GENERALIZED WIENER ESTIMATION ALGORITHMS BASED ON A FAMILY OF HEAVY-TAIL DISTRIBUTIONS

Guang Deng

Department of Electronic Engineering, La Trobe University
Bundoora, Victoria 3086, Australia
d.deng@latrobe.edu.au

ABSTRACT

A fundamental problem in signal processing is to estimate signal from noisy observations. When some prior information about the statistical models of the signal and noise is available, the estimation problem can be solved by using the maximum a posteriori (MAP) principle. In this paper, we develop an EM algorithm for the MAP estimate of signals modeled by a family of heavy-tail prior distributions: Laplacian, student-t and slash. We establish links between the EM algorithm and the Wiener estimation. We then modify the EM algorithm and propose two generalized Wiener estimation algorithms for image denoising. Experimental results show that the performance of the proposed algorithms is better than that of the bi-shrinkage algorithm which is arguably one of the best in recent publications.

1. INTRODUCTION

Estimating signal from noisy observations is a fundamental task in signal processing. The simplest Gaussian observation model is given by

$$\mathbf{y} = \mathbf{x} + \mathbf{e} \tag{1}$$

where $\mathbf{x}$ is the true signal vector, $\mathbf{y}$ and $\mathbf{e}$ are vectors of observations and Gaussian noise, respectively. When $\mathbf{x}$ is regarded as a realization of a random vector, the maximum a posteriori (MAP) principle is well established as a powerful tool for estimating the signal. The Wiener estimate is obtained when $\mathbf{x}$ is modeled by a Gaussian distribution [1]. When it is modeled by a Laplacian distribution, the soft-thresholding rule [2] is obtained. Compared to the Gaussian distribution, heavy-tail distributions such as the Laplacian usually fit real-world signals such as the transform domain representation of images better that the Gaussian. A further advantage of using heavy-tail distributions is that they usually result in a sparse wavelet representation of the signal [3–5].

Heavy-tail distributions including student-t [6] and slash distributions [7] have found many successful applications in robust statistical data analysis. Certain members from a family of heavy-tail distributions, which can be expressed as the scale mixtures of Gaussian distributions [8], have also been recently used to solve signal processing [4, 9] and machine learning problems [10].

In practical signal processing problems, parameters (called hyper-parameters) of the heavy-tail distributions are usually unknown. They can be estimated by using the evidence-based method or simply integrated out [11, 12]. They can also be handled by using the EM algorithm [13].

In this paper, we study MAP estimation of the signal with Gaussian additive noise. We study three scale mixtures of Gaussian distributions - Laplacian, student-t and slash as the prior signal distributions. It is assumed that hyper-parameters of the three prior distributions are random variables with their own hyper-prior distributions. We develop a unified EM algorithm to solve the MAP estimation problem under these three prior distributions. We show that the EM algorithm can be regarded as a generalized Wiener estimation. Based on the EM algorithm, we propose two algorithms for image denoising. One is a non-iterative algorithm which has a parameter to account for the heavy-tail characteristics of the signal. The other is an EM algorithm that use local statistical information. Experimental results show that the performance of the proposed algorithms is better than that of the bi-shrinkage algorithm [14] which is arguably one of the best in recent publications.

2. THE GENERALIZED WIENER ESTIMATION ALGORITHM

2.1. Problem formulation

We consider an signal observation model in equation (1), where $\mathbf{y}$ is the observed signal vector with N entries, $\mathbf{x}$ is the true signal vector with its *i*th entry x_i modeled as a zero mean random variable following one of the heavy-tail distributions specified in the next section, $\mathbf{e}$ is the noise vector with its *i*th entry e_i modeled as a random variable following a zero mean Gaussian distribution with variance σ_e^2. We also assume both signal and noise are uncorrelated and modeled as independent identically distributed according to their respective models. Given these model settings, the problem is to develop an EM algorithm to determine an MAP estimate of the signal based on the observations:

$$\mathbf{x}_{opt} = \arg\max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}) \tag{2}$$

2.2. Model specification

A family of heavy-tail distribution [13] for a zero-mean scaler random variable x is defined as a scale mixtures of Gaussian

$$p(x|\sigma^2, \nu) = \int_0^\infty p(x|\sigma^2, u)p(u|\nu)du \tag{3}$$

where $p(x|\sigma^2, u) = \frac{\sqrt{u}}{\sqrt{2\pi}\sigma} e^{-\frac{u}{2\sigma^2}x^2}$ and $p(u|\nu)$ is the prior distribution of u ($0 \le u < \infty$). The two parameters σ^2 and ν can be assumed fixed or can be random variables with their own prior distributions $p(\sigma^2)$ and $p(\nu)$. Different settings for $p(u|\nu)$ result in a family of heavy-tail distributions. The Gaussian distribution is a special case when u is not a random variable but is a constant

	$p(u\|\nu)$	$p(x\|\sigma^2,\nu)$
Laplacian	$\frac{1}{4}u^{-2}e^{-1/4u}$	$\frac{1}{2\sqrt{2}\sigma}e^{-\|x\|/\sqrt{2}\sigma}$
Student-t	$\mathcal{G}(u\|\nu/2,\nu/2)$	$\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi\sigma}}\left(1+x^2/\nu\sigma^2\right)^{-(\nu+1)/2}$
Slash	$\nu u^{\nu-1}$	$\frac{\nu}{\sqrt{2\pi\sigma}}\left(x^2/2\sigma^2\right)^{-(\nu+1/2)}\Gamma(\nu+1/2,x^2/2\sigma^2)$

Table 1. Three heavy-tail distributions and the three settings of $p(u|\nu)$. $\mathcal{G}(u|\nu/2,\nu/2)$ is the gamma distribution. $\Gamma(a)$ and $\Gamma(a,b)=\int_0^b t^{a-1}e^{-t}dt$ are the gamma function and incomplete gamma function, respectively.

$u=1$. In this paper, we consider the following three heavy-tail distributions: Laplacian, student-t and slash. They are shown in Table 1. We note that the Laplacian distribution does not depend on ν. It can be expressed in compact form by simply changing of variable $\alpha=\sqrt{2}\sigma$. However, we use its definition given in Table 1 to develop an EM algorithm which is applicable to the three distributions. Both the student-t and the slash distributions depend on two parameters: the scaling parameter σ and the degrees of freedom ν.

2.3. The EM algorithm

To simplify the development of the EM algorithm, we assume the noise variance σ_e^2 and the degrees of freedom ν are known. Using the above representation for the heavy-tail distribution and the model settings, we can regard the parameters $\mathbf{u}=\{u_i\}$ as the missing data which is denoted $\gamma=\{\mathbf{u}\}$ and the signal $\mathbf{x}$ and the scaling factor σ^2 as the data to be estimated, which is denoted $\phi=\{\mathbf{x},\sigma^2\}$.

To develop an EM algorithm, we have to determine the following function [13]

$$\begin{aligned}&Q(\phi,\phi^{old})\\&=\int\log p(\phi,\gamma|\mathbf{y})p(\gamma|\phi^{old},\mathbf{y})d\gamma\\&=-\frac{\sum(x_i-y_i)^2}{2\sigma_e^2}-\frac{\sum\bar{u}_i x_i^2}{2\sigma^2}+\log p(\sigma^2)-\tfrac{N}{2}\log\sigma^2+C\end{aligned}\tag{4}$$

where C represent unrelated terms and constants that can be omitted, and

$$\bar{u}_i=E\left[u_i|\phi^{old},\mathbf{y}\right]\tag{5}$$

Equation (5) states the calculation required in the E-step. Let $\hat{x}_i$ and $\hat{\sigma}$ represent results from the previous iteration, then the E-step is calculated as the following

$$\bar{u}_i\quad=\int_0^\infty u_i p(u_i|\hat{x_i},\hat{\sigma}^2)du_i\tag{6}$$

where

$$p(u_i|\hat{x}_i,\hat{\sigma}^2)=\frac{p(\hat{x}_i|u_i,\hat{\sigma}^2)p(u_i)}{\int_0^\infty p(\hat{x}_i|u_i,\hat{\sigma}^2)p(u_i)du_i}\tag{7}$$

The E-steps for the three heavy-tail distributions are listed in Table 2.

	Laplacian	student-t	slash
$\bar{u}_i$	$\frac{\hat{\sigma}}{\sqrt{2}\|\hat{x_i}\|}$	$\frac{\nu+1}{\nu+\hat{x}_i^2/\hat{\sigma}^2}$	$\frac{2\hat{\sigma}^2\,\Gamma(\frac{3}{2}+\nu,\hat{x}_i^2/2\hat{\sigma}^2)}{\hat{x}_i^2\,\Gamma(\frac{1}{2}+\nu,\hat{x}_i^2/2\hat{\sigma}^2)}$

Table 2. The E-steps for the three heavy-tail distributions.

The M-steps for the update the signal and the scaling parameter σ^2 are the same for the three distributions. This is evident from equation (4). The update for x_i is given by

$$x_i=\frac{\sigma^2}{\sigma^2+\bar{u}_i\sigma_e^2}y_i\tag{8}$$

The update of the scaling parameter depends on its prior distribution. We consider a conjugate prior for σ^2 given by the inverse-chi-square (Inv-χ^2) distribution and an informative prior-Jeffreys' prior ($p(\sigma^2)\propto 1/\sigma^2$) distributions [6]. The Inv-χ^2 distribution is given by

$$p(\theta|\eta)=\frac{2^{-\eta/2}}{\Gamma(\eta/2)}\theta^{-(\eta/2+1)}e^{-1/(2\theta)},\theta>0,\,\eta>0\tag{9}$$

where η is the degrees of freedom. For $\eta>2$, the mean of $\theta$is given by $E[\theta]=1/(\eta-2)$. Therefore, if we have prior knowledge about the mean of σ^2, say σ_0^2, then $\eta=1/\sigma_0^2+2$. With these considerations, we can determine the update for the scaling parameter using the Inv-χ^2prior and Jeffreys' prior as the following:

$$\sigma^2=\frac{\sum_{i=1}^N\bar{u}_i x_i^2+1}{N+\eta+2}\tag{10}$$

and

$$\sigma^2=\frac{1}{N+2}\sum_{i=1}^N\bar{u}_i x_i^2\tag{11}$$

It is also easy to show that for a uniform prior $p(\sigma^2)\propto\text{constant}$, the update is simply

$$\sigma^2=\frac{1}{N}\sum_{i=1}^N\bar{u}_i x_i^2\tag{12}$$

In summary, we see that for the three heavy-tail distributions the M-steps are the same. The E-step depends on the distribution. Due to space limitation, we have assumed that the two parameters σ_e^2 and ν are fixed. We note that they can be treated as random variables and Incorporated into the EM algorithm.

2.4. The generalized Wiener estimation

We recall that when the signal is modeled i.i.d. Gaussian with zero mean and variance σ^2, the MAP estimate of x is a Wiener estimate given by

$$x_i = \frac{\sigma^2}{\sigma^2 + \sigma_e^2} y_i \tag{13}$$

To link the above EM algorithm to the Wiener estimation, we compare the two equations (8) and (13). When $u_i = 1$, indicating a Gaussian model for the signal, equation (8) is the same as equation (13). Clearly, equation (13) is a special case of equation (8). We regard the EM algorithm as a generalized Wiener estimate, because (1) the variable u_i in equation (8) is a scaling factor that accounts for the heavy-tail characteristic of the distribution, and (2) it is an iterative algorithm.

To gain further insight into the EM algorithm, we study the student-t distribution with the degrees of freedom $\nu > 2$. The relationship between the variance of the signal σ_s and the scaling factor σ is given by

$$\sigma_s^2 = \frac{\nu}{\nu - 2}\sigma^2 \tag{14}$$

Using this relationship, we can rewrite equation (8) as

$$x_i = \frac{(\nu-2)\hat{\sigma}_s^2 + \hat{x}_i^2}{(\nu-2)\hat{\sigma}_s^2 + \hat{x}_i^2 + (\nu+1)\sigma_e^2} y_i \tag{15}$$

We can further rewrite equation (15) as

$$x_i = \frac{\sigma_L^2}{\sigma_L^2 + \sigma_e^2} y_i \tag{16}$$

where

$$\sigma_L^2 = \hat{\sigma}_s^2 + \frac{1}{\nu+1}(\hat{x}_i^2 - 3\hat{\sigma}_s^2) \tag{17}$$

It is now clear that the result given by equation (15) can be regarded as a generalized Wiener estimate of the signal, where a localized signal variance, denoted $\hat{\sigma}_L^2$, is estimated by taking a weighted average of the signal variance and the local signal energy. It can be easily seen that when $\nu \to \infty$, the Student-t distribution approaches the Gaussian distribution and equation (16) is the same as equation (13) because $\sigma_L^2 = \hat{\sigma}_s^2$.

3. APPLICATIONS IN IMAGE DENOISING

3.1. Two image denoising algorithms

Direct application of the above EM algorithm for image denoising does not necessarily lead to satisfactory results. This is because in developing the algorithm, we have ignored that image signals are generally non-stationary. Recognizing the EM algorithm as a generalized Wiener estimate, that uses localized information, leads naturally to the following two algorithms for image denoising.

3.1.1. A non-iterative generalized Wiener estimation algorithm

We consider a non-iterative algorithm given by

$$x_i = \frac{\sigma_i^2}{\sigma_i^2 + \alpha\sigma_e^2} y_i \tag{18}$$

where σ_i^2 is a localized estimate of the signal energy at the ith location and α is constant to be determined for a particular class of signals. When $\alpha = 1$, this algorithm is a Wiener estimate using local statistics. The heavy-tail distribution of the signal is accounted for by setting $\alpha \neq 1$. We can see that this algorithm uses the M-step of the EM algorithm and replace the E-step by a predetermined constant.

A robust estimation of the variance of the noise is given by

$$\sigma_e = \text{median}(|\mathbf{y}|)/0.6745 \tag{19}$$

A simple method to estimate the signal variance is the following

$$\sigma_i^2 = \begin{cases} S_i - \sigma_e^2, & S_i > \sigma_e^2 \\ 0, & otherwise \end{cases} \tag{20}$$

where $S_i = \frac{1}{2M+1}\sum_{k=-M}^{M} y_{i-k}^2$. The underlining principle for this estimation is that the signal is uncorrelated with noise. With the above results, we can see that the proposed algorithm (equation (18)) is actually a combination of shrinkage and hard-thresholding.

3.1.2. The EM algorithm using local statistics

This algorithm is motivated by using the local statistics discussed in section 2.4. Specifically, in the M-step, we replace the global scaling parameter σ^2 with a localized scaling parameter q_i^2 which is given by

$$q_i^2 = \frac{1}{2M+1}\sum_{k=-M}^{M} \bar{u}_{i-k}\hat{x}_{i-k}^2. \tag{21}$$

In the E-step, we can see from Table 2 that $\bar{u}_i$ is a function of $\hat{x_i}^2/\hat{\sigma}^2$ for the student-t and slash distribution. We replace it with z_i^2/q_i^2, where

$$z_i^2 = \frac{1}{2M+1}\sum_{k=-M}^{M} \hat{x}_{i-k}^2. \tag{22}$$

Similarly, in the Laplacian case, we replace $|\hat{x_i}|/\hat{\sigma}$ with t_i/q_i, where

$$t_i = \frac{1}{2M+1}\sum_{k=-M}^{M} |\hat{x}_{i-k}|. \tag{23}$$

The estimate of the signal is then given by

$$x_i = \frac{q_i^2}{q_i^2 + \bar{u}_i\sigma_e^2} y_i \tag{24}$$

3.2. Experimental results

The non-iterative and the iterative algorithm will be referred to as GWE and IGWE, respectively. For the GWE algorithm, extensive experiments using different images have shown that setting $\alpha = \sqrt{2}$ has led to the best results in terms of the peak-signal-to-noise ratio (PSNR) of the denoised image. For the IGWE algorithm, we use a letter L, T and S to indicate the Laplacian, student-t and slash the distribution is being used, respectively. Experimental results

Barbara

σ_e	GWE1	GWE2	IGWE_T	IGWE_S	bishrink[14]
10	33.10	32.84	33.05	33.03	32.25
15	30.69	30.62	30.78	30.76	29.97
20	28.97	29.07	29.22	29.20	28.36
25	27.65	27.89	28.03	28.01	27.16
30	26.58	26.98	27.11	27.09	26.28
35	25.71	26.19	26.32	26.31	-
40	24.92	25.56	25.67	25.67	-

Lena

σ_e	GWE1	GWE2	IGWE_T	IGWE_S	bishrink[14]
10	34.56	34.63	34.81	34.79	34.36
15	32.37	32.81	32.92	32.89	32.51
20	30.72	31.48	31.54	31.51	31.19
25	29.50	30.45	30.48	30.44	30.15
30	28.37	29.61	29.60	29.56	29.41
35	27.43	28.88	28.84	28.80	-
40	26.62	28.28	28.20	28.16	-

Table 3. PSNR (dB) results using two noisy images with different levels of additive noise. GWE1 and GWE2 are the GWE algorithms with $\alpha = 1$ and $\alpha = \sqrt{2}$, respectively.

show that the IGWE_L algorithm does not work well. From a Bayesian point of view, this simply indicates that the Laplacian distribution does not fit the wavelet coefficients of the image as well as the other two distributions. Experimental results also show that best results are obtained for 3 to 4 iterations for the IGWE_T ($\nu = 3$) and IGWE_S ($\nu = 15$) algorithms. In all experiments, we decomposed of an image into 6 levels using the `sym12` wavelet. Each subband of the signal is then denoised independently.

We can see from the experimental results that for the Barbara image, the iterative algorithms performs better than the non-iterative algorithms. For the Lena image, their performance is about the same. We can also see that using the GWE algorithm, the PSNR associated with the setting $\alpha = \sqrt{2}$ is generally higher than that with the setting $\alpha = 1$. The difference in PSNR is significant in images with high noise levels.

Next we compare the performance of the proposed algorithms with that of the bi-shrinkage [14] which is arguably one of the best in recent publications. We can see that the performance of proposed algorithms (GWE2, IGWE_T and IGWE_S) is consistently better than that of the bi-shrinkage algorithm.

4. CONCLUSIONS

In this paper, we have studied EM algorithms for the MAP estimate of a class of signals which can be modeled by heavy-tail distributions. In developing the EM algorithm for a particular family of heavy-tail distributions (Laplacian, student-t and slash), we take the advantage that these distributions can be expressed as a Gaussian scaled mixture. We show that the EM algorithm can be regarded as a generalization of the classical Wiener estimate. As such, based on the EM algorithm, we propose two algorithms for image denoising. Experimental results show that the performance of the proposed algorithm is better than that of the bi-shrinkage algorithm which is arguably one of the best in recent publications. We expect that when coupled with an over-complete wavelet representation, the proposed algorithms should produce even better results.

5. REFERENCES

[1] S. M. Kay, *Fundamentals of Statistical Signal Processing Estimation Theory*. Englewood Cliffs, New Jersey: Prentice Hall, 1993.

[2] D. Donoho and I. Johnstone, “Adapting to unknown smoothness via wavelet shrinkage,” *Biometrika*, vol. 81, pp. 425–455, 1995.

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

[4] J. Dias, “Fast GEM wavelet-based deconvolution algorithm,” in *Proc. IEEE International Conference on Image Processing*, vol. 2, Sept. 2003, pp. 961–964.

[5] 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.

[6] A. Gelman, H. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*, 2nd ed. Chapman & Hall/CRC, 2004.

[7] K. L. Lange and J. S. Sinsheimer, “Normal/independent distributions and their applications in robust regression,” *Journal of Computational and Graphical Statistics*, vol. 2, pp. 175–198, 1993.

[8] D. F. Andrews and C. L. Mallows, “Scale mixtures of normal distributions,” *Journal of the Royal Statistical Society. Series B*, vol. 36, pp. 99–102, 1974.

[9] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, “Image denoising using scale mixtures of Gaussian in the wavelet domain,” *IEEE Trans. Image Processing*, vol. 12, pp. 1338–1351, Nov. 2003.

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

[11] D. J. C. MacKay, “Comparison of approximate methods for handling hyperparameters,” *Neural Computation*, vol. 11, pp. 1035–1068, 1999.

[12] G. Deng, “Iterative learning algorithms for linear Gaussian observation models,” *IEEE Trans. Signal Processing*, vol. 52, pp. 2286–2297, 2004.

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

[14] L. Sendur and I. W. Selesnick, “Bivariate shrinkage with local variance estimation,” *IEEE Signal Processing Letters*, vol. 9, no. 12, pp. 438–441, Dec 2002.