Image Restoration using Gaussian Scale Mixtures in Overcomplete Oriented Pyramids

Javier Portilla

Visual Information Processing Group
Dept. of Computer Science and Artificial Intelligence
Universidad de Granada, Spain
javier@decsai.ugr.es

ABSTRACT

Gaussian Scale Mixtures (GSMs) in overcomplete oriented pyramids are, arguably, one of the most powerful available tools for image denoising: 1) they provide a new mathematical frame for modelling the variance-adaptation problem, an approach used in image denoising for the last 25 years; 2) they are applicable to contaminating sources of any spectral density; 3) they yield the smallest L2-norm distortion results in simulations under white Gaussian noise, up to this date; and 4) they allow for a solution, for the first time, to the problem of denoising images affected by unknown covariance noise. In this work, we focus first on the general properties of the GSMs. Then, we review the different ways GSMs have been used in overcomplete oriented pyramids (MAP-z-GSM, BLS-GSM, spatially variant GSM), and their applications: classical denoising, signal-dependent noise removal, unknown covariance noise removal and deblurring.

Keywords: image denoising, Gaussian Scale Mixture, overcomplete multiscale representation, local variance adaptation, higher-order image statistics, Bayesian estimation, image restoration

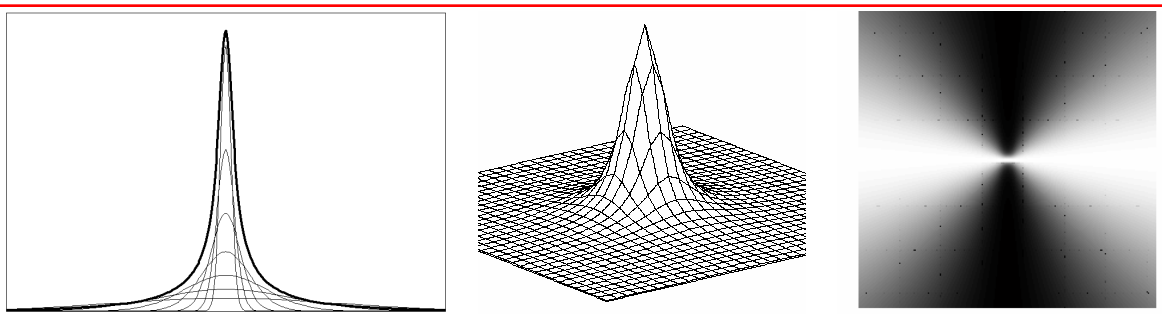
1. INTRODUCTION

The increase in computational power of digital devices for on-line and off-line image processing is making accessible to the public some high-performance applications that before were considered computationally too costly for being used outside the laboratory. One of the most extended applications is digital random noise removal. In recent years, a wide variety of methods have been proposed, and substantial progress has been made in terms of methodology and objective performance.

Examining some of the most successful recent denoising algorithms (e.g.[1–5]), we realize that they fulfill the two following features: 1) They use translation invariant, overcomplete representations, with local kernels selective to scale and orientation (such as curvelets,[4] complex wavelets,[6] translation invariant versions of classical separable wavelets, like undecimated wavelet transforms[7] or trapezoidal pyramids,[8–10] the steerable pyramid and its versions,[5,11–14] and other tight frames[15]); 2) They apply a multidimensional *shrinkage* function (possibly a thresholding) based on joint observations of the coefficients in neighborhoods (e.g.[2,3,5,16]). These two features seem to be the result of an emerging consensus about the importance, on the one hand, of using representations adapted to key image properties such as scale invariance and the existence of locally oriented features. And, on the other hand, of considering higher-order statistics in the new domain, not only marginally for every subband, but jointly for neighbor coefficients across space, scale and/or orientation.[17–19]

Some of these methods (e.g.[1,5,9,10,13,14,20,21]) can be interpreted as extensions of the classical Wiener estimate, which assumes a global Gaussian behavior of both signal and noise. These extensions go into two simultaneous directions: 1) from the pixel domain to multi-orientation pyramids; and 2) from the rigid description of the signal as globally Gaussian, to the flexibility of a local Gaussian model in the new domain, which allows the parameters of the Gaussian to get locally and selectively (in orientation and scale) adapted.

This work has been funded by TIC2003-1504 grant and the "Ramón y Cajal" Program (Spanish government)



(a) (b) (c)

Fig. 1. (a) A Gaussian Scale Mixture is obtained by adding up a continuum of zero-mean Gaussian densities, each one with a variance proportional to z, and with a weight given by $p_z(z)$. The result is always leptokurotic (kurtosis ≥ 3). (b) When we take rows or columns of a joint 2-D GSM, we see that the corresponding conditional densities are smoother and wider as we move from zero. (c) The columns of the joint GSM of (b) in gray-level and normalized to their maxima, showing the characteristic *bow-tie* conditional dependency between GSM components.

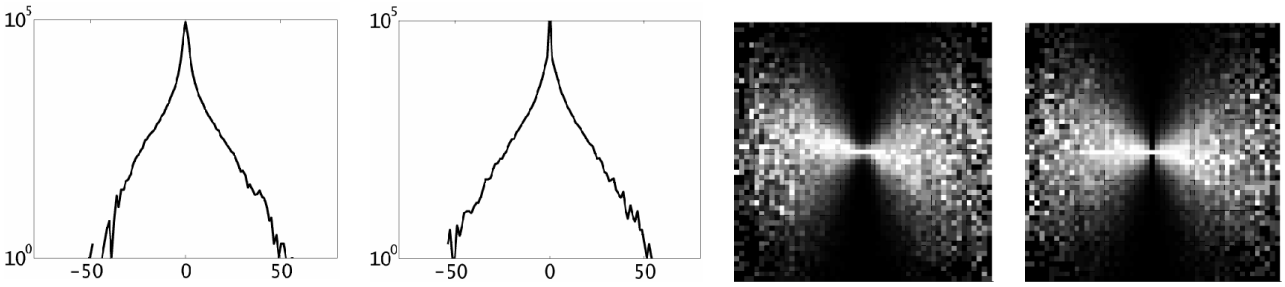
2. GAUSSIAN SCALE MIXTURES AND IMAGE STATISTICS

A Gaussian scale mixture (GSM) is the density of the product of a hidden positive scalar random variable ($\sqrt{z}$) times a zero mean Gaussian vector ($\mathbf{u}$), $\mathbf{x} = \sqrt{z}\mathbf{u}$.[22] Compared to classical Gaussian mixtures (GMs), GSMs are different in two senses. First, z is a continuous random variable (as opposed to the discrete variable selecting the "classes", in the GM case). As a consequence, the density of the mixture is an integral of a continuum of Gaussian functions, instead of a sum. Second, the involved Gaussian functions have fixed zero mean and the same covariance, up to the scale factor z (mixtures of two zero-mean Gaussians have also been successfully used for image denoising[23]). Therefore, same as Gaussian densities, they are elliptically symmetric distributions, i.e., their density depends uniquely on a quadratic function of $\mathbf{x}$. The GSM density is:

$$p_{\mathbf{x}}(\mathbf{x}) = \int_0^\infty p_{\mathbf{x}|z}(\mathbf{x}|z)\, p_z(z)\, dz = \int_0^\infty \frac{\exp\left(-\mathbf{x}^T(z\mathbf{C}_{\mathbf{u}})^{-1}\mathbf{x}/2\right)}{(2\pi)^{N/2}|z\mathbf{C}_{\mathbf{u}}|^{1/2}}\, p_z(z)\, dz \tag{1}$$

where $\mathbf{C}_{\mathbf{u}}$ is the covariance matrix of $\mathbf{u}$, and $p_z(z)$ is the multiplier density. Figure 1(a) shows, in 1-D, the effect of cumulating a continuum (discrete in the illustration) of zero-mean Gaussian functions each with a variance controlled by the multiplier z. The overall effect is that we obtain a much more "peaky" density (in fact, GSMs can never have kurtosis lower than 3). Another feature of GSMs is the particular shape they give raise for the conditional density of one component of the vector given another one. Figures 1(b) and (c) illustrate how cutting slices in a GSM density and normalizing we obtain at zero a very concentrated density, which becomes more and more spread as we go farther from zero. What this means, in terms of the information that a vector component carries about the others, is that when one component's amplitude is small, then very likely the others will also be small (because, very likely, the common multiplier will be small), but if its amplitude is large, then there is a fair amount of uncertainty about the other components (i.e., their conditional variance is high).

The reason for using GSMs in the wavelet domain for denoising is that GSMs are excellent descriptors of local clusters of wavelet coefficients responding to natural images. It turns out that the two features illustrated in Fig. 1 are precisely two key properties of the natural image statistics when seen through an oriented pyramid. First, the high sparseness of the wavelet responses to images (which corresponds to a "peaky" shape of the histogram, with heavy tails and high kurtosis) has been observed long ago,[24,25] and it is at the basis of dozens of important contributions. Second, the particular shape of the conditional density of neighbor wavelet coefficients is equally striking and important (the characteristic *bow-tie* shape[18]), and it also fits perfectly into the GSM model, taking the vector $\mathbf{x}$ as a neighborhood of coefficients (neighbor in space, scale or orientation) around a certain reference coefficient x_c. The use of GSMs for modelling image statistics in the wavelet domain was proposed by Wainwright and Simoncelli in 1999,[26] and it has been fruitfully used since then for denoising[9,10,13,14,27,28] and deblurring.[29] Figure 2 compares 1-D (up) and 2-D (bottom, conditional) histograms from a real wavelet subband responding to the image "boats" (left) and the result of a simulation using a GSM with parameters fitted to the observation.



(a) *Observed* (b) *Simulated* (c) *Observed* (d) *Simulated*

Fig. 2. Comparison of coefficient statistics from an example image subband (a vertical subband of the *Boats* image) with those arising from simulation of a local GSM model. Model parameters (covariance matrix and the multiplier density) are estimated by maximizing the likelihood of the observed set of wavelet coefficients. (a,b) log marginal histograms. (c,d): conditional histograms of two adjacent coefficients. Brightness corresponds to probability, except that each column has been independently rescaled to fill the range of display intensities. Figure taken from Ref. 14.

The two statistical features described above occur in any type of wavelet. However, overcomplete wavelets are more appropriate for image restoration than critically sampled representations, because the latter are not translation invariant.[7,30] All the GSM-denoising methods described in this review are based on overcomplete oriented pyramids, some of them using versions of the steerable pyramid,[5,11] some others a trapezoidal redundant Haar pyramid.[8–10]

3. ESTIMATION UNDER GAUSSIAN NOISE OF KNOWN COVARIANCE

The basic idea of the GSM-based restoration methods is to exploit the Gaussianity of both the noise and the local observation given the hidden multiplier z, for expressing the estimation in terms of the Wiener solutions that we would obtain for fixed z values. We have followed the usual denoising approach for wavelets: 1) decompose the image into subbands; 2) denoise each subband; and 3) recompose the subbands into the estimated image. We have used a local GSM model to represent small neighborhoods of coefficients belonging to each subband (3×3 size), sometimes including also a *parent*, i.e., a coefficient at the same spatial location as the reference coefficient from the next coarser subband with the same orientation. Each neighborhood provides the *context* for locally estimating the reference coefficient.

As we will do calculations numerically, it is convenient to sample z in a reduced number of discrete values (8-10 are usually enough), because this allows us to approximate the integrals in z with sums. We have also experienced that it is more efficient to sample z logarithmically than uniformly. Note that the effect of this non-uniform sampling translates into an implicit $1/z$ factor when computing the integrals (due to $d(\log z) = (1/z)dz$). For some of these models we have considered a prior of the hidden multiplier $p_z(z) \propto 1/z$ (non-informative prior[29,31]). In these cases, the implicit $1/z$ factor plays the role of the $p_z(z)$ term in the integrals involving the posterior density.

3.1. Additive independent noise

Let's consider first an additive independent source of zero mean Gaussian noise of known autocovariance. Because of the linearity of the pyramidal representation this translates into an added zero mean correlated Gaussian noise vector $\mathbf{w}$ in our wavelet GSM vector local model: $\mathbf{y} = \mathbf{x} + \mathbf{w} = \sqrt{z}\mathbf{u} + \mathbf{w}$. Note that we will use a different pair of covariance matrices $(\mathbf{C_u}, \mathbf{C_w})$ for modelling the neighborhoods corresponding to each subband. The density of every observed neighborhood is still an infinite mixture of Gaussian densities, but now it includes the noise term, so it is no longer a Gaussian scale mixture:

$$p_{\mathbf{y}}(\mathbf{y}) = \int_0^\infty p_{\mathbf{y}|z}(\mathbf{y}|z)\, p_z(z)\, dz = \int_0^\infty \frac{\exp\left(-\mathbf{y}^T(z\mathbf{C_u} + \mathbf{C_w})^{-1}\mathbf{y}/2\right)}{(2\pi)^{N/2}|z\mathbf{C_u} + \mathbf{C_w}|^{1/2}} p_z(z)\, dz. \quad (2)$$

We can compute $z\mathbf{C_u} + \mathbf{C_w}$ as $z\mathbf{C_y} + (1-z)\mathbf{C_w}$, because $\mathbf{C_y} = \mathbf{C_x} + \mathbf{C_w}$, and we assume without loss of generality that $\mathbf{C_u} = \mathbf{C_x}$ (which implies, in turn that $\mathbb{E}\{z\} = 1$). The covariance matrix of $\mathbf{w}$, $\mathbf{C_w}$, can be easily computed as the sample covariance, for every subband, of a deterministic input image with the same sample autocovariance as the noise (e.g., a delta for white noise)*. Once the noise covariance is computed, we estimate the signal covariance as $\hat{\mathbf{C_u}} = \hat{\mathbf{C_y}} - \mathbf{C_w}$, setting to zero any possible negative eigenvalue.

3.1.1. Double diagonalization and Wiener estimate

In order to deal with vectors having correlated samples (in our case, both signal and noise vectors), it is computationally convenient to express our observations for each subband into a new basis for which the noise vector is spherical (uncorrelated components all with the same variance) and the signal vector density is elliptical and aligned with the axes (uncorrelated components, but with different variance each one).[28] We will first apply this to make the terms $\mathbf{y}^T(z\mathbf{C_u}+\mathbf{C_w})^{-1}\mathbf{y}$ and $|z\mathbf{C_u}+\mathbf{C_w}|$ in Eq. 2 easy to compute, without requiring a matrix inversion or computing a determinant for each different z value. We can write:

$$z\mathbf{C_u} + \mathbf{C_w} = \mathbf{S}(z\mathbf{S}^{-1}\mathbf{C_u}(\mathbf{S}^T)^{-1} + \mathbf{I})\mathbf{S}^T = \mathbf{S}(z\mathbf{Q}\Lambda\mathbf{Q}^T + \mathbf{I})\mathbf{S}^T = \mathbf{SQ}(z\Lambda + \mathbf{I})\mathbf{Q}^T\mathbf{S}^T, \tag{3}$$

where $\mathbf{SS}^T = \mathbf{C_w}$ and $\{\mathbf{Q}, \Lambda\}$ are the eigenvectors and eigenvalues matrices of $\mathbf{S}^{-1}\mathbf{C_u}(\mathbf{S}^T)^{-1\dagger}$. Thus, Λ is diagonal and $\mathbf{Q}$ is orthogonal. Being this diagonalization independent of z, it only needs to be computed once per subband. Now, if we call $\mathbf{v}$ the observed vector in the new base, $\mathbf{v} = \mathbf{Q}^T\mathbf{S}^{-1}\mathbf{y}$, $\mathbf{y}^T(z\mathbf{C_u} + \mathbf{C_w})^{-1}\mathbf{y}$ can be simply computed as $\sum_{n=1}^{N} \frac{v_n^2}{z\lambda_n+1}$, where N is the size of the neighborhood, and $\lambda_n, n = 1...N$ are the eigenvalues in Λ. Also, by using Eq. 3 the determinant $|z\mathbf{C_u} + \mathbf{C_w}|$ simplifies to $|\mathbf{C_w}||z\Lambda + \mathbf{I}| = |\mathbf{C_w}| \prod_{n=1}^{N} (z\lambda_n + 1)$.

For a given hidden multiplier z both signal and noise are Gaussian, and thus we can write the optimal estimate conditioned on z as the classical Wiener solution, which is easier to compute in the new base:

$$\mathbb{E}\{\mathbf{x}|\mathbf{y}, z\} = z\mathbf{C_u}(z\mathbf{C_u} + \mathbf{C_w})^{-1}\mathbf{y} = z\mathbf{SQ}\Lambda(z\Lambda + \mathbf{I})^{-1}\mathbf{v} = \mathbf{M}z\Lambda(z\Lambda + \mathbf{I})^{-1}\mathbf{v}, \tag{4}$$

where $\mathbf{M} = \mathbf{SQ}$. Because we are only interested in the estimate of the reference coefficient x_c in the neighborhood, we particularize the previous expression:

$$\mathbb{E}\{x_c|\mathbf{y}, z\} = \sum_{n=1}^{N} m_{c,n} \frac{z\lambda_n}{z\lambda_n + 1} v_n, \tag{5}$$

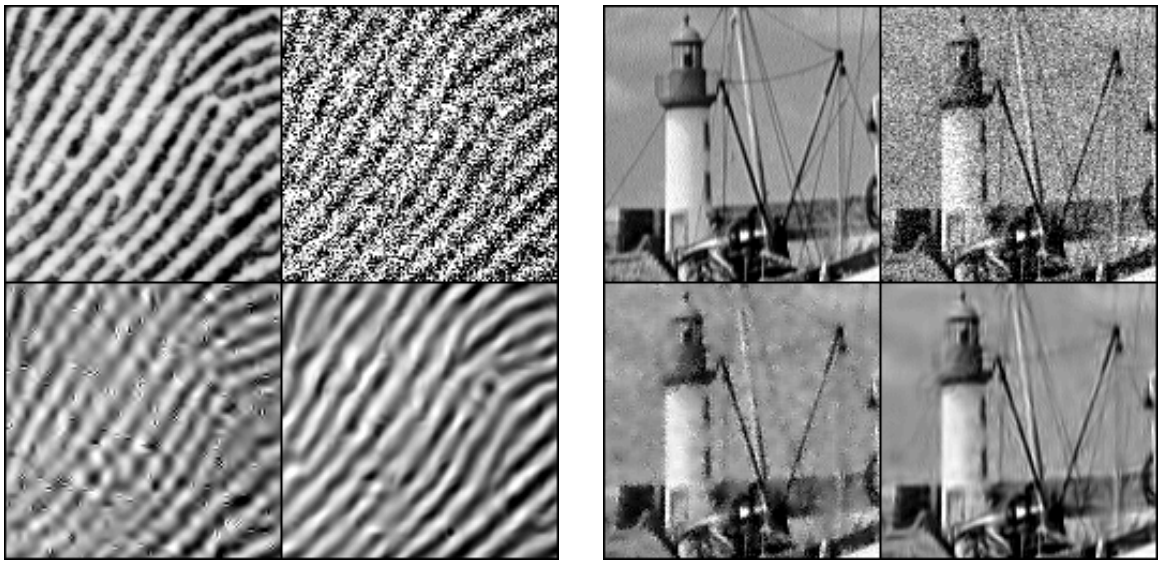
where $m_{c,n}$ is the element of matrix $\mathbf{M}$ at the c (reference) row and the n column.

3.1.2. Empirical Bayes Approach: MAP-z-GSM

One way to simplify the estimation of $\mathbf{x}$ given the observation $\mathbf{y}$ is the so called *empirical Bayes* approach: we first estimate, as an intermediate step, an unknown parameter or hidden variable of the model (in our case, the hidden multiplier z). Then, disregarding the uncertainty associated to that estimate, we take it as if it was a true value, and estimate $\mathbf{x}$ accordingly. This has been the approach followed in many denoising methods: they first estimated the local signal variance, and then applied Wiener locally.[1, 20, 21] If, in addition, we consider a prior model for the local variance (e.g.,[21]), then a reasonable option is to maximize the posterior density $p(z|\mathbf{y})$ (MAP-z estimate). In Ref.13 we chose a log-normal prior for the hidden multiplier z (precursors of this work are Refs. 28, 32). Using the method of moments we estimated the mean and variance of the logarithm of z, for every subband. When differentiating w.r.t. z and equating to zero the corresponding log-posterior, we obtained a non-linear equation in z for each neighborhood, that required a line search of the solution, which was computationally costly. Once we have an estimate for z at every neighborhood, the problem becomes Gaussian, and the classical Wiener solution is applied.

*The reason for using a deterministic function instead of directly a noise sample is to obtain an exact covariance computation, free from stochastic fluctuations.

†This calculation requires that $\mathbf{C_w}$ is invertible, which it is in practical cases.



(a) (b)

Fig. 3. For (a) and (b), from left to right and top to bottom: original, contaminated with white Gaussian noise, comparison method and our result. (a) Noise standard deviation: 100. Comparison method: optimal hard-thresholding on an undecimated wavelet representation.[7] (b) Noise standard deviation: 20. Comparison method: local Wiener in the spatial domain.[20] Figure taken from Ref. 5

3.1.3. Bayesian Least-Squares estimate: BLS-GSM

The previous method is appealing in the sense of dividing an estimation problem involving a hidden random variable, into two supposedly simpler problems: first estimating the local z, and then estimating $\mathbf{x}$ given both the observation $\mathbf{y}$ and the previous z estimate. However, the price of dividing a single optimization into two independent sequential steps is losing the overall optimality. In Refs. 5, 14 a different approach was proposed which allowed a single-step estimation of $\mathbf{x}$ given $\mathbf{y}$. The idea is to express the Bayesian Least Squares (BLS) solution for a neighborhood in terms of the BLS solutions conditioned on z and the posterior $p(z|\mathbf{y})$:

$$\mathbb{E}\{\mathbf{x}|\mathbf{y}\} = \int_{\mathbf{x}} \mathbf{x} p(\mathbf{x}|\mathbf{y}) d\mathbf{x} = \int_{\mathbf{x}} \mathbf{x} \int_{z} p(\mathbf{x}, z|\mathbf{y}) dz d\mathbf{x} = \int_{z} \int_{\mathbf{x}} \mathbf{x} p(\mathbf{x}|\mathbf{y}, z) p(z|\mathbf{y}) d\mathbf{x} dz = \int_{z} \mathbb{E}\{\mathbf{x}|\mathbf{y}, z\} p(z|\mathbf{y}) dz, \qquad (6)$$

i.e., the local estimate is expressed as a weighted average of the Wiener solutions corresponding to each hidden multiplier value, according to their posterior density. This new approach keeps the global optimality for each subband[‡], and it also increases the computational efficiency of the method, compared to MAP-z-GSM. In Ref. 5 we used a flat prior in $\log z$ for computing the posterior $p(z|\mathbf{y})$ ($p_z(z) \propto 1/z$, a *non-informative* prior[31]), for the considered range, and a fully oriented version of the steerable pyramid[11] with 5 scales and 8 orientations. The results (see Fig. 3, Ref. 5 and http://decsai.ugr.es/∼javier/denoise, including denoising software) confirm the superiority in performance of this approach w.r.t. previous methods, including MAP-z-GSM.

3.2. Spatially-variant Gaussian Scale Mixtures

The biggest advantage of using a local GSM model with respect to classical Wiener is that it allows for adaptation to the local variance. And, the main reason for using an overcomplete multi-scale multi-oriented representation is to provide selectivity to scale and orientation. However, by decomposing the image into subbands we obtain another desirable side-effect: the local covariance is much more homogenous in a subband than in the original image. This homogeneity makes possible the high performance of the BLS-GSM denoising method, because it

[‡]However, BLS-GSM does not achieve global optimality in the image domain, due to the lack of equivalence between the Euclidean norm in the coefficient space (overcomplete) and in the image domain.

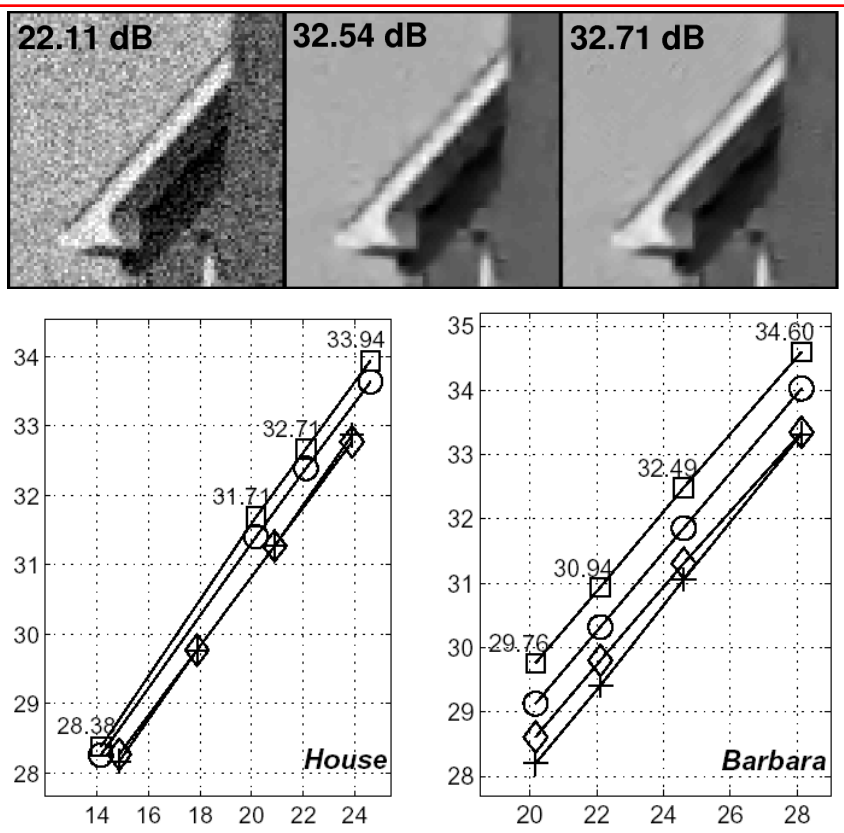
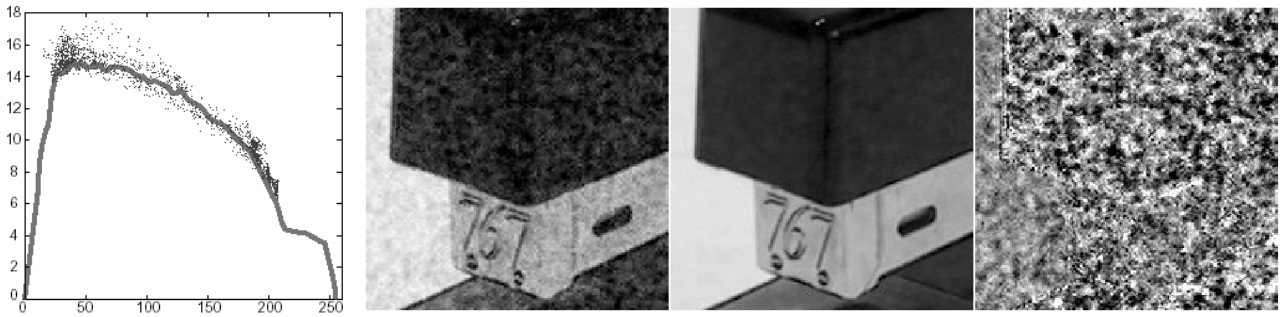


Fig. 4. Comparison between Spatially Variant BLS-GSM and previous methods. **Top**: Visual comparison results on House image (cropped to 80×80). From left to right: Noisy ($\sigma_w = 20$); BLS-GSM[5]; Spatially variant BLS-GSM.[10] **Bottom**: PSNR output vs. PSNR input, in dB. We used the Trapezoidal Haar pyramid[10] for House and the Full Steerable Pyramid[5] for Barbara. Comparison with three state-of-the-art methods: Squares: Spatially variant BLS-GSM,[10] indicating values. Circles: BLS-GSM.[5] House: Diamonds[2] and Crosses.[33] Barbara: Diamonds[3] and Crosses.[1] Figure taken from Ref. 10.

relies on the assumption that the covariance matrix of the signal in a subband is constant up to a local scale factor z. However, if there are clearly differentiated areas in our image (like 2 different textures), we may add some extra accuracy to the representation allowing the signal covariance to get locally adapted not only in amplitude, but also tuning to the local spectral features. In Ref. 10 we proposed a two-level adaptive scheme, using relatively wide areas (16×16, 32×32 or 64×64 blocks of coefficients) for estimating the local covariance, and using small GSM neighborhoods within them (3×3) for adapting also to finer changes in local variance, as before. Mathematically, the model only changes w.r.t. the original BLS-GSM in that now only the neighborhoods within the same block of coefficients in every subband share the covariance of the signal $\mathbf{C_u}^{(n,m)}$, where(n, m) are the spatial indices of the block within the subband. The noise covariance $\mathbf{C_w}$ is considered constant for every neighborhood of the subband. However it is straightforward to generalize this scheme to allow for a spatially variant noise as well, either modulating its strength, as in,[14] or even allowing spatial changes in its spectral density. Same as previous models, we used the Bayesian Least Squares approach and assumed $p_z(z) \propto 1/z$ in the considered range for all the blocks. Figure 4(top) shows an example of the visual improvement obtained when using a two-level adaptive BLS-GSM compared to classical BLS-GSM. Figure 4(bottom) presents a graphical performance comparison between the methods providing lowest quadratic error on the estimation.



(a) (b)

Fig. 5. (a) Calibrated curve of noise standard deviation as a function of local mean, for blue channel. Superimposed to the curve are dots representing empirical measurements of local noise standard deviation and local mean (using the image in (b)), showing a good agreement. (b) Left: blue channel taken at 400 ISO; Middle: same at 50 ISO; Right: difference, signal dependent noise. Figures taken from Ref. 14.

3.3. Signal-dependent noise

Many devices produce output images with significant dependency between local signal and noise. It may also happen that the noise is independent from the signal, but not additive (e.g., multiplicative). When this happens, the noise is interpreted as signal-dependent, under an additive-noise model. Another possibility is that the noise statistics depend on the spatial location. For all these situations it is convenient to refine the additive independent global Gaussian noise model, to allow for spatially variant noise statistics. In Ref.14 we showed examples of digital cameras generating signal-dependent noise. We calibrated in laboratory the noise generation pattern of a particular digital photo camera at a given ISO level (Canon Powershot G1, 400 ISO), using homogeneous gray-level patches. We obtained a measurement of the noise standard deviation in the image domain as a function of the signal level (see Fig 5(a)). For applying this to an observed image we used the local mean (a slightly blurred version of the observation) to provide the signal level, from which we applied the curve of figure 5(a) to obtain an estimate of the noise standard deviation at every spatial location, $\sigma_w(x,y)$. We used this information for estimating 1) the average noise variance in the image domain, $\sigma^2_{w,T} =< \sigma^2_w(x,y) >_{x,y}$, and 2) a local noise multiplier $\alpha(x,y) = \sigma^2_w(x,y)/\sigma^2_{w,T}$, the same for all subbands at a given location [§]. This translates into our signal-plus-noise GSM neighborhood model as:

$$\mathbf{y} = \sqrt{z}\mathbf{u} + \sqrt{\alpha(x,y)}\mathbf{w}, \tag{7}$$

where the spatial coordinates (x,y) correspond to the spatial location of the reference coefficient in the neighborhood. Note that $< \alpha(x,y) >_{x,y} = 1$, and, as before, $\mathbf{w} \sim N(\mathbf{0}, \mathbf{C_w})$. In order to compute the noise covariance $\mathbf{C_w}$ for each subband we need to know the noise autocorrelation in the image domain, as in previous models. For that purpose, in Ref. 14 we estimated the noise for several images in the lab, using for the noise-free reference the same images taken at long exposure and low ISO level (see figure 5(b) for an example). Note also that, unlike for the signal, now we know the scaling factor α for the noise vector, which, for this reason, is not a GSM, but just an amplitude modulated Gaussian vector. The effect of considering a local scaling factor for the noise in the resulting equations w.r.t. considering spatially invariant noise is simply to substitute the terms $(z\lambda_n + 1)$ in the corresponding equations by $(z\lambda_n + \alpha(x,y))$. In Fig. 6 we show a real picture denoised using this method.

4. ESTIMATION UNDER GAUSSIAN NOISE OF UNKNOWN COVARIANCE

In many practical situations we do not have a reliable model of the noise spectral behavior, and it is expensive, difficult or even impossible to obtain *noise-free* reference images from which to estimate the noise as the difference with the noisy observation (as we did in section 3.3). Some other times the spatial statistics of the noise change

[§]Because of the multirresolution, $\alpha(x,y)$ must be shrunk accordingly for some subbands.

Fig. 6. Denoising a digital camera photograph. **Left:** original image, shot at ISO=400 on a Canon PowerShot G1, cropped to 512×512 pixels. Estimated noise level is approximately 27.9dB. **Right:** denoised image, using the spatially variant noise model. Figure taken from Ref. 14.

over time and we have a single observation of a corrupted image. In these cases even an accurate spatio-temporal noise model is of little use. For this kind of situations the alternative is to estimate the noise statistics directly from the observation. When the noise is assumed to be white, the problem is relatively easy, because we just need to estimate its variance. There are several methods for noise variance estimation (e.g.,[34]). However, in many practical cases the noise is far from white, and more sophisticated methods are required to characterize its spectrum. In Refs. 8, 9 we proposed a generalized expectation-maximization algorithm for estimating the noise covariance for each subband of the pyramid, using our local GSM model.

4.1. Maximum Likelihood Estimate of the hidden multiplier density

We have seen in previous sections that when the noise covariance is known we can use parametric descriptions of the hidden multiplier density $p_z(z)$, as *priors*, like log-normal (section 3.1.2), or even non-informative[31] (flat in the logarithm, section 3.1.3). However, it turns out that when there are some other unknown parameters in the model, like the noise covariance $\mathbf{C_w}$ matrices for the subbands, an estimate of $p_z(z)$ adapted to the observation is indispensable for reliably estimating the other unknown parameters.

An Expectation-Maximization (E-M) approach can be used to derive an iterative calculation for the $p_z(z)$ that maximizes the likelihood of the observed neighborhoods for a given subband (assuming the neighborhoods are independent; for this estimation we use non-overlapping neighborhoods), along the same lines as for estimating the mass density function for classical Gaussian mixtures (see, e.g.,[35]). Here we first compute the maximum likelihood condition for $p_z(z)$, and then we propose a recursive computation whose fixed point is precisely that maximum likelihood condition for $p_z(z)$. We will show that this recursion is totally analogous as the E-M iterative solution for classical GMs.

First, in order to simplify the numerical optimization problem, we reduce the hidden multiplier density $p_z(z)$ to a finite number of dimensions. This allows us to differentiate the global likelihood with respect to each of these dimensions to find the optimum. Second, to impose positivity and unity area to $p_z(z)$, we express each discretized probability mass as $P(z_i) = \exp(t_i)/\sum_j \exp(t_j)$. This leaves one degree of freedom in vector $\mathbf{t}$, but this will have no effect on the final result, as we show next [¶]. The global log-likelihood for M neighborhoods can be written as:

[¶]Alternatively, a Lagrange multiplier can be used to impose the constraints on $P(z_i)$.

$C(\mathbf{t}) = \sum_{m=1}^{M} \log p_{\mathbf{y}}(\mathbf{y}_m) = \sum_{m=1}^{M} \log \sum_i p_{\mathbf{y}|z}(\mathbf{y}_m|z_i)P(z_i) = \sum_{m=1}^{M} \log \left[\sum_i p_{\mathbf{y}|z}(\mathbf{y}_m|z_i)\exp(t_i)/\sum_j \exp(t_j)\right]$. We differentiate $C(\mathbf{t})$ and equate to zero, for maximizing it:

$$\frac{\partial}{\partial t_k}\left[\sum_{m=1}^{M} \log \sum_i p_{\mathbf{y}|z}(\mathbf{y}_m|z_i)\exp(t_i) - \log \sum_j \exp(t_j)\right] = \sum_{m=1}^{M}\left(\frac{p_{\mathbf{y}|z}(\mathbf{y}_m|z_k)\exp(t_k)}{\sum_i p_{\mathbf{y}|z}(\mathbf{y}_m|z_i)\exp(t_i)} - \frac{\exp(t_k)}{\sum_j \exp(t_j)}\right) = 0$$

Operating in the previous expression, and noting that $\exp(t_k) \neq 0$, we can write the ML condition for $P(z_i)$:

$$\frac{1}{M}\sum_{m=1}^{M} \frac{p_{\mathbf{y}|z}(\mathbf{y}_m|z_k)}{\sum_i p_{\mathbf{y}|z}(\mathbf{y}_m|z_i)\exp(t_i)/\sum_j \exp(t_j)} = \frac{1}{M}\sum_{m=1}^{M} \frac{p_{\mathbf{y}|z}(\mathbf{y}_m|z_k)}{\sum_i p_{\mathbf{y}|z}(\mathbf{y}_m|z_i)P(z_i)} = 1, \forall k. \quad (8)$$

We check now that this condition is equivalent to the condition $P(z_k) =< p(z_k|\mathbf{y}_m) >_m, \forall k$ (whenever we impose that no $P(z_k)$ vanishes), which can be written as

$$P(z_k) = \frac{1}{M}\sum_{m=1}^{M} \frac{p_{\mathbf{y}|z}(\mathbf{y}_m|z_k)P(z_k)}{\sum_i p_{\mathbf{y}|z}(\mathbf{y}_m|z_i)P(z_i)}.$$

This suggests the iterative computation for $P(z_k)$:

$$P(z_k)^{(new)} = P(z_k)^{(old)}\frac{1}{M}\sum_{m=1}^{M} \frac{p_{\mathbf{y}|z}(\mathbf{y}_m|z_k)}{\sum_i p_{\mathbf{y}|z}(\mathbf{y}_m|z_i)P(z_i)^{(old)}}, \forall k \quad (9)$$

which is the same as the classical E-M algorithm to estimate the probability masses of Gaussian mixtures,[35] and whose fixed point, starting with $P(z_k) \neq 0, \forall k$, is the maximum likelihood condition. It is also easy to see that, under mild conditions, we can take this solution to the limit of infinite small sampling intervals for z to express the solution for the continuous form in z in a totally analogous way.

4.2. Maximum Likelihood Estimate of the noise covariance

In this section we propose an iterative approximate calculation of a maximum likelihood solution for the noise covariance $\mathbf{C_w}$ of the observed neighborhoods for a given subband of the pyramid representation. In this case we keep the original continuous notation for z, although working with z discretized is totally analogous. Naming $C(\mathbf{C_w})$ the log-likelihood of the observed neighborhoods as a function of $\mathbf{C_w}$, we can write:

$$C(\mathbf{C_w}) = \sum_{m=1}^{M} \log p_{\mathbf{y}}(\mathbf{y}_m) = \sum_{m=1}^{M} \log \left[\int_z p_{\mathbf{y}|z}(\mathbf{y}_m|z;\mathbf{C_w})p_z(z)dz\right].$$

Differentiating this scalar with respect to each element of $\mathbf{C_w}$ we obtain a matrix that we force to be nule. Calling $\mathbf{C}_z = z\mathbf{C_y} + (1-z)\mathbf{C_w}$, $\widehat{\mathbf{C}_z} = \sum_{m=1}^{M} p(z|\mathbf{y}_m)\mathbf{y}_m\mathbf{y}_m^T / \sum_{m=1}^{M} p(z|\mathbf{y}_m)$, $\hat{p}_z(z) = \frac{1}{M}\sum_{m=1}^{M} p(z|\mathbf{y}_m)$, and substituting in the Eq. 21 (see Appendix A), we obtain after some manipulation:

$$\frac{\partial C(\mathbf{C_w})}{\partial \mathbf{C_w}} = \frac{M}{2}\int_z (1-z)\hat{p}_z(z)\mathbf{C}_z^{-1}\left[\mathbf{I}_{N\times N} - \widehat{\mathbf{C}_z}\mathbf{C}_z^{-1}\right]dz = \mathbf{0_{N\times N}}. \quad (10)$$

Solving this equation exactly for $\mathbf{C_w}$ is not easy. However, we note that 1) $\widehat{\mathbf{C}_z}$ is the classical solution for Gaussian mixtures (see, e.g., [35]), when there is no coupling between the parameters of the Gaussian components of the mixture; 2) despite of that, $\widehat{\mathbf{C}_z}$ may be a reasonable estimate of $\mathbf{C}_z$; 3) the noise covariance $\mathbf{C_w} = \mathbf{C}_z|_{z=0}$. Therefore, it does not seem unreasonable to choose $\mathbf{C_w} = \widehat{\mathbf{C}_z}|_{z=0}$ as an approximation to the solution of Eq. 10. As the posteriors $p(z=0|\mathbf{y}_m)$ must, in turn, have been computed using a previous estimate of $\mathbf{C_w}$, this is clearly an iterative procedure (note that $\hat{p}_z(z)$ must also be updated at each iteration):

$$\mathbf{C}_{\mathbf{w}}^{(new)} = \widehat{\mathbf{C}_z}|_{z=0} = \frac{\sum_{m=1}^{M} p(0|\mathbf{y}_m;\mathbf{C}_{\mathbf{w}}^{(old)})\mathbf{y}_m\mathbf{y}_m^T}{\sum_{m=1}^{M} p(0|\mathbf{y}_m;\mathbf{C}_{\mathbf{w}}^{(old)})}. \quad (11)$$

If we were able to solve exactly for Eq. 10 this would be an Expectation-Maximization algorithm.[9] But, as we have substituted the maximization step by a simple increase in likelihood, it is instead a Generalized Expectation-Maximization algorithm. We have tested that this method provides satisfactory results consistently, increasing significantly the overall likelihood at each step in a vast majority of cases [‖], which quickly takes the solution close to a local maximum in likelihood.

As an interpretation of Eq. 11, we see that the noise covariance is estimated as a weighted average of local external products of the observation vectors according to the posterior probability of these observations having been produced *exclusively* by noise ($z = 0$ condition). Observe that the overall approach relies heavily on the non-Gaussianity of the signal: when the signal is Gaussian, the observation can be *explained* without the GSM term, and all the observation can be considered Gaussian "noise". In general, what the method does is to separate the leptokurtotic part of the observation (the GSM term), which we identify with the signal, from the rest, which is assigned to the Gaussian noise term, but that can even be platokurtotic. This has the effect of removing any homogeneous texture, for instance, highly-colored interferences. Figures 7 show some results obtained with different types of real images presenting correlated noise. These figures demonstrate the wide practical applicability of the method. Typical execution times for a non-optimized implementation using Matlab(c) under a Pentium IV 2 GHz machine are around 30 seconds for 256×256 images.

5. ESTIMATION UNDER BLUR AND GAUSSIAN NOISE

As we have seen, the local Gaussian Scale Mixture model is very flexible, in the sense of being applicable to noise with any spectral density, known or unknown. But we can go further and consider that there is also linear distortion of the image. The important feature to preserve here is that the observation can be locally expressed using a Gaussian density conditioned on the multiplier z, and this feature is preserved by a linear filtering. Our estimate, then, will not only remove noise, but will also compensate for blur.[29] The same BLS-GSM approach of Eq. 6 is applicable, but now the Wiener solution includes a linear distortion term:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w} = \sqrt{z}\mathbf{H}\mathbf{u} + \mathbf{w}, \tag{12}$$

where $\mathbf{x}$ is a vector of length M, large enough to include all the original samples that are involved in the computation of the N samples of $\mathbf{y}$ (see Fig.8), $\mathbf{w}$ is also $N \times 1$, and $\mathbf{H}$ is an $N \times M$ matrix implementing the convolution with the kernel h. The density of the observed neighborhood vector conditioned on z is zero-mean Gaussian, with covariance $\mathbf{C}_{\mathbf{y}|z} = z\mathbf{C}_{\mathbf{u}'} + \mathbf{C}_{\mathbf{w}}$:

$$p(\mathbf{y}|z) = \frac{\exp\left(-\mathbf{y}^T(z\mathbf{C}_{\mathbf{u}'} + \mathbf{C}_{\mathbf{w}})^{-1}\mathbf{y}/2\right)}{\sqrt{(2\pi)^N |z\mathbf{C}_{\mathbf{u}'} + \mathbf{C}_{\mathbf{w}}|}}, \tag{13}$$

where $\mathbf{C}_{\mathbf{u}'} = \mathbf{H}\mathbf{C}_{\mathbf{u}}\mathbf{H}^T$ is the $N \times N$ covariance matrix of $\mathbf{u}' = \mathbf{H}\mathbf{u}$. Naming the filtered clean signal $\mathbf{x}' = \mathbf{H}\mathbf{x}$, we can express the Wiener estimate, for a given z, as:

$$\mathbb{E}\{\mathbf{x}|\mathbf{y}, z\} = z\mathbf{C}_{\mathbf{x}\mathbf{x}'}\left(z\mathbf{C}_{\mathbf{x}'} + \mathbf{C}_{\mathbf{w}}\right)^{-1}\mathbf{y}, \tag{14}$$

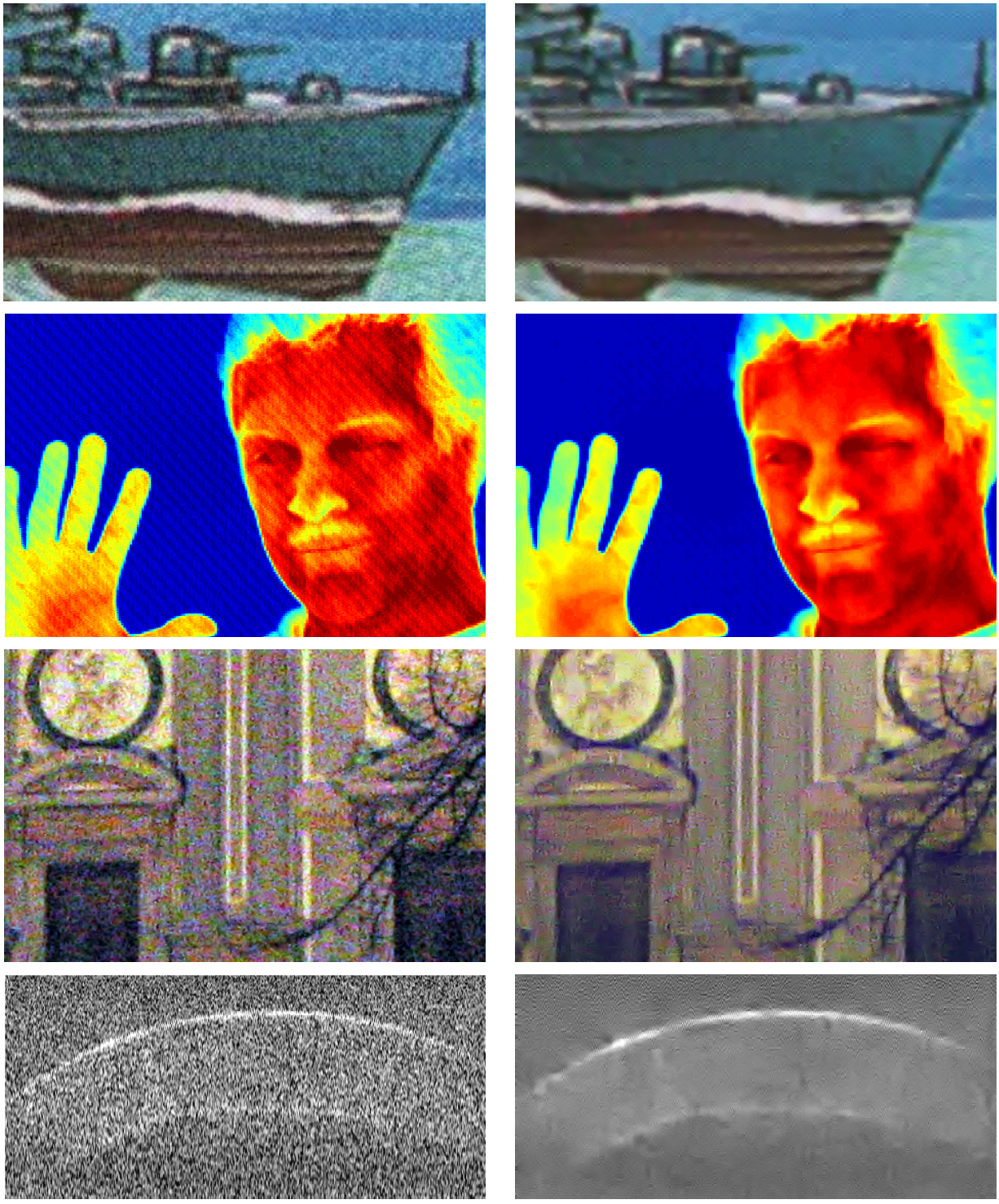
where $\mathbf{C}_{\mathbf{x}\mathbf{x}'} = \mathbf{C}_{\mathbf{x}}\mathbf{H}^T$ is the $M \times N$ cross-covariance matrix of $\mathbf{x}$ and $\mathbf{x}'$, the coefficients from the original image and those from its blurry version. We explain below a method for estimating this matrix. As done in section 3.1.1, we can simplify the dependence of this expression on z by diagonalizing the matrix $z\mathbf{C}_{\mathbf{x}'} + \mathbf{C}_{\mathbf{w}}$. We do the calculations analogously, except that now we call $\{\mathbf{Q}, \Lambda\}$ to the eigenvector/eigenvalue expansion of the matrix $\mathbf{S}^{-1}\mathbf{C}_{\mathbf{x}'}\mathbf{S}^{-T}$. Then (14) becomes:

$$\mathbb{E}\{\mathbf{x}|\mathbf{y}, z\} = z\mathbf{C}_{\mathbf{x}\mathbf{x}'}\mathbf{S}^{-T}\mathbf{Q}(z\Lambda + \mathbf{I})^{-1}\mathbf{Q}^T\mathbf{S}^{-1}\mathbf{y} = z\mathbf{M}(z\Lambda + \mathbf{I})^{-1}\mathbf{v}, \tag{15}$$

where now $\mathbf{M} = \mathbf{C}_{\mathbf{x}\mathbf{x}'}\mathbf{S}^{-T}\mathbf{Q}$, and $\mathbf{v} = \mathbf{Q}^T\mathbf{S}^{-1}\mathbf{y}$. Same as before, we restrict the estimate to the reference coefficient:

$$\mathbb{E}\{x_c|\mathbf{y}, z\} = \sum_{n=1}^{N} \frac{z m_{cn} v_n}{z\lambda_n + 1}, \tag{16}$$

[‖] For the few cases when this does not happen, we can use directly Eq. 10 to correct the estimation following the gradient direction.



Observed *After Blind BLS-GSM Denoising*

Fig. 7. Results obtained with four real noisy images. From top to bottom: **(a)** a printed image captured with a scanner; **(b)** an infrared video frame suffering from electronic interference; **(c)** a digital photography captured under poor light conditions; **(d)** an optical coherence tomography (OCT). Images have been zoomed and/or cropped for visibility of the artifacts. Figure taken from Ref. 8.

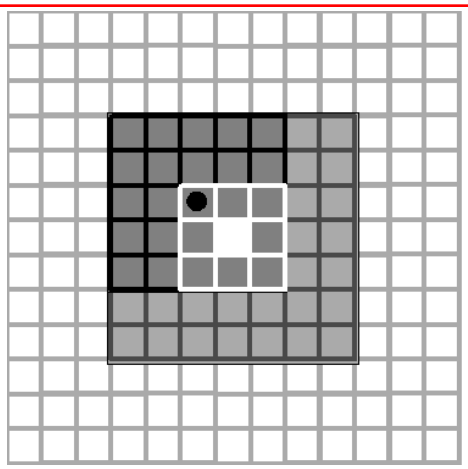


Fig. 8. Graphical explanation of the meaning of the neighborhoods in the GSM model including linear distortion. The background grid represents the coefficients in a subband. The light gray area represents $\mathbf{x}$, the coefficients of the original that have an influence on the observed neighborhood of coefficients, corresponding to $\mathbf{y}$ (in white), through the convolution with kernel h (darker gray). The white square is the reference coefficient.

5.1. Estimation of the cross-covariance matrix

Up to now we have followed a parallel approach for the deblurring case as for the denoising. However, the real difficulty for applying in practice the local GSM model for deblurring is the estimate of the cross-covariance matrix between the original coefficients and the filtered (not yet noisy) coefficients, $\mathbf{C_{xx'}}$. In Ref. 29 we followed a heuristic approach, that in future versions of this work should be replaced by a more rigorous estimation. In the cited reference we used the relation between the power spectrum of the observed image and that of the original image, $P_y(u,v) = P_x(u,v)|H(u,v)|^2 + P_w(u,v)$, to first obtain an approximation of the power spectrum of the filtered image:

$$P_{x'}(u,v) \simeq \left\lfloor |Y(u,v)|^2 * G(u,v) - P_w(u,v) \right\rfloor_+ , \tag{17}$$

where $G(u,v)$ is a Gaussian convolving window for removing sampling fluctuations. Then, we estimated an approximation of the power spectrum of the original using:

$$P_x(u,v) \simeq P_{x'}(u,v) / \max(|H(u,v)|^2, G_{max}^{-1}), \tag{18}$$

where G_{max} is the maximal allowed gain for the inverse filter. For the results shown in this paper (white noise case) we chose $G_{max} = \min(2400/\sigma_0^2, 180)$ with σ_0 the standard deviation of the noise in the image domain. Now $\mathbf{C_{xx'}}$ can be estimated as the sample cross-covariance of the coefficients of the inverse Fourier transforms of $\sqrt{P_x(u,v)}$ and $\sqrt{P_{x'}(u,v)}$. The resulting estimate $\widehat{\mathbf{C_{xx'}}}$ can be significantly improved by modeling its bias using an $N \times N$ linear transform: $\mathbb{E}\{\widehat{\mathbf{C_{xx'}}}\} \simeq \mathbf{C_{xx'}B}$. In order to estimate $\mathbf{B}$, we assume that we would incur the same proportional bias when estimating $\mathbf{C_{x'}}$ from $P_{x'}(u,v)$: $\mathbb{E}\{\widehat{\mathbf{C_{x'}}}\} \simeq \mathbf{C_{x'}B} = (\mathbf{C_y} - \mathbf{C_w})\mathbf{B}$. Solving this for $\mathbf{B}^{-1}$ yields the following improved estimator:

$$\widehat{\widehat{\mathbf{C_{xx'}}}} = \widehat{\mathbf{C_{xx'}}} \left[(\mathbf{C_y} - \mathbf{C_w}) \widehat{\mathbf{C_{x'}}}^{-1} \right] \tag{19}$$

Figure 9 shows some results obtained with simulated degraded images, compared to the classical Wiener method.

6. CONCLUSIONS AND FUTURE WORK

We have reviewed different image restoration methods based on a local Gaussian Scale Mixture model describing neighborhoods of coefficients in overcomplete pyramids. Starting from the classical empirical Bayes approach,

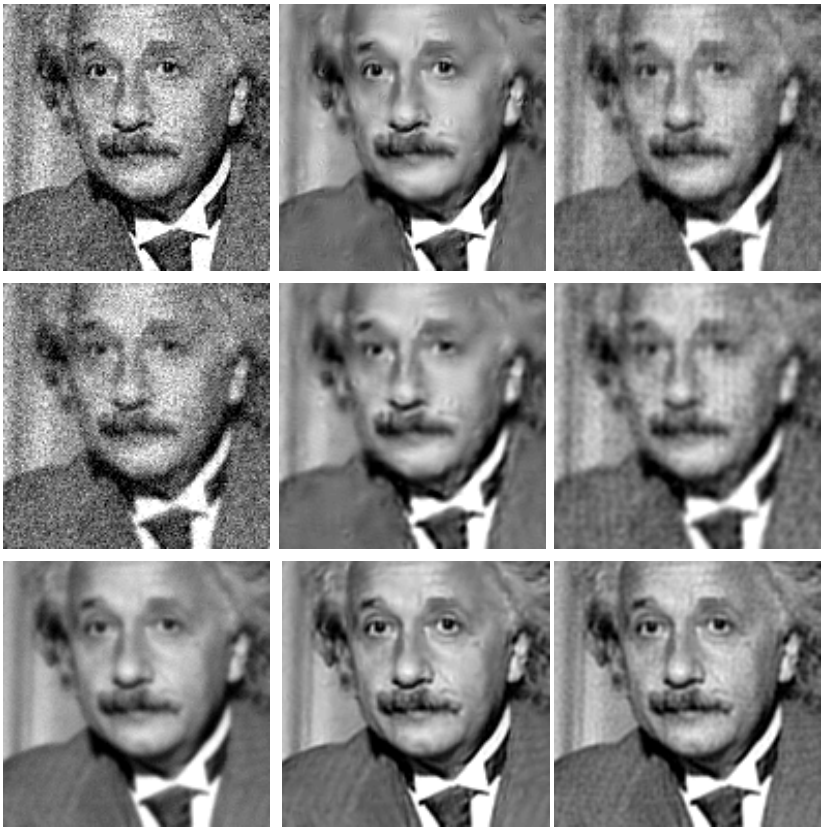


Fig. 9. BLS-GSM deblurring results. **Up**: High noise ($\sigma = 20$) and low blur ($\sigma_b = 0.1$); PSNR values are (dB): 22.1, 29.9, 28.3; **Middle**: High noise ($\sigma = 20$) and high blur ($\sigma_b = 1.0$), (21.5, 27.9, 27.0); **Bottom**: Low noise ($\sigma = 2$) and high blur ($\sigma_b = 1.0$), (29.7, 32.5, 31.8). Figure taken from Ref. 29.

which is clearly suboptimal, we have described the evolution towards a more powerful Bayesian Least Squares approach (BLS-GSM approach). This latter approach has been applied to different situations: independent additive Gaussian noise, signal dependent noise, and independent noise of unknown covariance. The method can also be refined to adapt locally not only to fluctuations in the energy of clusters of coefficients, but also to spatial changes in their local covariance, by using spatially variant GSMs. Finally, the model allows for a generalization of the degradation model, through the inclusion of a linear distortion term, besides the additive noise, which makes it applicable to image deblurring. We have shown that the flexibility, performance and practical applicability of the local GSM model in overcomplete pyramids is remarkable, compared to most previous models. However, we should also point to its limitations, which come, most of them, from its local nature. The model assumes that there is a single multiplier affecting simultaneously to a neighborhood of coefficients. The downside of this approach is that it is very rigid: either a coefficient belongs to a neighborhood or not, but there are no intermediate grades in the statistical coupling of the neighbors. This causes that best results are obtained with very small neighborhoods (3×3, possibly plus a *parent*), when we know that there is a fair amount of statistical coupling between coefficients two or three spatial positions apart, or at different orientations, or simultaneously at a spatially neighbor position at an adjacent scale, etc (see e.g.,[19]). A future model could consider a smoothly varying z, which would be applied individually to all the coefficients. Enlarging the effective coupling region between coefficients through a globally consistent model will improve the quality of the estimation, at the likely price of increasing the computational cost. Another pending issue is to improve the estimation of the cross-covariance between the original and the blurred original, for the deblurring application. This applies to any technique requiring the estimation of the power spectral density of the original from the degraded observation. Finally, it is also pending a BLS-optimization for the whole image, instead of for each subband.

ACKNOWLEDGMENTS

This paper presents a review of previous works, most of them co-authored by other people besides myself: Eero P. Simoncelli (New York University), Vasily Strela (Drexel University), Martin J. Wainwright (California University, Berkeley) and Jose A. Guerrero-Colón (Universidad de Granada). I am grateful for having had the opportunity of learning from them and with them. I also thank Carlos Dorronsoro (Centro de Investigación y Desarrollo de la Armada, Madrid) for having provided me with real images for applying the blind denoising method.

REFERENCES

1. X. Li and M. T. Orchard, "Spatially adaptive image denoising under overcomplete expansion," in IEEE Int'l Conf on Image Proc, **3**, pp. 300–303, IEEE, (Vancouver), September 2000.
2. A. Pižurica, W. Philips, I. Lemahieu, and M. Acheroy, "A joint inter- and intrascale statistical model for Bayesian wavelet based image denoising," *IEEE Trans. Image Proc.* **11**, pp. 545–557, May 2002.
3. L. Şendur and I. W. Selesnick, "Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency," *IEEE Trans. Signal Proc.* **50**, pp. 2744–2756, November 2002.
4. J. Starck, E. J. Candes, and D. L. Donoho, "The curvelet transform for image denoising," *IEEE Trans. Image Proc.* **11**, pp. 670–684, June 2002.
5. J. Portilla, V. Strela, M. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of Gaussians in the wavelet domain," *IEEE Trans. Image Proc.* **12**, pp. 1338–1351, November 2003.
6. N. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals," *Applied and Computational Harmonic Analysis* **10**, pp. 234–253, May 2001.
7. R. R. Coifman and D. L. Donoho, "Translation–invariant de–noising," in *Wavelets and statistics*, A. Antoniadis and G. Oppenheim, eds., Springer-Verlag lecture notes, San Diego, 1995.
8. J. Portilla, "Blind non-white noise removal in images using gaussian scale mixtures in the wavelet domain," in *Proc. of the IEEE Benelux Signal Processing Symposium*, pp. 17–20, (Hilvarenbeek), Apr 2004.
9. J. Portilla, "Full blind denoising through noise covariance estimation using gaussian scale mixtures in the wavelet domain," in *IEEE Int'l Conf. on Image Proc.*, pp. 1217–1220, (Singapore), Oct 2004.
10. J. A. Guerrero-Colon and J. Portilla, "Two-level adaptive denoising using gaussian scale mixtures in overcomplete oriented pyramids," in *IEEE Int'l Conf on Image Proc*, (Genoa), Sep 2005.
11. E. P. Simoncelli and W. T. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," in *Second Int'l Conf on Image Proc*, **III**, pp. 444–447, (Washington, DC), Oct 1995.
12. E. P. Simoncelli and E. H. Adelson, "Noise removal via Bayesian wavelet coring," in *Third Int'l Conf on Image Proc*, **I**, pp. 379–382, IEEE Sig Proc Society, (Lausanne), Sep 1996.
13. J. Portilla, V. Strela, M. Wainwright, and E. Simoncelli, "Adaptive Wiener denoising using a Gaussian scale mixture model in the wavelet domain," in *Proc 8th IEEE Int'l Conf on Image Proc*, pp. 37–40, (Thessaloniki), Oct 2001.
14. J. Portilla, V. Strela, M. Wainwright, and E. Simoncelli, "Image denoising using Gaussian scale mixtures in the wavelet domain," Tech. Rep. TR2002-831, Courant Inst. of Math. Sci., N.Y. Univ., Sep 2002.
15. L. Shen, M. Papadakis, I. A. Kakadiaris, I. Konstantinidis, D. Kouri, and D. Hoffman, "Image denoising using a tight frame," in *Proc. of ICASSP*, IEEE, ed., **II**, pp. 641–644, 2005.
16. S. G. Chang, B. Yu, and M. Vetterli, "Spatially adaptive wavelet thresholding with context modeling for image denoising," *IEEE Trans. Image Proc.* **9**, pp. 1522–1531, Sep 2000.
17. J. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans Sig Proc* **41**, pp. 3445–3462, Dec 1993.
18. E. P. Simoncelli, "Statistical models for images: Compression, restoration and synthesis," in *31st Asilomar Conf on Signals, Systems and Computers*, pp. 673–678, (Pacific Grove, CA), Nov 1997.
19. R. W. Buccigrossi and E. P. Simoncelli, "Image compression via joint statistical characterization in the wavelet domain," *IEEE Trans Image Proc* **8**, pp. 1688–1701, Dec 1999.
20. J. S. Lee, "Digital image enhancement and noise filtering by use of local statistics," *IEEE Pat. Anal. Mach. Intell.* **PAMI-2**, pp. 165–168, March 1980.

21. M. K. Mihçak, I. Kozintsev, K. Ramchandran, and P. Moulin, "Low-complexity image denoising based on statistical modeling of wavelet coefficients," *IEEE Trans. on Signal Processing* **6**, pp. 300–303, Dec 1999.
22. D. Andrews and C. Mallows, "Scale mixtures of normal distributions," *J. Royal Stat. Soc.* **36**, pp. 99–102, 1974.
23. M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Proc.* **46**, pp. 886–902, Apr 1998.
24. S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Pat. Anal. Mach. Intell.* **11**, pp. 674–693, Jul 1989.
25. B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature* **381**, pp. 607–609, 1996.
26. M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of Gaussians and the statistics of natural images," in *Adv. Neural Information Processing Systems*, S. A. Solla, T. K. Leen, and K.-R. Müller, eds., **12**, pp. 855–861, MIT Press, May 2000.
27. M. Wainwright, E. Simoncelli, and A. Willsky, "Random cascades on wavelet trees and their use in modeling and analyzing natural imagery," *Applied and Computational Harmonic Analysis*, 2000.
28. V. Strela, J. Portilla, and E. Simoncelli, "Image denoising using a local gaussian scale mixture model in the wavelet domain," in *Proc SPIE, 45th Annual Meeting*, (San Diego), Jul 2000.
29. J. Portilla and E. P. Simoncelli, "Image restoration using Gaussian scale mixtures in the wavelet domain," in *Proc IEEE Int'l Conf on Image Proc*, **2**, pp. 965–968, (Barcelona), Sep 2003.
30. E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multi-scale transforms," *IEEE Trans Information Theory* **38**, pp. 587–607, March 1992.
31. M. Figueiredo and R. Nowak, "Wavelet-based image estimation: An empirical Bayes approach using Jeffrey's noninformative prior," *IEEE Trans. Image Proc.* , Sep 2001.
32. V. Strela, "Denoising via block Wiener filtering in wavelet domain," in *3rd European Congress of Mathematics*, Birkhäuser Verlag, (Barcelona), Jul 2000.
33. M. Malfait and D. Roose, "Wavelet-based image denoising using a Markov random field a priori model," *IEEE Trans. Image Proc.* **6**, pp. 549–565, Apr 1997.
34. D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," *Biometrika* **81**, pp. 425–455, 1994.
35. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Recognition*, ch. Unsupervised Learning and Clustering. Wiley Interscience, 2001.

APPENDIX A.

$$\frac{\partial C(\mathbf{C_w})}{\partial \mathbf{C_w}} = \sum_{m=1}^{M} \frac{\partial}{\partial \mathbf{C_w}} \log\left[\int_z p_{\mathbf{y}|z}(\mathbf{y}_m|z;\mathbf{C_w})p_z(z)dz\right] = \sum_{m=1}^{M} \frac{\int_z p_z(z)\frac{\partial p_{\mathbf{y}|z}(\mathbf{y}_m|z;\mathbf{C_w})}{\partial \mathbf{C_w}}dz}{\int_{z_0} p_{\mathbf{y}|z}(\mathbf{y}_m|z_0;\mathbf{C_w})p_z(z_0)dz_0} \tag{20}$$

$$= \sum_{m=1}^{M} \frac{\int_z p_z(z)p_{\mathbf{y}|z}(\mathbf{y}_m|z;\mathbf{C_w})\frac{\partial \log p_{\mathbf{y}|z}(\mathbf{y}_m|z;\mathbf{C_w})}{\partial \mathbf{C_w}}dz}{\int_{z_0} p_{\mathbf{y}|z}(\mathbf{y}_m|z_0;\mathbf{C_w})p_z(z_0)dz_0} = \sum_{m=1}^{M} \int_z p(z|\mathbf{y}_m)\frac{\partial}{\partial \mathbf{C_w}} \log p_{\mathbf{y}|z}(\mathbf{y}_m|z;\mathbf{C_w})dz.$$

$$\frac{\partial}{\partial \mathbf{C_w}} \log p_{\mathbf{y}|z}(\mathbf{y}_m|z;\mathbf{C_w}) = -\frac{1}{2}\frac{\partial}{\partial \mathbf{C_w}}\left[\mathbf{y}_m^T\left(z\mathbf{C_y}+(1-z)\mathbf{C_w}\right)^{-1}\mathbf{y}_m + \log\left|z\mathbf{C_y}+(1-z)\mathbf{C_w}\right|\right].$$

$$\frac{\partial}{\partial \mathbf{C_w}} \log p_{\mathbf{y}|z}(\mathbf{y}_m|z;\mathbf{C_w}) = \frac{1}{2}(z-1)\left(z\mathbf{C_y}+(1-z)\mathbf{C_w}\right)^{-1}\left[\mathbf{I}_{N\times N} - \mathbf{y}_m\mathbf{y}_m^T\left(z\mathbf{C_y}+(1-z)\mathbf{C_w}\right)^{-1}\right].$$

Substituting into Eq. 20 it yields:

$$\frac{\partial C(\mathbf{C_w})}{\partial \mathbf{C_w}} = \frac{M}{2}\int_z (1-z)\left(z\mathbf{C_y}+(1-z)\mathbf{C_w}\right)^{-1}\frac{1}{M}\sum_{m=1}^{M} p(z|\mathbf{y}_m)\left[\mathbf{I}_{N\times N} - \mathbf{y}_m\mathbf{y}_m^T\left(z\mathbf{C_y}+(1-z)\mathbf{C_w}\right)^{-1}\right]dz. \tag{21}$$