

FULL BLIND DENOISING THROUGH NOISE COVARIANCE ESTIMATION USING GAUSSIAN SCALE MIXTURES IN THE WAVELET DOMAIN

Javier Portilla

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

We describe an efficient generalized expectation maximization algorithm for estimating the spectral features of a noise source corrupting an observed image. We use a statistical model for images decomposed in an overcomplete oriented pyramid. Each neighborhood of clean pyramid coefficients is modelled as a Gaussian scale mixture, whereas the noise is assumed Gaussian. Combining this GEM technique with a previous Bayesian denoise estimator, we obtain a full blind denoising algorithm, able to deal with homogeneous, Gaussian or mesokurtotic, noise sources of arbitrary covariance. Results demonstrate the high performance of the method for a wide range of corruption sources.

1. INTRODUCTION

Characterizing the noise from a single observation of a corrupted signal is a challenging task that strongly conditions the performance of any subsequently applied restoration method. Traditionally, most estimation methods assume additive Gaussian white noise. The whiteness assumption allows to estimate the noise variance using a decorrelating linear transform (e.g., Fourier, wavelet, PCA) for decoupling signal and noise, according to their different spectral features. To this kind of methods belongs the classical approach of estimating the noise variance as the smallest eigenvalue of the sample covariance matrix, or, in a more advanced version, the use of a robust statistic at the output of a high-pass wavelet subband [1]. However, a flat noise spectral response is a fairly unrealistic assumption in most practical cases. Electronic engineers know that most additive random perturbations present significant correlation patterns. Two examples are noise in digital cameras and interference artifacts in video reception and reproduction. Therefore, we can only expect high-performance denoising with real images after having included the noise covariance in the degradation model. Note that in order to distinguish

This work is supported by the Ministerio de Ciencia y Tecnología (Spain) through grant TIC2003-1504.

a signal from arbitrarily colored noise we must rely on differences in their higher-order statistics. Particularly, if we consider that the noise is Gaussian, then the image model must be non-Gaussian.

In previous works, we have described a model for vectors of coefficients using scale mixtures of Gaussians in overcomplete oriented pyramids [2]. Related models have been developed by other groups (e.g. [3, 4]). We have applied it to estimate images in the presence of independent additive Gaussian noise of known covariance [5, 6]. The referred image-plus-degradation model has the two required features pointed out above: 1) it captures significant higher-order statistics present in natural images; and 2) it considers noise with arbitrary covariance. However, assuming prior knowledge of the noise covariance results in severe limitations when applying the algorithm in many practical situations, where this information is not available.

In this paper we propose a generalized expectation maximization (GEM) method within the image-plus-degradation model frame described in [6]. This method provides good approximations to the most likely noise covariance matrices for every observed pyramid subband. Incorporating these estimates into the method described in [6], results in a full blind denoising method, applicable to any Gaussian or mesokurtotic, homogeneous additive perturbation, regardless of its power spectral density. We have tested its performance with both simulations (low-pass, high-pass and band-pass noise) and with real corrupted images, obtaining very satisfactory results. Besides blind denoising, another possibility is to use the noise spectral characterization for automatic quality assessment (a "noise meter").

2. A MODEL FOR NOISY IMAGES

Multi-scale oriented subband representations are widely spread tools in the image processing field. For noise removal, overcomplete pyramids (e.g. [7, 6]) have demonstrated higher performance than orthogonal wavelets. This

is a consequence of being translation-invariant and redundant [8, 9]. Any overcomplete oriented pyramid can be used to apply the method described here (see implementation in section 5).

For each coefficient belonging to a subband in the pyramid, we consider a *neighborhood* of coefficients around it. We have used a 3×3 neighborhood within the same subband for the results shown in this paper¹. For each subband we use a Gaussian scale mixture (GSM) to model the statistics of the neighborhoods of clean coefficients grouped in vectors. A Gaussian scale mixture [10] $\mathbf{x} = \sqrt{z}\mathbf{u}$ is the product of a zero-mean Gaussian vector \mathbf{u} and an independent positive scalar random variable \sqrt{z} . Marginal and joint statistics of GSM distributions are qualitatively similar to those of neighbor coefficients in oriented pyramids responding to real images, respect to their high kurtosis and the positive correlation in amplitude among neighbors [2]. For the degradation, we consider zero-mean independent additive Gaussian noise of arbitrary autocovariance \mathbf{a}_w . This produces zero-mean additive Gaussian noise (\mathbf{w}) in each subband of the pyramid, but with a different noise covariance matrix \mathbf{C}_w^j for each subband j ². The vector of a neighborhood of observed coefficients is, then:

$$\mathbf{y} = \mathbf{x} + \mathbf{w} = \sqrt{z}\mathbf{u} + \mathbf{w}. \quad (1)$$

Being \mathbf{u} and \mathbf{w} Gaussian and independent, $p_{\mathbf{x}}^j(\mathbf{x}|z)$ and $p_{\mathbf{y}|z}^j(\mathbf{y}|z)$ are also Gaussian and both the original \mathbf{x} and the observation \mathbf{y} are infinite mixtures of Gaussian vectors:

$$p_{\mathbf{y}}^j(\mathbf{y}) = \int_0^\infty p_{\mathbf{y}|z}^j(\mathbf{y}|z) p_z^j(z) dz \quad (2)$$

(analogous for $p_{\mathbf{x}}^j(\mathbf{x})$). Without loss of generality we set $\mathbb{E}\{z\} = 1$, which implies $\mathbf{C}_{\mathbf{x}}^j = \mathbf{C}_{\mathbf{u}}^j$. Note also that $\mathbf{C}_{\mathbf{x}}^j = \mathbf{C}_{\mathbf{y}}^j - \mathbf{C}_{\mathbf{w}}^j$. Then the covariance of \mathbf{y} for a given z is $\mathbf{C}_{\mathbf{y}|z}^j = z\mathbf{C}_{\mathbf{u}}^j + \mathbf{C}_{\mathbf{w}}^j = z\mathbf{C}_{\mathbf{y}}^j + (1-z)\mathbf{C}_{\mathbf{w}}^j$, and

$$p_{\mathbf{y}|z}^j(\mathbf{y}|z) = \frac{\exp(-\mathbf{y}^T(z\mathbf{C}_{\mathbf{y}}^j + (1-z)\mathbf{C}_{\mathbf{w}}^j)^{-1}\mathbf{y}/2)}{(2\pi)^{N/2}|z\mathbf{C}_{\mathbf{y}}^j + (1-z)\mathbf{C}_{\mathbf{w}}^j|^{1/2}} \quad (3)$$

where N is the number of coefficients in the neighborhood (9 in our case). The unknown features of $p_{\mathbf{y}}^j(\mathbf{y})$, then, are the noise covariance matrix $\mathbf{C}_{\mathbf{w}}^j$ and the mixing density $p_z^j(z)$.

3. ESTIMATING THE MODEL PARAMETERS

At this point, we have a model able to represent significant statistics of a wide range of degraded images. We have previously developed a Bayesian procedure to estimate \mathbf{x} from

¹It is also possible to include neighbors from different subbands [6].

²There is a simple relationship between \mathbf{a}_w and the $\mathbf{C}_{\mathbf{w}}^j$'s [6], although when doing blind denoising we do not know \mathbf{a}_w a priori. Instead, we may be interested in obtaining \mathbf{a}_w from $\{\mathbf{C}_{\mathbf{w}}^j\}$.

each observation \mathbf{y} [6]. The only piece missing to complete a full blind denoising algorithm is a method for estimating, from a single corrupted image, the unknown features $\{\mathbf{C}_{\mathbf{w}}^j, p_z^j(z)\}$. We will use initial estimates on these parameters and will update them iteratively, providing that the likelihood increases at each step (generalized expectation-maximization). From now on, we will focus on the estimates for a single subband, and we will drop the superindices j . We will fit the model to M^j non-overlapping observations, $\{\mathbf{y}_m, m = 1 \dots M\}$, considering them as independent realizations of $p_{\mathbf{y}}(\mathbf{y})$.

3.1. Mixing density $p_z(z)$

Previous works using a hidden multiplier for describing the observations have typically used a parametric prior for describing its density (e.g. a log-normal function [5], an exponential function [4], or a fixed non-informative prior [6, 11]). However, we have experienced that a more accurate estimation of $p_z(z)$ is necessary when additional unknown parameters are to be estimated (\mathbf{C}_w , in our case).

Although we do not show it here for lack of space, it is easy to demonstrate that the classical EM solution used to ML-estimate the probability mass of each index k for a finite mixture of K densities (see, e.g., [12]) also applies for infinite Gaussian scale mixtures, each z value playing the same role as the index k in the discrete mixture:

$$\begin{aligned} p_z^{new}(z) &= \frac{1}{M} \sum_{m=1}^M p_{z|\mathbf{y}}(z|\mathbf{y}_m; p_z^{old}(z)) \\ &= p_z^{old}(z) \frac{1}{M} \sum_{m=1}^M \frac{p_{\mathbf{y}|z}(\mathbf{y}_m|z)}{\int_0^\infty p_{\mathbf{y}|z}(\mathbf{y}_m|\alpha) p_z^{old}(\alpha) d\alpha}, \end{aligned} \quad (4)$$

where the Bayes rule has been applied to express the posteriors in terms of the known densities $p_{\mathbf{y}|z}(\mathbf{y}|z)$ and $p_z^{old}(z)$.

3.2. Noise covariance matrix \mathbf{C}_w

Starting with an initial guess, we look for an updating rule such that the likelihood of the observations according to the model increases. We use the $Q(\mathbf{C}_w, \mathbf{C}_w^{old})$ criterion [13]

$$Q = \sum_{m=1}^M \int_0^\infty p_{z|\mathbf{y}}(z|\mathbf{y}_m; \mathbf{C}_w^{old}) \log(p(\mathbf{y}_m, z; \mathbf{C}_w)) dz,$$

where we have made explicit the dependencies on \mathbf{C}_w and \mathbf{C}_w^{old} . Using $\log(p(\mathbf{y}_m, z; \mathbf{C}_w)) = \log(p_{\mathbf{y}|z}(\mathbf{y}_m|z; \mathbf{C}_w)) + \log(p_z(z; \mathbf{C}_w))$ and noting that $p_z(z)$ does not depend on \mathbf{C}_w , we express the gradient of $Q(\mathbf{C}_w, \mathbf{C}_w^{old})$ w.r.t. \mathbf{C}_w as:

$$\frac{\partial Q}{\partial \mathbf{C}_w} = \sum_{m=1}^M \int_0^\infty p_{z|\mathbf{y}}(z|\mathbf{y}_m; \mathbf{C}_w^{old}) \frac{\partial \log(p_{\mathbf{y}|z}(\mathbf{y}_m|z; \mathbf{C}_w))}{\partial \mathbf{C}_w} dz.$$

After operating in the previous equation [14] we obtain:

$$\frac{\partial Q}{\partial \mathbf{C}_w} = \frac{M}{2} \int_0^\infty p_z^{new}(z)(1-z)\mathbf{C}_z^{-1}(\mathbf{I} - \widehat{\mathbf{C}}_z\mathbf{C}_z^{-1})dz, \quad (5)$$

where

$$\mathbf{C}_z = \mathbf{C}_{y|z} = z\mathbf{C}_y + (1-z)\mathbf{C}_w \quad (6)$$

$$\widehat{\mathbf{C}}_z = \frac{\sum_{m=1}^M p_{z|y}(z|\mathbf{y}_m; p_z^{old}(z), \mathbf{C}_w^{old})\mathbf{y}_m\mathbf{y}_m^T}{\sum_{m=1}^M p_{z|y}(z|\mathbf{y}_m; p_z^{old}(z), \mathbf{C}_w^{old})}. \quad (7)$$

\mathbf{I} is the identity matrix of dimensions $N \times N$ and $p_z^{new}(z)$ is estimated through Eq. 4 (where the posteriors are obtained using \mathbf{C}_w^{old}). Unfortunately, equating the r.h.s. of Eq. 5 to zero and solving for \mathbf{C}_w , as the maximization step of an EM method, is a difficult task. Instead we just search for a \mathbf{C}_w^{new} accomplishing $Q(\mathbf{C}_w^{new}, \mathbf{C}_w^{old}) > Q(\mathbf{C}_w^{old}, \mathbf{C}_w^{old})$. An obvious possibility for updating the estimate is following the gradient direction:

$$\mathbf{C}_w^{new} = \mathbf{C}_w^{old} + \eta \frac{\partial Q(\mathbf{C}_w, \mathbf{C}_w^{old})}{\partial \mathbf{C}_w} \Big|_{\mathbf{C}_w = \mathbf{C}_w^{old}}, \quad (8)$$

where η represents a suitable scale factor. However, as the evaluation of $Q(\mathbf{C}_w, \mathbf{C}_w^{old})$ is computationally costly, gradient ascent is not an efficient solution in this case. Instead, we observe that: 1) $\mathbf{C}_w = \mathbf{C}_z|_{z=0}$; and 2) $\widehat{\mathbf{C}}_z$ could be regarded as ML-estimates of the covariance matrices \mathbf{C}_z (Gaussian mixture EM solutions [12]), if only those matrices were not coupled to each other through Eq. 6. Thus, $\widehat{\mathbf{C}}_z|_{z=0}$ is a reasonable (but not optimal) estimate of \mathbf{C}_w . According to this, we choose the new updating rule:

$$\mathbf{C}_w^{new} = \frac{\sum_{m=1}^M p(0|\mathbf{y}_m; p_z^{old}(z), \mathbf{C}_w^{old})\mathbf{y}_m\mathbf{y}_m^T}{\sum_{m=1}^M p(0|\mathbf{y}_m; p_z^{old}(z), \mathbf{C}_w^{old})}. \quad (9)$$

Note that the true posteriors $p(0|\mathbf{y}_m)$ would only contribute significantly to this estimate at the image locations where the noise clearly dominates over the signal. This explains why we can expect a good estimate of the noise covariance after a few iterations. Actually, we have found that Eq. 9, together with Eq. 4, provides a very fast fitting of the model to the data. However, Eq. 9 does not guarantee a monotonic increase in Q , and thus, it must be replaced by Eq. 8 for the rare cases where it decreases Q .

4. BLIND DENOISING

The Bayes Least Square solution we proposed in [6] for estimating x_c (central sample of \mathbf{x}) given each \mathbf{y} is a weighted average of the Wiener solutions for every z , according to their posterior density:

$$\hat{x}_c = \int_0^\infty [\mathbb{E}\{\mathbf{x}|\mathbf{y}, z\}]_c p_{z|y}(z|\mathbf{y}) dz \quad (10)$$

We have fused our GEM estimation technique with the original (non-blind) denoising method [6], yielding the full blind denoising scheme below. Note that for estimating the clean coefficients x_c we use all possible (overlapping) neighborhoods, whereas for estimating \mathbf{C}_w and $p_z(z)$ we use non-overlapping neighborhoods.

1. Decompose the image into subbands
2. For each subband:
 - (a) Organize the coefficients into vectors
 - (b) Estimate \mathbf{C}_y from the vectors
 - (c) Use initial estimates for \mathbf{C}_w and $p_z(z)$
 - (d) While not convergence
 - i. Compute $\mathbf{C}_z, p_{y|z}(\mathbf{y}_m|z), p_{z|y}(z|\mathbf{y}_m), \widehat{\mathbf{C}}_z$
 - ii. Update $p_z(z)$ (Eq.4)
 - iii. Update \mathbf{C}_w (Eq.9 / Eq.8)
 - (e) Estimate x_c from \mathbf{y} (Eq.10) [6]
3. Reconstruct the image from the denoised subbands

5. IMPLEMENTATION

For the results in this work we have used an overcomplete version of the Haar wavelet, with an oversampling factor of 2×2 for the highest frequency subbands, and 4×4 for the rest (overall redundancy factor of 7). This structure makes intra-subband aliasing negligible, whereas, unlike the *a trous* scheme, it maintains a pyramidal structure from the second level upwards, which is necessary for this multi-scale method.

As in previous implementations of the GSM model, we have used a finite number of scales for the Gaussian scale mixture (which becomes in fact a finite mixture). We have done that by sampling z on a finite interval. Noting that $p_z(z)$ is fairly concentrated around zero [4, 5, 6], we have chosen a logarithmic sampling, ranging $\log z$ from -5.5 to 5.0, in intervals of 1.5, for a total of just 7 scale samples. As the value $z = 0$ does not actually belong to that set, we have used in Eq. 9 the smallest value in our set ($\exp(-5.5) \approx 0.004$). We have used for initial guesses of the model parameters $p_z^0(z) = 1/z$ (uniform when sampled in the logarithm), and $\mathbf{C}_w^0 = 0.9\mathbf{C}_y$. We have set $\Delta Q < 10^{-4}$ as the stop criterion in the estimation of \mathbf{C}_w and $p_z(z)$.

To accelerate computations, we have applied a double diagonalization [5, 6] to \mathbf{C}_y and \mathbf{C}_w , and the corresponding whitening of the observations: $\mathbf{y}^T[z\mathbf{C}_y + (1-z)\mathbf{C}_w]^{-1}\mathbf{y} = \mathbf{y}^T\mathbf{S}^{-T}[z\mathbf{S}^{-1}\mathbf{C}_y\mathbf{S}^{-T} + (1-z)\mathbf{I}]^{-1}\mathbf{S}^{-1}\mathbf{y} = \mathbf{y}^T\mathbf{S}^{-T}\mathbf{Q}[z\Lambda + (1-z)\mathbf{I}]^{-1}\mathbf{Q}^T\mathbf{S}^{-1}\mathbf{y} = \sum_{n=1}^N v_n^2 / (z(\lambda_n - 1) + 1)$, where $\mathbf{C}_w = \mathbf{S}\mathbf{S}^T$, $[\mathbf{Q}, \Lambda]$ are the eigenvectors-eigenvalues of $\mathbf{S}^{-1}\mathbf{C}_y\mathbf{S}^{-T}$, and $\mathbf{v} = \mathbf{Q}^T\mathbf{S}^{-1}\mathbf{y}$. This transformation also gives $|\mathbf{C}_z| = |\mathbf{C}_w| \prod_{n=1}^N (z(\lambda_n - 1) + 1)$. Note that a new diagonalization is required every time \mathbf{C}_w is updated. In our non-optimized MATLAB© implementation on a Pentium© IV 2.0 Ghz system, we have obtained denoising times around 0.5 min for 256² images and 1.5 min for 512² images.

6. RESULTS AND DISCUSSION

Figure 1 shows some results (cropped, from the test image *Einstein*) obtained simulating three different Gaussian noise sources with the same variance ($\sigma^2 = 625$), but different spectral features (low-pass, high-pass and narrow-band-pass). Left column shows the simulated noisy images, middle column shows the results obtained with our full blind denoising method, and right column are the results obtained assuming white noise of the correct variance (using [6]). We observe the robustness and versatility of the model, and how negative can be for the estimation to assume white noise when there is significant covariance. We have also performed many experiments with a wide range of sources of real noisy images (digital photography, infrared, LADAR, optical coherence tomography, etc.³), obtaining very satisfactory results (not included here for lack of space). We have experienced that even when the noise is not Gaussian, the algorithm performs well, whenever 1) it is spatially homogeneous; and 2) it has kurtosis ≤ 3 . Otherwise the noise is not identified as such, because it becomes compatible with the GSM image model. Another issue is that the kurtosis of the subbands responding to natural images usually decreases as we go up into the pyramid levels, making more difficult the discrimination of signal and noise. Fortunately, the subbands of the higher levels usually keep high SNR ratios, so they may be left unprocessed.

7. REFERENCES

- [1] D L Donoho and I M Johnstone, "Ideal spatial adaptation via wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.
- [2] M J Wainwright and E P Simoncelli, "Scale mixtures of Gaussians and the statistics of natural images," in *Adv. Neural Inform. Proc. Sys.*, May 2000, vol. 12, pp. 855–861.
- [3] M S Crouse, R D Nowak, and R G Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Proc.*, vol. 46, pp. 886–902, April 1998.
- [4] M K Mihçak, I Kozintsev, K Ramchandran, and P Moulin, "Low-complexity image denoising based on statistical modeling of wavelet coefficients," *IEEE Trans. Sig. Proc.*, vol. 6, no. 12, pp. 300–303, December 1999.
- [5] 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, October 2001.
- [6] J Portilla, V Strela, M J Wainwright, and E P Simoncelli, "Image denoising using scale mixtures of Gaussians in the wavelet domain," *IEEE Trans. Image Proc.*, vol. 12, pp. 1338–1351, November 2003.
- [7] J Starck, E J Candes, and D L Donoho, "The curvelet transform for image denoising," *IEEE Trans. Image Proc.*, vol. 11, no. 6, pp. 670–684, June 2002.

³We thank Carlos Dorronsoro (CIDA, Spain), for facilitating us those images, and for useful comments.

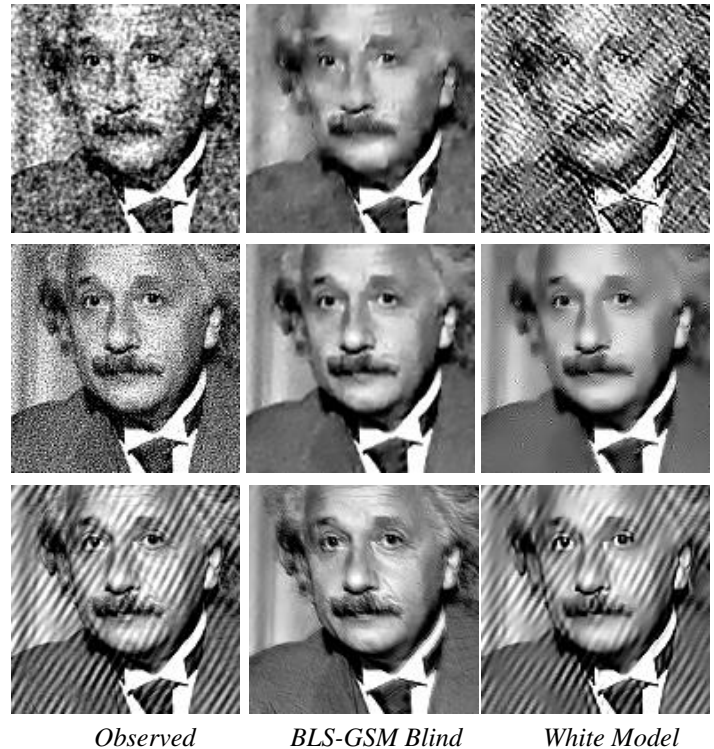


Fig. 1. A comparison of results obtained with simulated noise sources with common variance (625) but different spectral densities. PSD functions and PSNR values are: **1st row** (low pass noise $P_w(u, v) \propto \exp(-20(u^2 + v^2))$): 20.17, 26.84, 16.37; **2nd row** (high pass noise $P_w(u, v) \propto 1 - \exp(-2(u^2 + v^2))$): 20.17, 31.10, 28.20; **3rd row** (narrow-band pass noise $P_w(u, v) \propto \text{Re}\{2 \exp(-3000((u - 0.17)^2 + (v - 0.10)^2))\}$): 20.17, 32.48, 21.04. u and v are in cycles/pixel. Images have been cropped to 128^2 for visibility of the artifacts.

- [8] E P Simoncelli, W T Freeman, E H Adelson, and D J Heeger, "Shiftable multi-scale transforms," *IEEE Trans Information Theory*, vol. 38, no. 2, pp. 587–607, March 1992.
- [9] R R Coifman and D L Donoho, "Translation-invariant denoising," in *Wavelets and statistics*. Springer-Verlag lecture notes, San Diego, 1995.
- [10] D Andrews and C Mallows, "Scale mixtures of normal distributions," *J. Royal Stat. Soc.*, vol. 36, pp. 99–, 1974.
- [11] M Figueiredo and R Nowak, "Wavelet-based image estimation: An empirical Bayes approach using Jeffrey's noninformative prior," vol. 10, pp. 1322–1331, September 2001.
- [12] R O Duda, P E Hart, and D G Stork, *Pattern Classification, 2nd Edition*, chapter Unsupervised Learning and Clustering, Wiley Interscience, 2001.
- [13] G J McLachlan and T Krishnan, *The EM algorithm and extensions*, Wiley, New York, 1996.
- [14] J Portilla, "Full blind Bayesian denoising using scale mixtures of Gaussians in the wavelet domain," Tech. Rep. in preparation, DECSAI, Universidad de Granada, Spain, 2004.