DENOISING WITH INFINITE MIXTURE OF GAUSSIANS

Teodor Iulian Alecu, Sviatoslav Voloshynovskiy and Thierry Pun

Computer Vision and Multimedia Laboratory, University of Geneva, 24 Rue Général-Dufour, 1204 Geneva, Switzerland
phone: + (41) 22 379 1084, fax: + (41) 22 379 7780, email: Teodor.Alecu@cui.unige.ch

ABSTRACT

We show in this paper how an Infinite Mixture of Gaussians model can be used to estimate/denoise non-Gaussian data with local linear estimators based on the Wiener filter. The decomposition of the data in Gaussian components is straightforwardly computed with the Gaussian Transform, previously derived in [2]. The estimation is based on a two-step procedure, the first step consisting in variance estimation, and the second step in data estimation through Wiener filtering. We propose new generic variance estimators based on the Infinite Gaussian Mixture prior such as the cumulative estimator or the local-global estimator, as well as more classical Bayesian estimators. Results are presented in terms of distortion for the case of Generalized Gaussian data.

1. INTRODUCTION

Gaussian distributions are extensively used in the (broad sense) signal processing community, mainly for computational benefits. For instance, in an estimation/denoising problem Gaussian priors yield quadratic functionals and linear solutions. However, real data is most often non–Gaussian distributed, and is best described by other types of distribution (e.g in image processing, the most commonly used model for the wavelet coefficients distribution is the Generalized Gaussian distribution [6]).

In order to preserve the computational advantages provided by Gaussian modelling, in a number of papers (such as [4] and [5]), non-Gaussian distributions are approximated with finite Gaussian mixtures obtained through iterative numerical optimization techniques. We extend this approach and propose in this paper new generic denoising algorithms for non-Gaussian data based on the description of non-Gaussian distributions as an Infinite Mixture of Gaussian (IMG), following the work from [2]. Indeed, the Gaussian Transform, as presented therein, permits straightforward derivation of exact Infinite Gaussian Mixtures for a wide range of symmetric distributions, including the Generalized Cauchy and the Generalized Gaussian distributions.

The paper is organized as follows. Section 2 presents the denoising framework used throughout the paper, whereas Section 3 introduces the specific cumulative estimation scheme. The last sections propose Gaussian Transform based estimators in both point to point and local estimation schemes, and exemplify them for the Generalized Gaussian Distribution (GGD) family. The obtained results are compared in terms of distortion with the state-of-the-art (for the GGD) denoising method of Moulin & Liu [1].

2. DENOISING FRAMEWORK

We are interested in the generic denoising problem of estimating the original vector data **x** from the measured data **y**, degraded by additive noise **z**:

$$\mathbf{y} = \mathbf{x} + \mathbf{z} \,. \tag{1}$$

We consider the components of the random vector variables to be independent identically distributed (i.i.d.), and we denote them by the index k (e.g. $x[k]$). We omit the index for simplicity reasons whenever the context allows it.

We assume in this paper that the noise random variable Z (whose successive realizations form the vector **z**) is zero-mean Gaussian distributed with variance σ_Z^2:

$$p_Z(z) = \frac{1}{\sqrt{2\pi\sigma_Z^2}} e^{-\frac{z^2}{2\sigma_Z^2}} . \tag{2}$$

We further assume that the probability density function $p_X(x)$, which describes the source random variable X, is a zero-mean symmetric distribution, and that its Gaussian Transform $G_X(\sigma^2) = \mathcal{G}(p_X(x))$ exists [2]. We recall that the Gaussian Transform $\mathcal{G}$ describes a symmetric distribution as an infinite mixture of Gaussians, and that G_X satisfies:

$$\int_0^\infty G_X(\sigma^2)\mathcal{N}(x \mid \sigma^2) d\sigma^2 = p_X(x), \tag{3}$$

where $\mathcal{N}(x \mid \sigma^2)$ is the zero-mean Gaussian distribution:

$$\mathcal{N}(x \mid \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-x^2/2\sigma^2} .$$

For exemplification of the proposed techniques we consider $p_X(x)$ to be the Generalized Gaussian distribution:

$$p^G_{X|\gamma,\sigma_\gamma}(x \mid \gamma, \sigma_\gamma) = \frac{\gamma\eta(\gamma)}{2\Gamma\left(1/\gamma\right)} \frac{1}{\sigma_\gamma} e\left(-\left(\eta(\gamma) \left| \frac{x}{\sigma_\gamma} \right| \right)^\gamma \right), \tag{4}$$

where $\eta(\gamma) = \sqrt{\Gamma\left(3/\gamma\right)\Gamma\left(1/\gamma\right)^{-1}}$. For $\gamma = 1$ the GGD particularizes to the Laplacian distribution, while for $\gamma = 2$ one obtains the Gaussian distribution. The Gaussian Transform of the GGD is computed accordingly to [2]:

$$G_{X|\gamma,\sigma_\gamma}(\sigma^2) = \frac{1}{\sigma^2}\sqrt{\frac{\pi}{2\sigma^2}} \left(\mathcal{F}^{-1}\left(p^G_{X|\gamma,\sigma_\gamma}\left(\sqrt{i\omega}\right)\right)\right)_{t=\frac{1}{2\sigma^2}} ,$$

$$p^{G}_{X|\gamma,\sigma_\gamma}\left(\sqrt{i\omega}\right)=\frac{\gamma\eta(\gamma)}{2\Gamma\left(1/\gamma\right)}\frac{1}{\sigma_\gamma}e\left(-\left(\frac{\eta(\gamma)}{\sigma_\gamma}\right)^{\gamma}(i\omega)^{\gamma/2}\right),$$

where $\mathcal{F}^{-1}$ is the inverse Fourier Transform, *i* is the complex $\sqrt{-1}$ and ω is the variable in the Fourier space.

We base our estimation on the *Maximum a Posteriori* (MAP) principle:

$$\hat{x}=\arg\max_x p_{X|Y}(x\,|\,y),$$

which yields for i.i.d. zero-mean Gaussian data and noise (with variances σ_X^2 and σ_Z^2) the classical Wiener filter:

$$\hat{x}=\frac{\sigma_X^2}{\sigma_X^2+\sigma_Z^2}y. \tag{5}$$

We set the IMG estimation scheme into a local estimation paradigm. Let $\mathbf{y}_k$ be the local vector of samples with cardinality *m* and centred on the point *k*, where the estimation is performed (e.g. in image processing $\mathbf{y}_k$ would be the samples collected from a square window around the point of interest). According to the IMG model, each data sample is interpreted to be drawn from a Gaussian distribution with variance $\sigma_X^2[k]$ distributed accordingly to the Gaussian Transform $G_X\left(\sigma^2\right)$. Consequently, we model the data as locally Gaussian, estimate the variance from $\mathbf{y}_k$ and use Wiener filtering on the local level (5). Thus, our estimator is given by:

$$\hat{x}[k]=\frac{\sigma_X^2(\mathbf{y}_k)}{\sigma_X^2(\mathbf{y}_k)+\sigma_Z^2}y[k]. \tag{6}$$

The initial problem is now replaced with a variance estimation problem. To this purpose, we begin by introducing the cumulative estimation technique.

3. CUMULATIVE ESTIMATION

We are looking for an estimator that provides variance estimates consistent with the original data distribution, in the sense that the description of the data as a Gaussian mixture should be identical to the non-Gaussian original description. Then the probability density function of the estimated variances needs to be identical to the mixing function that reproduces $p_X(x)$, which is the Gaussian Transform G_X.

Moreover, since the distributions $p_X(x)$, and $p_Z(z)$ are symmetric and zero-mean, we assume an estimator of the form:

$$\hat{\sigma}_X^2(\mathbf{y}_k)=\xi^{\mathcal{M}}_{\sigma_X^2}\left(\mathcal{M}(\mathbf{y}_k)\right),$$

where $\xi^{\mathcal{M}}_{\sigma_X^2}$ denotes the consistent estimator depending on the symmetric real non-negative function $\mathcal{M}(\mathbf{y}_k)$:

$$\mathcal{M}:\mathbb{R}^m\mapsto\mathbb{R}^{0+};\ \mathcal{M}(\mathbf{y}_k)=\mathcal{M}(-\mathbf{y}_k)\ .$$

As an example, a possible $\mathcal{M}$ function is the L_2 (Euclidean) norm of $\mathbf{y}_k$ (see Section 4). We further assume that this estimator is monotonically increasing with the function $\mathcal{M}$:

$$\mathcal{M}(\mathbf{y}_{k1})<\mathcal{M}(\mathbf{y}_{k2})\Rightarrow\xi^{\mathcal{M}}_{\sigma_X^2}\left(\mathcal{M}(\mathbf{y}_{k1})\right)<\xi^{\mathcal{M}}_{\sigma_X^2}\left(\mathcal{M}(\mathbf{y}_{k2})\right). \tag{7}$$

The assumption in (7) implies that the cumulative distribution function of the variance estimates is equal to the cumulative distribution function of $\mathcal{M}$:

$$p\left(\hat{\sigma}_X^2(\mathbf{y}_k)\le\hat{\sigma}_X^2(\mathbf{y}_{k0})\right)=p\left(\mathcal{M}(\mathbf{y}_k)\le\mathcal{M}(\mathbf{y}_{k0})\right). \tag{8}$$

For continuous probability density functions, and incorporating the consistency condition (the distribution of the estimates is G_X), (8) can be rewritten as:

$$\begin{aligned}&P_{\mathcal{M}}\left(\mathcal{M}(\mathbf{y}_k)\right)=PG_X\left(\xi^{\mathcal{M}}_{\sigma_X^2}\left(\mathcal{M}(\mathbf{y}_k)\right)\right),\\ &P_{\mathcal{M}}\left(\mathcal{M}(\mathbf{y}_k)\right)=\int_0^{\mathcal{M}(\mathbf{y}_k)}p_{\mathcal{M}}(u)du;\ PG_X\left(\sigma_X^2\right)=\int_0^{\sigma_X^2}G_X(u)du,\end{aligned} \tag{9}$$

where $p_{\mathcal{M}}$ is the probability density function describing $\mathcal{M}(\mathbf{y}_k)$, and $P_{\mathcal{M}}$ and PG_X are cumulative probability functions, as defined in (9). Then the estimator we are looking for is simply given by:

$$\xi^{\mathcal{M}}_{\sigma_X^2}\left(\mathcal{M}(\mathbf{y}_k)\right)=PG_X^{-1}\left(P_{\mathcal{M}}\left(\mathcal{M}(\mathbf{y}_k)\right)\right), \tag{10}$$

with PG_X^{-1} the mathematical inverse of the function PG_X:

$$\text{if } PG_X\left(\sigma^2\right)=u \text{ then } PG_X^{-1}(u)=\sigma^2\ .$$

We denote (10) as the *cumulative estimator*, simply because it uses cumulative probability functions for estimation.

We can further infer that the cumulative estimator is robust with respect to monotonic increasing transformations of $\mathcal{M}$.

Theorem 1. If $\mathcal{M}_1$ and $\mathcal{M}_2$ are two symmetric real non-negative functions $\mathcal{M}_1,\mathcal{M}_2:\mathbb{R}^m\mapsto\mathbb{R}^{0+}$, and if

$$\mathcal{M}_1(\mathbf{y}_{k1})<\mathcal{M}_1(\mathbf{y}_{k2})\Leftrightarrow\mathcal{M}_2(\mathbf{y}_{k1})<\mathcal{M}_2(\mathbf{y}_{k2}),$$

then the associated cumulative estimators are identical:

$$\xi^{\mathcal{M}_1}_{\sigma_X^2}\left(\mathcal{M}_1(\mathbf{y}_k)\right)=\xi^{\mathcal{M}_2}_{\sigma_X^2}\left(\mathcal{M}_2(\mathbf{y}_k)\right).$$

Theorem 1 can be proven using (10). Indeeed, since the inequalities are preserved (theorem hypothesis), then the cumulative distribution functions $P_{\mathcal{M}_1}$ and $P_{\mathcal{M}_2}$ are equal (8).

4. POINT TO POINT ESTIMATION

In point to point (scalar) estimation, the local vector of samples reduces to one data sample, which we simply denote as $\mathbf{y}_k=y[k]=y$. The cumulative estimator (10) can be directly used in (6) by setting $\mathcal{M}(y)=|y|$. The point to point estimator (IMG PP) is then simply given by:

$$\begin{aligned}&\hat{x}_{IMG\_PP}=\frac{\sigma^2_{X\_PP}(y)}{\sigma^2_{X\_PP}(y)+\sigma_Z^2}y \text{ with}\\ &\sigma^2_{X\_PP}(y)=PG_X^{-1}\left(P_{\mathcal{M}}(y)\right),\ P_{\mathcal{M}}(y)=2\int_0^{|y|}p_Y(u)du.\end{aligned} \tag{11}$$

As *x* and *z* are assumed to be independent, $p_Y(y)$ can be computed as the convolution of $p_X(x)$ and $p_Z(z)$:

$$p_Y=p_X*p_Z\ .$$

In the general case the operations in (11) have to be performed numerically. However, if the distribution $p_X(x)$ is Laplacian (GGD with $\gamma=1$), its Gaussian Transform is [2]:

$$G^G_{X|\gamma,\sigma_\gamma}\left(\sigma^2\right)=\frac{1}{\sigma_\gamma^2}e^{-\frac{\sigma^2}{\sigma_\gamma^2}}, \tag{12}$$

and the cumulative estimator from (11) reduces to:

$$\sigma^2_{X\_PP,\gamma=1}(y) = -\sigma_\gamma^2 \ln\left(1 - P_{\mathcal{M}}(y)\right). \quad (13)$$

Also, as γ tends to 2, the Gaussian Transform of the GGD asymptotically tends to a Dirac function centred in σ_γ^2 [2], and the estimator (11) tends to the classical Wiener filter (5).

In order to asses the performances of the proposed estimator, we measure the distortion of the estimated data as the quadratic error:

$$e(x,\hat{x}) = (x - \hat{x})^2 .$$

The global distortion induced by an estimator is therefore:

$$D = \iint e(x,\hat{x}) p_X(x) p_Z(z)\, dxdz . \quad (14)$$

We compare our estimator in terms of distortion (Fig. 2) with the estimators known as shrinkage functions from [1]. The idea of shrinking originates from [3], but [1] provides both improved results and statistical interpretation of the functions. In fact, the shrinkage functions from [1] are the result of direct MAP, which in the case of GGD data and Gaussian noise yields $\hat{x}$ as the solution of the equation in x:

$$y = x - \sigma_Z^2 \sigma_\gamma^{-\gamma} \gamma \eta(\gamma) x^{\gamma-1} . \quad (15)$$

Equation (15) has analytical solutions for $\gamma = 0.5$ (cubic equation), $\gamma = 1$ or 2 (linear equation) and $\gamma = 1.5$ (quadratic equation). The resulting shrinkage functions have thresholding characteristics for $\gamma \le 1$ and yield the Wiener filter for $\gamma = 2$. Graphic illustration of the shrinkage functions and of the equivalent IMG PP functions obtained via (11) are presented in Fig. 1.

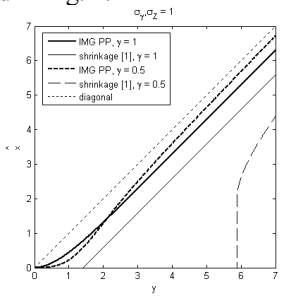

Figure 1. Shrinkage functions and IMG PP estimation curves

One can also compute the lower bound on the distortion induced by an estimation procedure based on IMG, which is given by the distortion of the data when the samples are generated as a Gaussian mixture and their variances are known perfectly. The ensuing distortion lower bound is:

$$D_{\text{lower bound}} = \int G_X(\sigma^2) \frac{\sigma^2 \sigma_Z^2}{\sigma^2 + \sigma_Z^2} d\sigma^2 , \quad (16)$$

where the ratio $\dfrac{\sigma^2 \sigma_Z^2}{\sigma^2 + \sigma_Z^2}$ is the theoretical distortion of the Wiener filter for Gaussian data and noise with known variances σ^2 and σ_Z^2. This result is superposed in Fig. 1, which displays the distortion of the estimation (14) as a function of the signal to noise ratio (SNR)

$$\text{SNR} = 10 \log_{10} \frac{\sigma_\gamma^2}{\sigma_Z^2} .$$

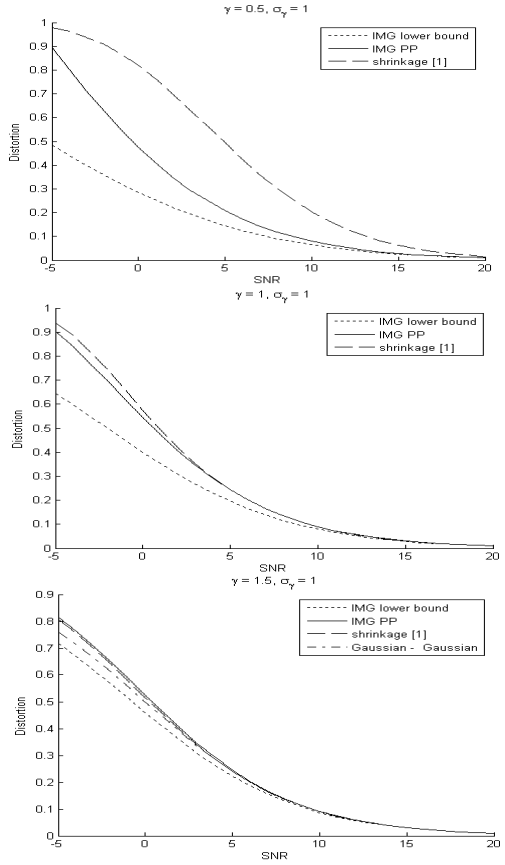

Figure 2. Distortion for GGD data as a function of SNR (dB)

A considerable difference in performance in the favour of IMG PP is observed for $\gamma = 0.5$. The difference tends to decrease for larger γ, and inverts for $\gamma = 1.5$, albeit with a very small relative module. All curves tend to the distortion of the Wiener filter when γ approaches 2, plotted dashed-dotted in Fig.2 on the graph corresponding to $\gamma = 1.5$.

5. LOCAL ESTIMATION

The simplest way of obtaining the variances within a local estimation paradigm ($m > 1$) is through *Maximum Likelihood* (ML) estimation:

$$\hat{\sigma}^2_{X\_ML}[k] = \arg\max_{\sigma^2} \left[p\left(\mathbf{y}_k \mid \sigma^2\right) \right], \quad (17)$$

which for zero-mean independent Gaussian source and noise data yields the solution:

$$\hat{\sigma}^2_{X\_ML}[k] = \max\left(\frac{\|\mathbf{y}_k\|^2}{m} - \sigma_Z^2, 0 \right). \quad (18)$$

(18) could be improved through MAP estimation, if the variance distribution were known. This is provided by the Gaussian Transform and the MAP variance estimate is:

$$\hat{\sigma}^2_{X\_MAP}[k] = \arg\max_{\sigma^2} \left[G_X(\sigma^2)\, p\left(\mathbf{y}_k \mid \sigma^2\right) \right]. \quad (19)$$

Generally the expression (19) requires numerical maximization, but an analytical form exists for $\gamma = 1$ and Gaussian noise (using (12) and differentiation of (19)):

$$\hat{\sigma}^2_{X\_MAP,\gamma=1}[k] = \max\left(\frac{2\|\mathbf{y}_k\|^2}{m + \sqrt{m^2 + 8\|\mathbf{y}_k\|^2 / \sigma_\gamma^2}} - \sigma_Z^2, 0 \right) . \quad (20)$$

We tested empirically both (18) and (20) by generating 1D Laplacian data of length N=10e6, adding independent white Gaussian noise, and performing denoising according to (5), using local windows of length m=5. The corresponding distortion was computed as the mean square error (Fig. 3). We also tested the point to point estimators from Section IV.

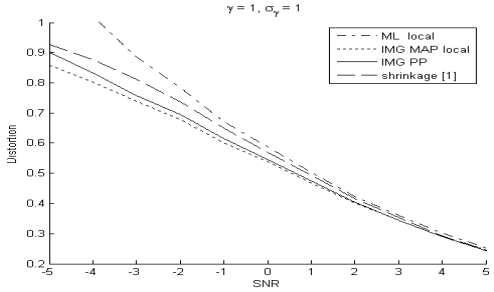

Figure 3. Empirical distortion

As expected, empirical results reproduce the theoretical results for the scalar estimators from the previous section (Fig. 3 zooms into the region [–5 5] dB), and while the ML estimation performs worst than the shrinkage functions, the IMG MAP technique improves even on the IMG PP estimator.

6. LOCAL-GLOBAL ESTIMATION

It is possible to go one step further and combine the local estimation approach with the cumulative estimation scheme presented in Section 3. It is sufficient to define the function $\mathcal{M}$ as the MAP estimates (20):

$$\mathcal{M}_{MAP}(\mathbf{y}_k) = \hat{\sigma}^2_{X\_MAP}[k], \quad (21)$$

compute empirically the distribution probability $P_{\mathcal{M}_{MAP}}$ (which is directly given by the histogram of $\hat{\sigma}^2_{X\_MAP}$), and perform cumulative estimation according to (10):

$$\hat{\sigma}^2_{X\_LG}[k] = PG_X^{-1}\left(P_{\mathcal{M}_{MAP}}\left(\hat{\sigma}^2_{X\_MAP}[k]\right)\right). \quad (22)$$

As the cumulative estimator is not sensitive to monotonic transformations (Theorem 1), in the case of Gaussian noise one can directly use as a norm the ML variance estimates (18), or simply the distribution of $\|\mathbf{y}_k\|^2$. This assertion, however, is not true in the general case.

We call this technique the *local-global (LG) estimation*, as it uses global information (empirical distribution of $\hat{\sigma}^2_{X\_MAP}$) to estimate the variance on a local level. The LG technique can be best described as local MAP estimation (choice of the $\mathcal{M}$ function) under the constraint of global consistency (cumulative estimation).

For Laplacian data the LG estimates are given by (12):

$$\hat{\sigma}^2_{X\_LG}[k] = -\sigma_\gamma^2 \ln\left(1 - P_{\mathcal{M}_{MAP}}\left(\hat{\sigma}^2_{X\_MAP}[k]\right)\right). \quad (23)$$

Distortion results are presented in Fig. 4 for the local ML, MAP and LG techniques.

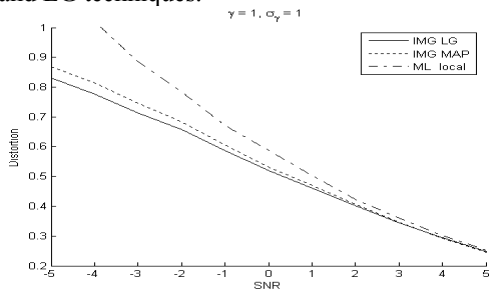

Figure 4. Distortion for local estimation techniques

As expected, the LG estimation further improves on the previous ML and MAP based estimates.

7. CONCLUSION

We presented in this paper various denoising algorithms based on the Infinite Mixture of Gaussians model. Their application to the denoising of Generalized Gaussian data showed significant improvements in terms of distortion, both theoretically and empirically, with respect to the state-of-the-art. However the potential of IGM based denoising is not yet fully explored, as the lower bound presented in this paper is still below current results. The local estimation methods proposed (MAP, LG) may reach this lower bound in the case where the data comes locally from the same Gaussian component, which would allow for perfect variance estimation. If this is not the case, the local window size should be adapted in order to optimally exploit local Gaussianity.

ACKNOWLEDGMENT

This work is supported in part by the Swiss NCCR (National Centre of Research) IM2 (Interactive Multimodal Information Management, http://www.im2.ch) and the EU Network of Excellence Similar (http://www.similar.cc).

The authors thank Dr. Oleksiy Koval for careful reading and fruitful discussions.

REFERENCES

1 P. Moulin and J. Liu, "Analysis of Multiresolution Image Denoising Schemes Using Generalized - Gaussian and Complexity Priors", *IEEE Trans. Info. Theory*, Special Issue on Multiscale Analysis, Vol. 45, No. 3, pp. 909-919, Apr. 1999

2 T. I. Alecu, S. Voloshynovskiy and T. Pun, "The Gaussian Transform", EUSIPCO2005, 13th European Signal Processing Conference, September 2005

3 D. L. Donoho and I. M. Johnstone, "Ideal Spatial Adaptation Via Wavelet Shrinkage", Biometrika, Vol. 81, pp. 425-455, 1994.

4 A. Bijaoui, "Wavelets, Gaussian mixtures and Wiener filtering", Signal Processing, Vol.82, No.4, April 2002, pp 709-712.

5 Benaroya, L; Bimbot, F; "Wiener-based source separation with HMM/GMM using a single sensor", Proc. 4th Int. Symp. on Independent Component Anal. and Blind Signal Separation, April 2003, Nara, Japan.

6 S.G. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation", in IEEE Transactions on Pattern Analysis and Machine Intelligence Volume 11 , Issue 7 (July 1989) p. 674-693.