

EM ALGORITHMS FOR ROBUST SIGNAL FILTERING AND PREDICTION

Guang Deng

Department of Electronic Engineering, La Trobe University
Bundoora, Victoria 3083, Australia.
email: d.deng@ee.latrobe.edu.au

ABSTRACT

Transform domain denoising, noise filtering based on data from a local neighborhood and linear prediction are three important signal processing tasks. In this paper we treat these tasks from a maximum a posteriori estimation (MAP) perspective and address the problem of robust estimation. The Student-t and Laplacian distributions are used to model the noise to permit robustness to outliers. Independent Gaussian distributions with different variances are used as the prior distributions for the parameters to be estimated. This provides a mechanism to incorporate into the solution certain desirable properties such as the sparseness constraint in transform domain denoising and regularization in linear prediction. EM algorithms are developed for the three signal processing tasks. Applications are demonstrated.

1. INTRODUCTION

Transform domain denoising, noise filtering based on data from a local neighborhood and linear prediction are three important signal processing tasks. In this paper we treat these tasks from a maximum a posteriori estimation (MAP) perspective and address the problem of robust estimation. Wavelet domain denoising has been treated as an MAP estimation problem in which the Gaussian distribution is used to model noise and a number of prior distributions have been used to represent the sparseness constraint on the wavelet coefficients [1, 4, 3, 14, 5]. A robust wavelet denoising scheme has been recently published [13]. In this scheme, the square-loss function (due to the Gaussian distribution) is replaced by the Huber's loss-function.

A typical noise filtering technique for image processing is to replace the current pixel gray level by a weighted average of its neighbouring pixels. A classical approach on this subject [8] uses Gaussian models for both signal and noise. Weights are adaptively changed according to the assumed statistical models.

In lossless image compression [9, 12], the least squares approach has been applied to design a linear predictor for each pixel. The assumption is that if the predictor performs well for the training image block, then it will perform reasonably well for the current pixel to be predicted. To improve the predictive performance, regularization can be applied [7]. The regularized least squares estimation is effectively an MAP estimation with Gaussian as the noise model and the prior model for the coefficients as regularization term[2].

In this paper, we formulate the above three problems in a unified MAP estimation perspective. Instead of using the Gaussian to model noise, we use two distributions with long tails: the Student-t distribution and the Laplacian distribution. The motivation is to develop new algorithms that are robust to outliers. We use independent Gaussian distributions with different variances as the prior distributions for the parameters to be estimated. This provides a mechanism to incorporate into the solution certain desirable properties such as the sparseness constraint in transform domain denoising [15, 1, 3] and regularization in linear prediction [10, 7]. The statistical modes for the three problems are listed in Table 1. In this paper, the model parameters are assumed unknown and we use the EM algorithm to solve an MAP problem with unknown parameters[11]. The EM algorithm has been recently used in wavelet based image estimation [4, 5]

2. PROPOSED EM ALGORITHMS

2.1 Decomposition of probability distributions

The Student-t distribution of a random variable x with a fixed degree of freedom v , can be represented as [11]

$$t_v(x|\mu, \sigma_0^2) = \int_0^\infty p(x|\mu, \frac{\sigma_0^2}{u})p(u)du \quad (1)$$

where μ is the mean, σ_0^2 is a scaling parameter, $p(x|\mu, \frac{\sigma_0^2}{u}) \sim N(\mu, \frac{\sigma_0^2}{u})$ and $p(u) = \Gamma(u|\frac{v}{2}, \frac{v}{2})$ is a gamma distribution. Similarly, the Laplacian distribution [5]

$$p(x|\mu, \sqrt{\eta}) = \frac{\sqrt{\eta}}{2} e^{-\sqrt{\eta}|x-\mu|} \quad (2)$$

can be represented as

$$p(x|\mu, \sqrt{\eta}) = \int_0^\infty p(x|\mu, u)p(u)du \quad (3)$$

where $p(x|\mu, u) \sim N(\mu, u)$ and $p(u) = \frac{\eta}{2} e^{-\frac{\eta u}{2}}$. To simplify our discussion, we assume that the noise variance (given by $\sigma_i^2 = \frac{v}{v-2}\sigma_0^2$, $v > 2$, in the Student-t distribution and $2/\eta$ in the Laplacian distribution) is known. If they are assumed unknown, they can be easily incorporated into the EM algorithm.

With these representations, we can regard the parameter u as the missing data and solve MAP estimation problem using the EM algorithm. Since procedures for solving the three problems mentioned in Section 1 are similar, we outline that for the transform domain signal estimation problem with noise being modelled by a Student-t distribution. We then present results for other problems.

2.2 Transform domain signal estimation

When noise is modelled by the Student-t distribution, the MAP estimation of the transform coefficient vector can be stated as¹

$$\hat{\phi} = \arg \max_{\phi} p(\phi|\mathbf{y}) \quad (4)$$

where $\phi = \{\mathbf{s}, \sigma_i^2 (i = 1 : N)\}$ represent parameters to be estimated. Let $\gamma = \{u_i (i = 1 : N)\}$ be the missing data. In the E-step, we evaluate the conditional mean for γ . This is equivalent to evaluate each $u_i^{(k)} = E[u_i|\phi^{(k-1)}, \mathbf{y}]$, where the superscripts (k) and ($k-1$) are used to indicate the iteration index. It can be shown that

$$p(u_i|\phi^{(k-1)}, \mathbf{y}) = \Gamma(\alpha, \beta) \quad (5)$$

where

$$\alpha = \frac{v+1}{2} \quad (6)$$

¹In this paper, we use a bold-face small letter to represent a column vector and a bold-face capital letter to represent a matrix. The i th entry of a vector \mathbf{s} is represented as s_i .

Problem	1	2	3
observation	$\mathbf{y} = \mathbf{s} + \mathbf{n}$	$\mathbf{y} = \mathbf{s} + \mathbf{n}$	$\mathbf{y} = \mathbf{A}\mathbf{w} + \mathbf{n}$
noise model-1	i.i.d, zero mean Student-t distribution		
noise model-2	i.i.d, zero mean Laplacian distribution		
signal model	$s_i \sim N(0, \sigma_i^2)$	$s_i \sim N(\mu, \sigma_i^2)$	$w_i \sim N(0, \sigma_i^2)$

Table 1: The three estimation problems and their associated statistical models studied in this paper. Problem-1 is the transform domain signal estimation problem, Problem-2 is the local signal estimation problem and Problem-3 is a linear prediction problem. \mathbf{y} is the observed signal vector, \mathbf{s} or \mathbf{w} is the vector to be estimated and \mathbf{n} is the noise vector. In all three problems, both noise and signal are model independent. $N(\mu, \sigma^2)$ represents a Gaussian distribution of mean μ and variance σ^2 .

	Student-t	Laplacian
\mathbf{D}	$\text{diag} \left[\frac{u_i^{(k)}}{\sigma_0^2} \right]$	$\text{diag} \left[\frac{1}{u_i^{(k)}} \right]$
\mathbf{R}	$\mathbf{R} = \text{diag} \left[(\sigma_i^{-2})^{(k-1)} \right]$	

Table 2: The definitions of the two diagonal matrices \mathbf{D} and \mathbf{R} for the two noise models.

and

$$\beta = \frac{1}{2}v + \frac{1}{2} \frac{(y_i - s_i^{(k-1)})^2}{\sigma_0^2}. \quad (7)$$

The conditional mean is thus given by $u_i^{(k)} = \alpha/\beta$. In the M-step, we maximize the following function² with respect to the parameters to be estimated

$$\begin{aligned} Q(\phi, \phi^{(k-1)}) &= \int (\log p(\phi, \gamma | \mathbf{y})) p(\gamma | \phi^{(k-1)}, \mathbf{y}) d\gamma \\ &= -\frac{1}{2}(\mathbf{y} - \mathbf{s})^T \mathbf{D}(\mathbf{y} - \mathbf{s}) \\ &\quad - \frac{1}{2} \sum \log \sigma_i^2 - \frac{1}{2} \mathbf{s}^T \mathbf{R} \mathbf{s} \end{aligned} \quad (8)$$

The definitions of the two diagonal matrices \mathbf{D} and \mathbf{R} for the two noise models are given in Table 2. Note that these definitions are valid for the three estimation problems studied in this paper. Following the same procedure, we can develop the EM algorithm for the case with a Laplacian noise model. The objective function to be optimized is in the same form as equation (8). However, the definitions for $u_i^{(k)}$ and \mathbf{D} are different. Results are summarized in Table 3.

We make the following observations. (1) The estimated signal is given by a shrinkage of the noise corrupted signal. For example, with the Student-t model, we have (see M-step 2 in Table 3)

$$s_i^{(k)} = \frac{1}{1 + \frac{(v\sigma_0^2 + (y_i - s_i^{(k-1)})^2)/(v+1)}{(s_i^{(k-1)})^2}} y_i \quad (9)$$

It is interesting to note that while $(s_i^{(k-1)})^2$ can be interpreted as an estimate of the signal variance, the quantity

$$\frac{v\sigma_0^2 + (y_i - s_i^{(k-1)})^2}{(v+1)} = \frac{(v-2)\sigma_i^2 + (y_i - s_i^{(k-1)})^2}{(v+1)}$$

is an estimate of the noise variance. Thus the shrinkage is a function of the noise-to-signal ratio. A higher ratio will lead to more

²Note that we have omitted constants and unrelated terms.

	Student-t	Laplacian
E-step	$u_i^{(k)} = \frac{v+1}{v + \frac{(y_i - s_i^{(k-1)})^2}{\sigma_0^2}}$	$u_i^{(k)} = \frac{ y_i - s_i^{(k-1)} }{\sqrt{\eta}}$
M-step 1	$\lambda = \frac{\sigma_0^2}{u_i^{(k)} (\sigma_i^2)^{(k-1)}}$	$\lambda = \frac{u_i^{(k)}}{(\sigma_i^2)^{(k-1)}}$
2	$s_i^{(k)} = \frac{y_i}{1 + \lambda}$	
3	$(\sigma_i^2)^{(k)} = (s_i^{(k)})^2$	

Table 3: EM algorithms for transform domain signal estimation with noise modelled by the Student-t and Laplacian distributions with known variances, respectively.

shrinkage that results in smaller value of $s_i^{(k)}$. Since the algorithm is operated iteratively, some signal samples will be effectively shrunk to zero. This has the same effect as the hard thresholding in wavelet based denoising. The thresholding function of the proposed algorithm comes from the sparseness constraint (expressed as an i.i.d. Gaussian distribution with different variances) on the signal. (2) On the other hand, unlike hard thresholding, equation (9) can also be regarded as a Wiener estimate of the signal. Therefore, the proposed algorithm is effectively a combination of Wiener filtering and hard thresholding. (3) The robustness to outliers is controlled by the degree of freedom v in the Student-t distribution and η in the Laplacian distribution. This can be easily seen from the estimated noise variance which is a weighted sum of the noise variance σ_i^2 and a local estimate $(y_i - s_i^{(k-1)})^2$. As v increases, the local estimate is de-emphasized. Thus the effect of outliers is suppressed.

2.3 Local signal estimation

The local signal estimation problem is similar to the transform domain estimation problem. We only need to change the prior model of s_i from zero mean to non-zero mean. As such, the objective function is given by (constants and unrelated terms are dropped)

$$\begin{aligned} Q(\phi, \phi^{(k-1)}) &= -\frac{1}{2}(\mathbf{y} - \mathbf{s})^T \mathbf{D}(\mathbf{y} - \mathbf{s}) \\ &\quad - \frac{1}{2} \sum \log \sigma_i^2 - \frac{1}{2}(\mathbf{s} - \boldsymbol{\mu})^T \mathbf{R}(\mathbf{s} - \boldsymbol{\mu}) \end{aligned} \quad (10)$$

Note that the objective function is in the same form for both noise models. As in the previous section, in the E-step, we calculate the conditional mean $u_i^{(k)}$ while in the M-step, we maximize the objective function with respect to the parameters to be estimated. Results are summarized in Table 4

We can make the following observations. (1) The algorithm produces two useful outputs—the estimated mean and the estimated signal. Both outputs can be used in subsequent processing. While the mean is a weighted average of the estimated signal, the signal is

	Student-t	Laplacian
E-step	$u_i^{(k)} = \frac{v+1}{v + \frac{(y_i - s_i^{(k-1)})^2}{\sigma_0^2}}$	$u_i^{(k)} = \frac{ y_i - s_i^{(k-1)} }{\sqrt{\eta}}$
M-step 1	$\lambda = \frac{\sigma_0^2}{u_i^{(k)}(\sigma_0^2)^{(k-1)} + \sigma_0^2}$	$\lambda = \frac{u_i^{(k)}}{u_i^{(k)} + (\sigma_0^2)^{(k-1)}}$
2	$s_i^{(k)} = y_i + \lambda(\mu^{(k-1)} - y_i)$	
3	$\mu^{(k)} = \frac{\sum s_i^{(k)} / (\sigma_i^2)^{(k)}}{\sum 1 / (\sigma_i^2)^{(k)}}$	
4	$(\sigma_i^2)^{(k)} = (s_i^{(k)} - \mu^{(k)})^2$	

Table 4: EM algorithms for local neighborhood signal estimation with noise modelled by Student-t and Laplacian distributions with known variances, respectively.

expressed as the noisy signal plus a scaled correction. The scaling factor λ (in the case of using the Student-t noise model) can be written as

$$\lambda = \frac{1}{1 + \frac{(s_i^{(k-1)} - \mu^{(k-1)})^2}{(v\sigma_0^2 + (y_i - s_i^{(k-1)})^2)/(v+1)}} \quad (11)$$

We recognize that while the term $(s_i^{(k-1)} - \mu^{(k-1)})^2$ is an estimate of the signal variance, the other term $(v\sigma_0^2 + (y_i - s_i^{(k-1)})^2)/(v+1)$ is an estimate of the noise variance. Therefore, λ is a function of the signal-to-noise ratio (SNR). A higher SNR results in a smaller λ and less correction is applied. Thus, the output is close to y_i . On the other hand, if the SNR is low, for example $SNR \rightarrow 0$, then $\lambda \rightarrow 1$ and the output is close to $\mu^{(k-1)}$. (2) Comments on the robustness to outliers in previous section also apply to algorithms presented in this section. (3) In our formulation of the problem, the signal samples in a local neighborhood are modelled by independent Gaussian distributions with identical mean, but possibly different variances. This model permits the incorporation of the structural information into the model. For example, we can easily modify the above EM algorithms such that σ_i^2 is regarded as fixed and is defined as a function of the distance between the center pixel and its neighbouring pixel.

2.4 Linear prediction

The linear prediction problem is very similar to the transform domain estimation problem. We only need to replace in the objective function s_i with $\mathbf{A}_i \mathbf{w}$, where \mathbf{A}_i is the i th row vector of the matrix \mathbf{A} . The objective function for both noise models is given by (constants and unrelated terms are dropped)

$$\mathcal{Q}(\phi, \phi^{(k-1)}) = -\frac{1}{2}(\mathbf{y} - \mathbf{A}\mathbf{w})^T \mathbf{D}(\mathbf{y} - \mathbf{A}\mathbf{w}) - \frac{1}{2} \sum \log \sigma_i^2 - \frac{1}{2} \mathbf{w}^T \mathbf{R} \mathbf{w} \quad (12)$$

As in the previous section, in the E-step, we calculate the conditional mean $u_i^{(k)}$ while in the M-step, we maximize the objective function with respect to the parameters to be estimated. Results are summarized in Table 5.

We can make the following observations. The vector of prediction coefficients is actually the result of a regularized (penalized) weighted least squares optimization. The regularization provided by the diagonal matrix \mathbf{R} effectively penalizes large coefficients. This is a desirable property that will lead to improved predictive (generalization) performance of the predictor [7]. The i th main di-

	Student-t	Laplacian
E-step	$u_i^{(k)} = \frac{v+1}{v + \frac{(y_i - \mathbf{A}_i \mathbf{w}^{(k-1)})^2}{\sigma_0^2}}$	$u_i^{(k)} = \frac{ y_i - \mathbf{A}_i \mathbf{w}^{(k-1)} }{\sqrt{\eta}}$
M-step 1	$\mathbf{D} = \text{diag} \left[\frac{u_i^{(k)}}{\sigma_0^2} \right]$	$\mathbf{D} = \text{diag} \left[\frac{1}{u_i^{(k)}} \right]$
2	$\mathbf{R} = \text{diag} \left[(\sigma_i^{-2})^{(k-1)} \right]$	
3	$\mathbf{w} = (\mathbf{A}^T \mathbf{D} \mathbf{A} + \mathbf{R})^{-1} \mathbf{A}^T \mathbf{D} \mathbf{y}$	
4	$(\sigma_i^2)^{(k)} = (w_i^{(k)})^2$	

Table 5: EM algorithms for the estimation of linear prediction coefficients with noise modelled by Student-t and Laplacian distributions with known variances, respectively.

agonal entry of the weight matrix is given by

$$d_i = \frac{1}{(v\sigma_0^2 + (y_i - \mathbf{A}_i \mathbf{w}^{(k-1)})^2)/(v+1)} \quad (13)$$

This is inversely proportional to an estimate of the local noise variance which is a weighted average of the global noise variance and the local prediction error. When v is not “too large”, the contribution of prediction error can not be neglected. The weight matrix thus de-emphasizes the role of the data pair (y_i, \mathbf{A}_i) that has large prediction error in determining the prediction coefficients. On the other hand, when v is large, the weight matrix becomes $\mathbf{D} \approx \tau \mathbf{I}$ where \mathbf{I} is an identity matrix and τ is a constant. The estimate given by the equation in M-step 3 reduces to a solution of a regularized least squares solution.

3. APPLICATIONS IN WAVELET DOMAIN DENOISING

Due to space limitation, we only present results in wavelet domain denoising using the proposed EM algorithm with noise being modelled by a Student-t distribution with the degree of freedom $v = 3$. Obviously, v can be treated as a free parameter that can be adjusted for a particular problem. In our experiments, we decompose an image into 3 levels. In each level, we treat each subband as a 1-D signal. These signals are denoised using the proposed EM algorithm. The convergence criteria are: (1) the difference between the denoised image from the previous iteration and that from the current iteration is less than a threshold or (2) the maximum number of iterations has reached.

We use a standard image denoising procedure in the Wavelet Toolbox³ and the ABE-rule [6] to process the same noisy image. The peak-signal-to-noise ratio (PSNR) shown Table 6 are obtained by averaging the results over 100 runs of the program.

We can see from this table that the results of the proposed EM algorithm and the ABE-rule are similar and they are consistently better than those of the denoise function (wden). As the PSNR value can only be used as an indication of the quality of the image, in Figure 1 we show the noisy ($\sigma = 20$) and denoised “Lena” images.

To demonstrate the dynamical characteristics of the EM algorithm that has a combined shrinkage and thresholding effect, we decompose a noisy block signal⁴ using a one-level decomposition

³The two Matlab functions we used are: [thr, sorth, keepapp] = ddencmp(‘den’, ‘wv’, xn) and xd = wdencmp(‘gbl’, xn, Filt, 3, thr, sorth, keepapp), where ‘xn’ is the noisy image, ‘Filt’ is the wavelet filter’s name. The noisy image is generated by using $xn = x + \sigma * \text{randn}(\text{size}(x))$.

⁴The noisy block signal is generated by the Matlab wavelet toolbox function [x,xn] = wnoise(1, 10, 3).

σ /PSNR noisy image	15/24.6	20/22.1	25/20.2	30/18.6
ABE	30.7	29.1	27.7	26.6
wden	27.3	26.6	26.1	25.6
proposed EM algorithm	30.8	29.4	28.1	26.9

Table 6: Average PSNR (dB) over 100 runs of the program using the ABE rule, the denoise function in the Wavelet Toolbox (wden) and the proposed EM algorithm. The top row shows the variance of the added noise and the resulting PSNR of the noisy image.

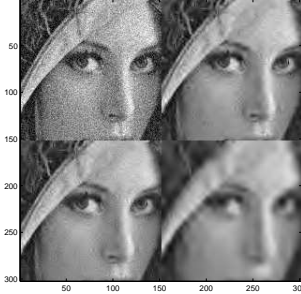


Figure 1: Noisy ($\sigma = 20$) and denoised of the center part of “Lena” images. Top left: noisy image, top right: the proposed EM algorithm denoised image, bottom left: ABE-rule denoised image, bottom right: wden function denoised image.

with the ‘sym4’ wavelet filter. Figure 2 shows the shrinkage factor $\theta_n^{(k)} = \frac{y_n^{(k)}}{y_n}$ in the k th iteration. We make the following observation from this figure. In the first iteration, most wavelet coefficients are scaled (shrunk) by a factor less than 0.5. The shrinkage factor $\theta_n^{(k)}$ ($k=1$) depends on the initial estimate of the noise energy and the squared-value of the wavelet coefficient. Coefficients of smaller values are shrunken to even smaller values. Comparing the sub-figures of iteration 1 with that of iteration 10, we see that as the iteration progresses from the starting stage to the convergent stage, the shrinkage factors for the large value wavelet coefficients are moving back from less than 0.5 to close to 1, while shrinkage factors for the smaller value coefficients are converged to zero.

4. CONCLUSIONS

In this paper, we have studied three important signal processing problems from a unified maximum a posteriori estimation perspective. We have shown that the three problems have very similar objective functions (please refer to equations (8), (10) and (12)) that need to be maximized. For the three problems, we have studied two combinations of statistical models and developed new EM algorithms which are robust and have certain desirable properties. The motivation for studying the Laplacian or the Student-t distribution as a model for the observation data is their inherent robustness. Desirable properties, such as sparseness in wavelet domain denoising, regularization in linear prediction and using the structural information in neighborhood-based filtering, are incorporated into the solution by using an individual Gaussian distribution as a prior model for the parameter to be estimated.

REFERENCES

[1] A. Antoniadis and J. Q. Fan, “Regularization of wavelet approximations,” *Journal of the American Statistical Association*, vol. 96, pp. 939–955, Sept. 2001.

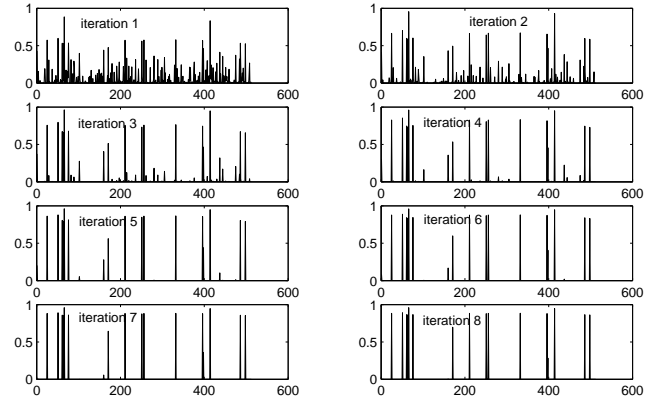


Figure 2: The effects of shrinkage and thresholding of the wavelet coefficients after each iteration of the proposed EM algorithm. These sub-figures (from top to bottom, iteration 1 to 8) show the shrinkage factor $\theta_n^{(k)} = \frac{y_n^{(k)}}{y_n}$ for each wavelet coefficient in each iteration. The horizontal axis is the index of the wavelet coefficient.

[2] C. M. Bishop, *Neural Networks for Pattern Recognition*. London: Oxford University Press, 1995.

[3] S. S. Chen, “Basis pursuit,” Ph.D. dissertation, Department of Statistics, Stanford University, Nov. 1995.

[4] J. Dias, “Fast GEM wavelet-based deconvolution algorithm,” in *Proc. IEEE ICIP’03, CDROM*, Sept. 2003.

[5] M. A. T. Figueiredo, “Adaptive sparseness for supervised learning,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, pp. 1150–1159, Sept. 2003.

[6] M. A. T. Figueiredo and R. D. Nowak, “Wavelet-based image estimation: an empirical bayes approach using jeffreys’ non-informative prior,” *IEEE Trans. Image Processing*, vol. 10, pp. 1322–1331, Sept. 2001.

[7] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag, 2001.

[8] J. S. Lee, “Digital image enhancement and noise filtering by use of local statistics,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 2, pp. 165–168, Mar. 1980.

[9] X. Li and M. Orchard, “Edge-directed prediction for lossless compression of natural images,” *IEEE Trans. Image Processing*, vol. 10, no. 6, pp. 813–817, June 2001.

[10] D. J. C. Mackay, “Bayesian interpolation,” *Neural Computation*, vol. 4, no. 3, pp. 415–447, 1992.

[11] G. J. McLachlan and T. Krishnan, *The EM algorithm and extensions*, ser. Wiley serise in Probability and Statistics. John Wiley & Sons, Inc., 1997.

[12] B. Meyer and P. Tischer, “TMW^{Lego} – an object oriented image modelling framework,” in *Proc. IEEE Data Compression Conference*, Snowbird, Utah, Mar. 2001.

[13] S. Sardy and A. B. P. Tseng, “Robust wavelet denoising,” *IEEE Trans. Signal Processing*, vol. 49, pp. 1146–1152, 2001.

[14] E. P. Simoncelli, “Bayesian denoising of visual images in the wavelet domain,” in *Bayesian Inference in Wavelet Based Models*, ser. Lecture Notes in Statistics. New York: Springer-Verlag, 1999, vol. 141, pp. 291–308.

[15] M. E. Tipping, “Sparse Bayesian learning and the relevance vector machine,” *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.