = (j_0, q_0, n_0, \alpha_0)$ denote the multi-index encoding a wavelet modulation $\overline{M}[\sigma_{z0}]$ whose envelope has the form

$$\sigma_{z_0}(j,u) = \begin{cases} e^{i\alpha_0}\Psi_{q_0}(u-2^{q_0}n_0) \ , & \text{if } |j-j_0| \leq \Delta_1 \ , \\ 0 & \text{otherwise} \ . \end{cases}$$

Here, $j_0 \in [1,J]$, $q_0 \in [j_0+1,J]$, $n_0 \in [1, N2^{-q_0}]$ and $\alpha_0 = 0, \pi$. Thus, the multi-index z belongs to a discrete space that we denote $\mathcal{Z}$. The subspace generated by small wavelet modulations is generated by the matrix of finite modulation differences

$$\nabla_{\overline{M}} S_J = \left[\cdots \ \frac{(S_J\overline{M}[\epsilon\sigma_z]x_m - S_Jx_m)}{\epsilon} \ \cdots \right]_{z\in\mathcal{Z}} \ ,$$

for a given small step $\epsilon > 0$.

The envelope of steepest descent σ_m is characterized via its wavelet decomposition coefficients Θ_m, by projecting the direction of steepest descent in the scattering domain, $S_Jx_m - y$, into the subspace generated by $\nabla_{\overline{M}} S_J$. This corresponds to the minimization

$$\min_\Theta \|(\nabla_{\overline{M}} S_{J x_m})\Theta - (S_Jx_m - y)\|^2 \ ,$$

whose solution is given by

$$\Theta_m = (\nabla_{\overline{M}} S_{J x_m})^\dagger (S_Jx_m - y) \ , \tag{4.12}$$

where $(\nabla_{\overline{M}} S_{J x_m})^\dagger$ is the pseudo-inverse, defined as

$$A^\dagger = (A^T A)^{-1} A^T \ .$$

The coefficients in (4.12) yield an envelope σ_m which defines the direction of descent. The step γ_m can be set to a fixed value, or it can also be adjusted dynamically with a line search strategy. The descent characterized by (4.12) adjusts all the coefficients in Γ simultaneously. In order to reduce the dimensionality and hence the risk of drifting towards a local minima, we choose to split the descent into the orthogonal subspaces $\Omega_j = \{(j, j_2); j < j_2 \leq J\} \subset \Gamma$, which pack scattering paths according to their first scale. The iterative descent thus projects the scattering modulation matrix $\nabla_{\overline{M}} S_{J x_m}$ into J smaller matrices

$$\nabla_{\overline{M}}[\Omega_{j_0}] S_J = \left[\cdots \ \frac{(S_J \overline{M}[\epsilon \sigma_z] x_m[\Omega_{j_0}] - S_J x_m[\Omega_{j_0}])}{\epsilon} \ \cdots \right]_{z \in \mathcal{Z}; j = j_0} ,$$

and iterates over these smaller subproblems.

The subspaces are visited from the coarsest scale Ω_J to the finest Ω_0, which corresponds to the increasing order defined in A.13. Although the multiscale modulations are well localized within the sets Ω_j, its influence on neighboring scales is not symmetric in general, and has a slower decay towards the finer scales. Besides, the multiscale modulations on coarse scales are likely to influence a larger number of signal coefficients than modulations which admit more irregular envelopes, which limits the risk of being stuck in local minima.

First order scattering coefficients $\mathcal{P}^1 = (j); j \leq J$ are adjusted with multiscale modulation operators having constant envelopes, and hence they can be adjusted seamlessly in the subspaces Ω_j. For many signals, however, first order coefficients concentrate an important fraction of the scattering energy, which means that the direction $S_J x - y$ is often aligned along the first order coordinates. In order to accelerate convergence within the set of second order paths, it is possible to adjust first order coefficients at the beginning of each loop on the subspaces Ω_j.

The scattering reconstruction gradient descent is summarized in algorithm 2. The reconstruction algorithm is tested numerically on discrete signals. When $J = \log N$, the averaged scattering representation estimates the expected scattering defined for stationary processes defined in Section 2.3.5.

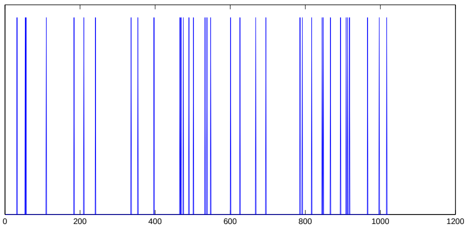
Figure 4.6 shows the reconstruction from a realization g of a white Bernoulli process, $y = S_J g$, with $J = \log N$. For illustration purposes, we show also the reconstructed signals when the scattering reconstruction is limited to paths of 0-th order (figures (c) and (d)) and first order (figures (e) and (f)). The scattering coefficient of order 0 is the signal average, hence the reconstructed signal is simply a white gaussian noise realization with the same mean as g. First order coefficients are wavelet amplitude averages, which give a measure of the average energy within each sub-band. Since the two signals come from white processes, they are not distinguished by their spectral density, which explains why the reconstruction from first order coefficients, in (e), is similar to (c). By adding

Algorithm 2 Scattering Reconstruction Algorithm

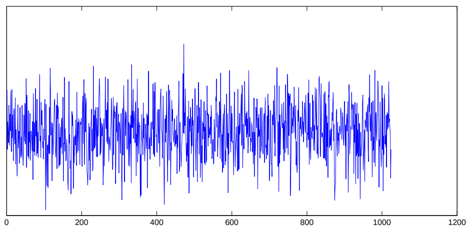
Initialize x with a sample of $\mathcal{N}(0,1)$.
while $r \leq r_{max}$ and $d_r \geq \rho \|y\|$ **do**
  Adjust first order scattering coefficients:
  $x \leftarrow \text{grad\_descent}(x, y, \mathcal{P}^1)$.
  for $j = J-1$ to 1 **do**
    Adjust coefficients in Ω_j:
    $x \leftarrow \text{grad\_descent}(x, y, \Omega_j)$.
  end for
  $d_r \leftarrow \|S_J x - y\|$.
  $r \leftarrow r + 1$.
end while

$x = \text{grad\_descent}(x_0, y, \Omega)$:
$x \leftarrow x_0$.
while $r \leq \tilde{r}_{max}$ **do**
  Compute $\nabla_{\overline{M}}[\Omega]S_J$ in x.
  $\Theta \leftarrow (\nabla_{\overline{M}}[\Omega]S_J)^\dagger (S_J x[\Omega] - y[\Omega])$.
  $\sigma(j,u) \leftarrow \sum_{(q,n,\alpha)} (\theta_j[q,n,0] + i\theta_j[q,n,\pi]) \Psi_q(u - 2^q n)$.
  $x \leftarrow \overline{M}[\gamma\sigma]x$.
end while

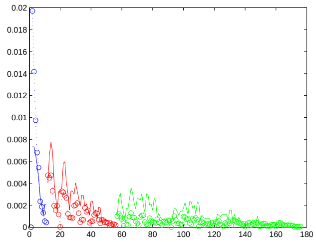
Parameters:
r_{max}: maximum number of outer loop iterations.
$\tilde{r}_{max}$: maximum number of inner loop iterations.
ρ: tolerance arrival scattering distance.
γ: gradient step.
ϵ: precision for the computation of $\nabla_{\overline{M}}[\Omega]S_J$.



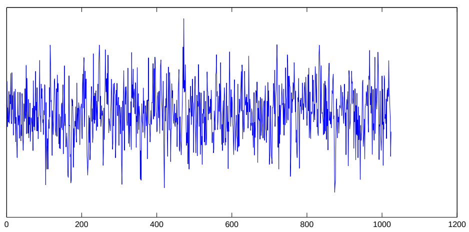
(a)



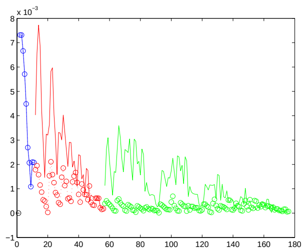
(b)



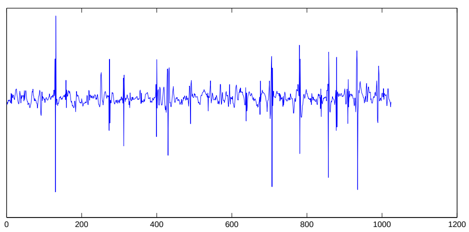
(c)



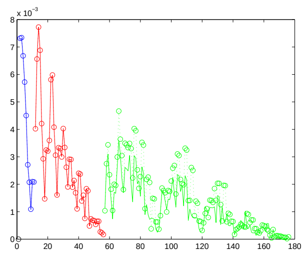
(d)



(e)



(f)



(g)

Figure 4.6: Reconstruction examples of a realization of a Bernouilli process. (a) Original signal. (b)-(c) Reconstruction obtained by adjusting only the mean of the process, and scattering representation. (d)-(e). Reconstruction obtained by adjusting first order coefficients. (f)-(g): Reconstruction obtained by adjusting first and second order scattering. First order coefficients are plotted in blue, second order coefficients are plotted in red and third order coefficients are plotted in green. Solid lines correspond to original signal. Circles and dotted lines correspond to reconstructed signals.

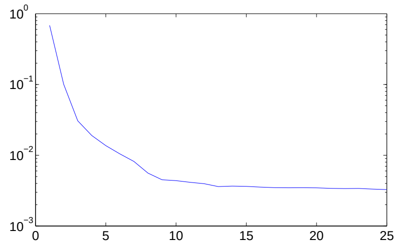


Figure 4.7: relative reconstruction error as a function of the number of iterations.

the $(\log N)^2$ second order coefficients, the reconstructed signal (g) shows the spikes which characterize Bernoulli processes.

Figure 4.7 shows the convergence of the gradient descent reconstruction algorithm as a function of the number of iterations, for the example in figure 4.6. The reconstruction error decays until it reaches a relative error floor of around $3 \cdot 10^{-3}$. While we have no theoretical guarantee of convergence, we observe that the asymptotic behavior is sensitive to the discretization parameters of the algorithm, especially the gradient step γ and the precision ϵ used to compute the partial derivatives $\epsilon^{-1}(S_J \overline{M}[\epsilon\sigma]x - S_J x)$.

As long as $(\log N)^2 < N$, reconstructing the scattering coefficients $S_J x = S_J x'$ does not imply that $x = x'$. Indeed, panels (g) and (h) from Figure 4.6 show that first and second order scattering coefficients are adjusted, but still $x \neq x'$. In particular, we notice that higher order scattering coefficients are not adjusted. If one considers a windowed scattering representation with Q wavelets per octave, such as in audio applications, then the number of first and second order coefficients obtained for a signal of size N is $N2^{-J}(QJ + \binom{QJ}{2})$, which might be larger than N for certain choices of J and Q. In that case, one might ask if $S_J x = S_J x'$ implies $x = x'$, or equivalently, whether x can be reconstructed from its scattering coefficients.

This question is related to the inversion of the wavelet modulus operator U. In [Wal12], the authors show that U can be inverted for appropriate wavelets by solving a convex linear program. The operator S_J is obtained by cascading U, followed by a convolution by the lowpass filter ϕ_J, which implies that a deconvolution step is needed prior to start inverting U. The reconstruction algorithm for S_J presented here is hence an alternative which exploits the differentiable structure of S_J with a family of multiscale modulations.

4.5.2 Auditory texture reconstruction

Section 4.4 showed that first and second order scattering coefficients have the capacity to discriminate non-gaussian audio textures. We may then ask whether they are sufficiently informative to reconstruct perceptually similar realizations.

For that purpose, we consider a 1 second example of several texture classes from the sound dataset of McDermott [McD]. For each example x_k, $k = 1 \dots K$ we estimate its expected scattering representation $\overline{S}X_k$ with the windowed scattering $y_k = S_J x_k$, and obtain new realizations $\tilde{x}_k$ satisfying

$$S_J \tilde{x}_k = y_k \ , \ k = 1 \dots K \ . \tag{4.13}$$

The constraints given by the scattering representation y_k thus define our Julesz ensemble, in the framework of [PS99; ZWM97]. We solve (4.13) with the gradient descent algorithm of the previous section.Similarly as in [MS11], the uniform sampling on the Julesz ensemble is approximated by initializing the gradient descent with random samples of gaussian white noise.

Figure 4.8 shows the scattering reconstruction results for $m_{max} = 1$ and $m_{max} = 2$, using dyadic spline wavelets with $Q = 1$ voices per octave. We notice that first order reconstruction does not capture the spurious behavior of sounds such as the hammer or the applause. The resulting reconstructions are perceived as equalized gaussian noise, far from the perception of the original texture. For $m = 2$, the reconstructed samples are perceptually similar to the original, especially for the examples of water, applause, helicopter, train and cocktail party. First and second order coefficients, when using a dyadic wavelet representation, represent the texture with $\log N + \binom{\log N}{2} \sim (\log N)^2$ coefficients, which gives for $N = 2^1 4$ a representation of less than 200 coefficients. The examples where this reconstruction gives best results are well modeled by an amplitude modulation of gaussian noise with a regular envelope. Contrarily to other audio textures such as the jackhammer, these examples have a regular spectral density, as shown in figure 4.9.

The frequency resolution of the scattering representation is controlled by the number of voices per octave Q. Figure 4.10 compares the reconstructed realizations using $Q = 1$ and $Q = 16$ on subset of examples. By increasing the frequency resolution of the wavelets, we are able to reconstruct realizations with more irregular spectral densities, but this increase in resolution comes at a cost. Indeed, if X is a stationary process, then $X \star \psi_\lambda$ and $X \star \psi_{\lambda'}$ are nearly decorrelated if ψ_λ and $\psi_{\lambda'}$ are supported in different frequency bands, but if X is not gaussian in general they are not independent, which creates a correlation on their respective envelopes $|X \star \psi_\lambda|$ and $|X \star \psi_{\lambda'}|$. In particular, this fact justifies why McDermott and Simoncelli include the correlations between the cochlea envelopes as part of their texture representation. By choosing wavelets with larger bandwidth (smaller Q) , these correlations are encoded in second order coefficients. Besides this trade-off, increasing the Q-factor impacts the size of the representation. Indeed, the second order representation with $Q = 16$ has $\sim Q(\log N)^2 \approx 3000$ coefficients.

We also explore the influence of adding third order scattering coefficients into the representation. For that purpose, we modify the gradient descent algorithm by incorporating on its objective function the third order coefficients. Although a priori the family of multiscale modulations described in subsection 4.5.1 does not generate the whole tangent space of third order coefficients, we numerically observe a convergence up to a relative error of $3 \cdot 10^{-2}$ on average for these considered examples. Third order coefficients significantly increase the size of the representation to $(\log N)^3 \sim 2500$, and produce a slight improvement on the reconstruction quality, barely noticeable.

Improving the reconstruction from scattering coefficients thus requires a good fre-

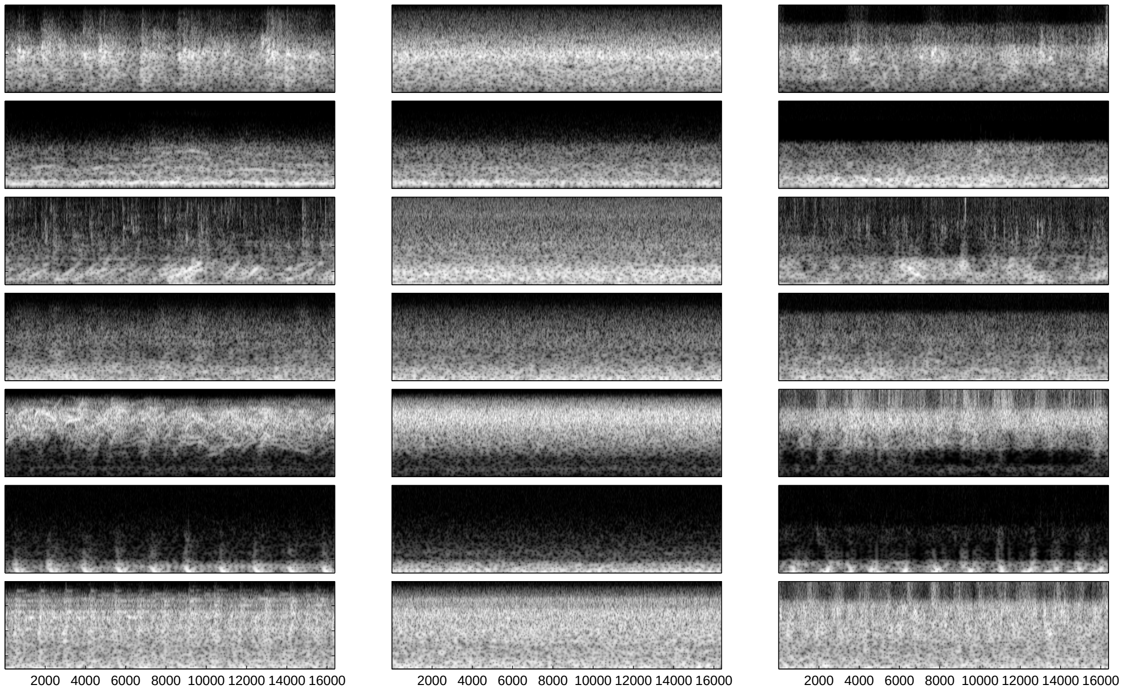


Figure 4.8: Scattering reconstruction of auditory examples from [McD]. Left column: Original sounds, Middle column: Reconstruction with $m_{max} = 1$, Right column: reconstruction with $m_{max} = 2$. We plot a scalogram obtained with a filter bank with logarithmicly spaced constant Q, from 1 KHz up to 10KHz, with $Q = 16$, completed with linearly spaced constant bandwidth filters covering the lower frequencies. The auditory examples correspond respectively to: water, jackhammer, applause, insect, helicopter, train, cocktail party, rusting paper.

quency resolution without losing the capacity to capture the correlation between neighbor frequency components. This flexibility can be obtained by generalizing the second layer of scattering coefficients. For a one-dimensional process X, the first scattering layer produces the envelopes $(U[\lambda]X)_{\lambda \in \Lambda_J}$, where λ encodes a frequency interval. One can then construct a second layer [AM12b] where the output coefficients are obtained by recombining several envelopes $U[\lambda]X$ with two-dimensional wavelets.

4.6 Scattering of Gaussian Processes

Gaussian processes are a fundamental family of stationary processes, for which the spectral density completely characterizes the full distribution. In this case, first order scattering coefficients are specified also from the spectral density, as well as the second moments of the processes defined by the propagators $U[\lambda_1, \lambda_2]$, $\lambda_1, \lambda_2 \in \Lambda_\infty^2$.

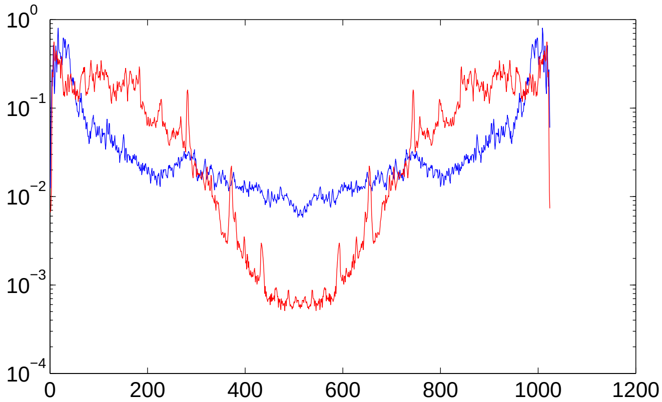


Figure 4.9: Comparison of power spectrum of two different audio textures: bubbling water (in blue), jackhammer (in red). The bubbling water texture has a regular power spectrum which is well captured by dyadic frequency intervals. However, the jackhammer shows harmonic frequency peaks that require a more selective wavelet in order to be reconstructed.

Let X be a centered, stationary Gaussian process and $\sigma^2 = E|X|^2$, and suppose that the scattering is computed with analytic wavelets having fast decay. We write the auto-covariance $R_X(\tau) = E\left((X(t) - E(X))(X(t+\tau) - E(X))^*\right)$, and we assume $R_X \in \mathbf{L}^1$. The first scattering layer starts by computing $X_\lambda = X \star \psi_\lambda$, which are also Gaussian and stationary. The complex modulus produces the collection $\forall \lambda \in \Lambda_\infty\,,\ U[\lambda]X = |X_\lambda|$, which are Rayleigh processes when the wavelets are analytic. Their auto-correlation can be obtained from the auto-correlation of X_λ, as shown by the following proposition from [DR87]:

Proposition 4.6.1 *Let $X_r(t)\,,\ X_i(t)$ be independent, stationary Gaussian processes with $\sigma^2 = 2E(|X_r(t)|^2)$. Then $Y(t) = |X_r(t) + iX_i(t)|$ is a stationary Rayleigh process. Its autocorrelation $\tilde{R}_Y(\tau) = E(Y(t)(Y(t+\tau)))$ is given by*

$$\tilde{R}_Y(\tau) = \frac{\pi}{2}\sigma^2\, {}_2F_1\left(-\frac{1}{2}, -\frac{1}{2}; 1; \frac{|R_X(\tau)|^2}{\sigma^2}\right) \,, \tag{4.14}$$

where ${}_2F_1$ is the hypergeometric function

$${}_2F_1(a, b; c, x) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n} \frac{x^n}{n!} \,, \tag{4.15}$$

defined with the Pochhammer symbol:

$$(q)_n = \begin{cases} 1 & \text{if } n = 0\,, \\ q(q+1)\dots(q+n-1) & \text{if } n > 0\,. \end{cases}$$

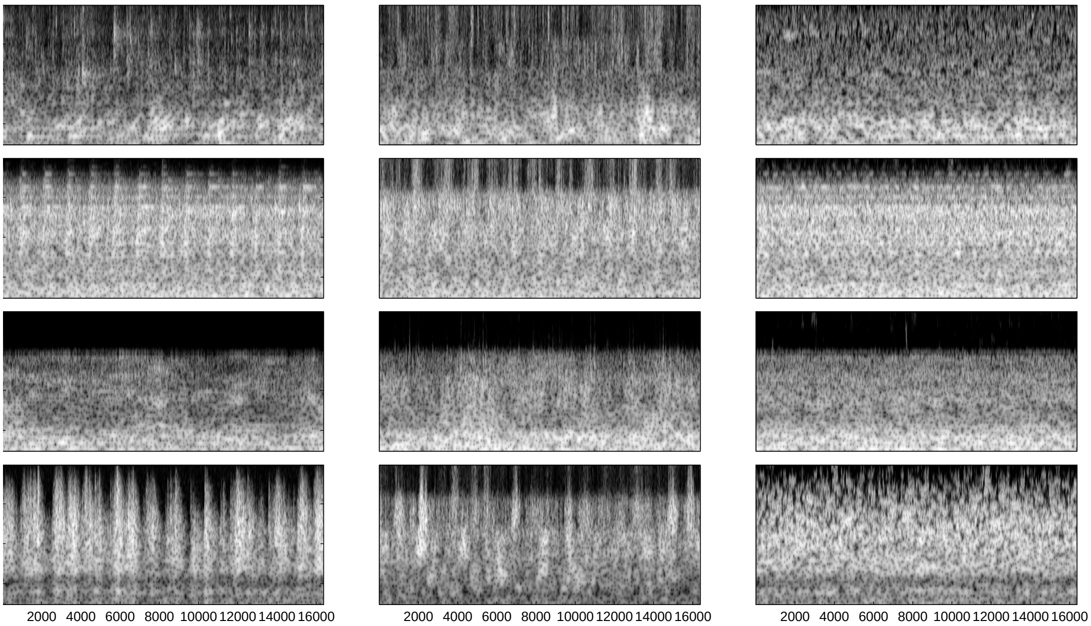


Figure 4.10: Scattering reconstruction of auditory examples from [McD]. Left column: Original sounds, Middle column: Reconstruction with $m_{max} = 2$ and $Q = 1$, Right column: reconstruction with $m_{max} = 2$ with $Q = 16$. We plot a scalogram obtained with a filter bank with logarithmicly spaced constant Q, from 1 KHz up to 10KHz, with $Q = 16$, completed with linearly spaced constant bandwidth filters covering the lower frequencies. The auditory examples correspond respectively to: water, jackhammer, cocktail party, rusting paper.

This result can be proved by decomposing $g(x) = |x|$ in the Hermite polynomials, which form an orthonormal basis of the space $L^2(\mathbb{R}, d\varphi)$, where $d\varphi$ is the Gaussian measure:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \ .$$

The n-th Hermite polynomial is given by

$$H_n(x) = \frac{(-1)^n}{\varphi(x)} \frac{d^n \varphi(x)}{dx^n} \ ,$$

and when tested on a gaussian random variable $N \sim \mathcal{N}(0, 1)$ they satisfy

$$E(H_n(N) H_m(N)) = n! \delta(n - m) \ .$$

As a result, any measurable function g satisfying $E(|g(N)|^2) < \infty$ can be expressed as

[Dou06]

$$g(x) = \sum_{n=0}^{\infty} \frac{g_n}{n!} H_n(x), \quad g_n = E(g(N)H_n(N)), \quad E(|g(N)|^2) = \sum_n \frac{|g_n|^2}{n!} .$$

The second moments of $g(N)$ can be computed similarly using the Mehler Formula:

Lemma 4.6.2 *(Mehler Formula, [Dou06]) Let X be a centered, finite energy Gaussian process with covariance $R_X(\tau) = E(X(t)X(t+\tau)^*)$. Then, if g is measurable and non-expansive, the process $g(X)$ has correlation*

$$E(g(X(t))g(X(t+\tau))^*) = \sum_{n=0}^{\infty} \frac{|g_n|^2}{n!} R_X(\tau)^n ,$$

and covariance

$$R_{g(X)}(\tau) = E(g(X(t))g(X(t+\tau))^*) - |E(g(X))|^2 = \sum_{n=1}^{\infty} \frac{|g_n|^2}{n!} R_X(\tau)^n .$$

This lemma can be used to obtain (4.14) by computing the Hermite Chaos expansion of $g(x) = |x|$. We obtain

$$g_n = \begin{cases} 0 & \text{if } n = 2k+1, \\ (-1)^{k-1}\sqrt{\frac{(2k-3)!!(2k)!}{2^{k-1}(k-1)!k(k-1)}} & \text{if } n = 2k. \end{cases}$$

This expansion reveals the singularity of the modulus, since $|g_{2k}| \sim k^{-3/4}$ exhibits a slow decay.

An important consequence of (4.14) is that when a centered Gaussian process X satisfies $R_X \in \mathbf{L}^1$, then the auto-covariance of $U[\lambda]X = |X \star \psi_\lambda|$ also satisfies $R_{U[\lambda]X} \in \mathbf{L}^1$ provided ψ is analytic:

Proposition 4.6.3 *If $X(t)$ is a stationary Gaussian process with $E(|X(t)|^2) = \sigma_0^2$, such that $R_X \in \mathbf{L}^1$, and ψ is an analytic wavelet with $\psi \in \mathbf{L}^2 \cap \mathbf{L}^1$, then the auto-covariance of $Y(t) = |X \star \psi(t)|$ satisfies $R_Y \in \mathbf{L}^1$.*

Proof: Since $X(t)$ is Gaussian and the convolution is a linear operator, it results that $Z(t) = X \star \psi$ is a complex Gaussian process and hence that $Y(t) = |Z(t)|$ is Rayleigh, thanks to the fact that ψ is analytic. By applying proposition 4.6.1 it results that its autocorrelation $\tilde{R}_Y(\tau) = E(Y(t)Y(t+\tau))$ equals

$$\tilde{R}_Y(\tau) = \frac{\pi}{2}\sigma^2 \, {}_2F_1\left(-\frac{1}{2}, -\frac{1}{2}; 1; \frac{|R_Z(\tau)|^2}{\sigma^2}\right) ,$$

where $\sigma^2 = E(|Z(t)|^2)$. Since the first term $n = 0$ of the hypergeometric series is 1 and $E(Y)^2 = \frac{\pi}{2}\sigma^2$ it results that the auto-covariance of Y is given by the series

$$R_Y(\tau) = \tilde{R}_Y(\tau) - E(Y)^2 = \frac{\pi}{2}\sigma^2 \sum_{n>0} \frac{(-1/2)_n(-1/2)_n}{(1)_n} \frac{\gamma^n}{n!} ,$$

with $\gamma = \frac{|R_Z(\tau)|^2}{\sigma^2}$. Let us see that there exists $C \geq 0$ such that

$$\forall \tau \ , \ |R_Y(\tau)| \leq C |R_Z(\tau)|^2 \ . \tag{4.16}$$

Indeed, from the definition of the Pochhammer symbol, one can verify that

$$\forall n > 0 \ , \left(\frac{(-1/2)_n}{(1)_n} \right)^2 \leq \left(\frac{1}{4n} \right)^2 \ ,$$

and, since $\gamma(\tau) \geq 0$,

$$\forall \tau \ , \ |R_Y(\tau)| \leq C \sum_{n>0} \frac{\gamma(\tau)^n}{n^2} = C F(\gamma(\tau)) \ , \tag{4.17}$$

where $F(z) = \sum_{n>0} z^n / n^2$ is analytic. Since $0 \leq \gamma(\tau) \leq 1$, $F(0) = 0$, $F(1) = \sum_{n>0} n^{-2} = \frac{\pi^2}{6} < \infty$, and $F''(\gamma) \geq 0$ for all $0 \leq \gamma \leq 1$, we conclude that $F(z)$ is convex in $(0,1)$, and hence that

$$\forall \tau \ , \ |R_Y(\tau)| \leq C F(\gamma(\tau)) \leq C F(1) \gamma(\tau) = \tilde{C} |R_Z(\tau)|^2 \ . \tag{4.18}$$

Now, since $Z = X \star \psi$, it results that

$$R_Z(\tau) = R_X \star \psi \star \tilde{\psi}(\tau) \ ,$$

with $\tilde{\psi}(u) = \psi(-u)^*$. By applying the Young inequality, and since $\psi \in \mathbf{L}^2 \cap \mathbf{L}^1$, it results that

$$\begin{aligned} \|R_Z\|_2 &= \|(R_X \star \psi) \star \tilde{\psi}\|_2 \leq \|R_X \star \psi\|_2 \|\tilde{\psi}\|_1 \\ &\leq \|R_X\|_1 \|\psi\|_2 \|\psi\|_1 \ , \end{aligned}$$

which, together with (4.18), implies that $\|R_Y\|_1 < \infty$ $\square$.

Proposition 4.6.3 will be used in the next chapter to study Gaussian fractal processes.

4.7 Stochastic Modulation Models

We study in this final section stochastic modulation models, and show that under some conditions, second order scattering coefficients separate the contributions of carrier and envelope, hence avoiding an ill-posed inverse demodulation step.

4.7.1 Stochastic Modulations in Scattering

Similarly as in the deterministic case, a stochastic modulation has a simple interaction with wavelet decompositions. We derive a stochastic commutation bound, which we then use in a stochastic modulation model to express second order scattering coefficients in terms of carrier and envelope. We consider in this section stationary processes in dimension 1.

Our analysis starts by studying how a stochastic modulation model $X(t) = X_s(t)X_f(t)$ is expressed in a wavelet decomposition. In this model, X_s and X_f are independent, stationary processes. One can interpret X as being the result of a stochastic modulation of X_f by X_s. In Section A.1, we showed that, thanks to the regularity of the envelope X_s, we could relate $X \star \psi_j$ with $X_s \cdot (X_f \star \psi_j)$. The following proposition derives the equivalent stochastic property.

Proposition 4.7.1 *Let $X = X_s X_f$ be a stationary process with X_s, X_f stationary and independent. Let $Y = X_s \cdot (X_f \star \psi_j) - X \star \psi_j$, and assume that $\int |u\psi(u)| < \infty$. Then Y is a stationary process satisfying*

$$E(|Y|^2) \leq E(|X_f|^2) \int \hat{R}_{X_s}(\omega)|A(2^j\omega)|^2 d\omega \ , \tag{4.19}$$

where $\hat{R}_{X_s}$ is the spectral density of X_s and $A(\omega) = \int |1 - e^{iu\omega}||\psi(u)|du$.

Proof: By definition, we have

$$\begin{aligned}
Y(t) &= X_s(t) \int X_f(u)\psi_j(t-u)du - \int X_s(u)X_f(u)\psi_j(t-u)du \\
&= \int (X_s(t) - X_s(u))X_f(u)\psi_j(t-u)du \\
&= -\int (X_s(t) - X_s(t-u))X_f(t-u)\psi_j(u)du \ ,
\end{aligned}$$

which yields

$$\begin{aligned}
E(|Y(t)|^2) &= \iint E\left((X_s(t) - X_s(t-u))(X_s(t) - X_s(t-u'))\right) E(X_f(t-u)X_f(t-u'))\psi_j(u)\psi_j(u')dudu' \\
&= \iint \left(R_{X_s}(0) + R_{X_s}(u-u') - R_{X_s}(u) - R_{X_s}(u')\right) R_{X_f}(u-u')\,\psi_j(u)\,\psi_j(u')dudu' \\
&= \iint \left(\int (1-e^{iu\omega})(1-e^{-iu'\omega})\hat{R}_{X_s}(\omega)d\omega\right) R_{X_f}(u-u')\,\psi_j(u)\,\psi_j(u')dudu' \ .
\end{aligned}$$

As a result, since $\hat{R}_{X_s}(\omega) \geq 0$ and $|R_{X_f}(\tau)| \leq R_{X_f}(0)$, we have

$$\begin{aligned}
E(|Y(t)|^2) &\leq \iiint |1-e^{iu\omega}|\,|1-e^{-iu'\omega}|\,\hat{R}_{X_s}(\omega)\,|\psi_j(u)|\,|\psi_j(u')|\,|R_{X_f}(u-u')|dudu'd\omega \\
&\leq R_{X_f}(0) \int \hat{R}_{X_s}(\omega) \left(\int |1-e^{iu\omega}|\,|\psi_j(u)|\,du\right)^2 d\omega \\
&= E(|X_f|^2) \int \hat{R}_{X_s}(\omega) \left(\int |1-e^{iu2^j\omega}|\,|\psi(u)|\,du\right)^2 d\omega \ ,
\end{aligned}$$

which proves (4.19). □.

Similarly as in the deterministic case, the bound is controlled by a spatial localization measure of the wavelet, contained in $A(\omega)$. Indeed, by considering a limited development of $|1 - e^{iu\omega}|$ we obtain

$$\begin{aligned} A(\omega) &= \int |1 - e^{iu\omega}||\psi(u)|du \\ &= \sqrt{2}\int \sqrt{1-\cos(u\omega)}|\psi(u)|du = \int (|u\omega| + o(|u\omega|))|\psi(u)|du \\ &\sim |\omega| \int |u\psi(u)|du \ , \end{aligned}$$

which yields

$$E(|Y|^2) \lesssim 2^j E(|X_f|^2) \left(\int |u\psi(u)|du\right) \left(\int \hat{R}_{X_s}(\omega)|\omega|d\omega\right) \ . \tag{4.20}$$

As a result, the more regular the envelope is, the more its spectral density is concentrated in the small frequencies, which reduces the term $\int \hat{R}_{X_s}(\omega)|\omega|d\omega$. In particular, if $\hat{R}_{X_s}(\omega)$ is negligible beyond a frequency 2^{-k}, then (4.20) shows that

$$E(|Y|^2) \lesssim C 2^{j-k} E(|X_f|^2) E(|X_s|^2) \ ,$$

which corresponds to the same asymptotic behavior as in the deterministic case.

Proposition 4.7.1 shows that stochastic modulations nearly commute with wavelet decompositions provided the envelope is regular with respect to the scale. We use this property to approximate second order scattering representations for modulated processes.

For that purpose, let us first introduce a regularity criteria for stationary processes $X(t)$, which asks its spectral density $\hat{R}_X$ to have uniform regularity across all dyadic frequency intervals.

Definition 4.7.2 *Let $X(t)$ be a stationary process with finite energy and let $\psi_s(t) = s^{-1}\psi(s^{-1}t)$. For a given exponent $\beta > 0$, let $(B_{inf}(X,\beta), B_{sup}(X,\beta))$ be defined as*

$$B_{inf}(X,\beta) = \inf_s s^{-\beta} \frac{E(|X \star \psi_s|^2)}{E(|X|^2)} \ , \quad B_{sup}(X,\beta) = \sup_s s^{-\beta} \frac{E(|X \star \psi_s|^2)}{E(|X|^2)} \ ,$$

We define the characteristic exponent of X as

$$\beta_X = \operatorname*{argmin}_{\beta > 0} |B_{sup}(X,\beta) - B_{inf}(X,\beta)|$$

and its dyadic bounds as

$$B_{inf}(X) = B_{inf}(X,\beta_X) \quad , B_{sup}(X) = B_{sup}(X,\beta_X) \ .$$

Dyadic regularity is related to a form of self-similarity of the spectral density of X. These dyadic bounds allow us to bound the second moments $E(|X \star \psi_j|^2)$ with

$$B_{inf}(X) 2^{-\beta_X j} E(|X|^2) \leq E(|X \star \psi_j|^2) \leq B_{sup}(X) 2^{-\beta_X j} E(|X|^2) \ .$$

The following proposition shows that when $X = X_s X_f$, and $p = (j_1, j_2)$ is a scattering path which separates the energy of the two components, then $E(|U[p]X|)$ can be approximated by a separable product of carrier and envelope.

Proposition 4.7.3 *Let $X = X_s X_f$, where X_s, X_f are stationary and with finite energy, with $X_s \geq 0$ and $E(X_s) = 1$. Suppose ψ has at least one vanishing moment. Let $p = (j_1, j_2)$ with $j_1 < j_2$, $A(\omega)$ be the modulus of continuity defined in Proposition (4.7.1), and let $B_{sup,0}, \beta_0$ be respectively the dyadic upper bound and the exponent of the process $Y = (X_s - E(X_s))(U[j_1]X_f - \overline{S}X_f(j_1))$. Then,*

$$|\overline{S}X(p) - \overline{S}X_f(j_1)\overline{S}X_s(j_2)| \leq \sqrt{b_1} + \sqrt{b_2} + \sqrt{b_3} \ , \tag{4.21}$$

with

$$\begin{aligned} b_1 &= E(|X_f|^2) \int \hat{R}_{X_s}(\omega) |A(2^{j_1}\omega)|^2 d\omega \ , \\ b_2 &= E(|U[p]X_f|^2) \ , \quad b_3 = B_{sup,0} 2^{-j_2 \beta_0} var(X_s) var(U[j_1]X_f) \ . \end{aligned}$$

Proof: From proposition 4.7.1 we have

$$E(|X \star \psi_{j_1} - X_s(X_f \star \psi_{j_1})|^2) \leq b_1 \ ,$$

which implies

$$X \star \psi_{j_1} = X_s(X_f \star \psi_{j_1}) + N_0 \ ,$$

where N_0 is an error term satisfying $E(|N_0|^2) \leq b_1$. Since the modulus is contractive, we have

$$|X \star \psi_{j_1}| = X_s |(X_f \star \psi_{j_1})| + N_1 \ , \tag{4.22}$$

with $E(|N_1|^2) \leq b_1$, since X_s is a positive process. We decompose

$$|(X_f \star \psi_{j_1})| = E(|(X_f \star \psi_{j_1})|) + \tilde{X} \ ,$$

with $E(\tilde{X}) = 0$, and hence (4.22) becomes

$$|X \star \psi_{j_1}| = \overline{S}X_f(j_1) X_s + X_s \tilde{X} + N_1 \ .$$

Now, a convolution with ψ_{j_2} produces

$$|X \star \psi_{j_1}| \star \psi_{j_2} = \overline{S}X_f(j_1) X_s \star \psi_{j_2} + X_s \tilde{X} \star \psi_{j_2} + N_2 \ , \tag{4.23}$$

with $E(|N_2|^2) \leq b_1$ since the convolution with ψ_j is also contractive. By denoting $X_1 = |X \star \psi_{j_1}| \star \psi_{j_2}$, $X_2 = \overline{S}X_f(j_1) X_s \star \psi_{j_2}$ and $X_3 = X_s \tilde{X} \star \psi_{j_2} + N_2$, it follows from

(4.23) that $E(|X_1 - X_2|) = E(|X_3|)$, and hence, since $f(x) = |x|$ is convex, thanks to the Jensen's inequality we obtain

$$|E(|X_1|) - E(|X_2|)| \leq E(||X_1| - |X_2||) \leq E(|X_3|) \ . \tag{4.24}$$

To conclude, we shall now bound $E(|X_3|)$. Since $E(|X|)^2 \leq E(|X|^2)$, we have

$$E(|X_3|) \leq \sqrt{E(|X_3|^2)} \ .$$

Finally, we decompose X_3 as the sum

$$\begin{aligned} X_3 &= (E(X_s) + \tilde{X}_s)\tilde{X} \star \psi_{j_2} + N_2 \\ &= \tilde{X} \star \psi_{j_2} + Y \star \psi_{j_2} + N_2 \ , \end{aligned}$$

since $E(X_s) = 1$ and by definition $Y = (X_s - E(X_s))(U[j_1]X_f - \overline{S}X_f(j_1))$. As a result, thanks to the fact that $\sqrt{E(|X+Y|^2)} \leq \sqrt{E(|X|^2)} + \sqrt{E(|Y|^2)}$, we obtain

$$\begin{aligned} \sqrt{E(|X_3|^2)} &\leq \sqrt{E(|\tilde{X} \star \psi_{j_2}|^2} + \sqrt{E(|Y \star \psi_{j_2}|^2)} + \sqrt{E(|N_2|^2)} \\ &\leq \sqrt{E(|U[p]X_f|^2)} + \sqrt{B_{sup,0}2^{-j_2\beta_0}E|Y|^2} + b_0 \ , \end{aligned}$$

which proves (4.21). □.

Proposition 4.7.3 thus approximates the second order coefficient $\overline{S}X(j_1, j_2)$ with a product of two first order coefficients, $E(U[j_1]X_f)$ and $E(U[j_2]X_s)$, which depend upon carrier and envelope respectively. The fidelity of the approximation is controlled by a commutator bound b_1, which satisfies $b_1 \sim 2^{j_1 - k}$ if the envelope X_s has its spectral density concentrated at scales $\geq k$.

The other two error terms b_2, b_3 measure the amount of interference $\tilde{X} = U[j_1]X_f - \overline{S}X_f(j_1)$ which remains after the first demodulation stage. This interference cannot in general be distinguished from the envelope X_s, but its influence on second order coefficients is bounded by b_2 and b_3. The first term $b_2 = E(|U[p]X_f|^2)^{1/2}$ corresponds to the residual energy of the carrier which escapes the first demodulation through the path (j_1, j_2). This energy is visible on $U[p]X$ since the spectral density of $X_s\tilde{X}$ in (4.23) contains a copy of $\hat{R}_{\tilde{X}}$, due to the fact that X_s is positive and hence its spectral density has a Dirac impulse at $\omega = 0$. The term b_3 carries the residual energy due to the interference Y. Its bound relies on the property that Y is a wideband process with regular spectrum, and hence that the fraction of energy captured by ψ_{j_2} is nearly proportional to the bandwidth of $|\hat{\psi}(2^{j_2}|\omega)|^2$, which is $\sim 2^{-j_2}$. This notion of wideband, regular spectrum is captured by the term $B_{sup,0}$. Since Y is the product of two independent noises, its spectrum is given by the convolution of the two densities and hence it inherits the best regularity amongst the two.

In particular, if X_f is Gaussian process, then $U[j_1]X_f$ is a Rayleigh process, and Section 4.6 shows that $\tilde{X} = U[j_1]X_f - \overline{S}X_f(j_1)$ has a spectral density which is well approximated by

$$\hat{R}_{\tilde{X}}(\omega) \approx C\hat{R}_{X_f \star \psi_{j_1}} \star \overline{\hat{R}}_{X_f \star \psi_{j_1}}(\omega) \ ,$$

where $\overline{\hat{R}}(\omega) = \hat{R}(-\omega)$. As a result, the regularity of $\hat{R}_{\tilde{X}}$ can be lower bounded from that of $\hat{R}_{X_f}$.

From the proof of proposition 4.7.3, we deduce also that first order scattering coefficients satisfy

$$|E(|X \star \psi_{j_1}|) - E(X_s)E(|X_f \star \psi_{j_1}|)| \leq \sqrt{b_1} \;, \tag{4.25}$$

and hence that for a sufficiently smooth envelope X_s with $E(X_s) = 1$ we have $\overline{S}X(j_1) \approx \overline{S}X_f(j_1)$. This suggests a renormalization of second order scattering coefficients, which eliminates the influence of the carrier:

$$TX(j_1, j_2) = \frac{\overline{S}X(j_1, j_2)}{\overline{S}X(j_1)} \approx \overline{S}X_s(j_2) \;. \tag{4.26}$$

This renormalized scattering is called scattering transfer, and will play a central role in Chapter 5 for the study of fractal processes.

Chapter 5

Multifractal Scattering

5.1 Introduction

Fractals are objects defined from a self-similarity property as one applies a dilation or a contraction. If $D_s X(t) = X(s^{-1}t)$ denotes a dilation operator on functions or processes, a fractal is characterized by a scaling law relating X with $D_s X$. Deterministic fractals such as the Cantor set correspond to functional relations $D_s X \equiv F(X, s)$ whereas fractal processes are defined through equalities in the sense of distributions or through their moments. Fractal processes are fundamental in the modeling of several physical systems, such as the study of turbulence, astronomical data, satellite imaging, and also in the field of finance. Self-similarity across scales implies that fractals are singular almost everywhere; different families of fractals are then obtained by describing their singularities.

A fundamental signature of a Multifractal is given by its spectrum of singularity. In presence of self-similarity, it is equivalent, by the Legendre transform, to a function $\zeta(q)$ giving the scaling law of its q-th order moments. Such quantities are in general difficult to estimate. Indeed, the estimation of high order moments requires a huge number of samples, due to the presence of rare large events which increase the variance of the estimators.

Scattering operators have proved effective in the discrimination of non-gaussian, fractal-like textures such as those in the Curet dataset [DVGNK99]. Figure 5.1 displays several examples from this dataset, showing that it contains many examples of self-similar textures with a multifractal behavior. The identification of the scaling laws which characterize multifractals is a particular instance of a texture discrimination problem. A natural question is then whether one can relate scattering coefficients to multifractal quantities. Many multifractals are characterized by rare, large jumps, which prevent them from having high order moments. As a consequence, the estimation of scaling laws from direct estimation of the moments is often unpractical due to lack of consistency. Expected scattering representations, on the other hand, are constructed as expected values defined from contractive operators, and exist as long as the process has finite first moments. We will study in this chapter which multifractal properties can be obtained from scattering coefficients, and how the resulting estimators compare to existing tools.

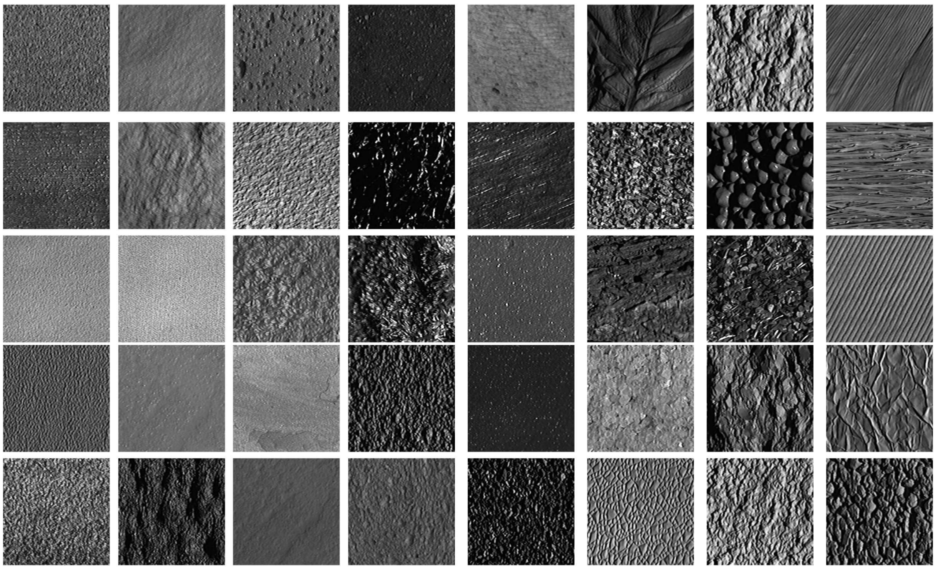


Figure 5.1: Examples taken from the CureT dataset studied in Chapter 4. Whereas some classes contain a characteristic scale and/or orientation, other examples exhibit multifractal behavior.

We shall concentrate in this chapter in one-dimensional multifractals. Self-similar fractal processes with stationary increments are characterized by their scaling laws. We introduce a new measure based on first and second order expected scattering coefficients, the scattering transfer:

$$TX(j_1, j_2) = \frac{\overline{S}X(j_1, j_2)}{\overline{S}X(j_1)} \ .$$

It is defined as a function of two path variables, but we show that the self-similarity of X implies that the scattering transfer is only a function of path increments: $TX(j_1, j_2) = \overline{T}X(j_2 - j_1)$. This normalized scattering measure, together with the expected first order scattering, defines a new tool to study fractal phenomena, capturing information which allows identification and discrimination of several self-similar fractal families. In particular, its asymptotic behavior as $l \to \infty$ yields a new characteristic exponent of the process. Moreover, the renormalization defining the scattering transfer brings near invariance with respect to fractional derivations; as a result, the scattering transfer can be interpreted as a form of geometric fractal descriptor, measuring the spatial or temporal dependencies between singularities.

The scattering transfer contains information from first and second order coefficients, but it also has the capacity to control the behavior of higher order scattering coefficients. For a wide class of fractals, we observe that high order scattering coefficients can be asymptotically predicted from the scattering transfer and its first order coefficients.

This asymptotic property, which can be interpreted as a Markov propagation across scattering paths, is proved for the Dirac measure, white Gaussian noise and Fractional Brownian motions.

Multifractal random measures are constructed with an integral scale, which restricts the self-similarity to a scale range of the form $(-\infty, 2^J)$. A fundamental signature of a multifractal X is given by its scaling exponents $\zeta(q)$, which control the asymptotic behavior of the moments of the increments $E(|X(t)-X(t-l)|^q)$ as l decreases. First order scattering measures the exponent $\zeta(1)$ since it is computed from first order moments of wavelet coefficients. Multifractals are characterized by a concave $\zeta(q)$, whose curvature gives an account of the degree of multifractality of the process. Therefore, its study requires access to information contained in high order moments. This information can be extracted consistently from the scattering transfer. In particular, thanks to the prediction of higher order scattering coefficients by the scattering transfer, we prove that the intermittency $2\zeta(1) - \zeta(2)$, measuring the curvature of $\zeta(q)$, is given by the logarithm of the smallest ρ satisfying the equation

$$\sum_{l \geq 1} \overline{T}X(l)^2 \rho^l = 1 \ .$$

Deterministic fractal measures can also be analyzed with the integral scattering transform defined in Section 2.3.4. Deterministic self-similarity is expressed as a periodicity property on the scattering transfer.

The chapter is structured as follows. Section 5.2 gives a background on stochastic and deterministic fractal analysis. Section 5.3 presents the scattering transfer and its associated transfer function, together with an estimation from windowed scattering coefficients. It also introduces an asymptotic Markov property which predicts higher order scattering coefficients from first and second order coefficients. Section 5.4 analyzes the scattering transfer for several self-similar fractal processes, showing that it captures important discriminative information, and we also show that they enjoy the asymptotic Markov property. We then focus, in Section 5.5, on the study of multifractal processes with an integral scale. We first obtain a characterization of the intermittency from the scattering transfer. Then we study the scattering transfer for several multifractal random cascades, and then we compare the resulting intermittency estimator with alternative estimates, showing a consistency in pair with state-of-the-art methods. Finally, in Section 5.7 we adapt the previous tools to the study of deterministic fractals, and show that the scattering transfer captures important information about the self-similarity of the fractal.

5.2 Review of Fractal Theory

This section reviews several fractal families and introduces the tools to measure their singularities, both for deterministic and stochastic fractals.

5.2.1 Fractals and Singularitites

Deterministic fractals are singular functions, characterized by their spectrum of singularity. Stochastic fractal processes are defined from a probabilistic scaling law.

From a deterministic point of view, point-wise singularities are studied with the Hölder exponents. A function f is $C^{\alpha}(u_0)$ if there exists a polynomial P_n of degree $n \leq \lfloor \alpha \rfloor$ and a constant $C > 0$ such that

$$\forall u \in \mathbb{R} \ , \ |f(u) - P_n(u - u_0)| \leq C |u - u_0|^{\alpha} \ .$$

One defines the Hölder exponent [Jaf00] of f at u_0 as

$$H_f(u_0) = \sup\{\alpha \,:\, f \in C^{\alpha}(u_0)\} \ .$$

A fractal f might contain singularities with different Hölder exponents. One can characterize the structure of singularities of f via its *spectrum of singularity* $\mathcal{D}(h)$, which is defined as the Hausdorff dimension of the sets

$$\Omega_h = \{u \,; H_f(u) = h\} \quad h \in \mathbb{R} \ .$$

A family $R_{\epsilon} = \{B_i\}_{i \in \mathbb{N}}$ is an ϵ-covering of A if $A \subset \bigcup_i B_i$ and $|B_i| \leq \epsilon \, , \forall i$. The Hausdorff dimension of a set A is defined as [Jaf00]

$$dim(A) = \sup\{\delta \,:\, \lim_{\epsilon \to 0} \inf_{R_{\epsilon} = \{B_i\}} \sum |B_i|^{\delta} = +\infty\} \ .$$

It measures the growth rate of an ϵ-covering of A as the maximum radius of the balls converges towards zero. As a result, for any covering of the support of f with balls of size at most s, the optimum cardinality of such covering is $\sim s^{-\mathcal{D}(h)}$. The spectrum of singularity thus measures the proportion of Holder h singularities visible at any scale s.

When all the singularities of a fractal f have the same Hölder exponent H, its spectrum of singularity $\mathcal{D}(h)$ is degenerate, i.e. $\forall \, h \neq H \, , \mathcal{D}(h) = 0$ and $\mathcal{D}(H) = 1$. Some authors [Jaf00; AFTV00] use this property to distinguish 'monofractals' from 'multifractals', which contain singularities of different Hölder exponents. We shall give a slightly different definition in Section 5.2.3.

5.2.2 Fractal Processes

Fractal processes are generally defined by a scaling law of the form

$$\forall \, s \in (S_-, S_+) \, , \ D_s X(t) \stackrel{l}{=} W_s X(t) \ ,$$

where $\stackrel{l}{=}$ means an equality in the probability law of the process for each s. The scaling law is valid for a range of scales (S_-, S_+) which can include $\pm\infty$. If W_s is a deterministic constant for any scale s, then the moments of $D_s X$ are uniformly scaled as s varies, and hence its probability distribution remains unchanged up to a dilation factor. On the other hand, if W_s is a random variable independent from X, then the probability distribution of $D_s X$ varies as the scale varies. Next section will show that these two cases correspond respectively to monofractal and multifractal processes.

The estimation of multifractal processes can be seen as a particular instance of a texture discrimination problem. Given a realization of a fractal, it is necessary to build consistent, informative estimators of the scaling laws which characterize the underlying physical processes.

5.2.3 Multifractal Formalism and Wavelets

We review in this section the main aspects of the multifractal formalism, originated by Collet et al. in [CLP87; GBP88] for the study of the statistical scaling properties of singular measures, and further developed by Parisi and Frisch [PF85].

The multifractal formalism originated in the framework of fully developed turbulence to explain the deviations of experimental data to the Kolmogorov theory of isotropic turbulence [Fri95]. Mandelbrot [Man74] introduced a family of multiplicative cascade measures in order to account for such deviations.

One can define [AJ04] the Holder exponent of a measure μ at u_0 as

$$H_\mu(u_0) = \lim_{\delta \to 0} \inf \frac{\log \mu([x_0 - \delta, x_0 + \delta])}{\log \delta} \ .$$

One can then associate a spectrum of singularity similarly as in the case of functions. Moreover, for one-dimensional signals, one can associate to each measure μ defined in the unit interval $[0, 1]$ the function

$$f_\mu(u) = \int_0^u d\mu(u) \ ,$$

which has the same Holder exponents as μ in the case $\sup_u H_\mu(u) < 1$ [AJ04]. Box-Counting (BC) methods [GBP88; MS91] obtain global singularity measures by defining suitable averages of local Hölder exponents. Originally they were defined as [GBP88]

$$\forall\, q \geq 0\,,\ \zeta_0(q) = \lim_{\delta \to 0} \frac{\log \sum_i \mu(b_i(\delta))^q}{\log \delta} \ ,$$

where the sum runs over a collection of boxes $\{b_i(\delta)\}$ of radius δ arranged over a regular grid. However, box-counting measures are dominated by strong Hölder singularities, which do not fully characterize the whole spectrum of singularity [MBA93a].

A multifractal analysis can be carried out by using a wavelet decomposition and studying the decay of the coefficients as the scale decreases. Jaffard and Meyer showed in [JM96] that if f has a Hölder exponent $H_f(u_0)$ at u_0 and ψ is a wavelet with enough vanishing moments, then under appropriate conditions on the wavelets, one has

$$|f \star \psi_s|(u_0) \sim s^{H_f(u_0)} \ ,$$

where $\psi_s(u) = s^{-1}\psi(s^{-1}u)$. By detecting wavelet modulus maxima, one can estimate the singularity spectrum by first defining a partition function [MBA93a]

$$Z(q,s) = \sum_{l\in\mathcal{L}(s)} \left(\sup_{(u,s')\in l} |f \star \psi_{s'}|(u) \right)^q , \tag{5.1}$$

where $\mathcal{L}(s)$ is the set of maxima lines crossing at scale s. The asymptotic behavior of $Z(q,s)$ as $s \to 0$ is given by the power law

$$Z(q,s) \sim s^{\zeta(q)} .$$

Bacry et al [MBA93b] and Jaffard [Jaf97] showed that using appropriate wavelets, the characteristic exponent $\zeta(q)$ can be related to the singularity spectrum by the Legendre transform, for self-similar signals with bounded spectrum:

$$\zeta(q) = \min_{h\in \mathrm{supp}(\mathcal{D}(h))} (qh - \mathcal{D}(h)) \ .$$

For signals with a concave spectrum of singularity, it is possible to recover $\mathcal{D}(h)$ by inverting the Legendre transform:

$$\mathcal{D}(h) = \min_q (qh - \zeta(q)) \ .$$

Self-similar functions have a concave spectrum of singularity and hence one can use the characteristic exponent $\zeta(q)$ to estimate $\mathcal{D}(h)$. However, in general the inverse Legendre transform only gives an upper bound of $\mathcal{D}(h)$.

Similarly, the characteristic exponents of a fractal measure can be estimated with a wavelet decomposition. Indeed,

$$\mu \star \psi_s(t) = s^{-1} \int \psi(s^{-1}(t-u)) d\mu(u)$$

defines a function in L^q for all q, which allows the computation of a partition function yielding the scaling law.

5.2.4 Multifractal Processes and Wavelets

We review now the theory of stochastic multifractals. We describe their scaling exponents in terms of wavelet decompositions and present some important families of fractal processes and measures.

Definition 5.2.1 *A stochastic process $\{X(t), t \geq 0\}$ with stationary increments is self-similar if it satisfies a scaling law of the form*

$$\forall s \geq 0 \ , \ \{X(st)\}_{t\geq 0} \stackrel{l}{=} \{s^H X(t)\}_{t\geq 0} \ ,$$

where H is a characteristic exponent [BKM08a].

Equivalently, one can show [BKM08b] that this is equivalent to a self-similarity of its increments:

$$\{X(st) - X(s(t-l))\}_t \stackrel{l}{=} \{s^H(X(t) - X(t-l))\}_t \ .$$

A change of time on the process thus recovers a scaled version of the same process, which leads to a power law behavior of its moments

$$\forall q \in \mathbb{R} \ , \ E(|X(t) - X(t-l)|^q) \simeq C_q l^{\zeta(q)} \ ,$$

where $\zeta(q) = qH$ is a linear function of q. We shall denote such processes as *monofractals*. In this case, the spectrum of singularity $\mathcal{D}(h)$ of X is degenerated, and supported in $h = H$. If a self-similar process $X(t)$ can be written as

$$X(t) \stackrel{l}{=} \int_0^t dX(s) \ , t \geq 0 \ ,$$

where dX is a stationary process, then the self-similarity of $X(t)$ induces a self-similarity on $dX(t)$:

$$\{D_{-s} dX\}_t \stackrel{l}{=} \{s^{H-1} dX\}_t \ .$$

Multifractal behavior is obtained by generalizing the notion of self-similarity. It can be defined as follows [BKM08a]:

Definition 5.2.2 *A process X with stationary increments has stochastic self-similarity if for a certain range of scales s we have*

$$\forall t \geq 0 \ , \forall 0 < l < t \ , X(st) - X(s(t-l)) \stackrel{l}{=} W_s(X(t) - X(t-l)) \ , \tag{5.2}$$

where W_s is a positive random variable independent of X.

This equivalence in distribution is stated for each fixed t, and in general does not imply that the two processes $X_1(t) = D_{s^{-1}} X(t) - D_{s^{-1}} X(t-l)$ and $X_2(t) = W_s(X(t) - X(t-l))$ have the same law. The stochastic self similarity yields a scaling law of the moments of its increments:

$$\forall q \in \mathbb{R} \ , \ E(|X(t) - X(t-l)|^q) \simeq C_q l^{\zeta(q)} \ . \tag{5.3}$$

The scaling exponent $\zeta(q)$ is now allowed to be a non-linear function of q. If (5.3) holds in the limit $l \to 0$, then $\zeta(q)$ is concave as a result of the convexity of the moments of a random variable, otherwise if it holds when $l \to \infty$, then $\zeta(q)$ is necessarily convex.

When $\zeta(q)$ is a non-linear function of q we say that the process is *multifractal*. In particular, if $\zeta(q)$ is strictly concave (resp stricly convex), it results [BKM08a] that the scaling law cannot hold at arbitrarily large scales (resp small scales).

Definition 5.2.3 *Assuming $\zeta(q)$ is concave, the integral scale of a multifractal X is the time scale T where the process ceases to be multifractal:*

$$T = \sup_l \ \{l \, s.t. \forall q \in \mathbb{R} \ , \ E(|X(t) - X(t-l)|^q) \simeq C_q l^{\zeta(q)}\} \ . \tag{5.4}$$

Equivalently, the integral scale also determines the range of stochastic self-similarity in (5.2): If $D_s X(t) = X(s^{-1}t)$, then

$$\forall s\,,\, l \geq 0\,\text{s.t.}\;\; l \leq T\,, st \leq T\;\;,\;\; D_{s^{-1}}X(t) - D_{s^{-1}}X(t-l) \stackrel{l}{=} W_s(X(t) - X(t-l))\;,t \geq 0\;. \tag{5.5}$$

One then generally assumes that the process becomes decorrelated beyond the integral scale. The degree of multifractality of a process, often referred as *intermittency*, is thus characterized by the curvature of its scaling exponents $\zeta(q)$. Some authors [BKM08a] define it as $\zeta''(0)$.

If ψ is a wavelet with at least a vanishing moment, then the scaling exponents can be recovered from the wavelet decompositions of the process [Mal12]:

$$\forall q \in \mathbb{R}\;,\;\; E(|X \star \psi_s|^q) \simeq C_q s^{\zeta(q)}\;. \tag{5.6}$$

In particular, $\zeta(1)$ measures the decay of the wavelet coefficients amplitude

$$E(|X \star \psi_s|) \simeq C_1 s^{\zeta(1)}\;,$$

which are estimated from realizations of X with $\mathbf{L}^1$ norms at each scale. Similarly, $\zeta(2)$ measures the decay of the variance of wavelet coefficients with the scale.

The integral scale of a multifractal can be also expressed in terms of wavelet coefficients. If $J = \log T$ is the log of the integral scale, then from (5.5), by setting $s = 2^j$ and $l = 2^k$ it results that

$$\forall\, j, k\,\text{s.t.}\;\; j + k \leq J\,, k \leq J\,,\;\; D_{2^{-j}}X \star \psi_k(t) \stackrel{l}{=} W_{2^j}(X \star \psi_k)(t)\;,t \geq 0\;. \tag{5.7}$$

In particular, monofractal processes satisfy (5.7) for all $j, k\, \in \mathbb{Z}^2$ with $W_{2^j} \equiv 2^{jH}$.

5.2.5 Cantor sets and Dirac Measure

We give in this section two examples of deterministic fractal measures: the Cantor measure and the Dirac measure.

Deterministic fractals can be constructed by giving a functional scaling law. In one dimension, a well known example is the triadic Cantor measure dC_∞ defined over the Cantor set. It is constructed recursively by subdividing a uniform measure into uniform measures on smaller intervals. The process starts with $dC_0(x) = dx$, where dx is the uniform Lebesgue measure of $[0, 1]$. This measure is then split into a piecewise uniform measure $dC_1(x)$ on the intervals $[0, 1/3]$, $[1/3, 2/3]$ and $[2/3, 1]$, with integrals respectively p_1, 0, and $1 - p_1$, such that $\int_{[0,1]} dC_1(x) = 1$. The process is repeated recursively, by updating each uniform measure of integral p on an interval $[a, b]$ into a piecewise uniform measure with integrals $p_1 p$, 0 and $(1 - p_1)p$ on the intervals $[a, (2a + b)/3]$, $[(2a+b)/3, (a+2b)/3]$ and $[(a+2b)/3, b]$ respectively. The limit measure dC_∞ is supported in the Cantor set and is self-similar, since

$$D_3 dC_\infty \equiv p_1 dC_\infty \quad \text{in } [0, 1]\;.$$

The Cantor measure is then invariant up to a normalisation constant to all dilations of the form 3^n. However, it is not self-similar when the dilation factor is not of that form. The Dirac measure δ is another example of self-similar measure. In this case, the self-similarity is observed for any dilation factor:

$$D_s \delta \equiv s\delta \quad , \forall s > 0 \; .$$

5.2.6 Fractional Brownian Motions

We briefly review an important class of self-similar processes, given by the fractional brownian motions.

An important family of fractal processes is given by the Fractional Brownian motion [Dou06]. It is defined as a centered Gaussian process $X_H(t)$ with covariance

$$E(X_H(t)X_H(t')) = |t|^{2H} + |t'|^{2H} - |t' - t|^{2H} \; .$$

The parameter $H \in (0, 1)$ is called the Hurst index and controls the average singularity of the process. The Brownian motion corresponds to a Hurst index $H = 1/2$. It is not a stationary process, but its increments $X_H \star \psi$ are stationary. $X_H(t)$ is a self-similar, monofractal process, since

$$X_H(st) \stackrel{l}{=} s^H X_H(t) \; , t \in \mathbb{R}, s > 0 \; .$$

Although X_H is not stationary, its increments $X_H(t) - X_H(t - l)$ are stationary, and thus one can define a generalized power spectrum [Wor95], given by the spectrum of the increment divided by the transfer function:

$$\widehat{R}_{X_H}(\omega) = \frac{\sigma_H^2}{|\omega|^{2H+1}} \; . \tag{5.8}$$

Each filtered process $X_H \star \psi_s$ has a power spectrum $\widehat{R}_{X,s}$ satisfying

$$\widehat{R}_{X,s}(\omega) = s^{2H+1} \widehat{R}_{X,1}(s\omega) \; ,$$

which, together with the fact that X_H is Gaussian, implies that

$$X_H \star \psi_s(st) \stackrel{l}{=} s^H X_H \star \psi_1(t) \; .$$

As a result, the characteristic exponent is the linear function $\zeta(q) = qH$, confirming the fact that X_H is a monofractal process. In particular, the Hurst exponent can be estimated from both $\zeta(1)$ and $\zeta(2)$.

5.2.7 α-stable Lévy Processes

Another important class of self-similar fractal processes comes from the class of stable Lévy processes.

A process $X(t)\,, t \geq 0$ is said to be a Lévy process [Kyp07] if its realizations are almost surely right continuous, with left limits, if $Prob(X(0) = 0) = 1$ and if its increments $X(t){-}X(t{-}\tau)$ are stationary in the strong sense and such that $X(t){-}X(s)$ is independent from $\{X(u)\ :u \leq s\}$.

Lévy processes necessarily originate from an infinitely divisible distribution, which ensures that $X(t)$ can be written as a sum of iid increments:

$$\forall t \geq 0\, n \in \mathbb{N}\,,\ X(t) \stackrel{l}{=} (X\,(t/n){-}X(0)){+}(X\,(2t/n){-}X\,(t/n)){+}\cdots{+}(X(t){-}X\,(t-t/n))\,.$$

The Lévy-Khintchine formula characterizes infinitely divisible distributions from their characteristic exponents. If Z is a random variable distributed along a probability law μ, and $\Psi(u) = -\log E(e^{iuZ})$ denotes the exponent of the characteristic function of Z, then μ is infinitely divisible if and only if

$$\forall\, u \in \mathbb{R},\ \Psi(u) = iau + \frac{1}{2}\sigma^2 u^2 + \int (1 - e^{iux} + iu\mathbf{1}_{|x|<1})\Pi(dx)\ , \tag{5.9}$$

where $a \in \mathbb{R}$, $\sigma \geq 0$ and Π is a measure, denoted the Lévy measure, concentrated on $\mathbb{R} \setminus \{0\}$ and satisfying $\int \min(1, x^2)\Pi(dx) < \infty$. A Lévy process can be decomposed as a sum of a Brownian motion, a compound Poisson point process, and a martingale [Kyp07].

Amongst infinitely divisible distributions, we are interested in those which are also *stable*. A random variable Z has a stable distribution if for all $n \geq 1$ we have

$$Z_1 + \cdots + Z_n \stackrel{l}{=} a_n Z + b_n\ ,$$

where Z_i are independent copies of Z, $a_n > 0$ and $b_n \in \mathbb{R}$. It results that necessarily $a_n = n^{1/\alpha}$ for $\alpha \in (0, 2]$. [Kyp07]. An α-stable Lévy process with centered increments satisfies

$$X(nt) \stackrel{l}{=} n^{1/\alpha} X(t)\ , t \geq 0, n \geq 1\ ,$$

and thus it is self-similar. The jump process $\Delta X(t)$ associated to the Lévy process $(X(t))_t$ is defined [Pap07] as

$$\forall\, 0 \leq t\,,\ \Delta X(t) = X(t) - X(t-)\ ,$$

where $X(t-) = \lim_{s \to t^-} X(s)$. It is a self-similar stationary process, satisfying

$$\forall 0 \geq t\ , \forall\, n > 0\ , \Delta X(nt) \stackrel{l}{=} n^{1/\alpha - 1} \Delta X(t)\ .$$

α-stable Lévy distributions are heavy tailed, with rare, large jumps. As a result, they only have moments strictly less than α [Kyp07].

5.2.8 Multifractal Random Cascades

Mandelbrot Cascade

The first example of a multifractal process is given by the Mandelbrot cascade [Man74]. It constructs a random measure dM_∞ in the unit interval from a binary tree. The root of the tree is initialized with a uniform measure dM_0 on the whole interval $I_0 = [0,1)$, $M_0(I_0) = \int_{I_0} dM_0(s) = Y_0$, where Y_0 is a random positive variable. The uniform measure dM_0 is then refined with dM_1, which is uniform in the two intervals $I_1 = [0,1/2)$, $I_2 = [1/2,1)$, and such that the measure of each of these intervals is set to $M_1(I_1) = Y_1 = Y_0X_1$, $M_1(I_2) = Y_2 = Y_0X_2$, where X_1, X_2 are iid, infinitely divisible positive random variables, independent from Y_0. The process is iterated along the binary tree which partitions each interval I_i in two adjacent intervals of equal length, yielding M_∞. This construction converges towards a non-trivial measure under appropriate conditions on the positive random variables X_i [Man74]. The resulting measure has stochastic self-similarity. If one defines a contraction of a measure as $D_s\mu(I) = \mu(D_s^{-1}I)$ for any Lebesgue measurable set I, then

$$D_sM_\infty(I) \overset{l}{=} X_sM_\infty(I) \ , s > 1 \ .$$

Multifractal Random Measure (MRM)

The construction of the Mandelbrot cascade defines a process which is not stationary, since it is constructed on a binary tree which is not translation invariant. However, one can generalize the construction with the so-called random multiplicative cascades [BKM08a]. One starts by defining the random measure

$$\forall\, l \geq 0 \ , \ MRM_{l,J}(dt) = e^{2\omega_l^J(t)}dt \ ,$$

in the sense that for all Lebesgue measurable sets I one has $MRM_{l,J}(I) = \int_I e^{2\omega_l^J(t)}dt$. Here, ω_l^J is a stationary Gaussian Process with mean and covariance

$$E(\omega_l^J) = -\frac{\lambda^2}{2}\ln(2^J/l) \ , \quad \mathrm{Cov}_\omega(\tau) = \lambda^2 \ln\left(\frac{2^J}{\sup(l,|\tau|)}\right)\mathbf{1}_{[-2^J,2^J]}(\tau) \ . \tag{5.10}$$

A log-normal Multifractal Random Measure (MRM) is defined as the weak limit when $l \to 0$ of the random measure $MRM_{l,J}$ [BKM08b]:

$$MRM(dt) = \lim_{l\to 0^+} e^{2\omega_l^J(t)}dt \ , \tag{5.11}$$

which exists and is non-trivial as long as $\lambda^2 < 1/2$.

An MRM is a positve random measure which models the stochastic volatility, and it has been successfully applied in the fields of finance and turbulence [BKM08a]. It defines a multifractal process. When ω is a Gaussian process, its scaling exponent is given by

$$\zeta(q) = \left(1 + \frac{\lambda^2}{2}\right)q - \frac{\lambda^2}{2}q^2 \ .$$

This process depends upon the intermittency coefficient λ^2, which controls the amount of non-linearity of the characteristic exponents $\zeta(q)$, and the integral scale 2^J, which defines the support of their correlation functions.

Multifractal Random Walk (MRW)

A Multifractal Random Walk (MRW) is constructed as the limit $l \to 0$ of

$$MRW(dt) = \lim_{l \to 0} e^{\xi_l^J(t)} dW(t) \ ,$$

where $dW(t)$ is a Wiener noise, and ξ_l^J is also a stationary Gaussian process with same covariance as in (5.10) and mean given by

$$E(\xi_l^J) = -\lambda^2 \ln(2^J/l) \ .$$

An MRW models the stock price fluctuations, and can be thought as the composition of an MRM with white gaussian noise. Similarly as in the MRM case, it is a multifractal process. When ξ follows a Gaussian distribution, its scaling exponents given by

$$\zeta(q) = \left(\frac{1}{2} + \lambda^2\right) q - \frac{\lambda^2}{2} q^2 \ .$$

Again, the intermittency coefficient λ^2, controls the amount of non-linearity of the characteristic exponents $\zeta(q)$, and the integral scale 2^J defines the decorrelation scale of the process.

5.2.9 Estimation of Fractal Scaling Exponents

The estimation of the parameters defining a multifractal process has been studied in [BKM08b; Jaf97]. A fundamental object to be estimated is the characteristic exponent $\zeta(q)$. In particular, the curvature of $\zeta(q)$ controls the intermittency of the process and characterizes a multifractal behavior, as opposed to the homogeneous or monofractal case. Some authors define the intermittency as $\zeta^{''}(0)$ [BKM08a]. Since $\zeta(0) = 0$, a measure of the curvature at $q = s_0$ is obtained with $\zeta(2s_0) - 2\zeta(s_0)$. In particular, for $s_0 = 1$, one obtains the intermittency measure given by $\zeta(2) - 2\zeta(1)$. In the case of a log-normal random cascade, since $\zeta(q)$ is a parabole, this finite difference coincides with $\zeta''(0)$.

A first strategy to obtain λ^2 estimates the moments $E(|X \star \psi_j|^q)$ and then performs a regression on the predicted scaling behavior

$$E(|X \star \psi_j|^q) \sim 2^{(J-j)\zeta(q)} \ .$$

Estimations based on wavelet leaders [ALJ04], which are used in texture discrimination [WAJZ09], are indirectly obtaining such scaling laws. However, the variance of this estimator converges very slowly with the sample size, as $N^{-1+\alpha}$ with $\alpha > 0$, as shown in [BKM08b; OW00]. High order moments are difficult to estimate due to the expansive nature of x^q for $q > 1$ and $x > 1$. In [BKM08a], another estimator for the intermittence is proposed, based on the covariance properties of the logarithm of the absolute value of the increments.

5.3 Scattering Transfer

Fractal processes are analyzed with scattering representations. We introduce the scattering transfer for self-similar fractals in Section 5.3.1 and also for non-stationary random measures in 5.3.2. It is computed with scattering operators, which are non-expansive. As a result, Section 5.3.3 defines a consistent estimator of the scattering transfer using the windowed scattering transform. Finally, in 5.3.4 we introduce an asymptotic prediction property of the transfer function, which we shall verify for several fractal families.

5.3.1 Scattering transfer for Processes with stationary increments

The scattering transfer is defined from first and second order expected scattering coefficients. Self-similarity yields a transfer function which gives a new signature of fractal processes.

If $X(t)$ is a process with stationary increments, then $X \star \psi_j(t)$ is stationary for all $j \in \mathbb{Z}$, and since convolutions and moduli do not affect stationarity, we have that $U[p]X(t)$ is stationary for all p. We recall from 2 that the expected scattering is then defined as

$$\overline{S}X(p) = E(U[p]X) \ .$$

Since we concentrate in this chapter on uni-dimensional functions and processes, we shall denote scattering paths as $p = (j_1, \dots, j_m)$ for sake of simplicity.

The scattering transfer is first defined for processes with stationary increments:

Definition 5.3.1 *Let $X(t)$ be a process with stationary increments. The Scattering transfer of $X(t)$ is defined for $(j_1, j_2) \in \mathbb{Z}^2$ by*

$$TX(j_1, j_2) = \frac{\overline{S}X(j_1, j_2)}{\overline{S}X(j_1)} \ . \tag{5.12}$$

The transfer is well defined as long as $\overline{S}X(j) > 0$ for all $j \in \mathbb{Z}$. This is guaranteed if in particular the wavelet ψ satisfies $\hat{\psi}(\omega) \neq 0$ almost everywhere in $\{\omega > 0\}$, and the generalized spectral density of $X(t)$ [Mal08] satisfies $\hat{R}_X(\omega) > 0$ on a set of positive measure . This last condition is automatically satisfied for stationary processes with auto-correlation $R_X \in \mathbf{L}^1$. Unless specified otherwise, we shall assume throughout the rest of the chapter that this admissibility condition is satisfied.

For any $j \in \mathbb{Z}$, for sake of simplicity let us denote the dyadic dilation $D_j X(t)$ of the process $X(t)$ by $D_j X(t) = X(2^{-j}t)$ for all t. Recall from definition 5.2.1 that self-similar processes satisfy $\{D_{-j}X(t)\}_t \stackrel{l}{=} \{2^{jH}\, X(t)\}_t$ for all $j \in \mathbb{Z}$.

The following proposition shows that self-similar processes have a scattering transfer which becomes a transfer function of a single scale variable .

Proposition 5.3.2 *If the stochastic process $X(t)$ satisfies $\{D_{-j}X(t)\}_t \stackrel{l}{=} \{A_j\, X(t)\}_t$ for $j \in \mathfrak{I}$, where A_j is independent from $X(t)$, then*

$$\forall j_1 \in \mathfrak{I} \ , \ \overline{S}X(j_1, j_2 ..., j_m) = E(A_{j_1})\, \overline{S}X(0, j_2 - j_1, ..., j_m - j_1) \ ,$$

and hence

$$\forall j_1 \in \mathfrak{I} \ , \ TX(j_1, j_2) = \overline{T}X(j_2 - j_1)\,. \tag{5.13}$$

The scattering transfer matrix is then called a transfer function. In particular, self-similar processes satisfy (5.13) for all $j_1 \in \mathbb{Z}$.

Proof: Since $\psi_j = 2^{-j} D_j \psi$, a change of variables yields $D_j|X \star \psi| = |D_j X \star \psi_j|$, and hence

$$|X \star \psi_j| = D_j |D_{-j} X \star \psi| \ . \tag{5.14}$$

If $p = (j_1 \dots j_m)$, it results that

$$\begin{aligned} U[p]X &= \ ||||X \star \psi_{j_1}| \star \psi_{j_2}| \star ...| \star \psi_{j_m}| \\ &= \ D_{j_1}||||D_{-j_1} X \star \psi| \star \psi_{j_2 - j_1}| \star ...| \star \psi_{j_m - j_1}| \ , \end{aligned} \tag{5.15}$$

If $X(t)$ is stationary, then $E(D_j X) = E(X)$, and thus from (5.15) we derive that

$$\overline{S}X(p) = E(|||\ |D_{-j_1} X \star \psi| \star \psi_{j_2 - j_1}| \star ...| \star \psi_{j_m - j_1}|) \ . \tag{5.16}$$

Now, if $j_1 \in \mathfrak{I}$, the self-similarity of X implies that

$$\{D_{-j_1} X(t)\}_t \stackrel{l}{=} \{A_{j_1} X(t)\}_t \ ,$$

where A_{j_1} is independent of $X(t)$. This implies that

$$\overline{S}X(p) = E(A_{j_1}) E((|||\ |X \star \psi| \star \psi_{j_2 - j_1}| \star ...| \star \psi_{j_m - j_1}|) = E(A_{j_1}) \overline{S}X(0, j_2 - j_1, \dots, j_m - j_1) \ ,$$

and hence

$$TX(j_1, j_2) = \frac{\overline{S}X(j_1, j_2)}{\overline{S}X(j_1)} = \frac{E(A_{j_1}) \overline{S}X(0, j_2 - j_1)}{E(A_{j_1}) \overline{S}X(0)} = \overline{T}X(j_2 - j_1) \ ,$$

which proves (5.13) $\square$.

The scattering transfer function thus defines a new measure for self-similar processes, computed from first and second order coefficients. We shall see that it contains highly discriminative information. Although it is computed from expected values of contractive operators $U[j_1]$, $U[j_1, j_2]$ applied to X, we shall see that it contains information about the scaling exponent $\zeta(q)$ for $q > 1$ of the process.

In particular, it defines a new characteristic exponent for $X(t)$. First order scattering coefficients $\overline{S}X(j)$ have a scaling law given by

$$\overline{S}X(j) = E(|X \star \psi_j|) \simeq 2^{j \zeta(1)} \ ,$$

provided the wavelets have at least a vanishing moment. The scattering transfer function defines a scaling law $\overline{T}X(l) \simeq 2^{l \alpha}$. The characteristic exponent α gives a new signature of $X(t)$, which complements the exponent $\zeta(1)$. Since

$$U[j]X(t) = \left| \int X(u) \psi_j(t - u) du \right| \leq \int |X(u)| |\psi_j(t - u)| du = |X| \star |\psi_j|(t) \ ,$$

it follows that $\overline{S}X(j) = E(U[j]X) \leq E(|X|)\|\psi\|_1$,$\forall j$ and hence that if $p = (j_1 \dots j_m)$, then

$$\overline{S}X(p) \leq E(|X \star \psi_{j_1}|)\|\psi\|_1^{m-1} \ , \tag{5.17}$$

which shows in particular that the scattering transfer is defined for any process with stationary increments having finite first moment.

The scattering transfer defines a normalization which brings further invariance properties to the scattering representation. The self-similarity used in proposition 5.3.2 is a particular case of a more general invariance produced by the renormalization. The following proposition shows that the scattering transfer generates invariance with respect to other linear, translation covariant operators satisfying a similarity property when applied to the wavelets ψ_j.

Proposition 5.3.3 *Let $X(t)$ be a process with stationary increments, and let L be a linear, translation covariant operator such that*

$$\forall j \ , \ |X \star L\psi_j| \stackrel{l}{=} C_{L,j}|X \star \psi_j| \ , \tag{5.18}$$

where $C_{L,j}$ is a random variable independent of X. Then, if $\tilde{X} = LX$, we have

$$TX(j_1, j_2) = T\tilde{X}(j_1, j_2) \ . \tag{5.19}$$

Proof: Since L is linear and translation covariant, it commutes with convolutions; it follows that

$$\tilde{X} \star \psi_j = LX \star \psi_j = X \star L\psi_j \ ,$$

and hence

$$|\tilde{X} \star \psi_j| = |X \star L\psi_j| \stackrel{l}{=} C_{L,j}|X \star \psi_j| \ .$$

As a result,

$$T\tilde{X}(j_1, j_2) = \frac{E(||\tilde{X} \star \psi_{j_1}| \star \psi_{j_2}|)}{E(|\tilde{X} \star \psi_{j_1}|)} = \frac{E(C_{L,j})E(||X \star \psi_{j_1}| \star \psi_{j_2}|)}{E(C_{L,j})E(|X \star \psi_{j_1}|)} = TX(j_1, j_2) \ \Box.$$

Propostion 5.3.3 gives sufficient, idealized conditions under which the transfer is invariant. These conditions are nearly satisfied by operators which are almost diagonalized by complex wavelets. Indeed, if $L\psi_j \approx C_{L,j}\psi_j$, then it follows that (5.18) is approximately verified. One can control the quality of this approximation with a supremum norm on the diagonalisation error, as shown by the following proposition.

Proposition 5.3.4 *Let $X(t)$ be a stationary process such that $E(|X|^2) < \infty$ and $R_X \in \mathbf{L}^1$, and let $j_1, j_2 \in \mathbb{Z}$. Suppose that L is a linear, translation covariant operator in $\mathbf{L}^2$, and let*

$$\delta = \inf_{c \in \mathbb{C}} \|L\psi_{j_1} - c\psi_{j_1}\|_2 \ . \tag{5.20}$$

Then, if $\tilde{X}(t) = LX(t)$, we have

$$\left|T\tilde{X}(j_1, j_2) - TX(j_1, j_2)\right| \leq \delta \, \frac{\sqrt{\|R_X\|_1}}{\overline{S}\tilde{X}(j_1)} \left(\|\psi\|_1 + TX(j_1, j_2)\right) \ . \tag{5.21}$$

Proof: Let us first approximate the numerator and the denominator of $T\tilde{X}(j_1, j_2)$. Since $\|L\psi_{j_1}\| < \infty$ and

$$|c|\|\psi_{j_1}\| - \|L\psi_{j_1}\| \leq \|L\psi_{j_1} - c\psi_{j_1}\|_2 \leq |c|\|\psi_{j_1}\| + \|L\psi_{j_1}\| \ ,$$

it follows that the infimum in (5.20) is attained in a compact set and hence that there exists $c_0 \in \mathbb{C}$ such that $\|L\psi_{j_1} - c_0\psi_{j_1}\|_2 = \delta$. We then have

$$\begin{aligned} \Big|E(|\tilde{X} \star \psi_{j_1}|) - |c_0|E(|X \star \psi_{j_1}|)\Big| &= |E(|X \star L\psi_{j_1}| - |X \star c_0\psi_{j_1}|)| \\ &\leq E(|X \star (L\psi_{j_1} - c_0\psi_{j_1})|) \ . \end{aligned} \tag{5.22}$$

But for a given $h \in \mathbf{L}^2$, we also have

$$\begin{aligned} E(|X \star h|)^2 &\leq E(|X \star h|^2) = \int \hat{R}_X(\omega)|\hat{h}(\omega)|^2 d\omega \\ &\leq \sup_\omega \hat{R}_X(\omega)\|\hat{h}\|_2^2 \leq \|R_X\|_1\|h\|_2^2 \ , \end{aligned} \tag{5.23}$$

where in the last inequality we have used the Plancherel identity. By applying (5.23) to $h = L\psi_{j_1} - c_0\psi_{j_1}$, (5.22) becomes

$$\Big|E(|\tilde{X} \star \psi_{j_1}|) - E(|X \star c_0\psi_{j_1}|)\Big| \leq \sqrt{\|R_X\|_1}\delta \ . \tag{5.24}$$

Similarly, the second order coefficients are approximated by

$$\begin{aligned} |E(||\tilde{X} \star \psi_{j_1}| \star \psi_{j_2}|) - E(||X \star c_0\psi_{j_1}| \star \psi_{j_2}|)| &\leq E(|(|\tilde{X} \star \psi_{j_1}| - |X \star c_0\psi_{j_1}|) \star \psi_{j_2}|) \\ &\leq E(||X \star L\psi_{j_1}| - |X \star c_0\psi_{j_1}||)\|\psi\|_1 \\ &\leq E(|X \star (L\psi_{j_1} - c_0\psi_{j_1})|)\|\psi\|_1 \\ &\leq \sqrt{\|R_X\|_1}\delta\|\psi\|_1 \ . \end{aligned} \tag{5.25}$$

Finally, by observing that if a, $\bar{a}$, b, $\bar{b} > 0$ and

$$|a - \bar{a}| \leq \delta_a \ , \ |b - \bar{b}| \leq \delta_b \ ,$$

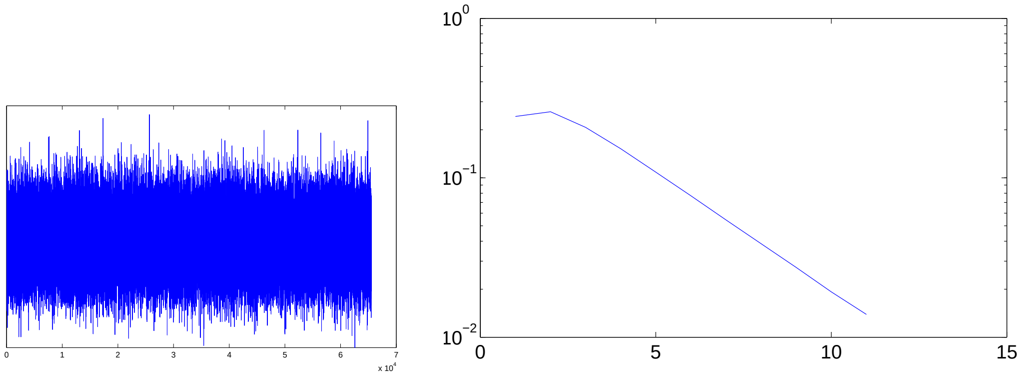
then

$$\left|\frac{a}{b} - \frac{\bar{a}}{\bar{b}}\right| \leq \frac{1}{b}\left(|a - \bar{a}| + |b - \bar{b}|\frac{\bar{a}}{\bar{b}}\right) \leq \frac{1}{b}\left(\delta_a + \delta_b\frac{\bar{a}}{\bar{b}}\right) \ , \tag{5.26}$$

the result (5.21) follows by applying (5.26) with $a = \overline{S}\tilde{X}(j_1, j_2)$, $\bar{a} = \overline{S}X(j_1, j_2)$, $b = \overline{S}\tilde{X}(j_1)$ and $\bar{b} = \overline{S}X(j_1)$ $\square$.

The scattering transfer is thus nearly invariant to linear operators which are nearly diagonalized by wavelets. Of particular importance are the fractional derivative operators D^α, defined in the Fourier domain as

$$\widehat{D^\alpha x}(\omega) = (i\omega)^\alpha \hat{x}(\omega) \ , \omega \in \mathbb{R} \ .$$



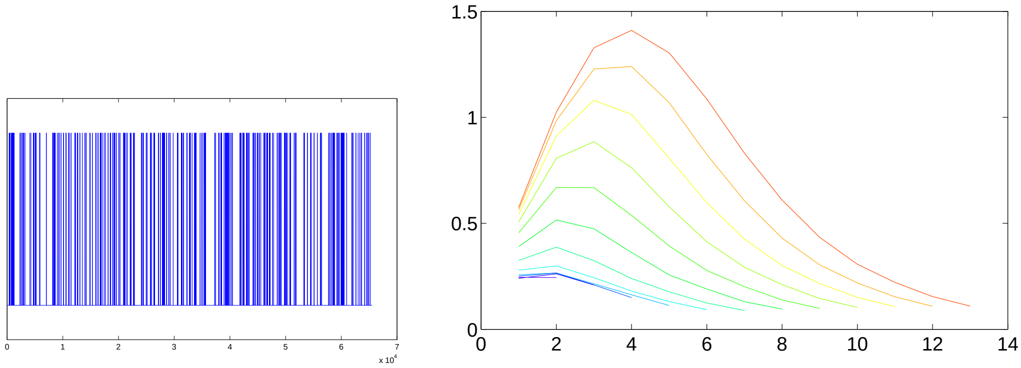
(a) (b)

Figure 5.2: (a) A realization of Gaussian white noise, (b) $\log \overline{T}X(l)$, estimated from 100 realizations of size 2^{16} points, following the procedure described in Section 5.3.3. We observe $\overline{T}X(l) \simeq 2^{-l/2}$. This asymptotic behavior will be obtained analytically in Section 5.4.1.

Appendix A.2 shows that wavelets with good frequency localization nearly diagonalize fractional derivatives. The near invariance of the scattering transfer with respect to derivation can be interpreted in geometric terms; a derivation operator modifies the Hölder singularity of all points in a uniform way. The renormalization of scattering coefficients creates a descriptor which is sensitive only to the local singularity differences, thus conveying geometric information.

Figure 5.2 shows an example of a self-similar stationary process, given by white Gaussian noise, and its associated transfer function, which only depends upon $j_2 - j_1$. This behavior contrasts with the one depicted in figure 5.3, which corresponds to a Bernouilli white noise. This noise is not self-similar, since dilations change the average density of the process. We observe that as the first scale j_1 increases, the scattering transfer $TX(j_1, j_1 + l)$ converges towards a transfer function. Indeed, in that case, the filtered process $X \star \psi_{j_1}(t)$ converges towards a Gaussian process thanks to the Central Limit theorem, which is self-similar.

Figure 5.4 shows examples of two processes and its corresponding jump processes, which are obtained in the discrete case by the linear, translation invariant derivative operator $\Delta X(t) = X(t) - X(t-1)$. The first row shows a realization of a Bernouilli process and its associated jump process, and the second row shows a realization of a MRW cascade with its jump process. First order scattering coefficients are sensitive to changes in the spectral density of the process, and are hence affected by derivative. On the other hand, as predicted, the scattering transfer remains nearly invariant to the derivation.



(a)

(b)

Figure 5.3: (a) A realization of Bernouilli white noise, with parameter $p = 2^{-8}$. (b) Scattering transfer. We plot the curves $T_{j_1}(l) = T(j_1, j_1 + l)$ for several j_1, estimated from 100 realizations of size 2^{16}. As j_1 increases, the curve converges towards a transfer function $\overline{T}(l) \simeq 2^{-l/2}$ of a Gaussian white noise, which is a the bottom of the plot.

5.3.2 Scattering transfer for non-stationary processes

We define the scattering transfer for non-stationary processes. We observe that self-similarity defines a transfer with similar behavior as in the stationary case.

An important class of fractal random processes are not stationary. Mandelbrot [Man74] constructed a multifractal process as a random measure defined from a multiplicative binary cascade. This cascade is not translation invariant and as a consequence the resulting measure is not stationary.

We can compute a scattering representation for a random measure μ with compact support using the integral scattering transform. For each $p = (j_1 \dots j_m)$, the integral

$$\widetilde{S}\mu(p) = \int (U[p]\mu)(u)du$$

is random variable. The scattering representation for a random measure is thus defined as

$$\overline{S}\mu(p) = E(\widetilde{S}\mu(p)) \ . \tag{5.27}$$

This representation is estimated from several realizations of the random measure by first computing the integral scattering for each realization and then averaging for each path p across the realizations.

The scattering transfer is then defined analogously as in (5.3.1), by

$$T\mu(j_1, j_2) = \frac{\overline{S}\mu(j_1, j_2)}{\overline{S}\mu(j_1)} \ .$$

A random measure of $[0, 1]$ is self-similar if for any open set $I \subset [0, 1]$

$$D_j\mu(I) \stackrel{l}{=} W_j\mu(I) \quad , j > 0 \ ,$$

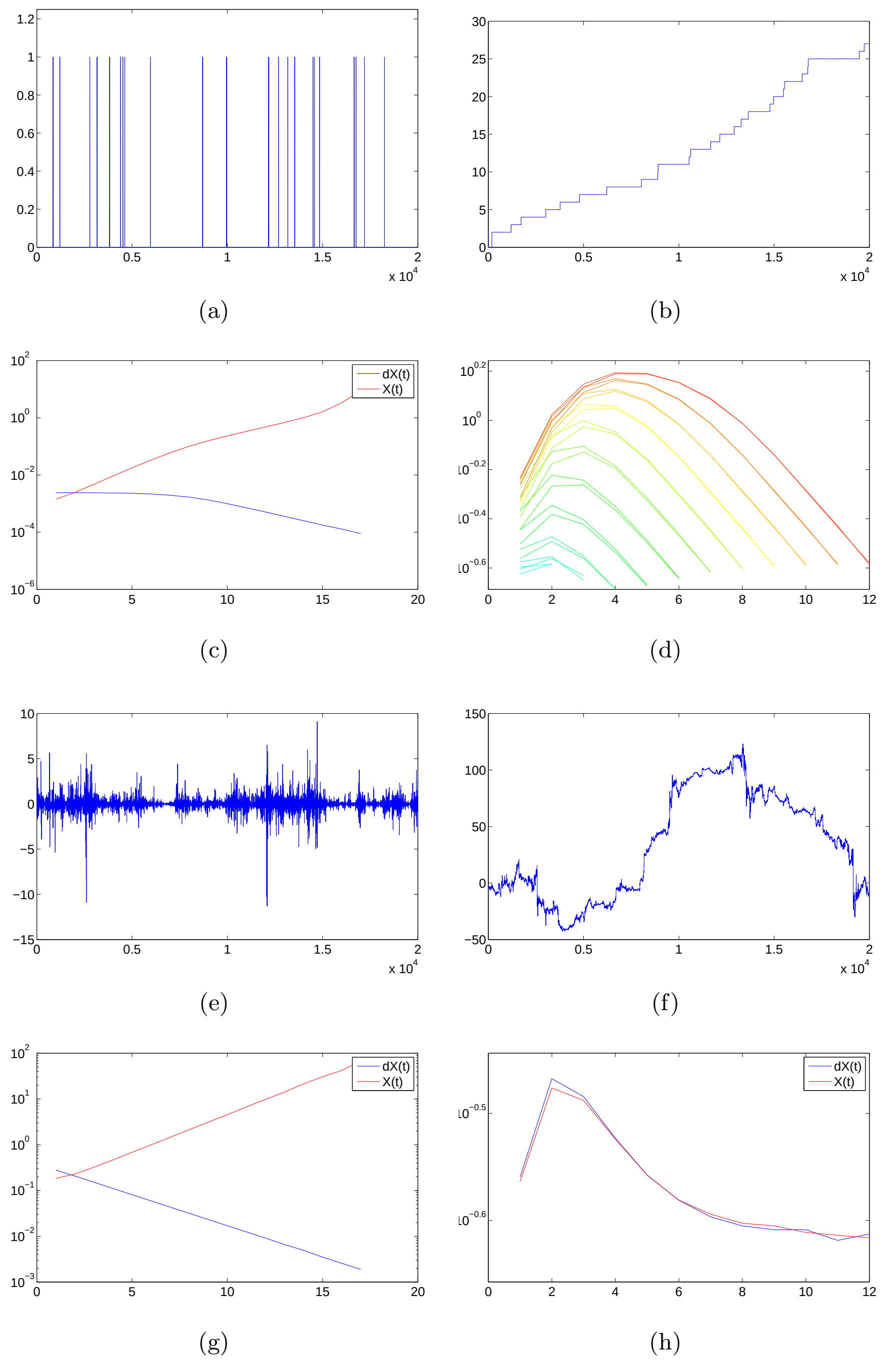


Figure 5.4: (a) A realization of Bernouilli white noise $dX(t)$. (b) A realization of the cumulated process $X(t) = \int_0^t dX(u)$. (c) First order scattering coefficients of $dX(t)$ and $X(t)$. (d) Scattering transfer of $dX(t)$ and $X(t)$. We plot the curves $T_{j_1}X(l) = TX(j_1, j_1 + l)$ for several j_1. Each color corresponds to a different value of j_1. Notice how the curves corresponding to $X(t)$ and $dX(t)$ are nearly overlapping. (e) and (f): Realizations of an MRW process $dX(t)$ and its associated cumulated process $X(t) = \int_0^t dX(u)$. (g): First order scattering coefficients $\overline{S}dX(j)$ and $\overline{S}X(j)$. (h): Scattering transfer $\overline{T}dX(l)$ and $\overline{T}X(l)$.

where W_j is a random variable independent of μ. The Mandelbrot cascade is self-similar for dyadic dilations. In that case, we recover a transfer function as in the stationary case before, before reaching an integral scale given by the support of the measure.

5.3.3 Estimation of Scattering transfer

The scattering transfer is estimated from averaged scattering coefficients. For self-similar processes, we construct an estimator of the transfer function which minimizes the mean squared error, by combining measures from all scales according to their covariance. For processes with finite second moments, one can approximate the scattering transfer covariance from the windowed scattering covariances.

If X is self-similar up to an integral scale 2^J, then the previous section showed that

$$\overline{T}X(l) = TX(j_1, j_1 + l) \quad , \; j_1 + l \leq J \; .$$

If one supposes that scattering coefficients can be measured for scales $j \geq 0$, an estimator of $\overline{T}X(l)$ can be obtained by aggregating estimators of each scattering transfer coefficient $TX(j_1, j_1+l)$ for $0 \leq j_1 \leq J-l$. Let us suppose that Y_j is an estimator of $TX(j, j+l) = \frac{E(U[j,j+l]X)}{E(|X \star \psi_j|)}$ with bias $b_j = E(Y_j - TX(j, j+l))$ and with finite variance. The aggregation of estimators given by the weights $\mathbf{h} = (h_0, \dots, h_{J-l})$

$$\sum_{0 \leq j \leq J-l} h_j Y_j \tag{5.28}$$

has a mean-squared error

$$F(\mathbf{h}) = E\left(\Big| \sum_j h_j Y_j - \overline{T}X(l) \Big|^2 \right) = \overline{T}X(l)^2 + \mathbf{h}^T E(\mathbf{Y}\mathbf{Y}^T)\mathbf{h} - 2\overline{T}X(l)\mathbf{h}^T((\overline{T}X(l))\mathbf{u} + \mathbf{b}) \; , \tag{5.29}$$

where $\mathbf{b} = (b_0, \dots, b_{J-l})$ and $\mathbf{u} = (1, \dots, 1)$. If we denote by $\Sigma = E(\mathbf{Y}\mathbf{Y}^T)$ the correlation matrix of the family of estimators, then this quadratic form is minimized by the linear combination $\mathbf{h}^*$ satisfying

$$\mathbf{h}^* = \overline{T}X(l)\,\Sigma^{-1}((\overline{T}X(l))\mathbf{u} + \mathbf{b}) = \overline{T}X(l)\,\Sigma^{-1}E(\mathbf{Y}) \; . \tag{5.30}$$

If the family of estimators Y_j is unbiased, then an unbiased aggregation $\bar{\mathbf{h}}$ must satisfy $\sum_j \bar{h}_j = 1$; its mean squared error expression is simplified to

$$F(\bar{h}) = E\left(\Big| \sum_j \bar{h}_j Y_j - \overline{T}X(l) \Big|^2 \right) = E\left(\Big| \sum_j \bar{h}_j (\overline{T}X(l) - Y_j) \Big|^2 \right) \; , \tag{5.31}$$

which implies that in this particular case, if $\bar{\Sigma} = E((\mathbf{Y} - \overline{T}X(l)\mathbf{u})(\mathbf{Y} - \overline{T}X(l)\mathbf{u})^T)$ denotes the centered covariance of the estimators Y_j, the MSE estimate is obtained by

$$\bar{\mathbf{h}}^* = \underset{\mathbf{v}, \sum_j v_j = 1}{\operatorname{argmin}} \; \mathbf{v}^T \bar{\Sigma} \mathbf{v} \; , \tag{5.32}$$

which corresponds to the normalized eigenvector of $\bar{\Sigma}$ with smallest eigenvalue.

The MSE estimator of the scattering transfer function is thus obtained from estimates of each transfer coefficient $TX(j, j+l)$, and require access to the covariance structure of these estimators. The coefficients defining these estimators are however not known, and need to be estimated from the data.

Let us now compute estimators of each scattering transfer coefficient $TX(j_1, j_2)$. The windowed scattering transform $S_J[p]X = U[p]X \star \phi_J$ is an unbiased estimator of $\overline{S}X(p) = E(U[p]X)$, since $\int \phi_J(u)du = 1$. For a wide range of ergodic processes, we observed that the variance of this estimator converges to zero as $2^J \to \infty$, and hence that $S_J[p]X(u)$ converges in probability to the constant $\overline{S}X(p)$ as $J \to \infty$ [Mal12]:

$$\forall \epsilon > 0\,, \lim_{2^J \to \infty} \mathrm{Prob}(|S_J[p]X(u) - \overline{S}X(p)| > \epsilon) = 0\ . \tag{5.33}$$

A first estimate for the scattering transfer is given by the ratio of two windowed scattering coefficients:

$$T_J[j_1, j_2]jX(u) = \frac{S_J[j_1, j_2]X(u)}{S_J[j_1]X(u)}\ . \tag{5.34}$$

We shall denote $T_J[j_1, j_2]$ the windowed scattering transfer, by analogy with the windowed scattering estimator. For a given J, this estimator is biased, since $S_J[j_1, j_2]X(u)$ and $S_J[j_1]X(u)$ are not independent random variables, and hence

$$E\left(\frac{S_J[j_1, j_2]X(u)}{S_J[j_1]X(u)}\right) \neq \frac{E(S_J[j_1, j_2](u))}{E(S_J[j_1]X(u))}$$

in general.

However, since $S_J[j_1, j_2]X(u) \to \overline{S}X(j_1, j_2)$ in probability and $S_J[j_1]X(u) \to \overline{S}X(j_1)$ in probability, then it results that $T_J[j_1, j_2]X(u) = \frac{S_J[j_1, j_2]X(u)}{S_J[j_1]X(u)}$ also converges in probability to $\frac{\overline{S}X(j_1, j_2)}{\overline{S}X(j_1)} = \overline{T}X(j_1, j_2)$ as long as $\overline{S}X(j_1) > 0$. As a result, the estimators $T_J[j_1, j_2]X(u)$ for $(j_1, j_2) \in \mathbb{Z}^2$ are asymptotically unbiased as the scale $2^J \to \infty$, for the class of processes having a mean squared consistent scattering.

If X has stationary increments and $X \star \psi_j$ have finite second moments, then one can approximate the covariance of $T_J[j_1, j_2]X$ using truncated Taylor approximations, as in the Delta method [KM00]. Indeed, let us first suppose that $S_J[j_1, j_2]X$ has finite energy and that X satisfies the consistency condition (5.33). If the random variables X and Y concentrate around their respective means μ_X and μ_Y, we can consider a second order approximation of the function $g(X, Y) = \frac{X}{Y}$:

$$\begin{aligned} g(X, Y) \quad &\approx \quad g(\mu_X, \mu_Y) + \nabla g(\mu_X, \mu_Y) \cdot (X - \mu_X, Y - \mu_Y) \\ &\quad + \frac{1}{2}(X - \mu_X, Y - \mu_Y)^T Hg(\mu_X, \mu_Y)(X - \mu_X, Y - \mu_Y)\ , \end{aligned} \tag{5.35}$$

where Hg is the Hessian matrix $Hg = (\frac{\partial^2 g}{\partial x_i \partial x_j})_{i,j}$. By applying the approximation (5.35) to $X = S_J[j_1, j_2]X(u)$, $Y = S_J[j_1]X(u)$ and taking the expected value, we obtain an approximation of the bias of $T_J[j_1,j_2]X(u)$:

$$
\begin{aligned}
E(T_J[j_1,j_2]X(u)) &= E\left(\frac{S_J[j_1,j_2]X(u)}{S_J[j_1]X(u)}\right)\\
&\approx \frac{E(S_J[j_1,j_2]X(u))}{E(S_J[j_1]X(u))} - \frac{\mathrm{Cov}(S_J[j_1,j_2]X(u),S_J[j_1]X(u))}{E(S_J[j_1]X(u))^2}\\
&\quad +E(S_J[j_1,j_2]X(u))\frac{\mathrm{Var}(S_J[j_1]X(u))}{E(S_J[j_1]X(u))^3}\\
&= TX(j_1,j_2) \qquad (5.36)\\
&+ \frac{1}{\overline{S}X(j_1)^2}\left(TX(j_1,j_2)\mathrm{Var}(S_J[j_1]X(u)) - \mathrm{Cov}(S_J[j_1,j_2]X(u),S_J[j_1]X(u))\right)\ .
\end{aligned}
$$

In particular, the leading bias term converges to 0 for processes having a mean squared consistent scattering, since in that case both $\mathrm{Var}(S_J[j_1]X(u))$ and $\mathrm{Cov}(S_J[j_1,j_2]X(u),S_J[j_1]X(u))$ converge to 0 as $J\to\infty$.

Similarly, the covariance of $(T_J[j_1,j_2]X)_{(j_1<j_2)}$, which determines the MSE estimates defined in (5.30, 5.32), is approximated by a second order development of the function $g(X_1,X_2,Y_1,Y_2)=\frac{X_1}{Y_1}\frac{X_2}{Y_2}$, yielding

$$
\begin{aligned}
&E(T_J[j,j+l]X\,T_J[j',j'+l]X) \approx TX(j,j+l)TX(j',j'+l) +\\
&\frac{\mathrm{Cov}(S_J[j,j+l]X,S_J[j',j'+l]X)}{\overline{S}X(j)\overline{S}X(j')} - \frac{\overline{S}X(j',j'+l)\mathrm{Cov}(S_J[j,j+l]X,S_J[j]X)}{\overline{S}X(j)^2\overline{S}X(j')}\\
&-\frac{\overline{S}X(j',j'+l)\mathrm{Cov}(S_J[j,j+l]X,S_J[j']X)}{\overline{S}X(j)\overline{S}X(j')^2} - \frac{\overline{S}X(j,j+l)\mathrm{Cov}(S_J[j',j'+l]X,S_J[j]X)}{\overline{S}X(j)^2\overline{S}X(j')}\\
&-\frac{\overline{S}X(j,j+l)\mathrm{Cov}(S_J[j',j'+l]X,S_J[j']X)}{\overline{S}X(j)\overline{S}X(j')^2} + \frac{\overline{S}X(j,j+l)\overline{S}X(j',j'+l)\mathrm{Cov}(S_J[j]X,S_J[j']X)}{\overline{S}X(j)^2\overline{S}X(j')^2}\\
&+\frac{\overline{S}X(j,j+l)\overline{S}X(j',j'+l)\mathrm{Var}(S_J[j]X)}{\overline{S}X(j)^3\overline{S}X(j')} + \frac{\overline{S}X(j,j+l)\overline{S}X(j',j'+l)\mathrm{Var}(S_J[j']X)}{\overline{S}X(j)\overline{S}X(j')^3}\ . \qquad (5.37)
\end{aligned}
$$

The approximation (5.37) contains cross-correlation terms between first and second order scattering across different scales. If X is a Gaussian process, then $X\star\psi_j$ is nearly decorrelated from $X\star\psi_{j'}$ for $j\neq j'$, which implies that the covariance approximation (5.37) can be simplified by removing the cross-scale correlation terms.

Figure 5.5 shows the scattering transfer estimators $T_J[j_1,j_2]X$ corresponding to a MRM cascade, together with the scattering transfer function estimator (5.28) with aggregation coefficients given by (5.30), where all the covariances have been estimated with the empirical covariances and the formulas (5.36, 5.37). We plot confidence intervals for $TX(j,j+l)$ for several values of j using an interval given by the mean of $T_J[j,j+l]X$ and its standard deviation. We observe that the variance of $T_J[j,j+l]X$ increases with j for all values of l. This phenomena can be explained by approximating $\mathrm{Var}(T_J[j,j+l]X)$

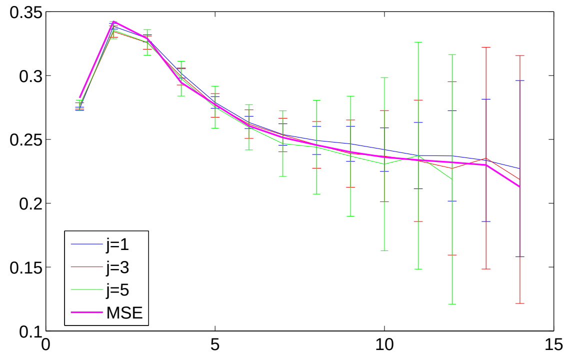


Figure 5.5: Estimation of transfer function for an MRM process with intermittency $\lambda^2 = 0.1$. We plot confidence intervals for scattering transfer coefficients $T_J[j, j+l]X(u)$ for several values of j. In bold magenta we plot the estimated transfer function, using the MSE estimation (5.30).

with the covariance approximation (5.37):

$$\begin{aligned} &\mathrm{var}\Big(T_J[j_1, j_2]X(u)\Big) \approx \\ &\frac{1}{\overline{S}X(j_1)^2}\Big(\mathrm{var}(S_J[j_1, j_2]X) + 2TX(j_1, j_2)\sqrt{\mathrm{var}(S_J X[j_1, j_2])\mathrm{var}(S_J[j_1]X)} \\ &+TX(j_1, j_2)^2\mathrm{var}(S_J[j_1]X)\Big)\ , \end{aligned} \tag{5.38}$$

Section 5.5.4 will show that MRM have first order scattering coefficients $\overline{S}X(j_1)$ which converge towards a constant value. On the other hand, the variances $\mathrm{var}(S_J[j_1, j_2]X)$ and $\mathrm{var}(S_J[j_1]X)$ increase with j_1, since the number of decorrelated samples in an interval of size 2^J is proportional to 2^{J-j_1}. The MSE estimate takes these variances into account by assigning lower aggregation weights to the estimates $T_J[j_1, j_1+l]X$ with large j_1.

There are a variety of multifractals which do not have finite second moments, such as α-stable Lévy processes with $\alpha < 2$. In this case, the covariance of $S_J[j_1, j_2]$ does not exist, and hence the approximations derived in (5.36, 5.37) cannot be applied. The large, rare jumps of such processes are also visible in the propagated processes $|X \star \psi_j|$ and $||X \star \psi_{j_1}| \star \psi_{j_2}|$.

However, the scattering transfer can mitigate the influence of these jumps, thanks to cancellation due to the renormalization. This suggests a more general estimator of the scattering transfer $TX[j_1, j_2]$:

$$T_{J,J'}[j_1, j_2]X(u) = \left(\frac{U[j_1, j_2]X \star \phi_{2^J}}{U[j_1]X \star |\psi_{2^{j_2}}| \star \phi_{2^J}}\right) \star \phi_{2^{J'}}(u)\ . \tag{5.39}$$

The first scale parameter J averages the wavelet modulus coefficients $U[j_1, j_2]X$ and $U[j_1]X$, but now we adjust the lowpass scales in such a way that the numerator and the

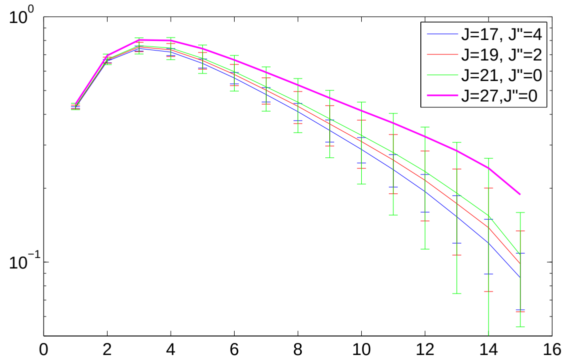


Figure 5.6: Estimation of scattering transfer $TX(1,l)$ for an α-stable Lévy process with $\alpha = 1.3$ with windowed transfer estimators for different values of J, J', using a total of 2^{21} points. In purple, an "oracle" estimator $T_J X$ using 2^{27} points. We plot confidence intervals using the empirical average and standard deviation for each estimator.

denominator have the same spatial resolution. The second scale parameter J' averages the windowed scattering transfer over a support proportional to $2^{J'}$. The available samples are thus combined in two separate stages. The scale J determines the amount of averaging on windowed scattering coefficients before they are combined to obtain an estimate of the scattering transfer. The scale J' is then responsible for averaging the normalized coefficients. The previous windowed scattering transfer corresponds to the case where $2^{J'} = 0$ and $J \to \infty$.

Figure 5.6 compares windowed scattering transfer estimates (5.39) for different values of J, J' for an equal amount of samples. We observe that adjusting the scale 2^J of the windowed scattering trades-off the bias and the variance of the estimator. The numerical results from Figure 5.6 seem to indicate that the bias produced by a small J dominates the risk of the estimator $T_{J,J'}$.

5.3.4 Asymptotic Markov Scattering

The scattering transfer has the capacity to predict a particular asymptotic of high order scattering coefficients. As the scattering path increments grow, we observe that the prediction error converges to 0 for a wide class of fractal processes.

A given path $p = (j_1, \dots, j_m)$ defines a nonlinear operator $U[p] = \prod_{1 \le i \le m} U[j_i]$ which is built from a cascade of operators $U[j_i]$. Since these operators are non-commutative, the resulting scattering coefficients $\overline{S}X(p) = E(U[p]X)$ in general depend upon the whole path p.

The scattering transfer can be used to predict high order scattering coefficients by assuming a Markov propagation across scattering paths:

Definition 5.3.5 *Let X be a process with stationary increments. The Markov scattering propagation of X is defined by $\overline{S}^M X(j_1) = \overline{S}X(j_1)$ and for $m \ge 2$ and progressive paths*

$(j_1 \dots j_m)$ by

$$\overline{S}^M X(j_1, ..., j_m) = \overline{S}X(j_1) \prod_{k=2}^{m} TX(j_{k-1}, j_k) \ . \tag{5.40}$$

A process X has an asymptotic Markov Scattering if

$$\lim_{\min(j_k - j_{k-1}) \to \infty} \left| \overline{S}X(j_1, ..., j_m) - \overline{S}^M X(j_1, \dots, j_m) \right| = 0 \ , \tag{5.41}$$

This property means that if the jumps $j_k - j_{k-1}$ of $p = (j_1 \dots j_m)$ are large enough, the scattering coefficient $\overline{S}X(p)$ can be obtained from its ancestor $p_0 = (j_1 \dots j_{m-1})$ using the scattering transfer:

$$\lim_{j_m \to \infty} |\overline{S}X(p) - \overline{S}X(p_0) TX(j_{m-1}, j_m)| = 0 \ .$$

This property is important since it ensures that the asymptotic behavior of scattering coefficients is captured by first and second order scattering. This phenomena is in accordance with the previous chapters, where we saw that first and second order coefficients captured most of the discriminative information in most of the encountered situations.

For a fractal X, the asymptotic Markov property is verified by computing the average relative error

$$\overline{e}_X(l) = \frac{1}{|\mathcal{P}^l|} \sum_{p \in \mathcal{P}^l} \frac{|\overline{S}X(p) - \overline{S}^M X(p)|^2}{|\overline{S}X(p)|^2} \ , \tag{5.42}$$

where

$$\mathcal{P}^l = \{ p = (j_1 < \dots < j_m) \in \mathcal{P}_J \ ; \ \min_k j_k - j_{k-1} = l \}$$

denotes the set of progressive paths whose smallest jump is equal to l. The asymptotic first order Markov property predicts that $\lim_{l \to \infty} \overline{e}_X(l) = 0$.

Next section shows that several monofractal processes have the asymptotic Markov scattering property.

5.4 Scattering Analysis of Monofractal Processes

This section computes the Scattering transfer for several monofractal processes. We derive the asymptotic behavior of the scattering transfer for the white gaussian noise and Fractional Brownian Motions. The scattering transfer yields a new signature which allows identification and discrimination of different fractal families. In particular, the exponent α in the scaling law $\overline{T}X(l) \simeq 2^{l\alpha}$ reveals information about the fractal family and its associated parameter.

5.4.1 Gaussian White Noise

We consider in this section the white Gaussian noise, which is a self-similar noise with monofractal behavior. We compute the first order scattering coefficients as well as the asymptotics of its scattering transfer.

Theorem 5.4.1 *Let $dX(t)$ be a Gaussian white noise of unit variance, and suppose the scattering is computed with an analytic wavelet ψ with fast decay. Then the following properties hold:*

1. $\overline{S}dX(j) = \|\psi\|_2 \sqrt{\frac{\pi}{4}} 2^{-j/2}$.
2. *Its scattering transfer satisfies* $TdX(j_1, j_2) = \overline{T}dX(j_1 - j_2)\,,\ \forall (j_1, j_2) \in \mathbb{Z}\times\mathbb{Z}$, *and*

$$\lim_{l\to\infty} 2^{l/2}\overline{T}dX(l) = \sqrt{\int R_{|dX\star\psi|}(\tau)d\tau}\ . \tag{5.43}$$

Proof: Let us first prove (i). Since $dX(t)$ follows a Gaussian distribution, so does $dX \star \psi_j$, and hence $|dX \star \psi_j|(t)$ is a Rayleigh random variable, since ψ_j is analytic. Its mean does not depend upon t and is given by $E(|dX \star \psi_j|) = \sqrt{\frac{\pi}{4}}\sqrt{E(|dX \star \psi_j|^2)}$. Since dX is a white noise of unit variance, its spectral density is given by

$$\widehat{R}_{dX}(\omega) = 1\ ,$$

and hence

$$E(|dX \star \psi_j|^2) = \int \widehat{R}_{dX}(\omega)|\hat{\psi}_j(\omega)|^2 d\omega = 2^{-j}\|\psi\|_2^2\ ,$$

which implies that

$$\overline{S}dX(j) = E(|dX \star \psi_j|) = \|\psi\|_2\sqrt{\frac{\pi}{4}}2^{-j/2}\ .$$

We shall now prove (ii). The first statement, namely that the scattering transfer satisfies $TdX(j_1, j_2) = \overline{T}dX(j_2 - j_1)$, follows immediately from Proposition 5.3.2, since dX is self-similar for all scales, with $D_j dX(t) \overset{l}{=} 2^{j/2}dX(t)$.

Let us now study the scattering transfer. Since dX is self-similar,

$$||dX \star \psi_{j_1}| \star \psi_{j_2}| \overset{l}{=} D_{j_1}||D_{-j_1}dX \star \psi| \star \psi_{j_2-j_1}| \overset{l}{=} 2^{-j_1/2}D_{j_1}||dX \star \psi| \star \psi_{j_2-j_1}\ .$$

If we write $l = j_2 - j_1$, we will prove (5.43) by first proving that the sequence of stationary processes $Z_l = 2^{l/2}|dX \star \psi| \star \psi_l$ converges in distribution to a Gaussian process, and then showing that this convergence yields a convergence of the first and second moments of Z_l. The following lemma, proved in Appendix B.1, shows that the second moments of Z_l are uniformly bounded and converge towards a non-zero constant:

Lemma 5.4.2 *Let $Z_l(t) = 2^{l/2}|dX \star \psi| \star \psi_l(t)$, $l \in \mathbb{N}$, and let $\gamma_0 = \|\psi\|_2^2(\int R_{|dX\star\psi|}(\tau)d\tau)$, $\gamma_1 = \|\psi\|_2^2\|\int R_{|dX\star\psi|}\|_1$. Then $\gamma_0\,,\gamma_1 < \infty$, and the sequence of random variables $Z_l(t)$ satisfies*

$$\forall l\,,\ E(|Z_l(t)|^2) \le \gamma_1\ ,\ \textit{and}\ \lim_{l\to\infty} E(|Z_l(t)|^2) = \gamma_0\ . \tag{5.44}$$

The convergence towards a gaussian distribution is given by the following lemma, proved in Appendix B.2:

Lemma 5.4.3 *The sequence of stationary processes $Z_l = 2^{l/2}|dX \star \psi| \star \psi_l$, $l \in \mathbb{N}$, satisfy*

$$\forall\ t\ , \lim_{l\to\infty} Z_l(t) \xrightarrow{d} Z = Z_{(r)} + iZ_{(i)}\ , \tag{5.45}$$

where $Z_{(r)}$, $Z_{(i)} \sim \mathcal{N}(0, \sigma^2/2)$, $\sigma^2 = \lim_{l\to\infty} E(|Z_l|^2)$, and $\xrightarrow{d}$ denotes convergence in distribution.

Let us first see how (5.45) implies (5.43). Since the complex modulus $F(X(t)) = |X(t)|$ is a continuous function, it follows from the Continuous Mapping Theorem on metric spaces ([Pol84], Theorem 12) that (5.45) implies

$$\forall t\ ,\ \lim_{l\to\infty} |Z_l(t)| \xrightarrow{d} R\ , \tag{5.46}$$

where R follows a Rayleigh distribution.

We can exploit the convergence in distribution to obtain the asymptotic behavior of the moments of Z_l. The following lemma, proved in Appendix B.3, shows that the convergence in distribution in such conditions implies the convergence of the first and second moments:

Lemma 5.4.4 *The sequence $Z_l = 2^{l/2}|dX \star \psi| \star \psi_l$ satisfies*

$$\forall\, t\,,\ \lim_{l\to\infty} E(|Z_l(t)|^r) = E(R^r)\ ,\ \textit{for}\ 1 \leq r \leq 2\ , \tag{5.47}$$

where R follows a Rayleigh distribution.

Using these lemmas, from (5.47) it follows that

$$\frac{\lim_{l\to\infty} E(|Z_l|^2)}{\lim_{l\to\infty} E(|Z_l|)^2} = \frac{E(|R|^2)}{E(|R|)^2} = \frac{4}{\pi}\ ,$$

implying that

$$\lim_{l\to\infty} E(|Z_l|) = \lim_{l\to\infty} \sqrt{\frac{\pi}{4}}\sqrt{E(|Z_l|^2)} = \frac{\pi}{4}\|\psi\|_2\sqrt{\int R_{|dX\star\psi|}(\tau)d\tau}\ , \tag{5.48}$$

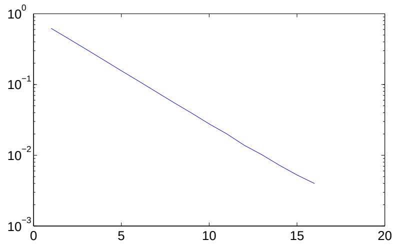
and hence

$$\lim_{l\to\infty} 2^{l/2}\overline{T}dX(l) = \lim_{l\to\infty} \frac{E(|Z_l|)}{E(|dX \star \psi|)} = \sqrt{\int R_{|dX\star\psi|}(\tau)d\tau}\ , \tag{5.49}$$

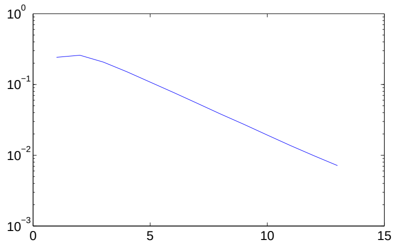
which finishes the proof $\square$.

Figure 5.7 shows the first order scattering and the transfer function estimated from the windowed scattering transform using 100 realizations of size $N = 2^{16}$. We display the logarithms $\log \overline{S}dX(j)$ and $\log \overline{T}dX(l)$ in order to reveal the power law behavior predicted by proposition 5.4.1:

$$\overline{S}dX(j) \simeq 2^{-j/2}\quad ,\quad \overline{T}dX(l) \simeq 2^{-l/2}\ .$$



(a)



(b)

Figure 5.7: (a) plot of $\log \overline{S}dX(j)$ as a function of j, estimated from 100 realizations of $N = 2^{16}$ points. (b) plot of $\log \overline{T}dX(l)$ as a function of l, estimated from the same number of realizations. We verify that $\overline{S}dX(j) \sim 2^{-j/2}$ and $\overline{T}dX(l) \sim 2^{-l/2}$.

For small l, we notice that the scattering transfer of dX does not follow the scaling law. Indeed, in that case $U[j_1]dX \star \psi_{j_1+l}$ has a distribution which is not well approximated by a Gaussian distribution. However, convergence is observed relatively early, at $l \approx 3$.

The proof of proposition 5.4.1 showed that the second order wavelet modulus process $U[1,l]dX(t) = ||dX \star \psi| \star \psi_l|(t)$ converges for all t in distribution towards a Rayleigh distribution, as $l \to \infty$. Thus, for sufficiently large l, $U[1,l]dX(t)$ has approximately the same distribution as $U[l]dX(t) = |dX \star \psi_l|(t)$, up to a normalization factor:

$$\forall t\,,\ ||dX \star \psi| \star \psi_l|(t) \overset{d}{\approx} C_l |dX \star \psi_l|(t)\ , l \gg 1\ . \tag{5.50}$$

It follows that when $j_2 \gg j_1$, then third order scattering coefficients satisfy

$$\begin{aligned} \overline{S}dX(j_1,j_2,j_3) &= \overline{S}dX(j_1)\,\frac{\overline{S}dX(j_1,j_2)}{\overline{S}dX(j_1)}\,\frac{\overline{S}dX(j_1,j_2,j_3)}{\overline{S}dX(j_1,j_2)} \\ &\approx \overline{S}dX(j_1)\,\overline{T}dX(j_2-j_1)\,\frac{C_{j_1,j_2}E(|dX\star\psi_{j_2}|\star\psi_{j_3}|)}{C_{j_1,j_2}E(|dX\star\psi_{j_2}|)} \\ &= \overline{S}dX(j_1)\,\overline{T}dX(j_2-j_1)\,\overline{T}dX(j_3-j_2)\ , \end{aligned}$$

and, by repeating the argument on higher order scattering, we obtain the asymptotic Markov property. Table 5.1 shows the relative approximation error

$$\overline{e}_X(l) = \frac{1}{|\mathcal{P}^l|}\sum_{p\in\mathcal{P}^l}\frac{|\overline{S}X(p) - \overline{S}^M X(p)|^2}{|\overline{S}X(p)|^2}\ ,$$

introduced in (5.42), which measures the asymptotic Markov scattering property. We verify that for large enough minimum jump, the Markov approximation reaches a relative approximation error of 10^{-3}.

The proof of the asymptotic Markov property requires showing that the approximation in distribution (5.50) is sufficient to invoke a central limit theorem as the jumps between scales increase. The proof of lemma 5.4.3 shows that having an integrable autocorrelation is a nearly sufficient condition. This motivates the following conjecture:

Table 5.1: Average Markov approximation error $\overline{e}_{dX}(l)$ as a function of minimum path jump l. The error is computed over paths of order $m \leq 4$.

l	1	2	3	4	5	6
$\overline{e}_{dX}(l)$	$1.5 \cdot 10^{-2}$	$8 \cdot 10^{-3}$	$5 \cdot 10^{-3}$	$5 \cdot 10^{-3}$	$2 \cdot 10^{-3}$	$1 \cdot 10^{-3}$

Conjecture 5.4.5 *Let X be a stationary process with finite energy $\sigma^2 = E(|X(t)|^2)$, and such that $R_X \in \mathbf{L}^1$. Then*

$$\lim_{l\to\infty} 2^{l/2} E(|X \star \psi_\lambda|) = \|\psi\|^2 \sqrt{\frac{\pi \int R_X(\tau)d\tau}{4}} \ . \tag{5.51}$$

5.4.2 Fractional Brownian Motion and FGN

Fractional Brownian Motions are an important class of monofractal processes. They are not stationary but do have stationary increments, which allows us to compute its expected scattering representation. Its associated increment processes are Fractional Gaussian Noises (FGN), which are stationary.

The following proposition shows that the Hurst exponent of a Fractional Brownian Motion is captured by first order scattering coefficients, while its scattering transfer has the same asymptotic behavior as the White Gaussian noise.

Proposition 5.4.6 *Let $X_H(t)$ be a Fractional Brownian Motion with Hurst exponent $0 < H \leq 1$. Let ψ be analytic, with fast decay and such that the zeros of $\hat{\psi}(\omega)$ in $\omega \in (0, \infty)$ form a discrete set. Then the following properties hold:*

1. *$\overline{S}X_H(j) = C2^{jH}$, with $C = \sqrt{\frac{\pi E(|X_H \star \psi|^2)}{4}}$.*
2. *Its scattering transfer satisfies $TX_H(j_1, j_2) = \overline{T}X_H(j_1 - j_2)$ and*

$$\lim_{l\to\infty} 2^{l/2} \overline{T}X_H(l) = \sqrt{\int R_{|X_H \star \psi|}(\tau)d\tau} \ . \tag{5.52}$$

Proof: The process $X_j(t) = X_H \star \psi_j(t)$ is a stationary Gaussian process [Mal08], with power spectrum given by:

$$\widehat{R}_{X_j}(\omega) = \sigma_0^2 |\hat{\psi}(2^j\omega)|^2 |\omega|^{-(2H+1)} \ . \tag{5.53}$$

As a result, a change of variables shows that

$$E(|X_j|^2) = \int \widehat{R}_{X_j}(\omega)d\omega = 2^{2jH}C \ ,$$

with $C = E(|X_H \star \psi|^2)$. Since X_j is Gaussian, we conclude that $\overline{S}X_H(j) = E(|X_j|) = \sqrt{\frac{\pi C}{4}} 2^{jH}$.

The second statement can be derived directly from theorem 5.4.1. Indeed, X_H is obtained from a white Gaussian noise dX with a fractional linear, translation operator, acting on the wavelet:

$$\forall t\,,\; X_H \star \psi_j(t) \stackrel{l}{=} dX \star \psi_j^H(t)\ ,$$

where

$$\forall \omega\ ,\ \hat{\psi}^H(\omega) = \hat{\psi}(\omega)\omega^{-(2H+1)/2}\ .$$

If ψ is analytic and with fast decay, it results that ψ^H is analytic and with fast decay as well. Besides, the zeroes of $\hat{\psi}^H(\omega)$ in $(0,\infty)$ are the same as the zeroes of $\hat{\psi}$. Thus, ψ^H satisfies the admissibility conditions of proposition 5.4.1, and hence we can replace $Z_l = 2^{l/2}|dX \star \psi| \star \psi_l$ by $Z_l^H = 2^{l/2}|dX \star \psi^H| \star \psi_l$ in (5.43) to obtain

$$\lim_{l\to\infty} 2^{l/2}\overline{T}X_H(l) = \left(\int R_{|dX\star\psi^H|}(\tau)d\tau\right)^{\frac{1}{2}} \quad \square.$$

Figure 5.8 shows the first order scattering and the transfer function of fractional brownian motions for $H = 0.2, 0.4, 0.6, 0.8$, estimated from the windowed scattering transform using 100 realizations of size $N = 2^{16}$. We identify the first scaling exponent $\overline{S}X^H(j) \simeq 2^{j\zeta(1)}$, with $\zeta(1) = H$, and we verify that the scattering transfer satisfies $\overline{T}X_H(l) \simeq 2^{-l/2}$.

Fractional Gaussian noises $dX_H(t)$ are defined as the increment processes of Fractional Brownian Motions. They are self-similar stationary processes, satisfying

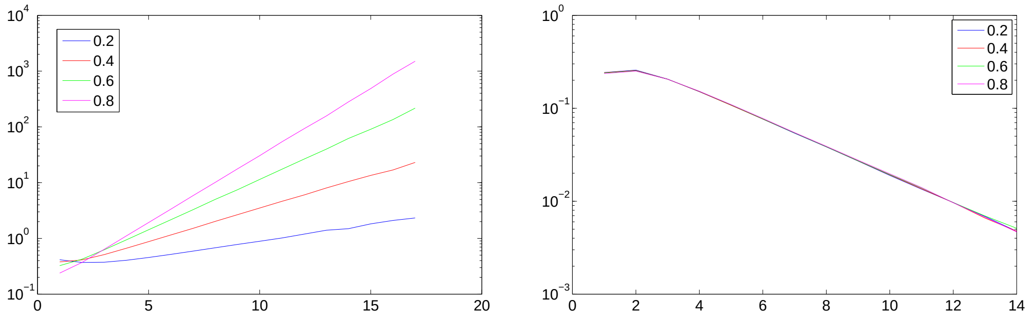
$$\forall t \geq 0\ ,\ dX_H(st) \stackrel{l}{=} s^{H-1}dX_H(t)\ .$$

The proof of proposition 5.4.6 shows that the same arguments can be applied to dX_H to obtain $\overline{S}dX_H(j) \simeq 2^{j(H-1)}$ and $\overline{T}dX_H(l) \simeq 2^{-l/2}$. Figure 5.9 displays the scattering results for Fractional Gaussian Noises $dX_H(t)$ for $H = 0.2, 0.4, 0.6, 0.8$, which confirm the predicted asymptotic behavior.

5.4.3 Lévy Processes

We consider in this section α-stable Lévy processes. Self-similar Lévy processes, described in section 5.2.7, have heavy tailed distributions, and its realizations contain rare, large, events which critically influence its moments. This phenomena appears frequently in nature in a variety of multifractal processes. In particular, an α-stable Lévy process only has moments strictly smaller than α. Figure 5.10 shows realizations of the increments of Lévy processes for $\alpha = 1.1\,, 1.3\,1.5$, revealing the presence of rare, large jumps.

If X_α is a α-stable Lévy process, its associated jump process dX_α is distributed according to the Lévy measure $\Pi(x)$, defined in (5.9), and satisfies $E(|dX_\alpha|) < \infty$. It results from the first order bound (5.17) that its second order scattering representation



(a) (b)

Figure 5.8: (a) plot of $\log \overline{S}X_H(j)$ estimated from 100 realizations of $N = 2^{16}$ points. (b) plot of $\log \overline{T}X_H(l)$, estimated from the same number of realizations. We verify that $\overline{S}X_H(j) \sim 2^{jH}$ and $\overline{T}X_H(l) \sim 2^{-l/2}$.

is well defined. We estimate its first order scattering and its scattering transfer using the estimation described in Section 5.3.3. Figure 5.11 shows the log-plot of first order scattering coefficients and the transfer function. For each value of α, we plot the estimated transfer $\overline{T}dX_\alpha(l)$ together with the confidence interval $\pm\sigma$ for each l.

Since by definition X_α and dX_α are self-similar with

$$\forall t \;\; X_\alpha(st) \stackrel{l}{=} s^{1/\alpha} X(t) \;, \quad dX_\alpha(st) \stackrel{l}{=} s^{1/\alpha - 1} dX_\alpha(t) \;,$$

the first order scattering coefficients of dX_α recover the scaling law of the first moments of its increments:

$$\lim_{j \to \infty} 2^{-j(\alpha^{-1}-1)} \overline{S} dX_\alpha(j) = \lim_{j \to \infty} 2^{-j(\alpha^{-1}-1)} E(|dX_\alpha \star \psi_j|) = C \;.$$

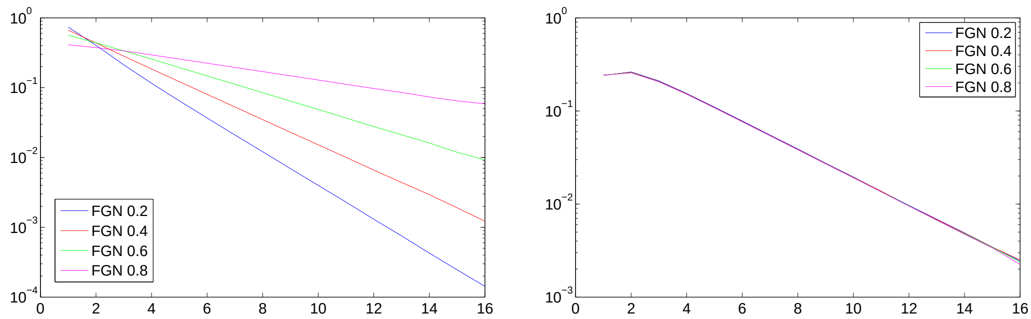
Figure 5.11-(a) confirms this scaling law.

The analysis of the scattering transfer, however, reveals that $\overline{T}dX_\alpha(l)$ behaves very differently than in the Brownian case. Figure 5.11 and Table 5.2 show that $\overline{T}dX_\alpha(l)$ has approximately the same scaling law as $\overline{S}dX_\alpha(j)$. These results can be interpreted as follows. Lévy processes with $\alpha < 2$ contain rare, large jumps, as seen in the examples of figure 5.10, and whose distribution is dictated by the Lévy measure $\Pi(x)$. These sparse, large events are transmitted to the processes $U[j]dX_\alpha(t) = |dX_\alpha \star \psi_j|(t)$, in the sense that $U[j]dX_\alpha$ also contains rare, large events which dominate its moments, for each scale j. The sparsity of $U[j]dX_\alpha$ means that the interaction between jumps is not statistically significant, and thus that we can approximate each $U[j]dX_\alpha$ by

$$\forall t \;\; |dX_\alpha \star \psi|(t) \stackrel{l}{\approx} |dX_\alpha \star |\psi||(t) \;,$$

which smoothes the jumps with the envelope $|\psi|$. As a result, it follows that

$$||dX_\alpha \star \psi_{j_1}| \star \psi_{j_2}| \stackrel{l}{\approx} |dX_\alpha \star (|\psi_{j_1}| \star \psi_{j_2})| \stackrel{l}{\approx} C_{j_1} |dX_\alpha \star \psi_{j_2}| \;,\; j_2 \gg j_1 \;, \qquad (5.54)$$



(a) (b)

Figure 5.9: (a) plot of $\log \overline{S} dX_H(j)$ estimated from 100 realizations of $N = 2^{16}$ points. (b) plot of $\log \overline{T} dX_H(l)$, estimated from the same number of realizations. We verify that $\overline{S} dX_H(j) \sim 2^{jH}$ and $\overline{T} dX(l) \sim 2^{-l/2}$.

Table 5.2: Estimated scaling laws from the first order scattering coefficients and the scattering transfer for different α-stable Lévy processes

α	$H = \alpha^{-1} - 1$	H_1 from $S_J dX_\alpha(j) \simeq 2^{jH_1}$	H_2 from $T_J dX_\alpha(l) \simeq 2^{lH_2}$
1.1	-0.1	-0.14	-0.13
1.3	-0.23	-0.21	-0.2
1.5	-0.33	-0.34	-0.33

which implies that

$$\overline{T} dX_\alpha(l) = \frac{E(||dX_\alpha \star \psi| \star \psi_l|)}{E(|dX_\alpha \star \psi|)} \approx \frac{C_0 E(|dX_\alpha \star \psi_l|)}{E(|dX_\alpha \star \psi|)} = C \overline{S} dX_\alpha(l) \ , \ l \gg 1 \ .$$

Thus, contrarily to the Brownian case, where the filtering $X_H \star \psi_j$ removes the long range dependencies of the process which determine its characteristic exponent, the wavelet coefficients of Lévy processes are still influenced by the large excursions in amplitude characterizing the jumps. The conjectured behavior (5.54) predicts that the amplitude of wavelet coefficients varies as a function of the scale following the same scaling law as the original jump process.

5.5 Scattering of Multifractal Processes

Multifractal processes are characterized by a concave scaling exponents function $\zeta(q)$, which implies the existence of an integral scale. By introducing an asymptotic scattering energy propagation condition, we give a measure of the curvature of $\zeta(q)$ from the scattering transfer function.

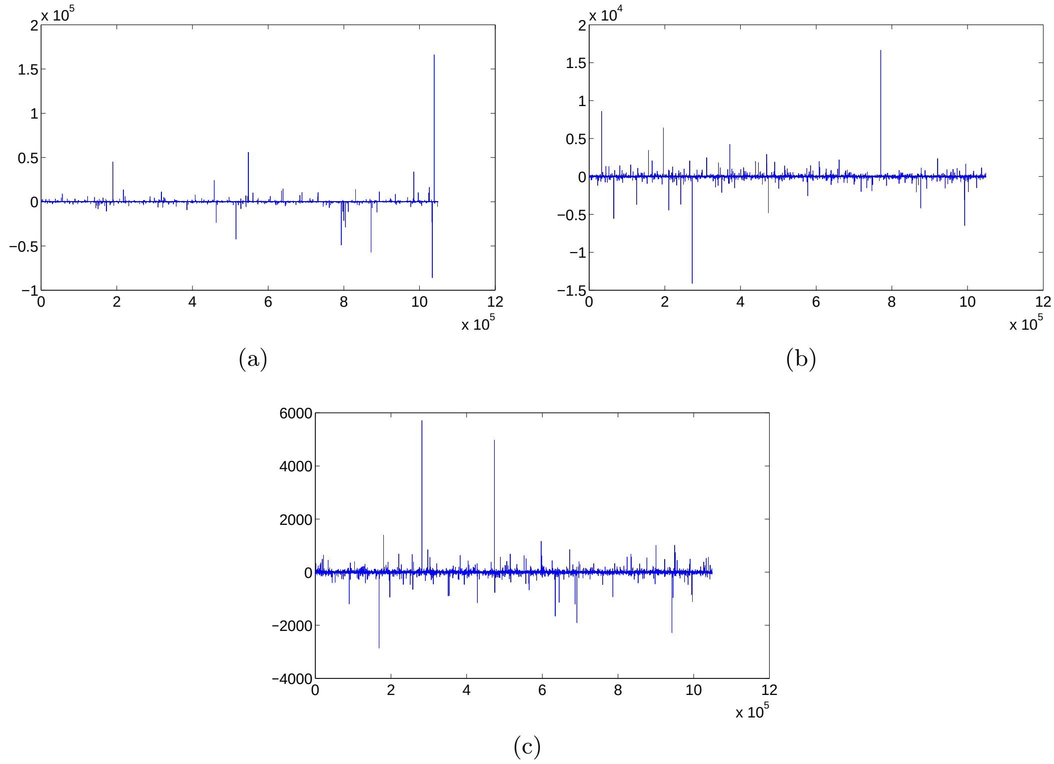


Figure 5.10: Realizations of α-stable Lévy process. (a) $\alpha = 1.1$, (b) $\alpha = 1.3$, (c) $\alpha = 1.5$. Notice that the intensity of the jumps is reduced by an order of magnitude from one panel to the next.

5.5.1 Multifractal Scattering transfer

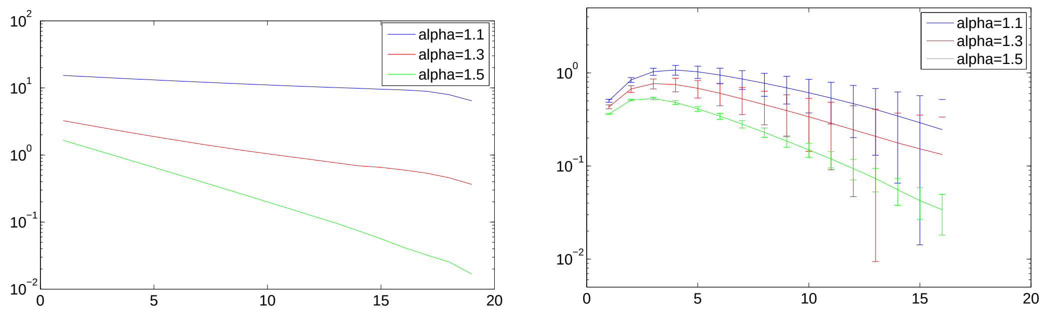
A transfer function is defined for multifractal processes having an integral scale, thanks to their stochastic self-similarity.

As explained in Section 5.2.4, a multifractal process $X(t)$ is characterized by a stochastic self-similarity defined in (5.2), which yields a non-linear scaling exponent $\zeta(q)$. The curvature of $\zeta(q)$ thus measures the degree of multifractality of the process, also referred as intermittency. We will now develop a tool which measures the curvature of $\zeta(q)$ from first and second order scattering coefficients. In particular, we will indirectly measure $\zeta(q)$ at $q = 1$ and $q = 2$, which yields a measure of the curvature with $\zeta(2) - 2\zeta(1)$.

The intermittency $\zeta(2) - 2\zeta(1)$ is usually difficult to estimate for a fractal process, since the estimation of

$$\frac{E(|X \star \psi_j|^2)}{E(|X \star \psi_j|)^2} \simeq 2^{j(\zeta(2)-2\zeta(1))}$$

requires to evaluate a second order moment $E(|X \star \psi_j|^2)$ at large scales 2^j and the estimation of such a moment has a high variance.



(a) (b)

Figure 5.11: (a) plot of $\log \overline{S}dX_\alpha(j)$ for α-stable Lévy processes, estimated from 64 realizations of $N = 2^{20}$ points. (b) plot of $\log \overline{T}dX_\alpha(l)$, estimated from the same number of realizations.

Let us start by defining a scattering transfer function for processes with stochastic self-similarity, up to an integral scale $T = 2^J$. We consider multifractal processes which decorrelate beyond this integral scale, an assumption justified in physical applications such as turbulence.

The stochastic self-similarity of definition 5.2.2 induces a decomposition of the multifractal $X(t)$ as a random cascade [BKM08a]:

$$X(t) = \prod_{j=-\infty}^{J} Y_j(t) , \tag{5.55}$$

where $Y_j(t)$ are independent processes with $E(Y_j) = 1$, and whose spectral density is concentrated in the interval $(-2^{-j}, 2^{-j})$. Let us show that this decomposition defines a scattering transfer having the same invariance as self-similar processes on a subset of scattering paths determined by the integral scale J. Section 4.7 showed that the wavelet coefficients of a process of the form $Y_s Y_f$, where Y_s is smooth with respect to Y_f, satisfies

$$(Y_s Y_f) \star \psi_j(t) \approx Y_s (Y_f \star \psi_j)(t) ,$$

with a mean squared error bound which depends upon the amount of energy of Y_s in the frequencies where $|\hat{\psi}_j|$ is non-negligible. Since each term Y_j in (5.55) has its energy concentrated on frequencies in the dyadic interval $(-2^j, 2^j)$, it results that

$$X \star \psi_{j_1} \approx \prod_{j_1 < j \leq J} Y_j(t) \left(\prod_{j=-\infty}^{j_1} Y_j \star \psi_{j_1} \right) ,$$

and hence

$$|X \star \psi_{j_1}| \approx \prod_{j_1 < j \leq J} Y_j(t) \left| \prod_{j=-\infty}^{j_1} Y_j \star \psi_{j_1} \right| .$$

It follows that if $j_2 \leq J$,

$$||X \star \psi_{j_1}| \star \psi_{j_2}| \approx \prod_{j_2 < j \leq J} Y_j \left| \prod_{j_1 < j \leq j_2} Y_j \right| \prod_{j \leq j_1} Y_j \star \psi_{j_1} \left| \star \psi_{j_2} \right| ,$$

and, since $E(Y_j) = 1$ for all j,

$$TX(j_1, j_2) \approx \frac{E\left(\left|\prod_{j_1 < j \leq j_2} Y_j \right| \prod_{j \leq j_1} Y_j \star \psi_{j_1} \left| \star \psi_{j_2} \right|\right)}{E\left(\left|\prod_{j=-\infty}^{j_1} Y_j \star \psi_{j_1}\right|\right)} . \tag{5.56}$$

The scattering transfer coefficient $TX(j_1, j_2)$ can thus be approximated by $T\tilde{X}_{j_2}(j_1, j_2)$, where $\tilde{X}_{j_2}(t) = \prod_{j \leq j_2} Y_j(t)$ is the truncated cascade. But $\tilde{X}_{j_2}$ satisfies the stochastic self-similarity relation

$$\{D_s \tilde{X}_{j_2}(t)\}_t \stackrel{l}{=} \{W_s \tilde{X}_{j_2+s}(t)\}_t ,$$

where W_s is independent from $\tilde{X}_j$. It results that if j_1', j_2' satisfy $j_1' < j_2' \leq J$ and $j_2 - j_1 = j_2' - j_1'$, then

$$\begin{aligned} TX(j_1', j_2') &\approx T\tilde{X}_{j_2'}(j_1', j_2') \\ &= \frac{E(||\tilde{X}_{j_2'} \star \psi_{j_1'}| \star \psi_{j_2'}|)}{E(|\tilde{X} \star \psi_{j_1'}|)} = \frac{E(W_{j_1'-j_1}) E(||\tilde{X}_{j_2} \star \psi_{j_1}| \star \psi_{j_2}|)}{E(W_{j_1'-j_1}) E(|\tilde{X}_{j_2} \star \psi_{j_1}|)} \\ &= T\tilde{X}_{j_2}(j_1, j_2) \approx TX(j_1, j_2) , \end{aligned} \tag{5.57}$$

which shows that

$$TX(j_1, j_2) \approx \overline{T}X(j_2 - j_1) ,$$

for $j_1, j_2 < J$. For $j_2 > J$, the decorrelation at large scales induces a transfer function which converges to that of the Gaussian white noise, with an asymptotic behavior $C2^{-l/2}$ as seen in Section 5.4.1. The resulting scattering transfer is

$$TX(j_1, j_2) \approx \begin{cases} \overline{T}X(j_2 - j_1) & \text{if } j_1 < J \text{ and } j_2 < J \\ C\, 2^{(J-j_2)/2} & \text{if } j_1 < J \text{ and } j_2 \geq J \\ C\, 2^{(j_1-j_2)/2} & \text{if } J \leq j_1 < j_2 \end{cases} \tag{5.58}$$

The transfer function $\overline{T}X(l)$ of X is thus defined for all $l \geq 0$. Figure 5.12 shows the estimated scattering transfer for the random MRM with $\lambda^2 = 0.04$, with an integral scale $2^J \approx 2^{17}$. As predicted, paths (j_1, j_2) with $j_1 < j_2 < J$ satisfy the invariance induced by the self-similarity, whereas as soon as $j_2 \geq J$, the scattering transfer decays as in the white gaussian noise.

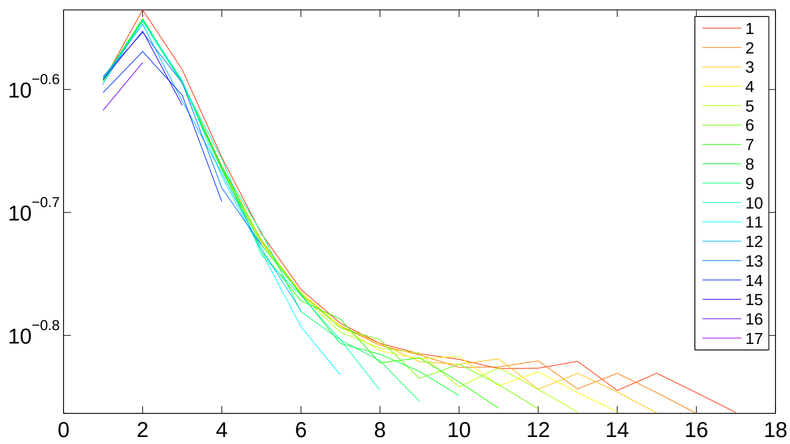


Figure 5.12: Estimation of the scattering transfer for an MRM random cascade with $\lambda^2 = 0.04$. We plot the curves $T_j(l) = TdX(j, j+l)$ as a function of l, and different colors stand for different values of j.

5.5.2 Energy Markov Property

We introduce an energy Markov property, which yields a measure based on the scattering transfer function with a scaling exponent given by $\zeta(2)$.

If $X(t)$ is a multifractal process with integral scale J, then its second moments satisfy

$$\forall j < J \ , \ E(|X \star \psi_j|^2) \simeq 2^{(j-J)\zeta(2)} \ .$$

Our objective is to define a measure based on first and second order scattering coefficients having a power law behavior given by $\zeta(2)$. Although the scattering representation is computed from first moments of $X(t)$ and its wavelet modulus decompositions, the conservation of energy described in Section 2.3.5 allows us to compute the second moments $E(|X \star \psi_j|^2)$ from its scattering representation. In particular, proposition 2.6.1 showed that if X is a process with stationary increments, such that $S_J X$ is mean square consistent, then

$$E(|X \star \psi_j|^2) = \sum_{p \in \overline{\mathcal{P}}_\infty} |\overline{S}X(j+p)|^2 \ . \tag{5.59}$$

The energy of the process $X \star \psi_j$ can thus be recovered by summing all the scattering coefficients starting by j. Section 2.6.5 also showed that most of the scattering energy is concentrated within the set of progressive paths $p \in \mathcal{P}_\downarrow$ satisfying $p = (j_1 \dots j_m)$ with $j_i > j_{i+1}$.

The scattering transfer can be used to predict high order scattering coefficients, using the markov scattering propagation defined in (5.40):

$$\overline{S}^M X(j_1, ..., j_m) = \overline{S}X(j_1) \prod_{k=2}^{m} TX(j_{k-1}, j_k) \ .$$

We can thus consider the quantity

$$\sum_{p \in \mathcal{P}_\downarrow} |\overline{S}^M X(j+p)|^2 \ , \tag{5.60}$$

which depends only upon first and second order scattering coefficients. Moreover, since X is decorrelated beyond an integral scale 2^J, most of the energy in (5.60) is captured by paths whose maximum scale is smaller than 2^J:

$$\sum_{p \in \mathcal{P}_\downarrow^J} |\overline{S}^M X(j+p)|^2 \ , \tag{5.61}$$

where $\mathcal{P}_\downarrow^J$ is the set of progressive paths with maximum scale given by 2^J.

We can then characterize the processes X for which this scattering measure gives information about $\zeta(2)$:

Definition 5.5.1 *A multifractal process X with stationary increments and with integral scale 2^J has the energy Markov property if there exist constants C_-, $C_+ > 0$ such that*

$$\forall j \leq J \ , \ C_- E(|X \star \psi_j|^2) \leq \sum_{p \in \mathcal{P}_\downarrow^J} |\overline{S}^M X(j+p)|^2 \leq C_+ E(|X \star \psi_j|^2) \ . \tag{5.62}$$

This property thus asks that, for all j, the energy captured by all scattering coefficients to be comparable to the energy captured by progressive, Markov scattering coefficients. It is a much weaker condition than the asymptotic Markov scattering, since we only demand an overall approximation of the energy instead of approximating each scattering coefficient. Section 5.5.4 shows numerical simulations showing that all tested multifractal processes satisfy the energy Markov property.

If a process $X(t)$ has the energy Markov property, then the following proposition shows that $2\zeta(1) - \zeta(2)$ can be recovered from the scattering transfer.

Proposition 5.5.2 *Let $X(t)$ be a process with stationary increments, with an integral scale 2^J, scaling exponents $\zeta(q)$, and satisfying the energy Markov property (5.62). Then there exist constants C_-, $C_+ > 0$ such that*

$$C_- \leq \lim_{j \to -\infty} 2^{(J-j)(\zeta(2)-2\zeta(1))} \sum_{\substack{p \in \mathcal{P}_\downarrow^J \\ p(1)=j}} \prod_{k=2}^{|p|} \overline{T}X(j_k - j_{k-1})^2 \leq C_+ \ , \tag{5.63}$$

and hence

$$\lim_{j \to -\infty} (J-j)^{-1} \log_2 \left(\sum_{\substack{p \in \mathcal{P}_\downarrow^J \\ p(1)=j}} \prod_{k=2}^{|p|} \overline{T}X(j_k - j_{k-1})^2 \right) = 2\zeta(1) - \zeta(2) \ . \tag{5.64}$$

Proof: By definition,

$$\overline{S}^M X(j_1 \dots j_m) = \overline{S} X(j_1) \prod_{k=2}^{m} TX(j_k, j_{k-1}) \; .$$

If $p \in \mathcal{P}_{\downarrow}^J$, then (5.58) shows that $TX(j_k, j_{k-1}) = \overline{T}X(j_k - j_{k-1})$. Then, property (5.62) implies that there exist $C_{1,-}$, $C_{1,+}$ such that

$$C_{1,-} \leq \lim_{j \to -\infty} \left(E(|X \star \psi_j|^2)\right)^{-1} \overline{S}X(j)^2 \sum_{\substack{p \in \mathcal{P}_{\downarrow}^J \\ p(1)=j}} \prod_{k=2}^{|p|} \overline{T}X(j_k - j_{k-1})^2 \leq C_{1,+} \; ,$$

which in turn implies that there exist $C_{2,-}$, $C_{2,+}$ satisfying

$$C_{2,-} \leq \lim_{j \to -\infty} 2^{(J-j)\zeta(2)} \overline{S}X(j)^2 \sum_{\substack{p \in \mathcal{P}_{\downarrow}^J \\ p(1)=j}} \prod_{k=2}^{|p|} \overline{T}X(j_k - j_{k-1})^2 \leq C_{2,+} \; .$$

Since $E(|X \star \psi_j|) = \overline{S}X(j)$ and there exists constants $C_{3,-}$, $C_{3,+}$ with

$$C_{3,-} \leq \lim_{j \to -\infty} 2^{(J-j)\zeta(1)} E(|X \star \psi_j|) \leq C_{3,+} \; ,$$

it follows that we can find constants $C_{4,-}$, $C_{4,+}$ so that

$$C_{4,-} \leq \lim_{j \to -\infty} 2^{(J-j)(\zeta(2)-2\zeta(1))} \sum_{\substack{p \in \mathcal{P}_{\downarrow}^J \\ p(1)=j}} \prod_{k=2}^{|p|} \overline{T}X(j_k - j_{k-1})^2 \leq C_{4,+} \; ,$$

which proves (5.63). As a result, by taking the limit of the logarithm as $j \to -\infty$, we obtain (5.64) $\square$.

This proposition shows that the scattering transfer is directly associated with the curvature of the scaling exponents $\zeta(q)$. Under the energy Markov property, the ratio between first and second moments, $\frac{E(|X \star \psi_j|^2)}{E(|X \star \psi_j|)^2}$, is expressed in terms of ratios between first and second order scattering coefficients, which are both computed with non-expansive operators applied to the multifractal.

The Markov energy property is numerically verified for several multifractal processes. Figure 5.13 compares the estimated second moments $E(|dX \star \psi_j|^2)$ with the prediction using the scattering transfer over progressive paths, given by (5.61). We confirm empirically that they yield the same scaling law, and hence that they verify the Markov energy property.

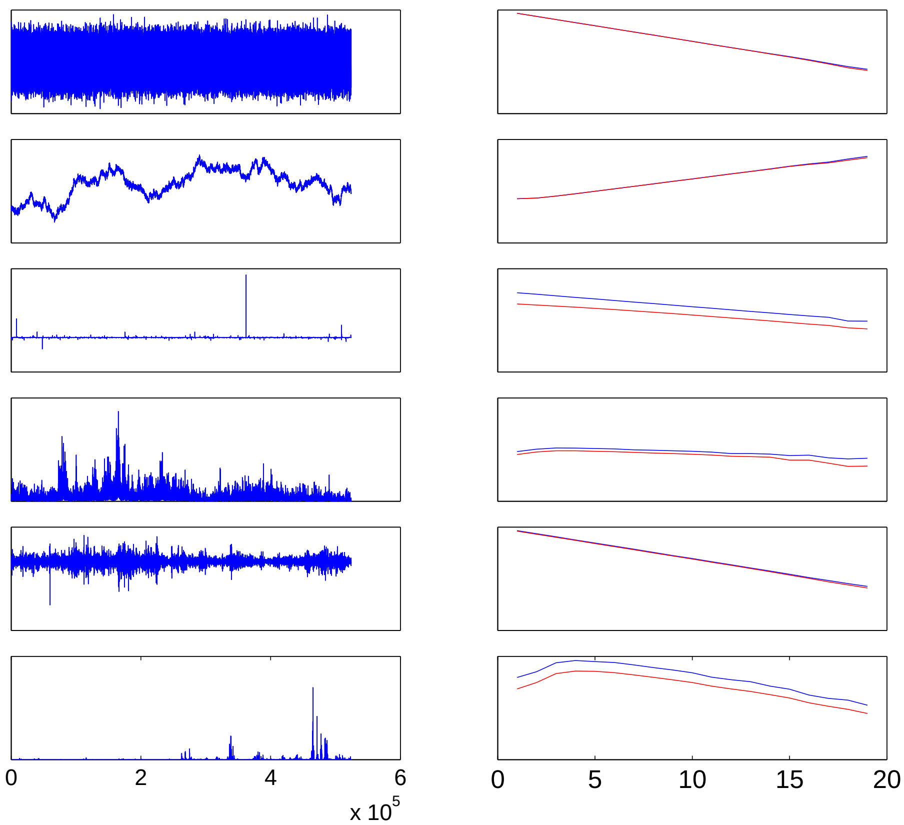


Figure 5.13: Numerical verification of Markov energy property for different multifractal families. From top to bottom: White gaussian noise, Fractional Brownian motion with $H = 0.4$, α-stable Lévy process with $\alpha = 1.3$, MRM cascade with $\lambda^2 = 0.04$, MRW cascade with $\lambda^2 = 0.04$, Mandelbrot cascade. In blue we display the estimated second moments $E(|dX \star \psi_j|^2)$ as a function of j. In red, the predicted second moments using the Markov energy model from (5.61), also as a function of j.

5.5.3 Intermittency characterization from Scattering transfer

This section proves that if the energy Markov property is satisfied, we can obtain $\zeta(2)-\zeta(1)$ from the smallest zero of a series defined by the scattering transfer.

We saw in (5.64) that thanks to the energy Markov property, we can obtain $2\zeta(1)-\zeta(2)$ as the limit of a quantity defined entirely from the scattering transfer. We shall see in this section that one can make the dependency explicit as the root of a series defined from $\overline{T}$.

Let us consider the following power series defined from the scattering transfer of X:

$$\mathcal{T}_X(z)=\sum_{l\geq 1}\overline{T}X(l)^2z^l\ , z\in\mathbb{C}\ . \tag{5.65}$$

The following theorem proves that the limit defining $2\zeta(1)-\zeta(2)$ is equal to the logarithm of the smallest real $x<1$ satisfying

$$\mathcal{T}_X(x)=1\ , \tag{5.66}$$

Theorem 5.5.3 *Let X be a self-similar process with an integral scale 2^J, and such that X satisfies the energy Markov property (5.5.1). Let $\|\overline{T}X\|^2=\sum_{l\geq 1}\overline{T}X(l)^2$, and $\mathcal{T}_X(z)=\sum_{l\geq1}\overline{T}X(l)^2z^l$. Then*

1. *$2\zeta(1)-\zeta(2)>0$ if and only if $\|\overline{T}X\|>1$, and*
2. *If ρ is the smallest $x\in\mathbb{R}^+$ satisfying $\mathcal{T}_X(x)=1$, then*

$$2\zeta(1)-\zeta(2)=\max(0,-\log_2(\rho))\ . \tag{5.67}$$

Proof: Let us first turn the limit $j\to-\infty$ in (5.64) into a limit where the integral scale $J\to\infty$. If X has integral scale J, then $D_{-j}X(t)=X(2^jt)$ has integral scale $J-j$. The ratio

$$\frac{E(|X\star\psi_j|^2)}{E(|X\star\psi_j|)^2}$$

can be rewritten as

$$\frac{E(|D_{-j}X\star\psi|^2)}{E(|D_{-j}X\star\psi|)^2}$$

where the integral scale of $D_{-j}X$ increases to $J-j$ as $j\to-\infty$.

The transfer function of $D_{-j}X$ only depends upon j through its domain of definition, which controls when it reaches its integral scale $J-j$. We define

$$\begin{aligned}F(J-j) &= (\overline{S}^MD_{-j}X(0))^{-2}\sum_{p\in\mathcal{P}_{J-j}}\overline{S}^MD_{-j}X(0+p)^2\\ &= \sum_{p\in\mathcal{P}_{J-j}}\prod_{j_1=0<j_2<..j_m\leq J-j}\overline{T}(j_k-j_{k-1})^2\ .\end{aligned}$$

and we write $\tilde{J} = J - j$. As a result, the energy Markov property (5.64) is equivalent to

$$\lim_{\tilde{J}\to\infty} \frac{\log_2 F(\tilde{J})}{\tilde{J}} = 2\zeta(1) - \zeta(2) \ . \tag{5.68}$$

We shall now compute the limit in terms of $F(\tilde{J})$.

Let us first prove that $\|\overline{T}X\| \leq 1$ implies $2\zeta(1) - \zeta(2) = 0$. We decompose $F(\tilde{J})$ in terms of the path orders:

$$F(\tilde{J}) = 1 + \sum_{m=2}^{\tilde{J}} F(\tilde{J})_m \ ,$$

where

$$F(\tilde{J})_m = \sum_{p\in\mathcal{P}_{\tilde{J}}, |p|=m} \prod_{p=(j_1\ldots j_m)} \overline{T}(j_k - j_{k-1})^2 \ .$$

The quantity $F(\tilde{J})_m$ is bounded by indexing the set of paths $B_{\tilde{J},m} = \mathcal{P}_{\tilde{J}} \cap \{|p| = m\}$ from $B_{\tilde{J},m-1}$ as

$$\begin{aligned}
F(\tilde{J})_m &= \sum_{p\in\mathcal{P}_{\tilde{J}}, |p|=m} \prod_{p=(j_1,..j_m)} \overline{T}(j_k - j_{k-1})^2 \\
&= \sum_{p\in\mathcal{P}_{\tilde{J}}, |p|=m-1} \sum_{\tilde{J}\geq j'>j_{m-1}} \overline{T}(j' - j_{m-1})^2 \prod_{p=(j_1,..j_{m-1})} \overline{T}(j_k - j_{k-1})^2 \\
&= \sum_{p\in\mathcal{P}_{\tilde{J}}, |p|=m-1} \prod_{p=(j_1,..j_{m-1})} \overline{T}(j_k - j_{k-1})^2 \left(\sum_{\tilde{J}\geq j'>j_{m-1}} \overline{T}(j' - j_{m-1})^2 \right) \\
&\leq F(\tilde{J})_{m-1} \|\overline{T}\|^2 \ .
\end{aligned}$$

As a result,

$$F(\tilde{J}) \leq 1 + \sum_{m=2}^{\tilde{J}} \|\overline{T}\|^{2m} \ . \tag{5.69}$$

This shows that $F(\tilde{J})$ grows at most as $C\|\overline{T}\|^{2\tilde{J}}$, which means that

$$\lim_{\tilde{J}\to\infty} \frac{\log_2 F(\tilde{J})}{\tilde{J}} \leq \log_2 \|\overline{T}\|^2 \ . \tag{5.70}$$

Since $\zeta(q)$ is concave, necessarily $2\zeta(1) - \zeta(2) \geq 0$, and we deduce from (5.70) that if $\|\overline{T}X\| \leq 1$ then $2\zeta(1) - \zeta(2) = 0$.

Let us now prove (5.67). For this purpose, we use an alternative decomposition of the set $\mathcal{P}_{\tilde{J}}$. Let

$$\mathcal{Q}_{\tilde{J}} = \{p = (j_1 \ldots j_p) \in \mathcal{P}_{\tilde{J}} \, ; \, j_p = \tilde{J}\}$$

denote the subset of paths in $\mathcal{P}_{\tilde{J}}$ which end by $\tilde{J}$, and

$$(S, j_0) = \{p \in \mathcal{P}_{\tilde{J}} \, : p = (p', j_0) \, : p' \in S\}$$

denote the set which extends S with a fixed scale j_0. For each $\Delta > 0$, we have the following disjoint decompositions:

$$\begin{aligned}\mathcal{P}_{\tilde{J}} &= \mathcal{P}_{\tilde{J}-1} \cup \mathcal{Q}_{\tilde{J}} ,\\ \mathcal{Q}_{\tilde{J}} &= (\mathcal{P}_{\tilde{J}-\Delta-1}, \tilde{J}) \cup (\mathcal{Q}_{\tilde{J}-\Delta}, \tilde{J}) \cup \ldots \cup (\mathcal{Q}_{\tilde{J}-1}, \tilde{J}) .\end{aligned} \tag{5.71}$$

Let $E(S) = \sum_{p\in S} \prod_{j_1\ldots j_k} \overline{T}(j_k - j_{k-1})^2$. The normalized energy corresponding to the set $(\mathcal{Q}_{\tilde{J}-l}, \tilde{J})$ is computed as

$$\begin{aligned}E((\mathcal{Q}_{\tilde{J}-l}, \tilde{J})) &= \sum_{p\in(\mathcal{Q}_{\tilde{J}-l},\tilde{J})} \prod_{j_1\ldots j_k} \overline{T}(j_k - j_{k-1})^2 \\ &= \sum_{p\in\mathcal{Q}_{\tilde{J}-l}} \prod_{j_1\ldots j_k} \overline{T}(j_k - j_{k-1})^2 \overline{T}(l)^2 \\ &= \overline{T}(l)^2 E(\mathcal{Q}_{\tilde{J}-l}) .\end{aligned}$$

If $Q(j) = E(\mathcal{Q}_j)$, then the relations (5.71) yield

$$\begin{aligned}F(\tilde{J}) &= E(\mathcal{P}_{\tilde{J}}) = E(\mathcal{P}_{\tilde{J}}) + E(\mathcal{Q}_{tj)} = F(\tilde{J}-1) + Q(\tilde{J}) ,\\ Q(\tilde{J}) &= \sum_{l=1}^{\infty} \overline{T}(l)^2 Q(\tilde{J}-l) .\end{aligned} \tag{5.72}$$

By substituting $Q(\tilde{J}) = F(\tilde{J}) - F(\tilde{J}-1)$ in (5.72) we obtain the following linear recursion on $F(\tilde{J})$:

$$F(\tilde{J}) = b_0 F(\tilde{J}-1) + b_1 F(\tilde{J}-2) + \ldots + b_\Delta F(\tilde{J}-\Delta-1) + \ldots , \tag{5.73}$$

where the coefficients are given by

$$\begin{cases} b_0 &= 1 + \overline{T}(1)^2 ,\\ b_l &= \overline{T}(l+1)^2 - \overline{T}(l)^2 ,\ l \geq 1 .\end{cases} \tag{5.74}$$

Since $F(\tilde{J}) = 0\,,\ \tilde{J} < 0$ and $b_l = 0\,,\ l < 0$, we define their causal Z-transforms as

$$\begin{aligned}\mathcal{F}(z) &= \sum_{n\geq 0} F(n) z^{-n} ,\\ \mathcal{B}(z) &= \sum_{n\geq 0} b_n z^{-n} .\end{aligned} \tag{5.75}$$

The linear recurrence (5.73) becomes

$$F(n+1) = \sum_{l\geq 0} b_l F(n-l) = (F \star b)(n) \quad , n \geq 0 . \tag{5.76}$$

Since $F(0) = 1$ and the Z-transform of a convolution $F \star b$ is $\mathcal{F}(z) \cdot \mathcal{B}(z)$, the linear relation (5.76) is expressed in the transformed domain as

$$
\begin{aligned}
z + \mathcal{F}(z)\mathcal{B}(z) &= zF(0) + \sum_{n \geq 0} \left(\sum_{n' \geq 0} F(n') b_{n-n'} \right) z^{-n} \\
&= zF(0) + \sum_{n \geq 0} F(n+1) z^{-n} \\
&= zF(0) + z \sum_{n > 0} F(n) z^{-n} \\
&= z\mathcal{F}(z) \ . \qquad (5.77)
\end{aligned}
$$

As a result, we have

$$\mathcal{F}(z)(1 - z^{-1}\mathcal{B}(z)) = \mathcal{F}(z) P_T(z) = 1 \ , \qquad (5.78)$$

where

$$P_T(z) = 1 - z^{-1} \sum_{l \geq 0} b_l z^{-l} \ . \qquad (5.79)$$

If z_0 is a zero of $P_T(z)$, then necessarily it must also be a pole of $\mathcal{F}(z)$.

On the other hand, the radius of convergence of the complex series $\mathcal{F}(z^{-1})$ is given by

$$R = \limsup_{\tilde{J} \to \infty} F(\tilde{J})^{-1/\tilde{J}} = \lim_{\tilde{J} \to \infty} F(\tilde{J})^{-1/\tilde{J}} = 2^{2\zeta(2) - 2\zeta(1)} \ ,$$

thanks to the Cauchy-Hadamard theorem. But the radius of convergence of $\mathcal{F}(z^{-1})$ is determined by its pole of smallest magnitude, which coincides with the zero of largest magnitude of $P_T(z)$ thanks to (5.78). $2\zeta(1) - \zeta(2)$ is thus characterized from the zero of largest magnitude of $P_T(z)$. Let us now factorize $P_T(z)$.

We define

$$T_2(l) = \begin{cases} 0 & \text{if } l < 0 \ , \\ -1 & \text{if } l = 0 \ , \\ \overline{T} X(l)^2 & \text{if } l > 0 \ . \end{cases} \qquad (5.80)$$

Then, if we write $P_T(z) = \sum_{l \geq 0} \tilde{b}_l z^{-l}$, by substituting from (5.74) we have that

$$\forall l \geq 0 \, , \ \tilde{b}_l = T_2(l-1) - T_2(l) \ ,$$

and hence

$$P_T(z) = (z^{-1} - 1)(\sum_{l \geq 0} T_2(l) z^{-l}) = (z^{-1} - 1)(\mathcal{T}_X(z^{-1}) - 1) \ .$$

The zeros of P_T of magnitude greater than 1 are thus contained in the zeros of $\mathcal{T}_X - 1$ inside the unit circle, and $2\zeta(1) - \zeta(2)$ is obtained from the zero of smallest magnitude of $Q_T(z) = \mathcal{T}_X(z) - 1$.

Let us finally show that $\|\overline{T}\| > 1$ necessarily gives a real root with those characteristics. Since $Q_T(0) = -1$ and $Q_T(1) = \|\overline{T}\| - 1 > 0$, then necessarily we have a real root $x_0 \in (0,1)$ satisfying $Q_T(x_0) = 0$, which shows in particular that $2\zeta(1) - \zeta(2) > 0$. Let us consider the smallest real root x_0 satisfying $Q_T(x_0) = 0$. This root is necessarily the one with smallest magnitude amongst the roots in the unit circle. Indeed, suppose z_0 is a complex root with nonzero imaginary part satisfying $Q_T(z_0) = 0$. Then, since $\mathcal{T}_X(z_0) = 1$, we have $|\mathcal{T}_X(z_0)| = 1$ and

$$Q_T(|z_0|) = \mathcal{T}_X(|z_0|) - 1 > |\mathcal{T}_X(z_0)| - 1 = 0 \ .$$

We thus have $Q_T(0) < 0$ and $Q_T(|z_0|) > 0$, which implies that there exists a real root with magnitude smaller than z_0. This concludes the proof. □.

This theorem thus characterizes the multifractal behavior of a process from its scattering transfer function. A measure of the intermittence is obtained analytically by finding the solutions of the equation

$$\mathcal{T}_X(x) = 1 \ .$$

We observe that this characterization depends upon all the values of $\overline{T}(l)$ and not only upon its asymptotic behavior. In fact, as l grows, the influence of $\overline{T}X(l)$ on the zero of $\mathcal{T}_X(x) - 1$ diminishes, since its associated power x^l tends to 0. This shows that the transient values of $\overline{T}X$ for small values of $l \geq 0$ contain critical information about the multifractal behavior of X. We shall see in the estimation section that this property ensures that estimates of $2\zeta(1) - \zeta(2)$ are based on the estimated values of $\overline{T}X(l)$ with smallest variance, corresponding to small values of l.

As a corollary, we can bound the value of $2\zeta(1) - \zeta(2)$ by bounding the scattering transfer function:

Corollary 5.5.4 *Under the same hypothesis of theorem 5.5.3, then*

$$\log_2(1 + \min_l \overline{T}X(l)^2) \leq 2\zeta(1) - \zeta(2) \leq \min(\log_2(\|\overline{T}\|^2), \log_2(1 + \max_l \overline{T}X(l)^2)) \ . \quad (5.81)$$

Proof: From (5.70) we know that $2\zeta(1) - \zeta(2) \leq \log_2(\|T\|^2)$. Let us now obtain the rest of the bounds in (5.81). Let $K = \min \overline{T}X(l)^2$. Then, we can write

$$\overline{T}X(l)^2 = K + d(l) \ , \quad (5.82)$$

where $d(l) \geq 0\,, \forall l$. If $\rho = 2^{\zeta(2) - 2\zeta(1)}$, then from theorem (5.5.3) we know that $\sum_{l \geq 1} \overline{T}(l)^2 \rho^l = 1$. By using (5.82) we have

$$\sum_{l \geq 1} (K + d(l))\rho^l = K \sum_{l \geq 1} \rho^l + \sum_{l \geq 1} d(l)\rho^l = 1 \ ,$$

and since $d(l) \geq 0$ it results that

$$K \sum_{l \geq 1} \rho^l = K \frac{\rho}{1 - \rho} \leq 1 \ .$$

By substituting $\rho = 2^{\zeta(2)-2\zeta(1)}$ we obtain

$$(K+1)2^{\zeta(2)-2\zeta(1)} \leq 1 \ ,$$

which implies $\log_2(1+K) \leq 2\zeta(1) - \zeta(2)$.

If we now replace $K = \min \overline{T}X(l)^2$ by $\tilde{K} = \max \overline{T}X(l)^2$, then $d(l) \leq 0$, $\forall l$, and the same reasoning yields $\log_2(1+\tilde{K}) \geq 2\zeta(1) - \zeta(2)$, which concludes the proof. □.

5.5.4 Analysis of Scattering transfer for Multifractals

Multifractal Random Cascades [BKM08a] construct stationary multifractal random measures from a log-infinitely divisible measure. The scattering transfer $\overline{T}X(l)$ function converges towards a constant value as $l \to \infty$. We compute numerically the scattering transfer for multifractal random cascades and verify empirically the energy Markov property.

The asymptotic scattering transfer for random cascades has been studied in collaboration with E. Bacry and J.F. Muzy. We conjecture that if dX is a self-similar random cascade, then the scattering transfer converges towards a nonzero constant as $l \to \infty$.

Conjecture 5.5.5 *Let $dX(t)$ be a random log-infinitely divisible cascade with an integral scale J. Then there exist two constants K, $K' > 0$ such that*

1. *$\lim_{j \to -\infty} \overline{S}dX(j) = K$, and $|\overline{S}dX(j) - K|^2 \sim 2^{j-J}$.*
2. *Its scattering transfer $TdX(j_1, j_2)$ converges towards a scattering transfer function $\overline{T}dX$ when $J - j_1 \to \infty$,$J - j_2 \to \infty$, and*

$$\lim_{l \to \infty} \overline{T}dX(l) = K \ .$$

This conjecture is explained from the particular behavior of wavelet decompositions of a random cascade. A random cascade with integral scale 2^J can be written as $dX(t) = \lim_{j \to -\infty} e^{\omega_j^J(t)}$, where $\omega_j^J(t)$ is an infinitely divisible process. This process satisfies a self-similarity property

$$\omega_{j-l}^{J-l}(2^{-l}t) \stackrel{l}{=} \omega_j^J(t) \ ,$$

and, thanks to the infinite divisibility, it also satisfies the cascade property

$$\omega_j^J(t) = \omega_j^{\tilde{j}}(t) + \omega_{\tilde{j}}^J(t) \ , \quad , \forall j \leq \tilde{j} \leq J \ ,$$

with $\omega_j^{\tilde{j}}$ and $\omega_{\tilde{j}}^J$ independent. This property implies that $dX(t)$ can be written as a product

$$dX(t) = e^{\omega_{j_0}^J(t)} \lim_{j \to -\infty} e^{\omega_j^{j_0}(t)}$$

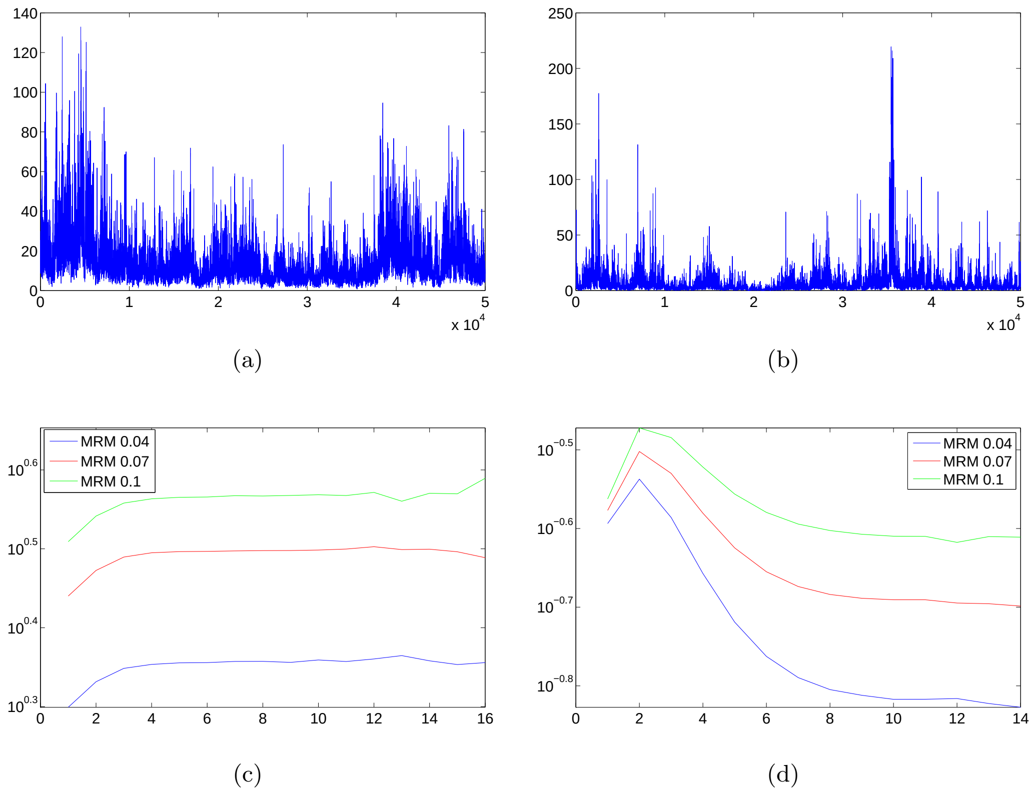


Figure 5.14: Scattering of MRMs with $\lambda^2 = 0.04,\ 0.07,\ 0.1$. (a) Example of a realization of an MRM with $\lambda^2 = 0.04$, (b) a realization with $\lambda^2 = 0.1$, (c) First order scattering $\log \overline{S}dX(j)$, (d) Scattering transfer $\log \overline{T}dX(l)$. The scattering is estimated from 64 realizations of 2^{20} points.

of two independent processes: a slow one which influences the long range correlations, and a fast one which dominates the increments of dX at small scales. Using this decomposition, one can approximate the filtered process $dX \star \psi_{j_0}(t)$ as the product

$$e^{\omega_{j_0}^{J}(t)} \left(\lim_{j \to -\infty} e^{\omega_j^{j_0}(t)} \star \psi_{j_0} \right) .$$

Second order scattering coefficients can thus be controlled by exploiting this decomposition.

Figures 5.14 and 5.15 display the estimated first order scattering and the scattering transfer for MRM and MRW cascades respectively, confirming the results predicted by conjecture 5.5.5. In the case of MRW, first order moments are controlled by the Wiener noise, which yields an asymptotic decay of $\overline{S}dX(j) \simeq 2^{-j/2}$. The influence of the noise composition disappears in the scattering transfer.

Figure 5.16 shows a realization of the binary Mandelbrot cascade, which is self-similar. We observe that the estimated scattering transfer $TX(j_1, j_1+l)$ does not depend

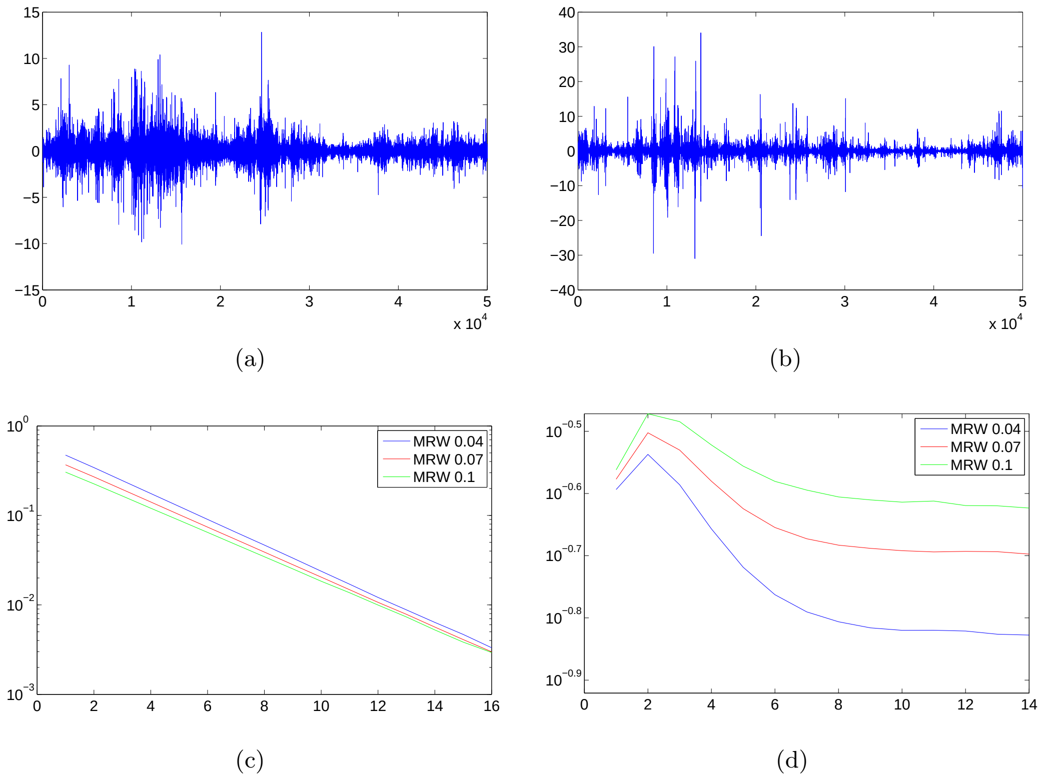
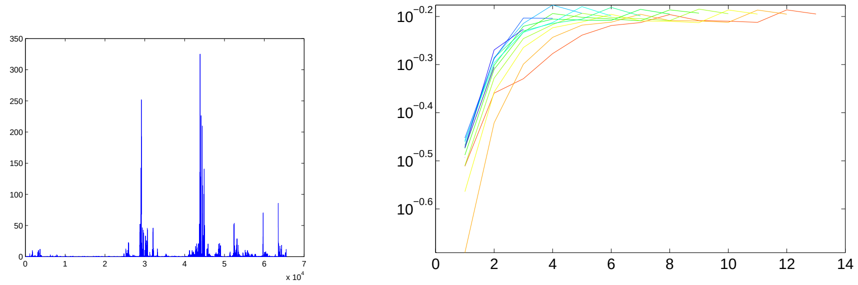


Figure 5.15: Scattering of MRWs with $\lambda^2 = 0.04,\ 0.07,\ 0.1$. (a) Example of a realization of an MRW with $\lambda^2 = 0.04$, (b) a realization with $\lambda^2 = 0.1$, (c) First order scattering $\log \overline{S}dX(j)$, (d) Scattering transfer $\log \overline{T}dX(l)$. The scattering is estimated from 64 realizations of 2^{20} points.



(a) (b)

Figure 5.16: (a) A realization of a Mandelbrot random measure, (b) Estimated Scattering transfer. We plot the curves $T_{j_1}X(l) = TX(j_1, j_1 + l)$ for several j_1. The curves define a transfer function $\overline{T}X(l)$, since the random cascade is self-similar.

upon j_1, thus defining a transfer function.

5.5.5 Intermittency Estimation for Multifractals

This section estimates $2\zeta(1) - \zeta(2)$ from the scattering transfer, using the result of theorem 5.5.3. Numerical experiments compare this estimate with a regression of the estimated moments and the covariance of the logarithm, on MRM and MRW multifractals.

Section 5.5.4 gave numerical evidence that MRM, MRW and the Mandelbrot cascade verify the Markov energy property. In such conditions, we can apply theorem 5.5.3 to estimate the intermittence $2\zeta(1) - \zeta(2)$ from an estimation of the scattering transfer.

Conjecture 5.5.5 predicts that as $l \to \infty$, $\overline{T}dX(l)$ converges towards a constant in the case of random cascades. As a result, it is only necessary to estimate the first terms of $\overline{T}dX(l)$, which are precisely those having the smallest variance.

The scattering transfer $\overline{T}dX$ is estimated with the windowed scattering transfer $\overline{T}_J$, using the procedure described in Section 5.3.3. Beyond a certain scale J_0, we regularize the estimation by assuming that the transfer function is a constant value. As a result, the estimated intermittence is obtained from the smallest solution of

$$\sum_{l\geq 1}^{J_0} \overline{T}X(l)^2 x^l + \overline{T}X(J_0)^2 \frac{x^{J_0+1}}{1-x} - 1 = 0 \ ,$$

which amounts to finding the smallest root of the polynomial

$$(1-x)\sum_{l\geq 1}^{J_0} \overline{T}X(l)^2 x^l + \overline{T}X(J_0)^2 x^{J_0+1} + x - 1 \ .$$

More powerful regularizations can be achieved by fitting the asymptotic decay of $\overline{T}(l)$ predicted in (5.5.5). Besides, the solutions $x \in (0,1)$ of $\mathcal{T}_X(x) - 1 = 0$ are stable with respect to the tail of $\overline{T}$. Indeed, if $A = \sup_l |\overline{T}X(l)^2 - \overline{T}X(J_0)^2|$, then

$$|\mathcal{T}_X(x) - \sum_{l=1}^{J_0} \overline{T}X(l)^2 x^l - \overline{T}X(J_0)^2 \frac{x^{J_0+1}}{1-x}| \leq A \frac{x^{J_0+1}}{1-x} , \quad (5.83)$$

which converges to 0 as $J_0 \to \infty$ for each $x \in (0,1)$.

The log-normal MRM and MRW are multifractal processes with $\zeta(q)$ given respectively by $\zeta(q) = \left(1 + \frac{\lambda^2}{2}\right) q - \frac{\lambda^2}{2} q^2$ and $\zeta(q) = \left(\frac{1}{2} + \lambda^2\right) q - \frac{\lambda^2}{2} q^2$. In the log-normal case, the scaling exponent is a parabole and hence $\zeta(2) - 2\zeta(1) = -\lambda^2$.

Table 5.3 reports the results of the intermittency estimation for MRM and MRW for several values of λ^2. We simulate cascades using $N = 2^{16}$ points. We estimate the expected scattering representation by averaging over 32 realizations, which is then used to estimate the intermittency. We repeat this experience over 8 runs in order to compute the standard deviation of the estimators. The estimate based on the scattering transfer is compared with the linear regression on the estimated first and second order moments

$$\frac{E(|dX \star \psi_j|^2)}{E(|dX \star \psi_j|)^2} ,$$

and also with the estimate from [BKM08b], resulting from

$$\mathrm{Cov}\left(\log |dX \star \psi_\tau|(t), \log |dX \star \psi_\tau|(t+l)\right) \sim -\lambda^2 \ln\left(\frac{l}{2^J}\right) + o\left(\frac{\tau}{l}\right) .$$

For the first method, the moments are estimated by averaging the empirical first and second order moments $|x \star \psi_j| \star \phi_J$, $|x \star \psi_j|^2 \star \phi_J$, whereas the second estimate is obtained by first estimating the covariance $\mathrm{Cov}\left(\log |dX \star \psi_\tau|(t), \log |dX \star \psi_\tau|(t+l)\right)$ for $\tau \leq J$ and $l \geq \tau$, and then performing a log regression to recover λ^2.

The intermittency estimate based on the scattering transfer outperforms the regression on the moments, and shows a variance comparable to the covariance of the logarithm. We observe a small bias, which might be due to the simplistic regularization based on predicting a threshold on the scattering transfer. The low variance is explained by the consistency of the scattering representation, in contrast with the estimation of higher order moments. Besides, the intermittency is dominated by the transient of the scattering transfer. Section 5.3.3 showed that the variance of the estimator of $\overline{T}X(l)$ is roughly proportional to 2^l, and hence that the transient state corresponds to the regime where the variance is smallest.

5.6 Scattering of Turbulence Energy Dissipation

Turbulent flows appear in a variety of dynamical systems. They contain random fluctuations across time and space, and are thus modeled as stochastic processes. Fully developed turbulence contains physical phenomena at different scales, from the fine dissipation scale where the turbulence flow is smooth, until the integral scale, which can be proportional to the size of the medium.

Table 5.3: Estimation of $2\zeta(1) - \zeta(2) = \lambda^2$ using the scattering transfer, the regression on first and second moments, and the log covariance from [BKM08b], for different values of λ^2. We report the mean and the standard deviation of each estimator.

Cascade	Intermittency	Regression moments	Regression Log-Cov	Scatt transfer
MRM	0.05	$0.05 \pm 8 \cdot 10^{-3}$	$0.05 \pm 6 \cdot 10^{-4}$	0.051 ± 10^{-3}
	0.1	$0.092 \pm 1 \cdot 10^{-2}$	$0.099 \pm 2 \cdot 10^{-3}$	0.099 ± 10^{-3}
	0.15	$0.142 \pm 1 \cdot 10^{-2}$	$0.151 \pm 4 \cdot 10^{-3}$	0.147 ± 210^{-3}
	0.2	0.23 ± 10^{-2}	$0.248 \pm 4 \cdot 10^{-3}$	$0.24 \pm 4 \cdot 10^{-3}$
MRW	0.05	$0.05 \pm 8 \cdot 10^{-3}$	$0.05 \pm 6 \cdot 10^{-4}$	0.05 ± 10^{-3}
	0.1	0.09 ± 10^{-2}	$0.1 \pm 2 \cdot 10^{-3}$	$0.1 \pm 2 \cdot 10^{-3}$
	0.15	$0.14 \pm 2 \cdot 10^{-2}$	$0.15 \pm 2 \cdot 10^{-3}$	0.15 ± 10^{-3}
	0.2	$0.23 \pm 2 \cdot 10^{-2}$	$0.2 \pm 3 \cdot 10^{-3}$	$0.24 \pm 3 \cdot 10^{-3}$

The dissipation of kinetic energy at a given time is given by

$$F(t) = \nu \left(\frac{\partial v}{\partial t} \right)^2 ,$$

where $v(t)$ is the velocity flow and ν is a viscosity constant. The Kolmogorov model [LMC86] predicts an isotropic energy dissipation of energy, which induces a power spectrum of $F(t)$ following the well-known Kolmogorov '$k^{-5/3}$' law:

$$\hat{R}_F(\omega) \propto |\omega|^{-5/3} . \tag{5.84}$$

This model predicts a dissipation process with self-similarity across scales. However, the Kolmogorov theory does not account for the intermittency observed in developed turbulent flows; as a result of the isotropy, the energy dissipation behaves as a Brownian motion with a Hurst parameter adjusted so that its spectral density decays according to (5.84). An alternative model which can account for the multifractality of turbulent flows was introduced also by Kolmogorov in 1962 [MS91]. It modeled the volatility of the dissipation with a log-normal distribution, which corresponds to a random multiplicative log-normal cascade.

The pertinence of these models can be elucidated using the scattering transfer tools developed in the previous sections. Figure 5.17-(a) shows the energy dissipation $F(t)$ as a function of time, from the velocity measurements of a fluid in a Helium jet, with a Reynolds number of $R_\lambda = 703$, which corresponds to the turbulent regime. The dissipation scale is observed at approximately 2^2 sample points, whereas the integral scale is approximately 2^{11} sample points.

Its first order scattering coefficients are displayed in panel (b) of Figure 5.17. We observe that between the diffusion and the integral scale $\log \overline{S}F(j) \simeq 2^{-jH}$, with $H \approx 1/3$.

This corresponds to the Kolmogorov $k^{-5/3}$ law, since from (5.8) we know that Fractional Brownian Motions have a generalized power spectrum which decays as $|\hat{R}_F(\omega) \propto |\omega|^{-2H-1}$.

However, the monofractal isotropic dissipation model can be discarded by computing the scattering transfer of the observed dissipation. Panel (c) of Figure 5.17 shows the estimated scattering transfer coefficients $TF(j_1, j_2)$. We observe that between the diffusion and the integral scales, the dissipation exhibits a form of stochastic self-similarity. The scattering transfer $\overline{T}F(l)$ corresponding to this regime can thus be used to discriminate between a monofractal model and multifractal one.

For that purpose, we first verify the Markov energy property of Section 5.5.2. Panel (d) of Figure 5.17 shows that second moments $E(|F \star \psi_j|^2)$ are well predicted by scattering coefficients. Thus, from Theorem 5.5.3 we use the test $\|\overline{T}F\| \geq 1$ to assess the multifractality of F. By fitting a power law $\overline{T}F(l) \simeq C2^{-\alpha l}$ in the regime between diffusion and integral scale, we obtain $\alpha = 0.19$ and $C = 0.45$, which yields $\|\overline{T}F\| = 1.02 > 1$, corresponding to the multifractal regime.

We also use Theorem 5.5.3 to estimate a measure of intermittence $2\zeta(1) - \zeta(2) \approx 0.01$, which we then use to simulate a log-normal random cascade with $\lambda^2 = 0.01$. Figure 5.18 shows the estimated scattering transfer $\overline{T}F(l)$, which is compared with the dissipation transfer corresponding to the isotropic dissipation model and the log-normal cascade model. The log-normal cascade transfer function predicts an asymptotic regime $\overline{T}F(l) \simeq C$, which does not correspond to the observed asymptotic regime of $\overline{T}F(l) \simeq 2^{-0.2l}$.

The scattering transfer is thus able to discriminate between different multifractal models with consistent estimators, thanks to the fact that it is computed with non-expansive operators. As opposed to first order scattering coefficients, which do not provide insight on the multifractal nature of the process, the scattering transfer contains enough information to discriminate between different multifractal models.

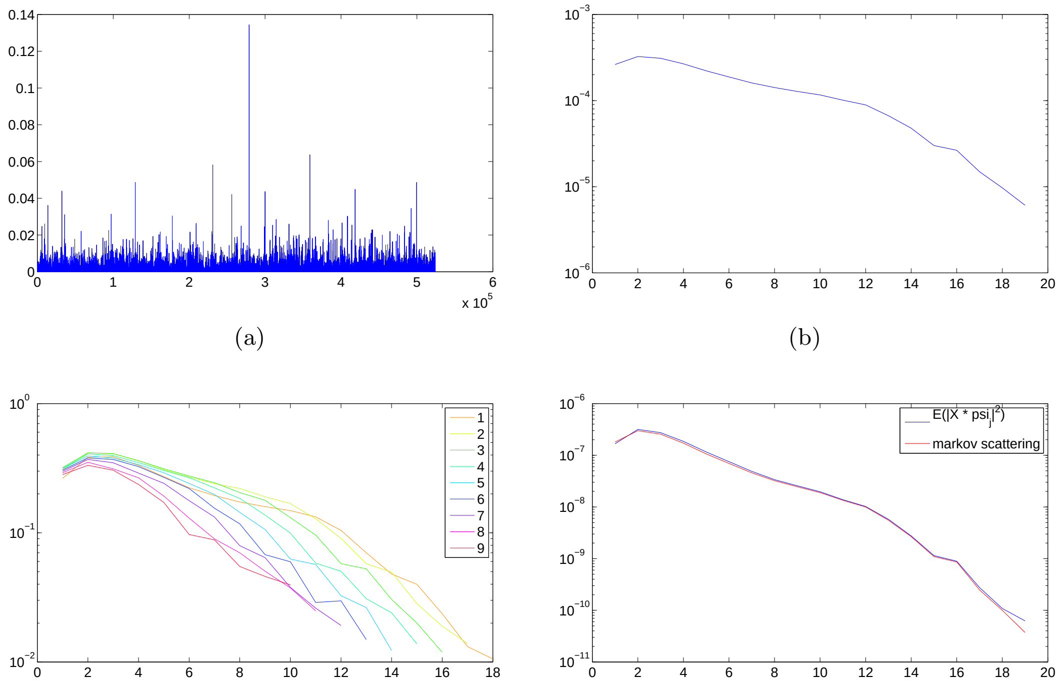
5.7 Scattering of Deterministic Multifractal Measures

The tools developed for stochastic fractals can be applied for the study of deterministic multifractals. We show that self-similarity can be detected and analyzed with the scattering transfer

5.7.1 Scattering transfer for Deterministic Fractals

One dimensional deterministic fractals are studied with the integral scattering transform introduced in Chapter 2. First order scattering coefficients yield a power law given by $\zeta(1)$ which characterizes monofractal singularity. The scattering transfer is defined from first and second order scattering coefficients. We show that self-similarity is expressed in terms of an invariance property on the scattering transfer.

Multifractal analysis requires global singularity measurements, such as the partition function in (5.1). The integral scattering transform for functions $f \in \mathbf{L}^2$ computes $\mathbf{L}^1$



(a) (b) (c) (d)

Figure 5.17: (a) Realization of dissipation $F(t) = \nu \left(\frac{\partial v}{\partial t}\right)^2$ in a turbulent flow. (b) First order scattering coefficients $\log \overline{S}F(j)$ as a function of j, estimated from 4 realizations of 2^{19} samples each. (c) Scattering transfer coefficients $\log TF(j_1, j_2)$ estimated from the same data. We plot curves $\log TF(j_1, j_1 + l)$ as a function of l for different values of j_1. (d) Verification of the energy markov property. In blue, we plot $E(|F \star \psi_j|^2)$, in red the energy predicted by markov scattering coefficients of (5.61), as a function of j.

norms of wavelet modulus decompositions $U[p]f$, given by

$$\widetilde{S}f(p) = \int U[p]f(u)du \quad , p \in \overline{\mathcal{P}}_\infty \ . \tag{5.85}$$

Here, p is a finite order path $p \in \overline{\mathcal{P}}_\infty$. Since $f \in \mathbf{L}^2$ and $\psi \in \mathbf{L}^1$, $U[p]f \in \mathbf{L}^1$ $\forall p \in \overline{\mathcal{P}}_\infty$ and the transform is well defined. Note that in this case we do not renormalize the integral by the Dirac scattering, since we do not consider the limit scattering as path order goes to infinity.

The same transform can be applied to a measure μ defined on a compact set $\Omega \subset \mathbb{R}$. In this case, the first scattering decomposition yields

$$\begin{aligned} \widetilde{S}\mu(j) &= \int |\mu \star \psi_j(u)| du \\ &= \int \left| \int \psi_j(u - u') d\mu(u') \right| du \ , \end{aligned}$$

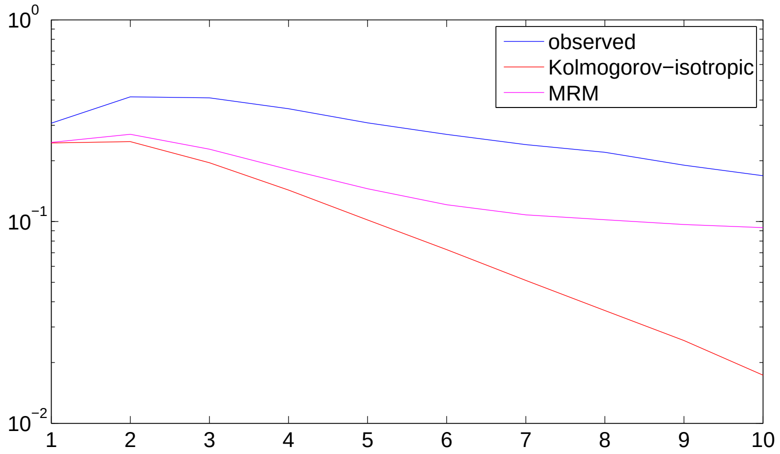


Figure 5.18: Comparison of the scattering transfer estimated from the observed turbulent flows with two energy dissipation models.

which is well defined since $\mu \star \psi_j \in \mathbf{L}^1$. Higher order scattering coefficients are obtained as in the functional case. Indeed, one can decompose any path p of order $|p| > 1$ as $p = j_0 + p'$, yielding

$$\widetilde{S}\mu(p) = \widetilde{S}(U[j_0]\mu)(p') \ ,$$

which is equivalent to the previous functional case since $U[j_0]\mu \in \mathbf{L}^2$ too.

Similarly as in the stochastic case, we introduce the scattering transfer as the main tool to study self-similarity.

Definition 5.7.1 *Let $f \in \mathbf{L}^2$. The Scattering transfer is defined for $(j_1, j_2) \in \mathbb{Z}^2$ by*

$$Tf(j_1, j_2) = \frac{\widetilde{S}f(j_1, j_2)}{\widetilde{S}f(j_1)} \ .$$

The Scattering transfer for a measure μ is defined analogously.

The Scattering transfer, together with the first order coefficients

$$\widetilde{S}f(j) = \|f \star \psi_j\|_1 \quad , j \in \mathbb{Z} \ ,$$

yield a descriptor for fractals, computed with a contractive operator, but which depends upon high order moments.

We shall now see how self-similarity is expressed on the scattering domain.

Let D_j be the operator $D_j f(u) = f(2^{-j}u)$ which dilates f by a factor 2^j.

Similarly, we can define a dilation of a measure μ of $\mathbb{R}$ by specifying its integral on any Borelian set S:

$$D_j\mu(S) = \mu(2^{-j}S) \ , \tag{5.86}$$

where $2^{-j}S = \{u\,; 2^j u \in S\}$.

The following proposition specifies how a dilation operator is seen in the scattering domain.

Proposition 5.7.2 *Let $f \in \mathbf{L}^2$ and $p = (j_1 \dots j_m)$. If $D_j f(u) = f(2^{-j}u)$ denotes a dilation of f by 2^j and $L_j p = (j_1 - j, j_2 - j \dots j_m - j)$ denotes a translation of p by j, then*

$$\widetilde{S}(D_j f)(p) = 2^j \widetilde{S} f(L_j p) \tag{5.87}$$

and hence

$$T(D_j f)(j_1, j_2) = Tf(j_1 - j, j_2 - j)\ .$$

Equivalently, if μ is a measure in a compact set of $\mathbb{R}$, then with the same definition of $L_j p$ we have

$$\widetilde{S}(D_j \mu)(p) = 2^j \widetilde{S}\mu(L_j p) \quad \text{and } T(D_j\mu)(j_1, j_2) = T\mu(j_1 - j, j_2 - j)\ .$$

Proof: Relation (5.15) applied to $D_j f$ yields

$$\begin{aligned} \widetilde{S} D_j f(p) &= 2^{j_1} \int ||||D_{-j_1} D_j f \star \psi| \star \psi_{j_2 - j_1}| \star ...| \star \psi_{j_m - j_1}|(u) du\ . \\ &= 2^{j_1} \int ||||D_{-(j_1 - j)} f \star \psi| \star \psi_{j_2 - j_1}| \star ...| \star \psi_{j_m - j_1}|(u) du\ . \end{aligned} \tag{5.88}$$

Since $L_j p$ is a translation in the path scale variables, the path increments $j_k - j_1$ of $L_j p$ are the same as those of p. It follows from (5.88) that

$$\widetilde{S} D_j f(p) = 2^j \widetilde{S} f(T_j p)\ ,$$

and in particular

$$T D_j f(j_1, j_2) = \frac{\widetilde{S} D_j f(j_1, j_2)}{\widetilde{S} D_j f(j_1)} = \frac{\widetilde{S} f(j_1 - j, j_2 - j)}{\widetilde{S} f(j_1 - j)} = Tf(j_1 - j, j_2 - j)\ .$$

The case of the measure is treated analogously. $\square$.

When a function or a measure is self-similar for a scale s, it is also self-similar for any scale of the form s^n, $n \in \mathbb{N}$. The wavelet decomposition which defines the scattering representation is composed of dilated versions of a mother wavelet $2^{-j} D_j \psi(u) = 2^{-j}\psi(a^{-j}u)$. Dyadic wavelets are obtained by setting $a = 2$. If the self-similarity of a function appears for $s = a$, then (5.7.2) shows that the scattering transfer is stationary: $Tf(j_1, j_1 + l) = \overline{T} f(l)$.

When $s \neq a$, then the scattering transfer is not stationary. However, the pairs $(n, m) \in \mathbb{N}^2$ for which $s^n \approx a^m$ produce a periodicity phenomena on the scattering transfer matrix,

$$Tf(j_1, j_2) \approx Tf(j_1 + m, j_2 + m)\ ,$$

as it will be shown with the Triadic Cantor Set.

5.7.2 Dirac measure

We start by the Dirac measure, which is self-similar for any dilation.

Proposition 5.7.3 *Let δ be the Dirac measure. Then the following properties hold:*

1. $\widetilde{S}\delta(j) = \|\psi\|_1$.
2. *Its scattering transfer satisfies* $T\delta(j_1, j_2) = \overline{T}\delta(j_2 - j_1)$ *and*
$$\lim_{l\to\infty} \overline{T}\delta(l) = \|\psi\|_1 \ . \tag{5.89}$$
3. *δ has an asymptotic first order Markov Scattering.*

Proof: The first order scattering gives $U[j]\delta = |\delta\star\psi_j| = |\psi_j|$. Since $\psi_j(u) = 2^{-j}\psi(2^{-j}u)$, a change of variables shows that

$$\widetilde{S}\delta(j) = \|\psi_j\|_1 = \int |\psi_j(u)|du = \|\psi\|_1 \ ,$$

and hence that first order scattering is constant and equal to $\|\psi\|_1$.

Let us now verify that $T\delta(j_1, j_2) = \overline{T}\delta(j_2 - j_1)$. The Dirac measure satisfies $D_{j_1}\delta = 2^{j_1}\delta$. By applying proposition 5.7.2 we have

$$2^{-j_1}\widetilde{S}D_{j_1}\delta(j_1, j_2) = \widetilde{S}\delta(j_1, j_2) = \widetilde{S}\delta(0, j_2 - j_1) \ ,$$

and since $\widetilde{S}\delta(j_1) = \|\psi\|_1$, we conclude that $T\delta(j_1, j_2)$ is only a function of the difference $j_2 - j_1$.

As $j \to -\infty$, the envelope $|\psi_j|$ is an approximation of the identity in $\mathbf{L}^1$. Since

$$\begin{aligned} U[j_1, j_2]\delta &= 2^{-j_1}D_{j_2}|D_{j_1-j_2}|\psi| \star \psi| \\ &= 2^{-j_2}D_{j_2}|2^{j_2-j_1}D_{j_1-j_2}|\psi| \star \psi| \\ &= 2^{-j_2}D_{j_2}||\psi_{j_1-j_2}| \star \psi| \ , \end{aligned}$$

it results that for each j_2,

$$\lim_{j_1\to-\infty} \|U[j_1, j_2]\delta - \|\psi\|_1 2^{-j_2}D_{j_2}|\delta \star \psi|\|_1 = \lim_{j_1\to-\infty} \|U[j_1, j_2]\delta - \|\psi\|_1 U[j_2]\delta\|_1 \tag{5.90}$$

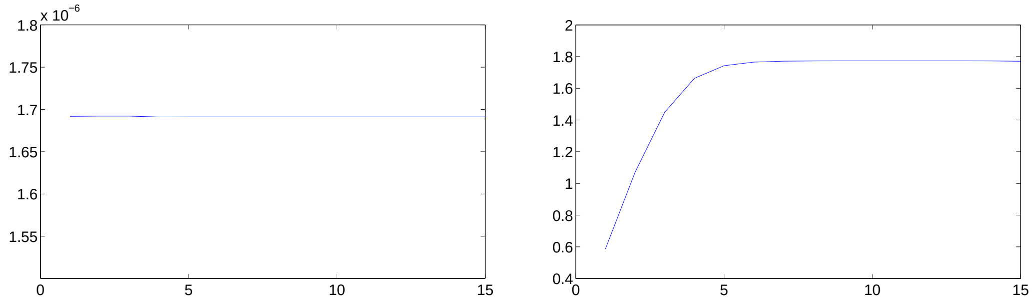
which yields

$$\lim_{j_1\to-\infty} \int U[j_1, j_2]\delta(u)du = \|\psi\|_1 \int U[j_2]\delta(u)du \ ,$$

and hence $\lim_{l\to\infty} \overline{T}\delta(l) = \|\psi\|_1$. By applying (5.90) for each $p = (j_1 \dots j_m)$ shows that

$$\lim_{l\to\infty} \widetilde{S}\delta(p, j_m + l) = \overline{T}\delta(l)\widetilde{S}\delta(p) \ ,$$

and hence that $X = \delta$ has an asymptotic Markov Scattering. $\square$.



(a) (b)

Figure 5.19: Scattering of Dirac measure. (a) First order scattering $\widetilde{S}\delta(j)$, (b) Scattering transfer $\overline{T}\delta(l)$.

Figure 5.19 displays the scattering first order coefficients, together with the scattering transfer of the Dirac measure, computed with spline wavelets. As expected, first order coefficients are constant, and the scattering transfer, after a transient state which depends upon the wavelet, reaches a constant value.

If one chooses Gabor wavelets, then $\widetilde{S}\delta$ can be computed analytically and hence one can compute the speed of convergence of $\overline{T}(l)$ towards the constant. In that case, $\psi(u) = e^{i\xi_0 u}e^{-u^2/(2\sigma_0^2)}$, and $|\psi_{j_1}|$ is a lowpass Gaussian window. The convolution $|\psi_{j_1}| \star \psi_{j_2}$ is again a complex Gaussian, characterized by a variance, a central frequency and an amplitude, which implies that $U[j_1, j_2]\delta$ is again a lowpass Gaussian characterized by a variance and a maximum amplitude. This means that for any path $p = (j_1 \dots j_m)$,

$$\widehat{U[p]\delta}(\omega) = e^{-\beta_p} e^{-\sigma_p^2 \omega^2/2} ,$$

which yields

$$\widetilde{S}\delta(p) = \widehat{U[p]\delta}(0) = e^{-\beta_p} .$$

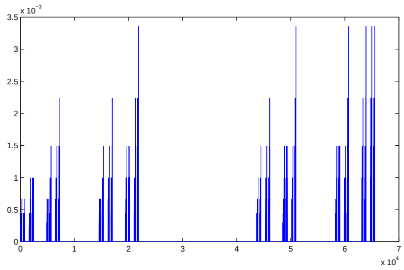
The parameters β_p and σ_p^2 are obtained from the previous path $p_0 = (j_1 \dots j_{m-1})$ by solving a linear system:

$$\begin{cases} \sigma_p^2 & = \sigma_{p_0}^2 + 2^{2j_m}\sigma_0^2 \ , \\ \beta_p & = \beta_{p_0} + \frac{\sigma_0^2\xi_0^2}{2}\left(1 + \frac{\sigma_0^2 2^{2j_m}}{\sigma_p^2}\right) \ . \end{cases} \tag{5.91}$$

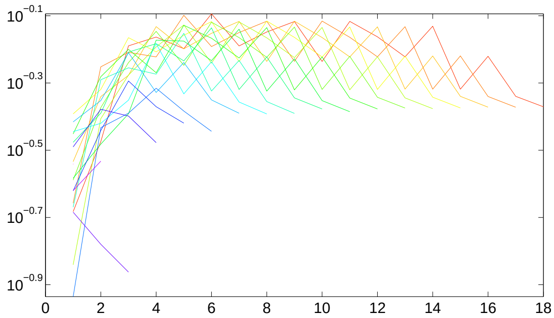
As a result,

$$\overline{T}\delta(j_m - j_{m-1}) = \frac{\widetilde{S}\delta(p)}{\widetilde{S}\delta(p_0)} = \exp\left\{\frac{\sigma_0^2\xi_0^2}{2}\left(1 + \frac{2^{2j_m}}{\sum_{l=0}^m 2^{2j_l}}\right)\right\} .$$

If p is a frequency decreasing path, the denominator $\sum_{l=0}^m 2^{2j_l}$ is dominated by the last terms, thus showing that in this case the asymptotic Markov is reached at exponential rate.



(a)



(b)

Figure 5.20: (a) Triadic Cantor fractal (b) Scattering transfer $T\mu(j_1, j_2)$. We plot the curves $T\mu(j_1, j_1 + l)$ as a function of l for different values of j_1. The curves do not converge towards a transfer function since the scattering is defined over dyadic wavelets, which does not match the scale self-similarity of the fractal.

5.7.3 Cantor Measures

Figure 5.20 shows a triadic Cantor fractal obtained with weights $\alpha_1 = 0.4$ and $\alpha_2 = 0.6$. As expected, its scattering coefficients are not self-similar since they are computed with dyadic wavelets at scales of the form $s_j = 2^j$, $j \in \mathbb{Z}$, and the self-similarity of the measure is observed for dilations of the form $\tilde{s}_k = 3^k$, $k \in \mathbb{Z}$. The oscillatory pattern observed in Figure 5.20 is explained by considering the continuous surface

$$\tilde{T}\mu(\log s_1, \log s_2) = \frac{E(||\mu \star \psi_{s_1}| \star \psi_{s_2}|)}{E(|\mu \star \psi_{s_1}|)} \ , \ s_1, s_2 \in \mathbb{R} \ , s_1, s_2 > 0 \ ,$$

where $\psi_s(u) = s^{-1}\psi(s^{-1}u)$ is a dilation by a factor s. $\tilde{T}\mu(\log s_1, \log s_2)$ is a continuous and periodic two-dimensional surface, where the period is determined by the self-similarity of μ. The scattering transfer $T\mu(j_1, j_2)$ corresponds to a discrete sampling of $\tilde{T}\mu$ at points of the form $(2^{j_1}, 2^{j_2})$, which creates oscillations as soon as the self-similarity factor does not correspond to the sampling step.

Appendix A

Wavelet Modulation Operators

A.1 Wavelet and Modulation commutation

This Section studies the commutation of a wavelet decomposition operator and a modulation. When the envelope is smooth with respect to the scale of the wavelet, the two operators nearly commute, thanks to the good spatial localization of wavelets.

We shall express the regularity of σ with a multi-resolution analysis. For that purpose, we consider the multi-resolution approximation spaces $\{\mathbf{V}_k\}_{k\in\mathbb{Z}}$ [Mal12] generated by the orthogonal basis

$$\{\phi_k(t-n2^k)\,,\, n\in\mathbb{Z}\}\ ,$$

where $\phi_k(u) = 2^{-kd/2}\phi(2^{-k}u)$ is a scaling function. Notice that in this case we use an $\mathbf{L}^2$ normalization. When the context requires it, we shall distinguish between the $\mathbf{L}^2$ normalized scaling functions and the $\mathbf{L}^1$ normalization used in the definition of the scattering operator. We say that $\sigma \in \mathbf{V}_k^\infty$ if

$$\sigma(t) = 2^{kd/2}\sum_n c[n]\phi_k(t-n2^k)$$

with $\sup_n |c[n]| < \infty$. In particular, if ϕ is C^2, then σ is C^2 too.

The following proposition computes a bound for the commutator $M[\sigma]W_\lambda - W_\lambda M[\sigma]$. As before, $|\nabla\sigma|_\infty = \sup_u |\nabla\sigma(u)|$ denotes the sup of the Euclidean norm of $\nabla\sigma$, and $|H\sigma|_\infty = \sup_u \|H\sigma(u)\|$, where $\|H\sigma(u)\|$ is the operator norm of the Hessian $H\sigma(u)$.

Proposition A.1.1 *Let j and k be two integers, and let $M[\sigma]x(u) = \sigma(u)x(u)$ be a modulation operator with $\sigma \in \mathbf{V}_k^\infty$. Then, if $\lambda = 2^j r$, there exists a constant $C > 0$ depending upon the scaling function ϕ such that*

$$\|M[\sigma]W_\lambda - W_\lambda M[\sigma]\| \leq C|\sigma|_\infty 2^{j-k}\Delta_u(\psi)\ , \tag{A.1}$$

with

$$\Delta_u(\psi) = \int |u|\,|\psi(u)|du\ .$$

Proof: The kernel of $M[\sigma]W_\lambda - W_\lambda M[\sigma]$ is

$$k(v,u) = \sigma(v)\psi_\lambda(v-u) - \sigma(u)\psi_\lambda(v-u) = (\sigma(v)-\sigma(u))\,\psi_\lambda(v-u)\ .$$

Since σ is twice differentiable, we can consider the following Taylor development:

$$k(v,u) = \left(\int_0^1 (v-u)\cdot\nabla\sigma(u+t(v-u))dt\right)\psi_\lambda(v-u)\ , \tag{A.2}$$

Let us now bound the operator using the Schur lemma. Since

$$k(u,v) = -\left(\sigma(v)-\sigma(u)\right)\tilde{\psi}_\lambda(v-u)\ ,$$

with $\overline{\psi}(u) = \psi(-u)$, we can exchange the roles of v and u and hence it is sufficient to bound

$$\sup_u \int |k(v,u)|dx\ .$$

Fix $u \in \mathbb{R}^d$. From (A.2) we have

$$\begin{aligned}\int |k(v,u)|dv &\leq |\nabla\sigma|_\infty \int |v-u||\psi_\lambda(v-u)|dv \\ &= |\nabla\sigma|_\infty 2^j \int |v|\,|\psi(v)|dv\end{aligned}$$

via a change of variables $\tilde{v} = 2^{-j}r^{-1}v$. As a result,

$$\|M[\sigma]W_\lambda - W_\lambda M[\sigma]\| \leq \sup_u \int |k(v,u)|dv \leq |\nabla\sigma|_\infty 2^j \int |v||\psi(v)|dx\ . \tag{A.3}$$

We shall now bound the gradient of an element of $\mathbf{V}_k^\infty$. Since

$$\sigma(u) = 2^{kd/2}\sum_n c[n]\phi_k(u-n2^k)\ ,$$

it follows that

$$\nabla\sigma(u) = 2^{kd/2}\sum_n c[n]\nabla\phi_k(u-n2^k),$$

and hence

$$\begin{aligned}|\nabla\sigma(u)| &\leq 2^{kd/2}\sum_n |c[n]||\nabla\phi_k(u-n2^k)| \\ &\leq 2^{-k}\sup_n |c[n]|\sum_n |\nabla\phi(2^{-k}u-n)| \\ &\leq 2^{-k}\sup_n |c[n]|F(\nabla\phi)\ ,\end{aligned} \tag{A.4}$$

where

$$F(\nabla\phi) = \sup_{\gamma\in[0,1]^d}\sum_n |\nabla\phi(\gamma+n)|$$

only depends upon ϕ, and is finite as soon as ϕ and its derivatives have fast decay.

Finally, by definition

$$2^{kd/2}c[n] = 2^{-kd/2} \int \sigma(t)\phi(2^{-k}(t - n2^k)dt \ ,$$

which implies that

$$\begin{aligned} \forall n \ , |c[n]| &= 2^{-kd} \left| \int \sigma(t)\phi(2^{-k}t - n)dt \right| \\ &\leq |\sigma|_\infty 2^{-kd} \int |\phi(2^{-k}t)|dt = |\sigma|_\infty \|\phi\|_1 \ . \end{aligned} \tag{A.5}$$

From (A.3, A.4, A.5), we obtain

$$\|M[\sigma]W_\lambda - W_\lambda M[\sigma]\| \leq 2^{j-k}|\sigma|_\infty \Delta_u(\psi)\left(F(\nabla\phi)\|\phi\|_1\right) \ ,$$

which concludes the proof. $\square$.

This proposition shows that when $k \gg j$, the modulation with an envelope $\sigma \in \mathbf{V}_k^\infty$ nearly commutes with the wavelet decomposition at scale j. In that case, we exploit the good spatial localization of wavelets. When $k \gg j$, then the envelope is smooth with respect to the wavelet, which means that, locally, the envelope is well approximated by a constant within the support of the wavelet.

A.2 Wavelet Near Diagonalisation Property

Lemma A.2.1 *Let $\lambda = 2^j r$ and $\nabla W_\lambda f = f \star \nabla\psi_\lambda$. Then, if ω_0 is the central frequency of ψ,*

$$\|\nabla W_\lambda - i2^{-j}(r^{-1}\omega_0)W_\lambda\| \leq d\,2^{-j}\Delta_\omega(\psi) \ , \tag{A.6}$$

where

$$\Delta_\omega(\psi) = \sup_{\omega'} |\omega' - \omega_0||\hat{\psi}(\omega')|$$

is the frequency spread of ψ.

Proof of lemma A.2.1: Since scalar products in $\mathbb{R}^d$ and gradients commute with orthogonal transformations, we can assume without loss of generality that $r = 1$ and that ω_0 is aligned along the first cartesian coordinate. Let us approximate the convolution $f \star \partial_k\psi_\lambda$ with another linear operator of the form $\alpha f \star \psi_\lambda$, and let

$$E_k(f) = f \star \partial_k\psi_\lambda - \alpha f \star \psi_\lambda = f \star (\partial_k\psi_\lambda - \alpha\psi_\lambda) \ .$$

Since E_k is a convolution operator, it is diagonalized in the Fourier basis and hence its operator norm is given by

$$\sup_\omega \left| i\omega_k\widehat{\psi}_\lambda - \alpha\widehat{\psi}_\lambda \right| \ . \tag{A.7}$$

We choose as approximation $\alpha = i2^{-j}(\omega_0)_k$, which corresponds to the k-th coordinate of the central frequency of ψ_λ. As a result, (A.7) yields

$$\begin{aligned} \|E_k\| &= \sup_\omega |\omega_k - 2^{-j}\omega_{0,k}||\widehat{\psi}(2^j\omega)| \\ &= 2^{-j}\sup_\omega |\omega_k - \omega_{0,k}||\widehat{\psi}(\omega)|^2 \leq 2^{-j}\sup_\omega |\omega - \omega_0||\widehat{\psi}(\omega)|^2 . \end{aligned} \quad (A.8)$$

By summing over all coordinates $k = 1 \ldots d$, we obtain (A.6). $\square$.

A.3 Local Scattering Analysis of Wavelet Modulation Operators

This Section concentrates in the study of the multiscale perturbations defined by (4.9). We restrict ourselves to the unidimensional case $d = 1$ and to first and second order scattering coefficients.

We first verify that the perturbations given by wavelet modulations define a stable operator in $\mathbf{L}^2(\mathbb{R}^d)$.

Proposition A.3.1 *Let W_λ and $\widetilde{W_\lambda}$ be the wavelet decomposition frame and dual frame respectively. If the wavelet frame satisfies the Littlewood-Paley condition (2.7) with $\delta > 0$, then $\overline{M}[\overline{\sigma}]$ is a bounded linear operator of $\mathbf{L}^2(\mathbb{R}^d)$, and its norm satisfies*

$$\|\overline{M}[\overline{\sigma}] - \mathbf{1}\| \leq \frac{|\sigma|_\infty}{1-\delta} . \quad (A.9)$$

In particular, the map $\sigma \mapsto \overline{M}[\overline{\sigma}]f - f$ is Lipschitz with respect to the norm $\sup_{\lambda,u}|\sigma(\lambda,u)|$.

Proof: The multiscale operator $\overline{M}[\overline{\sigma}]$ can be written as

$$\overline{M}[\overline{\sigma}] = \mathbf{1} + \widetilde{\mathcal{W}}\mathcal{M}\mathcal{W} ,$$

where $\mathcal{W}$ is the decomposition frame $\mathcal{W} = \{A_J, W_\lambda\}_{\lambda\in\Lambda_J}$, $\mathcal{M}$ is a point-wise modulation operator in the frame decomposition coefficients, and $\widetilde{\mathcal{W}}$ is the frame reconstruction operator. From the Littlewood-Paley condition (2.7) we know that the frame bounds are $[1-\delta, 1]$, which implies that $\|\mathcal{W}\| \leq 1$ and $\|\widetilde{\mathcal{W}}\| \leq \frac{1}{1-\delta}$. As for the diagonal operator $\mathcal{M}$, its norm is bounded by $|\sigma|_\infty = \sup_{\lambda,u}|\sigma(\lambda,u)|$. It follows that

$$\|\overline{M}[\overline{\sigma}] - \mathbf{1}\| \leq \|\mathcal{W}\|\,\|\mathcal{M}\|\,\|\widetilde{\mathcal{W}}\| \leq \frac{|\sigma|_\infty}{1-\delta} ,$$

which proves (A.9). Since for each $f \in \mathbf{L}^2(\mathbb{R}^d)$ we have

$$\|\overline{M}[\overline{\sigma}]f - f\| \leq \frac{\|f\|}{1-\delta}|\sigma|_\infty ,$$

it follows that $\sigma \mapsto \overline{M}[\overline{\sigma}]f - f$ is Lipschitz with respect to the supremum norm $\sup_{\lambda,u} |\sigma(\lambda, u)|$ $\square$.

The main result that we obtain is that one can find a family of wavelet envelopes such that they define scattering perturbations with an asymptotic triangular structure. Thanks to this property, a small scattering perturbation can be linearly inverted as a cascade of multiscale wavelet modulations.

Let us first define a collection of localized envelopes, which generate a family of "atomic" wavelet modulations in correspondence with scattering paths. We recall the definition of Γ:

$$\Gamma = \mathcal{P}_\downarrow^1 \cup \mathcal{P}_\downarrow^2 = \{j_1 \in \mathbb{Z}\,, j_1 \leq J\} \cup \{(j_1, j_2) \in \mathbb{Z}^2\,;\, j_1 < j_2 \leq J\}\ .$$

For convenience, we can identify first order paths $j_1 \in \mathcal{P}^1$ with $(j_1, J+1)$ and parametrize $\Gamma = \{(j_1, j_2) \in \mathbb{Z}^2\,;\, j_1 < j_2 \leq J+1\}$. A multiscale modulation operator $\overline{M}[\overline{\sigma}]$ is defined from a bi-dimensional envelope

$$\sigma(j_1, u)\,, j_1 \leq J\,,\, u \in \mathbb{R}\ .$$

We are now going to construct a family of envelopes $\{\sigma_p\}_{p=(j_1,j_2)\in\Gamma}$ in correspondence with Γ. Let us denote $\mathbf{W}_k$ the space of details from a multiresolution analysis [Mal12]: $\mathbf{V}_k = \mathbf{W}_k \oplus \mathbf{V}_{k+1}$.

Definition A.3.2 *Let $\Delta_1, \Delta_2 \geq 0$, and suppose that $\{\Psi_j(u - 2^j n)\}_n$ is a wavelet basis for the space of details $\mathbf{W}_j$ of $\mathbf{L}^2(\mathbb{R})$ generated by ψ_j. We define an envelope of type $p = (j_1, j_2)$, $j_1 < j_2$, as*

$$\sigma_p(j, u) = \begin{cases} \sum_{|j'-j_2|\leq\Delta_2, j'>j} \sum_n \theta[j', n]\Psi_{j'}(u - 2^{j'}n) & ,\ \mathit{if}\ |j - j_1| \leq \Delta\ , \\ 0\ \ ,\ \ \mathit{otherwise}\ , \end{cases} \tag{A.10}$$

for any complex vector of coefficients such that $\sup |\theta[j', n]| < \infty$.

An envelope of type $p = (j_1, j_2)$ thus has its energy well localized in the log-frequency plane (j, j') defined by the scales of f and the scales of the envelope, and illustrated in figure A.1.

In particular, when $j_2 = \infty$, the envelope has no spatial variations, $\sigma_p(j, u) \equiv C_0$ for $|j - j_1| \leq \Delta_1$, and hence the resulting operator $\overline{M}[\sigma]$ becomes

$$\overline{M}[\sigma]x = x + C_0 \sum_{|j-j_1|\leq\Delta_1} \tilde{W}_j W_j x\ , \tag{A.11}$$

which amplifies the energy of the wavelet subbands in $[j_1 \pm \Delta_1]$.

The operator $\overline{M}[\sigma]$ with an envelope of type $p = (j_1, j_2)$ modifies the wavelet coefficients of x at scales in the neighborhood of j_1. The wavelet representation is redundant, and the wavelet coefficients from different scales are related by a reproducing kernel

$$W_j x = \mathcal{W}\widetilde{\mathcal{W}} W_j x\ , \tag{A.12}$$

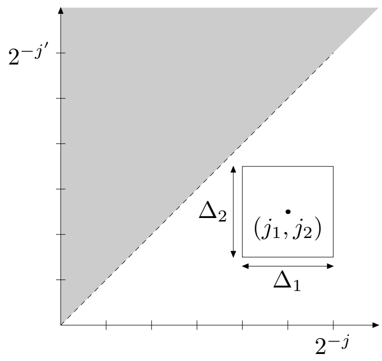


Figure A.1: Definition of an envelope of type $p = (j_1, j_2)$ in a plane (j, j') representing the scales 2^j of the $\mathbf{L}^2(\mathbb{R})$ wavelet decomposition W_j and the scales j' of each envelope $\sigma(j, \cdot)$. An envelope of type p is localized around (j_1, j_2) with a spread controlled by (Δ_1, Δ_2). The shaded region corresponds to envelopes more irregular than the wavelets they are acting upon. We shall see that the "valid" region is in correspondence with the set of progressive first and second order scattering paths.

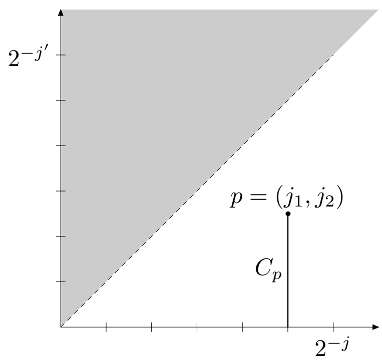


Figure A.2: Cone of influence assiociated with the path $p = (j_1, j_2)$.

where $\mathcal{W}$ and $\widetilde{\mathcal{W}}$ are respectively the forward and the dual wavelet decomposition. This implies that a perturbation $M[\sigma(j_1, \cdot)]W_{j_1}$ on a given scale j_1 won't in general satisfy the reproducing kernel equation (A.12). As a result, the dual wavelet projections $\tilde{W}_j$ will propagate the perturbation to other scales.

The complex coefficients $\{\theta[j', n] \in \mathbb{C}\}_n$ encode variations, both in phase and amplitude, of wavelet coefficients at scale 2^j, along envelopes whose own variations are localized at scale $2^{j'}$, with $j' > j$. We shall now see that an envelope of type p produces a perturbation concentrated along a band of influence $C_p \subset \Gamma$ centered at $p = (j_1, j_2)$. It is defined by

$$C_p = \{q = (l_1, l_2) \in \Gamma \, ; \, l_1 = j_1 \, ; \, l_2 \geq j_2\} \ .$$

Figure A.2 illustrates the band of influence C_p.

The scattering paths in Γ can be ordered along a lexicographic ordering induced from

$\mathbb{Z}^2$. For any p, $q \in \Gamma$ with $p = (j_1, j_2)$, $q = (j'_1, j'_2)$, we say that

$$p < q \Longleftrightarrow 2^{-j_1} < 2^{-j'_1} \text{ or } (j_1 = j'_1 \text{ and } 2^{-j_2} < 2^{-j'_2}) \ . \tag{A.13}$$

Then the band of influence C_p of a given path p is included in the set of older paths $\{q \in \Gamma; q \leq p\}$. The following theorem shows that for each x and small ϵ, we can find a family of envelopes of type p, $p \in \Gamma$, such that the operator $\nabla_{\overline{M}} S_{Jf}$ mapping each envelope of type $p = (j_1, j_2)$ to the corresponding scattering difference,

$$\nabla_{\overline{M}} S_{Jx,\epsilon} \, : \, p \longmapsto \frac{(S_J \overline{M}[\epsilon\sigma_p]x - S_J x)}{\epsilon} \ , \tag{A.14}$$

has its energy concentrated along a band lying in the lower triangular region of the scattering coordinates. More precisely, the theorem shows that the energy of the perturbation $S_J\overline{M}[\sigma]x - S_J x$ along a scattering path q, given by

$$|S_J\overline{M}[\sigma]x(q) - S_J x(q)|^2 \ ,$$

decays proportionally to a discrete distance $\mathrm{dist}(q, C_p)$ in Γ.

Theorem A.3.3 *Let $x \in \mathbf{L}^2(\mathbb{R})$ and $p = (j_1, j_2) \in \Gamma$. Suppose that S_J is computed with a unitary wavelet frame generated by ψ with fast frequency decay. The following properties hold:*

1. *If $j_0 \neq j_1$ and $\Omega_{j_0} = \{(q_1, q_2) \in \Gamma\, ;\, q_1 = j_0\}$, then for any envelope of type p*

$$\|S_J\overline{M}[\overline{\sigma}]x[\Omega_{j_0}] - S_J x[\Omega_{j_0}]\|^2 \leq C|\sigma|^2_\infty \|f\|^2 \sup_{\delta \in [-\Delta_1, \Delta_1]} K_1(j_1 - j_0 + \delta) \ , \tag{A.15}$$

   where K_1 depends only upon ψ and satisfies $K_1(n) \to 0$ as $|n| \to \infty$.

2. *Suppose Ψ and ϕ have compact support. Then there exists a non-zero envelope σ^* of type p, and C_0 depending only on the wavelets, such that if $j_2 + \log(j_2 - j_1) \leq J - C_0$, then for all paths $q = (q_1, q_2)$ satisfying $q_1 = j_1$ and $q_2 < j_2$, the discrete scattering vector*

$$\tilde{F}_q[k] = F[q](2^J k) = S_J\overline{M}[\overline{\sigma}][q]x(2^J k) - S_J[q]x(2^J k) \ , k \in \mathbb{Z} \ ,$$

   satisfies

$$\forall\, q_1 = j_1 \ , q_2 < j_2 \ , \ \|\tilde{F}_q\|^2 \leq C\|x\|^2 \left(K_2(\Delta_1) + |\sigma|^2_\infty 2^{-(j_2 - q_2)}\right) \ , \tag{A.16}$$

   where K_2 depends only upon ψ and satisfies $K_2(n) \to 0$ as $n \to \infty$, and C depends upon p and Δ.

Proof: Let us first prove (A.15). We define $F = S_J \overline{M}[\overline{\sigma}]x - S_J x$. Since Ω_{j_0} regroups the paths in Γ starting by j_0, for any $x, x' \in \mathbf{L}^2(\mathbb{R})$ we have

$$\begin{aligned} \|S_J[\Omega_{j_0}]x - S_J[\Omega_{j_0}]x'\| &\leq \|S_J U[j_0]x - S_J U[j_0]x'\| \\ &\leq \|U[j_0]x - U[j_0]x'\| \leq \|W_{j_0}(x - x')\| \ , \end{aligned} \tag{A.17}$$

thanks to the fact that S_J and U are contractive operators. Let us apply (A.17) with $x' = \overline{M}[\overline{\sigma}]x$. If σ is an envelope of type $p = (j_1, j_2)$, then

$$\overline{M}[\overline{\sigma}]x - x = Re(\sum_{|j-j_1|\leq \Delta_1} \tilde{W}_j M[\sigma(j,\cdot)]W_j x) \ ,$$

and hence

$$\begin{aligned} \|F[\Omega_{j_0}]\| &\leq \|W_{j_0}(x - \overline{M}[\overline{\sigma}]x)\| \\ &\leq \sum_{|j-j_1|\leq \Delta_1} \|W_{j_0}\tilde{W}_j M[\sigma(j,\cdot)]W_j x\| \\ &\leq |\sigma|_\infty \|x\| \sum_{|j-j_1|\leq \Delta_1} \|W_{j_0}\tilde{W}_j\| \ . \end{aligned} \tag{A.18}$$

Since the wavelet frame is unitary, the norm of the operator $W_{j_0}\tilde{W}_j$ is given by

$$\sup_\omega |\hat{\psi}_{j_0}\hat{\psi}_j^*(\omega)| = \sup_\omega |\hat{\psi}(2^{j_0}\omega)||\hat{\psi}(2^j\omega)| = K_1(j_0 - j) \ ,$$

where

$$K_1(j) = \sup_\omega |\hat{\psi}(2^j\omega)|\,|\hat{\psi}(\omega)| \ . \tag{A.19}$$

If the wavelet has fast frequency decay, then $K_1(j) \leq C_0 a^{|j|}$ for $a < 1$, and hence (A.18) is bounded by

$$\begin{aligned} \|F[\Omega_{j_0}]\| &\leq |\sigma|_\infty \|x\|^2 \sup_{\delta\in[-\Delta_1,\Delta_1]} K_1(j_1 + \delta - j_0) \left(\sum_{j=0}^{2\Delta_1} C_0 a^j\right) \\ &\leq C\|x\| \sup_{\delta\in[-\Delta_1,\Delta_1]} K_1(j_1 + \delta - j_0) \ , \end{aligned}$$

which proves (A.15).

We shall now prove (B.1). If $q = (j_1, q_2)$ is a progressive path with $q_2 < j_2$, we start by computing $W_{j_1}\overline{M}[\overline{\sigma}]x$ with an envelope of type $p = (j_1, j_2)$. If $g = Re(g_0)$ and ψ is a complex wavelet, then $2W_j g = W_j g_0$, and hence we can drop the real part in the definition of $\overline{M}[\overline{\sigma}]$ with just a constant impact on the bound. If $j \in [j_1 \pm \Delta_1]$, then by definition $\sigma(j,u)$ does not depend upon j, and we shall write $\sigma(j,u) = \sigma(u)$. Thus,

$$W_{j_1}\overline{M}[\overline{\sigma}]x = W_{j_1}f + W_{j_1} \sum_{|j-j_1|\leq\Delta_1} \tilde{W}_j M[\sigma]W_j f \ . \tag{A.20}$$

Let us approximate (A.20) with a modulated version of the original wavelet decomposition, $(\mathbf{1}+M[\sigma])W_{j_1}f$:

$$\begin{aligned}
W_{j_1}\overline{M}[\overline{\sigma}]f &= W_{j_1}f + W_{j_1}\left(\sum_{|j-j_1|\leq\Delta_1}\left(M[\sigma]\tilde{W}_j - [M[\sigma],\tilde{W}_j]\right)W_jf\right) \\
&= W_{j_1}f + M[\sigma]W_{j_1}\sum_{|j-j_1|\leq\Delta_1}\tilde{W}_jW_jf \ + \\
&\quad +[M[\sigma],W_{j_1}]\sum_{|j-j_1|\leq\Delta_1}\tilde{W}_jW_jf - W_{j_1}\sum_{|j-j_1|\leq\Delta_1}[M[\sigma],\tilde{W}_j]W_jf \\
&= (\mathbf{1}+M[\sigma])W_{j_1}f + (E_1+E_2+E_3)f \ ,
\end{aligned}$$

with

$$\begin{aligned}
E_1 &= M[\sigma]W_{j_1}\left(\sum_{|j-j_1|\leq\Delta_1}\tilde{W}_jW_j - \mathbf{1}\right) \ , \ E_2 = [M[\sigma],W_{j_1}]\sum_{|j-j_1|\leq\Delta_1}\tilde{W}_jW_j \ , \\
E_3 &= -W_{j_1}\sum_{|j-j_1|\leq\Delta_1}[M[\sigma],\tilde{W}_j]W_j \ .
\end{aligned}$$

We now bound the linear operators E_i, $i=1,2,3$. We have

$$\|E_1\| = |\sigma|_\infty \sup_\omega |\hat{\psi}(2^{j_1}\omega)|\left|1-\sum_{|j-j_1|\leq\Delta_1}|\hat{\psi}(2^j\omega)|^2\right| = K_2(\Delta_1) \ , \tag{A.21}$$

where

$$K_2(\Delta) = \sup_\omega |\hat{\psi}(\omega)|\left|1-\sum_{|j|\leq\Delta}|\hat{\psi}(2^j\omega)|^2\right| \tag{A.22}$$

satisfies $K_2(\Delta)\to 0$ as $\Delta\to\infty$, since the wavelet frame is unitary, and hence $\sum_j|\hat{\psi}(2^j\omega)| = 1$ for all $\omega>0$.

Since σ is an envelope of type $p=(j_1,j_2)$, the commutator $[M[\sigma],W_j]$ is bounded using proposition A.1.1 by $\|[M[\sigma],W_j]\|\leq C2^{j-j_2+\Delta_2}$, which implies

$$\|E_2\| \leq \|[M[\sigma],W_{j_1}]\| \ \|\sum_{|j-j_1|\leq\Delta_1}\tilde{W}_jW_j\| \leq C2^{j_1-j_2+\Delta_2} \ , \tag{A.23}$$

and

$$\|E_3\| \ \leq \sum_{|j-j_1|\leq\Delta_1}\|[M[\sigma],\tilde{W}_j]\| \leq C\sum_{|j-j_1|\leq\Delta_1}2^{j-j_2+\Delta_2} \leq C'2^{j_1+\Delta_1-j_2+\Delta_2} \tag{A.24}$$

By reassembling (A.21), (A.23) and (A.24) we have that

$$W_{j_1}\overline{M}[\overline{\sigma}]x = (\mathbf{1}+M[\sigma])W_{j_1}x + \overline{E}x \ , \tag{A.25}$$

where $\overline{E}$ satisfies $\|\overline{E}\| \leq K_2(\Delta_1) + C2^{j_1-j_2+\Delta_1+\Delta_2}$.

Second order scattering coefficients of the form (j_1, q_2) are computed from the complex modulus of $W_{j_1}\overline{M}[\overline{\sigma}]x$. If $M[\sigma]$ is a modulation operator, then $|M[\sigma]x| = M[|\sigma|]|x|$. As a result, $U[j_1]\overline{M}[\overline{\sigma}]x$ satisfies

$$\|U[j_1]\overline{M}[\overline{\sigma}]x - M[\tilde{\sigma}]U[j_1]x\| \leq \|\overline{E}\|\|x\| \ , \tag{A.26}$$

with

$$\tilde{\sigma}(u) = |1 + \sigma(u)| \ .$$

Now, we decompose $U[j_1]\overline{M}[\overline{\sigma}]x$ with ψ_{q_2}, which yields

$$\|W_{q_2}U[j_1]\overline{M}[\overline{\sigma}]x - W_{q_2}M[\tilde{\sigma}]U[j_1]x\| \leq \|\overline{E}\|\|x\| \ . \tag{A.27}$$

Since $|\sigma|_\infty < 1$, we have that $\tilde{\sigma}(u) > 0\,, \forall u$, which means that the modulus does not vanish, and hence that it preserves the regularity of σ. As a consequence, if σ is an envelope of type $p = (j_1, j_2)$, it follows that $\tilde{\sigma}$ is also of type p, which means that W_{q_2} and $M[\tilde{\sigma}]$ commute with an error $\sim 2^{q_2-j_2+\Delta_2}$. We can thus write

$$\|W_{q_2}U[j_1]\overline{M}[\overline{\sigma}]x - M[\tilde{\sigma}]W_{q_2}U[j_1]x\| \leq (\|\overline{E}\| + C2^{q_2-j_2+\Delta_2})\|x\| \ ,$$

which leads to

$$\|U[j_1, q_2]\overline{M}[\overline{\sigma}]x - M[\tilde{\sigma}]U[j_1, q_2]x\| \leq (\|\overline{E}\| + C2^{q_2-j_2+\Delta_2})\|x\| \ , \tag{A.28}$$

since $|\tilde{\sigma}| = \tilde{\sigma}$. Now, by decomposing $\tilde{\sigma}$ as

$$\tilde{\sigma} = |1 + \sigma(u)| = \sqrt{1 + 2Re(\sigma(u)) + |\sigma(u)|^2} = 1 + \sigma_0(u) \ ,$$

with $\sigma_0(u) = Re(\sigma(u)) + |\sigma(u)|^2/2 + o(|\sigma(u)|)$, we obtain that

$$\|U[j_1, q_2]\overline{M}[\overline{\sigma}]x - U[j_1, q_2]x - M[\sigma_0]U[j_1, q_2]x\| \leq (\|\overline{E}\| + C2^{q_2-j_2+\Delta_2})\|x\| \ . \tag{A.29}$$

The second order scattering differences $F[q]$ are obtained from $U[j_1, q_2]\overline{M}[\overline{\sigma}]x - U[j_1, j_2]x$ with

$$F[q] = (U[q]\overline{M}[\overline{\sigma}]x - U[q]x) \star \phi_J \ . \tag{A.30}$$

We now define

$$L(J) = \inf_{|\beta|_\infty = 1} \sup_{j_1 < q_2 < j_2} \left\| \left(\left(\sum_n \beta[n]\Psi_{j_2}(t - n2^{j_2}) \right) U[j_1, q_2]x/\|x\| \right) \star \phi_J \right\|^2 \ . \tag{A.31}$$

This quantity is minimized by envelopes of the form

$$\sigma_0 = \sum_n \beta[n]\Psi_{j_2}(t - n2^{j_2}) \tag{A.32}$$

which are orthogonal to the functions

$$\{U[j_1, q_2]x\phi_J(k2^J - t)\}_{j_1 < q_2 < j_2, k} \ . \tag{A.33}$$

Since both ϕ_J and Ψ have compact support, for each k, the size of the support of $\phi_J(u-k2^J)$ is $C2^J$, and those of Ψ_{j_2} is $C'2^{j_2}$. As a consequence, on an interval of size $C2^J$, the envelope σ_0 has $\alpha 2^{J-j_2}$ coefficients which influence the support of $\phi_J(u-k2^J)$. Thus, we can divide the constraints

$$\{U[j_1,q_2]x\phi_J(k2^J-t)\}_{j_1<q_2<j_2,k}$$

into disjoint sets of size $(j_2-j_1)N_0$

$$\mathcal{I}_l = \{U[j_1,q_2]x\phi_J(k2^J-t)\}_{j_1<q_2<j_2,k\in\mathcal{I}(l)}\ ,|\mathcal{I}(l)|=N_0\ ,$$

according to the spatial position k, in such a way that for each set $\mathcal{I}_l$ we can find a subset of basis elements

$$\mathcal{I}'_l = \{\Psi_{j_2}(t-l'2^{j_2})\}_{l'\in\mathcal{I}'(l)}$$

of size $\alpha' 2^{J-j_2}$, with the property that each element of $\mathcal{I}'_l$ only intersects scaling functions $\phi_J(k2^J-t)$ within the set $\mathcal{I}_l$.

As a result, by renaming the position index n in (A.32), the orthogonality constraints for each fixed group $\mathcal{I}_l$ in (A.33) become

$$\int\left(\sum_{n=0}^{\alpha'2^{J-j_2}}\beta[n]\Psi_{j_2}(t-n2^{j_2})\right)U[j_1,q_2]x(t)\phi_J(k2^J-t)dt=0\ ,j_1<q_2<j_2,\ k\in\mathcal{I}(l)\ ,$$

$$\sum_{n=0}^{\alpha'2^{J-j_2}}\beta[n]\left(\int\Psi_{j_2}(t-n2^{j_2})U[j_1,q_2]x(t)\phi_J(k2^J-t)dt\right)=0\ ,j_1<q_2<j_2\ ,\ k\in\mathcal{I}(l)\ ,$$

$$\sum_{n=0}^{\alpha'2^{J-j_2}}\beta[n]\gamma_{q_2,k}[n]=0, j_1<q_2<j_2\ ,\ k\in\mathcal{I}(l)\ \text{(A.34)}$$

The constraints in (A.34) on the envelope are thus equivalent to finding a nonzero vector β of dimension $\alpha' 2^{J-j_2}$ orthogonal to a collection of $(j_2-j_1-1)N_0$ vectors $\gamma_{q_2,k}$. As a result, we can find a non-zero envelope σ_0^* of the form (A.32) with an error $L(J)=0$ in (A.31) as long as $\alpha' 2^{J-j_2}\geq (j_2-j_1)N_0$, which yields

$$J\geq j_2+\log_2(j_2-j_1)+\log(N_0)-\log_2(\alpha)\ .$$

Since the mapping $\sigma\mapsto\sigma_0$ is onto, we can consider and envelope σ such that $|1+\sigma| = 1+\sigma_0^*$.

We then obtain from (A.29)

$$\begin{aligned}
\|\tilde{F}_q\|^2 &= \sum_k (U[q]\overline{M}[\overline{\sigma}]x-U[q]x)\star\phi_J(k2^J)^2\\
&\leq \sum_k (U[q]\overline{M}[\overline{\sigma}]x-U[q]x-M[\sigma_0]U[q]x)\star\phi_J(k2^J)^2\\
&\leq \left\|U[q]\overline{M}[\overline{\sigma}]x-U[q]x-M[\sigma_0]U[q]x\right\|^2\\
&\leq \|x\|(K_2(\Delta_1)+C|\sigma|_\infty^2 2^{j_1-j_2+\Delta_1+\Delta_2}+C|\sigma|_\infty^2 2^{q_2-j_2+\Delta_2})\ ,
\end{aligned}$$

which proves (B.1). □.

Theorem A.3.3 shows that by localizing a modulation both in the scale of the envelope and in the scale of the carrier, the effect of such perturbation in the scattering domain is limited to a band of influence C_p, centered at a point which corresponds to the coordinates of the complex wavelet envelope. Although the result of theorem A.3.3 controls the absolute scattering difference $F_p = S_J\overline{M}[\sigma_p]f - S_J f$, in practice we observe that the relative difference $F/\|F\|$ has an asymptotic band behavior. Figure A.3 shows the operator $\widetilde{\nabla_{\overline{M}} S_{Jf}}$ obtained by normalizing each column of $\nabla_{\overline{M}} S_{Jf}$, for two different examples of f and for $J = \log N$. We use the ordering of Γ described in (A.13). In order to optimize the relative energy concentrated in the cone of influence, we draw 20 random envelopes of type p, we project each of them in the orthogonal space

$$(\oplus_{j_1 < q \leq j_2} U[q]f)^{\perp} = \{h \in (\mathbf{L}^1(\mathbb{R}))' \,;\, \int h(u)U[q]f(u)du = 0 \,, j_1 < q_2 \leq j_2\} \;,$$

and we retain those achieving maximum concentration in the band of influence C_p. The first row corresponds to a realization of white gaussian noise, whereas the second row is obtained from a NASDAQ stock price signal. The middle column shows the obtained operator setting $\Delta_1 = 0$, whereas the right column corresponds to $\Delta_1 = 1$. As expected, the relative energy is concentrated around the cone of influence. When $\Delta_1 = 0$, the impact on scattering coefficients having different first scale is minimized, as predicted by (A.15). For a white gaussian noise realization, the operator is virtually diagonal, whereas for the stock price signal the operator is nearly triangular with its energy concentrated along a band. White noise yields a nearly diagonal operator because the propagator $U[q]$ of $\overline{M}[\sigma_p]f$ is well approximated by the wavelet decomposition modulus of σ_p, thanks to the concentration of $U[q]f$ along its mean. This phenomena is not observed on sparser signals, such as that of (d).

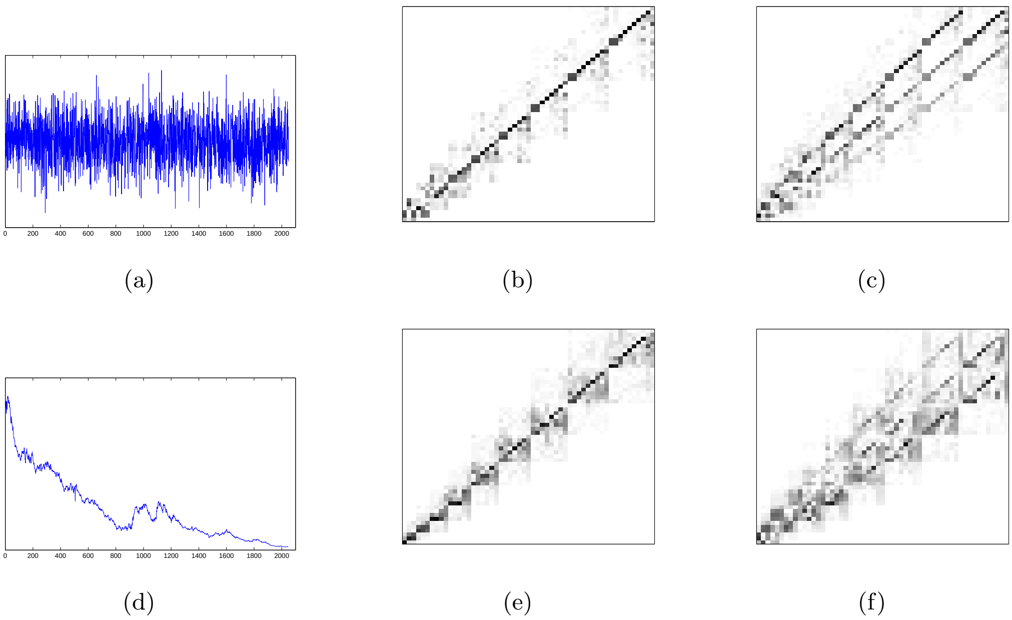


Figure A.3: Numerical simulations of the differential scattering operator $\widetilde{\nabla_{\overline{M}}} S_J f$ for two different signals. First row: (a) realisation of white gaussian noise; (b) $\widetilde{\nabla_{\overline{M}}} S_J f$ obtained with $\Delta_1 = 0$, (c) $\widetilde{\nabla_{\overline{M}}} S_J f$ obtained with $\Delta_1 = 1$. Second row: (d) stock price of NASDAQ:AAPL (e) $\widetilde{\nabla_{\overline{M}}} S_J f$ obtained with $\Delta_1 = 0$, (f) $\widetilde{\nabla_{\overline{M}}} S_J f$ obtained with $\Delta_1 = 1$.

Appendix B

Proof of Theorem 5.4.1

B.1 Proof of Lemma 5.4.2

Lemma B.1.1 *Let $Z_l(t) = 2^{l/2}|dX \star \psi| \star \psi_l(t)$, $l \in \mathbb{N}$, and let $\gamma_0 = \|\psi\|_2^2(\int R_{|dX\star\psi|}(\tau)d\tau)$, $\gamma_1 = \|\psi\|_2^2\| \int R_{|dX\star\psi|}\|_1$. Then γ_0 , $\gamma_1 < \infty$, and the sequence of random variables $Z_l(t)$ satisfies*

$$\forall l\,,\; E(|Z_l(t)|^2) \leq \gamma_1 \;,\; \textit{and} \; \lim_{l\to\infty} E(|Z_l(t)|^2) = \gamma_0 \;. \tag{B.1}$$

Proof: Let us compute the limit $\lim_{l\to\infty} E(|Z_l|^2)$ for $Z_l = 2^{l/2}|dX \star \psi| \star \psi_l$. If $dX(t)$ is a Gaussian white noise then we saw in Section 4.6 that $U[0]dX = |dX \star \psi|$ is a stationary Rayleigh process. Proposition 4.6.3 shows that its auto-correlation function $R_{U[0]dX}$ belongs to $\mathbf{L}^1$.

The spectral density of $Z_l = 2^{l/2}|dX \star \psi| \star \psi_l$ is given by

$$\widehat{R}_{Z_l}(\omega) = 2^l \widehat{R}_{U[0]dX}(\omega)|\hat{\psi}(2^l\omega)|^2 \;,$$

Since $R_{U[0]dX} \in \mathbf{L}^1$, its Fourier transform is continuous, which implies that

$$\begin{aligned}
\lim_{l\to\infty} E(|Z_l|^2) &= \lim_{l\to\infty} 2^l \int \widehat{R}_{Z_l}(\omega)d\omega = \lim_{l\to\infty} 2^l \int \widehat{R}_{U[0]dX}(\omega)|\hat{\psi}(2^l\omega)|^2 d\omega \\
&= \lim_{l\to\infty} \widehat{R}_{U[0]dX}(0) 2^l \int |\hat{\psi}(2^l\omega)|^2 d\omega \\
&= \widehat{R}_{U[0]dX}(0) \int |\hat{\psi}(\omega)|^2 d\omega = \|\psi\|_2^2 \int R_{U[0]dX}(\tau)d\tau \;,
\end{aligned} \tag{B.2}$$

since $|\hat{\psi}(2^l\omega)|$ concentrates towards the low frequencies as $l \to \infty$. In addition, we have

$$\begin{aligned}
E(|Z_l|^2) &= 2^l \int \widehat{R}_{U[0]dX}(\omega)|\hat{\psi}(2^l\omega)|^2 d\omega \\
&\leq \sup_{\omega} |\widehat{R}_{U[0]dX}(\omega)| 2^l \int |\hat{\psi}(2^l\omega)|^2 d\omega \\
&= \|R_{U[0]dX}\|_1 \|\psi\|_2^2 \;,
\end{aligned}$$

which completes the proof $\square$.

B.2 Proof of Lemma 5.4.3

Lemma B.2.1 *Let ψ be an analytic wavelet with fast decay and such that the zeros of $\hat{\psi}$ in $(0, \infty)$ form a discrete set. Then the sequence of stationary processes $Z_l = 2^{l/2}|dX \star \psi| \star \psi_l$, $l \in \mathbb{N}$, satisfy*

$$\forall\ t\ , \lim_{l \to \infty} Z_l(t) \xrightarrow{d} Z = Z_{(r)} + iZ_{(i)}\ , \tag{B.3}$$

where $Z_{(r)}$, $Z_{(i)} \sim \mathcal{N}(0, \sigma^2/2)$, $\sigma^2 = \lim_{l\to\infty} E(|Z_l|^2)$, and $\stackrel{d}{\to}$ denotes convergence in distribution.

Proof: This result will be proved using a version of the Central Limit theorem for strong mixing processes, that we briefly recall. Rosenblatt [Ros] introduced a notion of dependence of two σ-algebras of events $\mathcal{F}_1$, $\mathcal{F}_2$ where a probability measure is defined:

$$\alpha(\mathcal{F}_1, \mathcal{F}_2) = \sup_{A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2} |P(A_1 \cap A_2) - P(A_1)P(A_2)|\ .$$

If $X(t)$ is a stationary process, its α-mixing coefficient is defined as

$$\forall \tau \geq 0\,,\ \ \alpha_X(\tau) = \alpha(t, t + \tau) = \alpha(\mathcal{M}^t_{-\infty}, \mathcal{M}^{\infty}_{t+\tau})\ ,$$

where $\mathcal{M}^t_s$ denotes the σ-algebra of events generated by the quantities $X(u)$, $s \leq u \leq t$ [KR60]. If

$$\lim_{\tau \to \infty} \alpha_X(\tau) = 0\ , \tag{B.4}$$

then the process X has the strong mixing condition. We will use the following Central Limit theorem for strong mixing stationary sequences, first introduced by Rosenblatt [Ros], and later refined by several authors.

Lemma B.2.2 *([Ros]). Suppose $(X_k\,,\ k \in \mathbb{Z})$ is a strictly stationary sequence of random variables satisfying $EX_0 = 0$, $EX_0^2 < \infty$, having the strong mixing condition, and such that for some $\delta > 0$,*

$$E(|X_k|^{2+\delta}) < \infty\ ,\ \ \textit{and} \sum_{n \geq 1} \alpha_X(n)^{\delta/(2+\delta)} < \infty\ ,$$

then the partial sums $S_l = l^{-1/2} \sum_{0<k\leq l} X_k$ satisfy $\lim_{l\to\infty} Var(S_l) = \sigma^2 < \infty$. If moreover $\sigma^2 > 0$, then

$$S_l \stackrel{d}{\to} \mathcal{N}(0, \sigma^2)\ , (l \to \infty)\ , \tag{B.5}$$

where $\stackrel{d}{\to}$ stands for convergence in distribution.

We shall prove (B.3) by first considering for each l an approximation of the random variable $Z_l(t)$ obtained by discretizing $|dX \star \psi|$.

Let $m \in \mathbb{N}$, and $\Delta_m(X)$ denote the discretized process defined as

$$\forall t \ , \ \Delta_m(X)(t) = X\left(\frac{\lfloor tm + \theta \rfloor}{m}\right) \ , \tag{B.6}$$

where θ is a random offset. We consider the following approximation:

$$\forall t, \ Z_l(t) = Z_{l,m}(t) + \epsilon_{l,m}(t) \ , \tag{B.7}$$

where

$$Z_{l,m}(t) = 2^{l/2} \Delta_m(|dX \star \psi|) \star \psi_l(t) \ . \tag{B.8}$$

For each l, the approximation error $\epsilon_{l,m}(t)$ is a random variable which satisfies, for every $\delta > 0$,

$$\lim_{m \to \infty} \mathrm{Prob}(|\epsilon_{l,m}(t)| > \delta) = 0 \ . \tag{B.9}$$

Indeed, if we write

$$\epsilon_m(t) = |dX \star \psi|(t) - \Delta_m(|dX \star \psi|)(t) \ ,$$

then by definition of $\epsilon_{l,m}$ in (B.7,B.8) we have

$$\epsilon_{l,m}(t) = 2^{l/2} \epsilon_m \star \psi_l(t) \ .$$

We can write $\epsilon_m(t)$ as

$$\epsilon_m(t) = |dX \star \psi|(t) - \Delta_m(|dX \star \psi|)(t) \stackrel{l}{=} |X_1| - |X_2| \ , \tag{B.10}$$

where X_1 and X_2 are 0-mean Gaussian random variables, with a covariance given by

$$\Sigma = \begin{pmatrix} \Psi(0) & \Psi(\eta) \\ \Psi(\eta) & \Psi(0) \end{pmatrix} \ ,$$

where

$$\Psi(t) = \psi \star \tilde{\psi}(t) \ , \ \tilde{\psi}(t) = \psi(-t) \ , \tag{B.11}$$

and $\eta \leq m^{-1}$ is the distance from t to its closest quantifier $\frac{\lfloor tm \rfloor}{m}$. The random variable $\epsilon_m(t)$ satisfies

$$E(\epsilon_m(t)) = E(|dX \star \psi(t)|) - E\left(\left|dX \star \psi\left(\frac{\lfloor mt \rfloor}{m}\right)\right|\right) = 0 \ ,$$

since $|dX \star \psi|$ is stationary. Let us now compute its variance.

$$\begin{aligned} \mathrm{Var}(\epsilon_m(t)) &= E(|\epsilon_m(t)|^2) = 2(E(|X_1|^2) - E(|X_1|\,|X_2|)) \\ &= 2\Psi(0) - 2\Psi(0)\left(\frac{\pi}{2}\,{}_2F_1\left(-1/2, -1/2; 1; \frac{\Psi(\eta)}{\Psi(0)}\right)\right) \ , \end{aligned}$$

using again the correlation function of a Rayleigh process from Section 4.6. The hypergeometric function ${}_2F_1(-1/2, -1/2; 1, \cdot)$ is continuous with respect to its last argument, and such that ${}_2F_1(-1/2,-1/2;1,1) = 1$. It results that for all $\delta_0 > 0$ there exists M_0 such that for $m \geq M_0$

$$\forall t \ , \ \mathrm{Var}(\epsilon_m(t)) \leq \delta_0 \ . \tag{B.12}$$

We then use (B.12) to obtain a bound for the variance of $\epsilon_{l,m}(t)$. By making the offset θ of the sampling grid of Δ_m random, $\epsilon_m(t)$ is stationary, and hence

$$\mathrm{Var}(\epsilon_{l,m}(t)) = 2^l \int \hat{R}_{\epsilon_m}(\omega)|\hat{\psi}(2^l\omega)|^2 d\omega \ .$$

The autocorrelation $R_{\epsilon_m}(\tau)$ satisfies

$$R_{\epsilon_m}(\tau) \leq 2R_{|dX\star\psi|}(\tau) \leq C|R_{dX\star\psi}(\tau)|^2 \ , \tag{B.13}$$

thanks to the bound on the hypergeometric function of Propostion (4.6.3). As a result, $R_{\epsilon_m} \in \mathbf{L}^1$ and hence its spectral density is continuous and bounded by $\|R_{\epsilon_m}\|_1$. It results that

$$\begin{aligned} \mathrm{Var}(\epsilon_{l,m}(t)) &\leq \sup_\omega |\hat{R}_{\epsilon_m}(\omega)| 2^l \int |\hat{\psi}(2^l\omega)|^2 d\omega \\ &\leq \|R_{\epsilon_m}\|_1 \|\hat{\psi}\|_2^2 \ . \end{aligned}$$

Since $\mathrm{Var}(\epsilon_m(t)) = R_{\epsilon_m}(0) \leq Cm^{-1}$ and $\|R_{\epsilon_m}\|_1 \leq C\|R_{dX\star\psi}\|_2^2$, for each $\eta > 0$ we can find $T > 0$ such that

$$\int_{|\tau|\geq T} |R_{\epsilon_m}(\tau)| d\tau \leq \eta \ .$$

On the other hand,

$$\forall\, m \geq M_0 \, , \ \int_{|\tau|\leq T} |R_{\epsilon_m}(\tau)| d\tau \leq R_{\epsilon_m}(0) 2T \leq 2T\delta_0 \ .$$

It results that

$$\lim_{m\to\infty} \|R_{\epsilon_m}\|_1 = 0 \ ,$$

and hence that $\lim_{m\to\infty} \mathrm{Var}(\epsilon_{l,m}(t)) = 0$.

Thus, using Chebyshev's inequality, given $\delta > 0$, if $\sigma_m = \sqrt{E(|\epsilon_{l,m}(t)|^2)}$,

$$\mathrm{Prob}(|\epsilon_{l,m}(t)| > \delta) \ = \ \mathrm{Prob}\left(|\epsilon_{l,m}(t)| > \sigma_m \frac{\delta}{\sigma_m}\right) \leq \frac{\sigma_m^2}{\delta^2} \ ,$$

which converges to 0 as $m \to \infty$.

We will now prove that the sequence $Z_{l,m}(t)$, as $l \to \infty$, converges in distribution to a Normal random variable. For this purpose, the first step is to show that for each m, the discrete, stationary Gaussian process

$$Y_k = (dX \star \psi)(k\Delta_0 m^{-1}) \ , \ k \in \mathbb{Z} \ , \tag{B.14}$$

has the strong mixing condition. To see this, we use a result from Kolmogorov and Rozanov [KR60], which gives a sufficient condition for a discrete stationary Gaussian process to have the strong mixing condition:

Lemma B.2.3 *([KR60], th. 4) If X_k is a discrete, Gaussian, stationary process with finite energy, such that its spectral density $\hat{R}_X(e^{i\omega})$ is continuous and does not vanish for any $-\pi \leq \omega < \pi$, then X_k has the strong mixing condition (B.4).*

Since

$$R_Y[n] = E(Y_k Y_{k+n}^*) = E((dX \star \psi)(\Delta_0 k/m)\,(dX \star \psi)^*(\Delta_0(k+n)/m)) = R_{dX\star\psi}(\Delta_0 n/m)$$

and $R_{dX\star\psi}(\tau) = \Psi(\tau)$, using the definition from (B.11), it follows that the spectral density of Y_k is

$$\hat{R}_Y(\omega) \;=\; \sum_k \left|\hat{\psi}\left(\omega + \frac{mk}{\Delta_0 \pi}\right)\right|^2 \,. \tag{B.15}$$

Since $\psi \in \mathbf{L}^1$, its Fourier transform is continuous, which implies that $|\hat{\psi}|^2$ is also continuous. Moreover, since the zeros of $\hat{\psi}$ are isolated and $|\hat{\psi}(\omega)| \geq 0$, one can find Δ_0 such that the periodic spectrum $\hat{R}_Y(\omega)$ does not vanish for $-\pi \leq \omega < \pi$, which, by virtue of lemma B.2.3, implies that Y_k has the strong mixing condition.

In addition, [KR60] shows how to control the decay of the mixing coefficient $\alpha(\tau)$. If there exists an analytic function $\varphi(\omega)$ such that $\frac{|\hat{R}_Y(\omega)|}{\varphi(\omega)} \geq \epsilon > 0$ and the derivative $\left(\frac{|\hat{R}_Y(\omega)|}{\varphi(\omega)}\right)^{(l)}$ is uniformly bounded, then

$$\rho(\tau) \leq C\tau^{-l} \,.$$

Since ψ has fast decay, the spectrum $\hat{R}_Y(\omega)$ is C^∞, which implies that $\alpha(\tau)$ has fast decay by setting $\varphi = 1$.

Let us now see that if Y_k has the strong mixing condition, then its modulus $|Y|_k$ also enjoys the strong mixing condition. By definition, the mixing coefficient of Y_k is

$$\forall \tau \geq 0\,,\;\; \alpha_Y(\tau) = \alpha(\mathcal{M}_{-\infty}^t, \mathcal{M}_{t+\tau}^\infty) = \sup_{A_1 \in \mathcal{M}_{-\infty}^t, A_2 \in \mathcal{M}_{t+\tau}^\infty} |P(A_1 \cap A_2) - P(A_1)P(A_2)| \,,$$

where $\mathcal{M}_s^t$ denotes the σ-algebra of events generated by the quantities $Y(u)$, $s \leq u \leq t$. Since the σ-algebra of events $\tilde{\mathcal{M}}_s^t$ generated by $|Y(u)|$, $s \leq u \leq t$, is included in $\mathcal{M}_s^t$, it follows that

$$\begin{aligned}\alpha_{|Y|}(\tau) &= \alpha(\tilde{\mathcal{M}}_{-\infty}^t, \tilde{\mathcal{M}}_{t+\tau}^\infty) \\ &\leq \alpha(\mathcal{M}_{-\infty}^t, \mathcal{M}_{t+\tau}^\infty) = \alpha_Y(\tau) \,,\end{aligned}$$

which implies that $|Y|_k$ has the strong mixing condition.

Let us now see that for each m, the limit as $l \to \infty$ of $Z_{m,l}$ converges in distribution towards a gaussian random variable. For that purpose, we decompose the convolution by the wavelet ψ_l in (B.8) as a cascade of two convolutions, the first one implementing a low-pass filter which generates partial sums of $|Y|_k$. Since by definition

$$\Delta_m(|dX \star \psi|)(t) = \sum_n |Y|_n \mathbf{1}_{(-\Delta_0/m, \Delta_0/m)}(t - n\Delta_0/m) \,,$$

we have

$$Z_{m,l}(t) = 2^{l/2}\Delta_m(|dX \star \psi|) \star \psi_l(t) = \sum_n |Y|_n \beta_l(t - n\Delta_0/m) \ , \tag{B.16}$$

where $\beta_l(t) = 2^{l/2}(\mathbf{1}_{(\pm\Delta_0/m)} \star \psi_l)(t)$. Since $\hat{\psi}(0) = 0$, it results that

$$\forall t \ , |dX \star \psi| \star \psi_l(t) \stackrel{l}{=} (|dX \star \psi| - E(|dX \star \psi|)) \star \psi_l(t) \ ,$$

and we can replace $|Y|_k$ by the centered stationary sequence $\tilde{Y}_k = |Y|_k - E(|Y|_0)$ with the same mixing properties as $|Y|_k$. The lowpass filter g_n given by the partial sum

$$n^{-1/2} \sum_{|k-k'| \leq n/2} X_{k'} = X \star g_n[k]$$

has a Fourier transform $\hat{g}(e^{i\omega})$ given by the Dirichlet kernel

$$\hat{g}(e^{i\omega}) = n^{-1/2}\frac{\sin((n+1/2)\omega)}{\sin(\omega/2)} \ ,$$

which has its energy concentrated on the frequency interval $(-\frac{\pi}{n}, \frac{\pi}{n})$. Since the spectrum of β_l has its energy concentrated on the frequency range $(a_1 2^{-l-1}, a_2 2^{-l})$, then for any $n > 0$ and any $\epsilon_W > 0$, there exists a scale $l(n) > 0$ and a finite energy analytic filter h such that

$$Z_{m,l}(t) \stackrel{l}{=} (\tilde{Y} \star \beta_{l(n)})(t) = ((\tilde{Y} \star g_n) \star h)(t) + W \ , \tag{B.17}$$

with $E(|W|^2) \leq \epsilon_W$.

But $\tilde{Y}$ satisfies $E(\tilde{Y}) = 0$, $E(|\tilde{Y}|^2) < \infty$, $E(|\tilde{Y}|^4) < \infty$, and

$$\sum_n \alpha_{\tilde{Y}}(n)^{1/2} < \infty \ ,$$

since its mixing coefficients have fast decay. We can thus apply lemma B.2.2 with $\delta = 2$ to the partial sums $\tilde{Y} \star g_n$. The same argument we used in (B.2) can now be applied by replacing $\hat{\psi}_l$ with the Dirichlet kernel $\hat{g}_n$, to show that

$$\lim_{n\to\infty} \mathrm{Var}(\tilde{Y} \star g_n) = \left(\sum_\tau R_{\tilde{Y}}[\tau]\right) \|g\|_2^2 < \infty \ ,$$

and hence that, as $n \to \infty$, the partial sums $\tilde{Y} \star g_n(t)$ converge in distribution towards a Gaussian random variable.

Finally, let us see that this implies the convergence in distribution of Z_l towards a gaussian distribution. We use Slutsky's theorem, which states that

$$\left.\begin{array}{l} X_n \ \stackrel{d}{\to} X \\ Y_n \ \stackrel{P}{\to} 0 \end{array}\right\} \Longrightarrow X_n + Y_n \stackrel{d}{\to} X \ ,$$

where $\xrightarrow{d}$ and $\xrightarrow{P}$ stand respectively for convergence in distribution and convergence in Probability. Since the discretization error $\epsilon_{m,l}$ converges to 0 in probability as $m \to \infty$ uniformly in l, and (B.17) approximates each discretized $Z_{m,l}(t)$ with a linear transformation $(\tilde{Y} \star g_n) \star h$ of partial sums, with an error W converging to 0 in probability as $\epsilon_W \to 0$, we conclude by letting $m \to \infty$ and $\epsilon_W \to 0$ that $Z_l(t)$ converges in distribution to a complex Gaussian random variable. $\square$.

B.3 Proof of Lemma 5.4.4

Lemma B.3.1 *The sequence $Z_l = 2^{l/2}|dX \star \psi| \star \psi_l$ satisfies*

$$\forall\, t\,,\ \lim_{l\to\infty} E(|Z_l(t)|^r) = E(R^r)\ ,\ \textit{for } r = 1, 2\ , \tag{B.18}$$

where R follows a Rayleigh distribution.

Proof: This result is proved as a consequence of lemma 5.4.3, which lead to the convergence in distribution of (5.46):

$$\forall t\,,\ Z_l(t) \xrightarrow{d} R\ , (l \to \infty)\ ,$$

where R follows a Rayleigh distribution.

The proof of the Central limit theorem showed that $\lim_{l\to\infty} E(|Z_l|^2) = E(|R|^2)$. In order to see that the first moments also converge, we use a classic result on uniform integrability of sequences of random variables:

Lemma B.3.2 *([Das08], thm 6.1-6.2) Let X_l be a sequence of random variables. Suppose that for some $\delta > 0$,*

$$\sup_l E(|X_l|^{1+\delta}) < \infty\ ,$$

and that $X_l \xrightarrow{d} X$. Then X_l is uniformly integrable, and

$$\lim_{l\to\infty} E(X_l) = E(X)\ .$$

Lemma 5.4.2 showed in particular that $\sup_l E(|Z_l|^2) \leq \|\psi\|_2^2 \|R_{|dX\star\psi|}\|_1 < \infty$. Then, by setting $\delta = 1$ in lemma B.3.2, we conclude that $|Z_l|$ is uniformly integrable and hence that

$$\lim_{l\to\infty} E(|Z_l|) = E(|R|)\ \ \square.$$

References

[AAT07] Y. Amit, S. Allassonnière, and A. Trouvé. Towards a coherent Statistical framework for dense deformable template estimation. *Journal of the Royal Statistical Society*, 2007. 17

[AFTV00] P. Abry, P. Flandrin, M. Taqqu, and D. Veitch. Wavelets for the Analysis, estimation and synthesis of scaling data. *K. Park and W. Willinger eds*, 2000. 106

[AJ04] A. Arneodo and S. Jaffard. L'Analyse Multi-fractale des Signaux. *Journal du CNRS*, 2004. 107

[ALJ04] P. Abry, B. Lashermes, and S. Jaffard. Revisiting scaling, Multifractal and Multiplicative cascades with the wavelet leader lens. *SPIE*, 2004. 114

[AM12a] J. Anden and S. Mallat. Scattering Audio Representations. *submitted to IEEE Trans on Signal Processing*, 2012. 48

[AM12b] J. Anden and S. Mallat. Scattering Representations for Audio Recognition. *Manuscript in Preparation*, 2012. 92

[ANVV97] T.F. Andre, R.D. Nowak, and B.D. Van Veen. Low-rank estimation of higher order statistics. *Signal Processing, IEEE Transactions on*, 45(3):673 –685, mar 1997. 75

[AT07] Y. Amit and A. Trouvé. Pop: Patchwork of Parts Models for Object Recognition. *International Journal of Computer Vision*, 2007. 62, 65

[BBLP10] Y-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning Midlevel Features for Recognition. *IEEE CVPR*, 2010. 35, 44, 68

[Ber99] Ch. Bernard. *Wavelets and Ill-posed Problems: Optic Flow and Scattered Data Interpolation*. PhD thesis, CMAP, Ecole Polytechnique, 1999. 17

[BETVG08] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. SURF: Speeded Up Robust Features. *Computer Vision and Image Understanding*, 2008. 20

[BGV92] B.E. Boser, I.M. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. *Fifth Annual Workshop on Computational Learning Theory*, 1992. 55

[BJ03] R. Basri and D. Jacobs. Lambertian Reflectance and Linear Subspaces. *IEEE Trans on PAMI*, 2003. 79

[BKM08a] E. Bacry, A. Kozhemyak, and J-F. Muzy. Continuous Cascade Models for Asset Returns. *Journal of Economic Dynamics and Control*, 2008. 10, 108, 109, 110, 113, 114, 136, 147

[BKM08b] E. Bacry, A. Kozhemyak, and J-F. Muzy. Log-Normal Continuous Cascades: Aggregation properties and estimation. *Quantitative Finance*, 2008. 109, 113, 114, 151, 152

[BL08] P.J. Bickel and E. Levina. Covariance Regularisation by Thresholding. *Annals of Statistics*, 2008. 60

[BM97] L. Birge and P. Massart. From Model Selection to Adaptive Estimation. *Festchrift for Lucien Le Cam: research papers in Probability and Statistics*, 1997. 59

[Bro05] R.E. Broadhurst. Statistical Estimation of Histogram Variation for Texture Classification. *Proc Workshop on Texture Analysis and Synthesis*, 2005. 79, 80

[BRP09] J. Bouvrie, L. Rosasco, and T. Poggio. On Invariance in Hierarchical Models. *NIPS*, 2009. 44

[Bur98] C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 1998. 7, 54

[BZM07] A. Bosch, A. Zisserman, and X. Muñoz. Image Classification using Random Forests and Ferns. *ICCV*, 2007. 20

[CG10] M. Crosier and L. Griffin. Using Basic Image Features for Texture Classification. *International Journal of Computer Vision*, 2010. 79, 80

[CL11] C. Chang and C. Lin. Libsvm: a library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2011. 62, 63

[CLP87] P. Collet, J. Lebowitz, and A. Porzio. The dimension spectrum of some dynamical systems. *Journal of Statistical Physics*, 1987. 107

[CV95] C. Cortes and V. Vapnik. Support Vector Networks. *Machine Learning*, 1995. 54

[Das08] A. DasGupta. *Asymptotic Theory of Statistics and Probability*. Springer, 2008. 180

[DDS+09] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 67

[DMM01] A. Desolneux, L. Moisan, and J-M. Morel. Edge Detection by Helmholtz Principle. *Journal of Mathematical Imaging and Vision*, 14:271–284, 2001. 60

[Dou06] P. Doukhan. *Processus Empiriques et Séries Temporelles*. 2006. 95, 111

[DR87] W. B. Davenport and W. L. Root. *An Introduction to the Theory of Random Signals and Noise*. Piscatway, NJ, 1987. 93

[DT05] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. *IEEE CVPR*, 2005. 20

[DVGNK99] K.J. Dana, B. Van-Ginneken, S.K. Nayar, and J.J. Koenderink. Reflectance and Texture of Real World Surfaces. *ACM Transactions on Graphics (TOG)*, 18(1):1–34, Jan 1999. 8, 72, 103

[EF01] A. Efros and B. Freeman. Image Quilting for Texture Synthesis and Transfer. *SIGGRAPH*, 2001. 77

[EVGW+] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 67

[FFFP04] L. Fei-Fei, R. Fergus, and P. Perona. Learning Generative Visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. *IEEE CVPR*, 2004. 46, 67

[FKL10] C. Farabet, K. Kavukcuoglu, and Y. LeCun. Convolutional Networks and Applications in Vision. *Proc of ISCAS*, 2010. 2, 7, 13, 20, 44, 65

[Fri95] U. Frisch. *Turbulence*. Cambridge University Press, 1995. 107

[Gab46] D. Gabor. Theory of Communication. *Journal of the Institute of Electronic Engineers*, 1946. 78

[GBP88] P. Grassberger, R. Badii, and A. Politi. Scaling Laws for Invariance Measures on hyperbolic and non-hyperbolic attractors. *Journal of Statistical Physics*, 1988. 107

[GG84] S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on PAMI*, 1984. 75

[Gre93] U. Grenader. *General Pattern Theory*. Oxford Science Publications, 1993. 16

[HB95] D. Heeger and J. Bergen. Pyramid Based Texture Analysis/Synthesis. *SIGGRAPH*, 1995. 75

[HCFE04] E. Hayman, B. Caputo, M. Fritz, and J.O. Eklundh. On the significance of Real-World Conditions for Material Classification. *ECCV*, 2004. 79, 80

[HK76] P. Huber and B. Kleiner. Statistical Methods for investigating phase relations in stochastic processes. *IEEE Trans on Audio and Electroacoustics*, 1976. 75

[HK02] B. Haasdonk and D. Keysers. Tangent Distance Kernels for Support Vector Machines. *IEEE International Conference on Pattern Recognition*, 2002. 58, 65

[HS81] B. Horn and B. Schunck. Determining Optical Flow. *Artificial Intelligence*, 1981. 17

[Jaf97] S. Jaffard. Multifractal Formalism for functions, parts i and ii. *SIAM J. Math Anal.*, 1997. 10, 108, 114

[Jaf00] S. Jaffard. Wavelet Expansions, Function Spaces and Multifractal Analysis. 2000. 106

[JM96] S. Jaffard and Y. Meyer. Wavelet Methods for Pointwise Regularity and Local Oscillations of Functions. *Memoirs of the AMS*, 1996. 107

[Jul62] B. Julesz. Visual Pattern Discrimination. *IRE Trans. Info Theory*, 1962. 8, 71

[Jul81] B. Julesz. Textons, the elements of texture perception and their interactions. *Nature*, 1981. 71

[KDGH07] D. Keysers, T. Deselaers, C. Gollan, and N. Hey. Deformation Models for Image Recognition. *IEEE trans of PAMI*, 2007. 19, 62

[KEBK05] V. Kwatra, I. Essa, A. Bobick, and N. Kwatra. Texture Optimization for Example-based Synthesis. *SIGGRAPH*, 2005. 77

[KM00] M. Kayid Mohamed. *Convergence of Random Variables*. Chapman and Hall, 2000. 123

[KR60] A.N. Kolmogorov and Y.A. Rozanov. On Strong Mixing conditions for Stationary Gaussian Processes. *Theory of Probability and its Applications*, 5:204–208, 1960. 175, 177, 178

[KS04] Y. Ke and R. Sukthankar. PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. *CVPR*, 2004. 20

[KSH12] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet Classification with Deep Convolutional Neural Networks. *NIPS (to appear)*, 2012. 68

[Kyp07] A. Kyprianou. *An Introduction to the Theory of Lévy Processes.* 2007. 112

[LBBH98] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998. 20, 21

[LBLL09] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin. Exploring Strategies for Training Deep Neural Networks. *Journal of Machine Learning Research*, 2009. 66

[LH05] S. Lefebvre and H. Hoppe. Parallel Controllable Texture Synthesis. *SIGGRAPH*, 2005. 77

[LM01] T. Leung and J. Malik. Representing and Recognizing the Visual Appearance of Materials Using Three-dimensional Textons. *IJCV*, 2001. 8, 75, 79

[LMC86] M.T. Landahl and E. Mollo-Christensen. *Turbulence and Random processes in fluid Mechanics.* Cambridge University Press, 1986. 152

[Low04] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vision*, 60(2), November 2004. 1, 19, 20, 44

[LPT12] J. Linderstrauss, D. Preiss, and J. Tise. *Frechet Differentiability of Lipschitz Functions and Porous Sets in Banach Spaces.* Princeton University Press, 2012. 57, 84

[LSP06] S. Lazebnik, C. Schmidt, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. *IEEE CVPR*, 2006. 68

[Mal08] S. Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, New York, 2008. 43, 85, 115, 131

[Mal12] S. Mallat. Group Invariant Scattering. *Communications in Pure and Applied Mathematics (to appear)*, 2012. 2, 3, 5, 13, 14, 18, 21, 26, 28, 29, 35, 44, 59, 82, 110, 123, 161, 165

[Man74] B.B. Mandelbrot. Intermittent Turbulence in self-similar Cascades: divergence of high moments and dimension of the carrier. *Journal of Fluid Mechanics*, 1974. 107, 113, 120

[MBA93a] J-F. Muzy, E. Bacry, and A. Arneodo. Multifractal Formalism for fractal signals: The structure-function approach versus the wavelet-transform modulus-maxima method. *Phys. Rev*, 1993. 107, 108

[MBA93b] J-F. Muzy, E. Bacry, and A. Arneodo. Singularity Spectrum of Fractal signals from wavelet analysis: exact results. *J. Stat. Phys*, 1993. 108

[MBP10] J. Mairal, F. Bach, and J. Ponce. Task Driven Dictionary Learning. *IEEE Pami*, 2010. 62

[McD] J. McDermott. Auditory textures. 72, 82, 83, 90, 92, 94

[Mer76] P. Mermelstein. Distance Measures for Speech Recognition, psychological and instrumental. *Pattern Recognition and Artificial Intelligence*, 1976. 1

[MNY06] H. Minh, P. Niyogi, and Y. Yao. Mercer's Theorem, Feature Maps and Smoothing. *Proc. of Computational Learning Theory*, 2006. 16

[MP90] J. Malik and P. Perona. Preattentive Texture Discrimination with early Vision Mechanisms. *J Opt Soc Am*, 1990. 44, 75

[MS91] C. Meneveau and K.R. Sreenivasan. The Multifractal nature of Turbulent energy dissipation. *J. Fluid Mechanichs*, 1991. 107, 152

[MS11] J. McDermott and E. Simoncelli. Sound Texture Perception via statistics of the auditory periphery: Evidence from Sound Synthesis. *Neuron*, 2011. 9, 72, 76, 84, 91

[MY01] I. Miller and L. Younes. Group Actions, Diffeomorphisms and Matching: a general framework. *IJCV*, 2001. 16

[NJ02] A.Y. Ng and M. Jordan. On Discriminative vs Generative classifiers: A comparison of logistic regression and naive Bayes. *Advances in Neural Information Processing Systems (NIPS)*, 2002. 57, 61, 62, 63

[OW00] M. Ossiander and E.C. Waymire. Statistical Estimation for Multiplicative Cascades. *Annals of Statistics*, 2000. 114

[Pap07] A. Papapantoleon. *An Introduction to Lévy Processes with applications in Finance*. 2007. 112

[Per10] L. Perrinet. Role of Homeostasis in Learning Sparse Representations. *Neural Computation Journal*, 2010. 55

[Pet99] A. Petropulu. *Higher-Order Spectral Analysis*. CRC Press, 1999. 73

[Pey09] G. Peyré. Sparse Modelling of Textures. *Journal of Mathematical Imaging and Vision*, 2009. 77

[PF85] G. Parisi and U. Frisch. Fully Developed Turbulence and Intermittency. *Turbulence and Predictability in Geophysical Fluid Dynamics*, 1985. 107

[PL98] R. Paget and I.D. Longstaff. Texture Synthesis via a Noncausal nonparametric multiscale Markov Random Field. *IEEE Transactions on Image Processing*, 1998. 75

[Pol84] D. Pollard. *Convergence of Stochastic Processes*. Springer Verlag, 1984. 129

[PS99] J. Portilla and E. Simoncelli. Texture Modeling and Synthesis using Joint Statistics of Complex Wavelet Coefficients. *IEEE workshop on statistical and computational theories of vision*, 1999. 44, 76, 84, 91

[RACW10] I. Rekik, S. Allassonnière, T. Carpenter, and J. Wardlaw. A Review Of Medial Image Analysis Methods in mr/ct-imaged Acute-subacute Ischemic Stroke Lesion. *NeuroImage Clinical*, 2010. 17

[RBL07] M.A. Ranzato, Y-L. Boureau, and Y. LeCun. Sparse Feature Learning for Deep Belief Networks. *NIPS*, 2007. 21

[RHBL07] M.A. Ranzato, F. Huang, Y-L. Boureau, and Y. LeCun. Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition. *IEEE CVPR*, 2007. 62

[Ros] M. Rosenblatt. A Central limit Theorem and a Strong Mixing condition. 175

[Sch07] S.M. Schimmel. *Theory of Modulation Frequency Analysis and Modulation Filtering, with Applications to Hearing Devices*. PhD thesis, "University of Washington", 2007. 78

[SM12] L. Sifre and S. Mallat. Combined Scattering for Rotation Invariant Texture Analysis. *ESANN*, 2012. 44, 68, 81

[Soa09] S. Soatto. Actionable Information in Vision. *ICCV*, 2009. 19

[SS09] Arthur Szlam and Guillermo Sapiro. Discriminative k-metrics. In *ICML*, page 127, 2009. 57, 58

[STC04] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004. 15

[TLF10] E. Tola, V. Lepetit, and P. Fua. DAISY: An Efficient Dense Descriptor Applied to WIde-Baseline Stereo. *IEEE trans on PAMI*, 2010. 20, 44

[Tro95] A. Trouvé. Diffeomorphism Groups and Pattern Matching in Image Analysis, 1995. 17

[Tur10] R.E. Turner. *Statistical Models for Natural Sounds*. PhD thesis, "University of London", 2010. 78

[TY05] A. Trouvé and L. Younes. Local Geometry of Deformable Templates, 2005. 17

[UB05] I. Ulusoy and C. Bishop. Comparison of Generative and Discriminative Techniques for Object Detection and Classification. 2005. 61

[Vap79] V. Vapnik. Estimation of Dependences Based on Empirical Data. *Nauka, Moscow*, 1979. 54

[Vap98] V. Vapnik. *Statistical Learning Theory.* John Wiley and Sons, 1998. 61

[VMYT04] M. Vaillant, M. Miller, L. Younes, and A. Trouvé. Statistics on Diffeomorphisms via Tangent Space Representations. *NeuroImage*, 2004. 17

[VZ03] M. Varma and A. Zisserman. Texture Classification: are Filter banks necessary? *CVPR*, 2003. 75

[VZ09] M. Varma and A. Zisserman. A Statistical Approach to Material Classification using Image Patch Exemplars. *IEEE Trans on PAMI*, 2009. 79, 80

[WAJZ09] H. Wendt, P. Abry, S. Jaffard, and H.Ji Z.Shen. Wavelet leader multifractal analysis for texture classification. *IEEE ICIP*, 2009. 114

[Wal12] I. Waldspurger. Recovering the phase of a complex wavelet transform, 2012. 47, 90

[Wor95] G.W. Wornell. *Signal Processing with Fractals: A Wavelet based Approach.* Prentice-Hall, 1995. 111

[YGZ09] K. Yu, Y. Gong, and T. Zhang. Nonlinear Learning using Local Coordinate Coding. *NIPS*, 2009. 68

[YYGH09] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear Spatial Pyramid matching using Sparse Coding for Image Classification. *IEEE CVPR*, 2009. 68

[ZFCV02] A. Zalesny, V. Ferrari, G. Caenen, and L. VanGool. Parallel Composite Texture Synthesis. 2002. 75

[ZWM97] S. Zhu, Y. Wu, and D. Mumford. Minimax Entropy Principle and Its Application to Texture Modeling. *Neural Computation*, 1997. 77, 91