A MINORIZATION-MAXIMIZATION ALGORITHM FOR MAXIMUM A POSTERIORI SIGNAL ESTIMATION

Guang Deng

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

Wai-Yin Ng

Department of Information Engineering
The Chinese University of Hong Kong, Hong Kong
w.ng@acm.org

ABSTRACT

We develop an iterative algorithm based on minorization - maximization optimization to determine the maximum a posteriori estimate of the signal. We focus on linear Gaussian signal model with a family of heavy-tailed prior distributions which can be represented as scale mixture of Gaussian. We then modify the proposed algorithm for wavelet domain image denoising. Experimental results show that using complex wavelet representations, the performance of the proposed algorithm is very competitive with that of the state-of-the-art algorithms.

1. INTRODUCTION

Estimating signal from noisy observations is a fundamental task in signal processing. A linear observation model is given by

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

where $\mathbf{x}$ is a column vector of the true signal, $\mathbf{y}$ and $\mathbf{e}$ are vectors of observations and noise, respectively and $\mathbf{A}$ is a known matrix. When we assume the model for noise is i.i.d. Gaussian, the maximum a posteriori (MAP) estimation problem is essentially a penalized least squares problem. For example, the problem is known as a ridge regression or weight decay [1] problem when the prior for $\mathbf{x}$ is also i.i.d. Gaussian. When the prior for $\mathbf{x}$ is non-Gaussian, it is usually chosen to model the sparseness of the signal [2], to control the complexity of the model [1], or to perform model selection [3, 4]. A typical application that exploits the sparseness of the signal is in wavelet based image denoising [2, 5]. In this case, $\mathbf{A} = \mathbf{I}$ is the identity matrix and $\mathbf{x}$ is the wavelet coefficient vector of the signal.

Among the non-Gaussian priors, a particular family of heavy-tailed distributions, that has been widely used in statistics research and has been recently used in signal processing, is the scale mixture of Gaussian distribution [6]. These distribution functions, including generalized Gaussian, student-t and slash distributions [7], have found many successful applications in signal and image processing [2,8,9]. The EM algorithm has been a major computational tool [10].

Recently, the minorization-maximization (MM) algorithm [11] has been studied for many statistical problems including variable selection [12] and quantile regression [11]. The basic idea is that, instead of directly maximizing the cost function, another cost function that minorizes[1] the original cost function is iteratively maximized.

[1]The formal definition is presented in Section 2.1.

In this paper, we develop MM algorithms for MAP signal estimation problem stated in (1). More specifically, we assume that noise is modelled i.i.d. Gaussian with zero mean and known variance and that the prior is the scale mixture of Gaussian given by

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

where $\mathcal{N}(x|\sigma^2, u) = \frac{\sqrt{u}}{\sqrt{2\pi}\sigma} e^{-\frac{u}{2\sigma^2}x^2}$ and $p(u|\nu)$ is the prior distribution of u ($0 \le u < \infty$). Different settings for $p(u|\nu)$ result in a family of heavy-tailed distributions [6, 9, 13]. Table 1 shows the definitions for three distribution functions which can be represented as scale mixture of Gaussian (SMG). Here, we have included two parameters σ^2 and ν to account for certain distributions such as the student-t distribution which has a scaling parameter and a parameter for the degree-of-freedom. We further assume that ν is known and σ is to be estimated. With these model assumptions, the problem is essentially a penalized least squares problem which does not have a close form solution. A key step in developing the MM algorithm is to determine the minorizing function for the objective function to be optimized. We show that the logarithm of the SMG distribution function is convex. As such, minorizing function can be determined based on the convexity.

As an application, we modify the proposed algorithms for image denoising in the wavelet domain. We study two heavy-tailed priors for wavelet coefficients: student-t and slash distributions which are of interest as they have not yet been widely studied in image denoising. Experimental results show that using overcomplete wavelet representations, the performance of the proposed algorithm is competitive to that of the state-of-the-art image denoising algorithms.

2. MINORIZATION-MAXIMIZATION ALGORITHM

2.1. A brief review of the MM algorithm

We first give a brief review of the MM algorithm. Let $f(t)$ be a real-valued function and $g(t; t^{(k)})$ be another real-valued function with a known parameter $t^{(k)}$. The function $g(t; t^{(k)})$ is said to minorize $f(t)$ at the point $t^{(k)}$ provided

$$\begin{aligned} g(t; t^{(k)}) &\le f(t) \text{ for all } t \\ g(t^{(k)}; t^{(k)}) &= f(t^{(k)}) \end{aligned} \tag{3}$$

Generalized Gaussian	$\frac{1}{2\Gamma(1+1/\nu)\sigma}\exp\left[-\left(x^2/\sigma^2\right)^{\nu/2}\right],\ 0<\nu\leq 2$
Student-t	$\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}\sigma}\left(1+x^2/\nu\sigma^2\right)^{-(\nu+1)/2}$
Slash	$\frac{\nu}{\sqrt{2\pi}\sigma}\left(x^2/2\sigma^2\right)^{-(\nu+1/2)}\Gamma(\nu+1/2, x^2/2\sigma^2)$

Table 1. Three heavy-tail distributions, where $\Gamma(a)=\int_0^\infty t^{a-1}e^{-t}dt$ and $\Gamma(a,b)=\int_0^b t^{a-1}e^{-t}dt$ are the gamma function and incomplete gamma function, respectively. We assume ν is a fixed parameter.

Let $t^{(k)}$ represent the last iteration in a search for the maximum of $f(t)$ and denote $t^{(k+1)}$ as the maximizer of $g(t;t^{(k)})$, such that

$$t^{(k+1)}=\max_t g(t;t^{(k)}). \tag{4}$$

From the definition, we have

$$g(t^{(k+1)};t^{(k)})\geq g(t^{(k)};t^{(k)})=f(t^{(k)})$$

and

$$f(t^{(k+1)})\geq g(t^{(k+1)};t^{(k)}).$$

Therefore, maximizing $g(t;t^{(k)})$ results in a non-decreasing sequence of $f(t)$:

$$f(t^{(k+1)})\geq f(t^{(k)}) \tag{5}$$

Instead of directly maximizing $f(t)$, we can iteratively maximize the minorizing function $g(t;t^{(k)})$. The algorithm is thus called a minorization-maximization (MM) algorithm.

Now let's consider a function $F(x)=f[q(x)]$. We assume that $f(t)$ is convex and differentiable [14], such that

$$f(t)\geq f(t^{(k)})+f^{'}(t^{(k)})(t-t^{(k)}) \tag{6}$$

Substituting $t=q(x)$ into (6), we have the minorizing function for $F(x)$

$$F(x)\geq D(x,x^{(k)}) \tag{7}$$

where

$$D(x;x^{(k)})=f^{'}[q(x^{(k)})]q(x)+\text{constant} \tag{8}$$

Therefore, the MM algorithm that iteratively maximizes $D(x;x^{(k)})$

$$x^{(k+1)}=\max_x D(x;x^{(k)}) \tag{9}$$

leads to a non-decreasing sequence $F(x^{(k+1)})\geq F(x^{(k)})$.

2.2. The minorizing function for the log-SMG function

We study the logarithm of the scale-mixture of Gaussian (log-SMG) distribution function

$$\begin{aligned}&\log\textstyle\int_0^\infty \mathcal{N}(x|\sigma^2,u)p(u|\nu)du\\&=-\tfrac{1}{2}\log(2\pi)-\tfrac{1}{2}\log s+\log\textstyle\int_0^\infty\sqrt{u}p(u|\nu)\exp[-ux^2/2s]du\end{aligned} \tag{10}$$

where $s=\sigma^2$. Let us define a function $F(x)$ as

$$F(x)=\log\int_0^\infty\sqrt{u}p(u|\nu)\exp[-ux^2/(2s)]du \tag{11}$$

Changing the variable $t=q(x)=x^2/2s$, we have a new function $f(t)$

$$f(t)=\log\int_0^\infty\sqrt{u}p(u|\nu)\exp[-ut]du. \tag{12}$$

The function $f(t)$ and its first derivative for the three distributions are listed in Table 2.

Since $\sqrt{u}p(u|\nu)$ is non-negative for any u ($0\leq u<\infty$) and $f(t)$ is the mixture of the convex function $\exp[-ut]$, therefore $f(t)$ is convex [15]. We can also verify the convexity by recognizing that $\int_0^\infty\sqrt{u}p(u|\nu)\exp[-ut]du$ is the Laplace transform of $\sqrt{u}p(u|\nu)$. The result is log-convex [14]. Therefore, $F(x)$ is minorized by

$$F(x)\geq f^{'}[q(x^{(k)})]\frac{x^2}{2s}+\text{constant} \tag{13}$$

It is easy to prove that

$$D(\mathbf{x})+\sum_{n=1}^N F(x_n)\geq D(\mathbf{x})+f^{'}[q(x_n^{(k)})]\frac{x_n^2}{2s} \tag{14}$$

where $D(\mathbf{x})$ is a real function of the vector $\mathbf{x}=[x_1,x_2,...,x_N]$.

2.3. An MM algorithm for penalized least squares problem

For the penalized least squares problem with known noise variance σ_e^2 and unknown scaling parameter $s=\sigma^2$, the log-posterior, ignoring constants, is given by

$$\log p(\mathbf{x},s|\mathbf{y})=\log p(\mathbf{y}|\mathbf{x})+\log p(\mathbf{x}|s,\nu)+\log p(s) \tag{15}$$

where

$$\log p(\mathbf{y}|\mathbf{x})=-\frac{N}{2}\log(2\pi\sigma_e^2)-\frac{1}{2\sigma_e^2}\mathbf{e}^T\mathbf{e} \tag{16}$$

and

$$\log p(\mathbf{x}|s,\nu)=-\frac{N}{2}\log 2\pi-\frac{N}{2}\log s+\sum_{n=1}^N f(t_n) \tag{17}$$

The function $f(t)$ is given by (12).

We now define a function

$$\begin{aligned}&H(\mathbf{x},s;\mathbf{x}^{(k)},s^{(k)})\\&=\log p(\mathbf{y}|\mathbf{x})+\log p(s)-\frac{N}{2}\log s+\textstyle\sum_{n=1}^N f^{'}(t_n^{(k)})\frac{x_n^2}{2s}\\&=-\frac{1}{2\sigma_e^2}\mathbf{e}^T\mathbf{e}+\log p(s)-\frac{N}{2}\log s-\frac{1}{2s}\mathbf{x}^T\mathbf{W}^{(k)}\mathbf{x}\end{aligned} \tag{18}$$

	Generalized Gaussian	Student-t	Slash
$f(t)$	$-(2t)^{\nu/2}$	$-\frac{\nu+1}{2}\log(\nu+2t)$	$-(\nu+\frac{1}{2})\log t + \log\Gamma(\nu+\frac{1}{2},t)$
$f^{'}(t)$	$-\nu(2t)^{\frac{\nu-2}{2}}$	$-\frac{\nu+1}{\nu+2t}$	$-\frac{\Gamma(\nu+\frac{3}{2},t)}{t\Gamma(\nu+\frac{1}{2},t)}$

Table 2. The function $f(t)$ and its first derivative for the three heavy-tailed distributions. For generalized Gaussian, the parameter ν must be set $0 < \nu \leq 2$ such that $f(t)$ is convex.

where $\mathbf{W}^{(k)} = \text{diag}[-f^{'}(t_n^{(k)})]$ is a diagonal matrix.

Using (14), we can easily see that $H(\mathbf{x}, s; \mathbf{x}^{(k)}, s^{(k)})$ is the minorizing function for the log-posterior

$$\log p(\mathbf{x}, s|\mathbf{y}) \geq H(\mathbf{x}, s; \mathbf{x}^{(k)}, s^{(k)}).$$

Therefore, instead of directly optimizing the log-posterior, we can iteratively maximize $H(\mathbf{x}, s; \mathbf{x}^{(k)}, s^{(k)})$. Since a joint optimization for $\mathbf{x}$ and s is still difficult, we use the same idea of the generalized EM algorithm [10] which holds one fixed and optimizes for the other. As such, the update for $\mathbf{x}$ is obtained by

$$\mathbf{x}^{(k+1)} = \max_{\mathbf{x}} H(\mathbf{x}, s^{(k)}; \mathbf{x}^{(k)}, s^{(k)}) \quad (19)$$

$$= \left(\mathbf{A}^T\mathbf{A} + \frac{\sigma_e^2}{s^{(k)}}\mathbf{W}^{(k)}\right)^{-1}\mathbf{A}^T\mathbf{y} \quad (20)$$

The update for the scaling parameter s depends on the prior. To simplify presentation, we assume $p(s)$ is a uniform distribution. Its update is given by

$$s^{(k+1)} = \max_{\mathbf{s}} H(\mathbf{x}^{(k+1)}, s; \mathbf{x}^{(k)}, s^{(k)}) \quad (21)$$

$$= \frac{1}{N}[\mathbf{x}^{(k+1)}]^T\mathbf{W}^{(k)}\mathbf{x}^{(k+1)} \quad (22)$$

2.4. Connection with the EM algorithm

The penalized least squares problem can also be solved by using the EM algorithm [16]. It can be shown that the MM algorithm presented in the previous section is equivalent to the EM algorithm in that the two update steps are the M-steps and the E-step is given by

$$u_n^{(k)} = E\left[u_n|x_n^{(k)}, s^{(k)}, y\right] = -f^{'}(t_n^{(k)})$$

where $t_n^{(k)} = (x_n^{(k)})^2/(2s^{(k)})$.

3. APPLICATIONS IN IMAGE DENOISING

3.1. Generalized Wiener estimate

The denoising problem is a special case of the penalized least squares problem where $\mathbf{A} = \mathbf{I}$ is an identity matrix. From (19), we can easily derive the update for the signal and scaling parameter

$$x_n^{(k+1)} = \frac{s^{(k)}}{s^{(k)} + u_n^{(k)}\sigma_e^2}y_n. \quad (23)$$

and

$$s^{(k+1)} = \frac{1}{N}\sum_{n=1}^{N} u_n^{(k)}\left(x_n^{(k+1)}\right)^2 \quad (24)$$

where we have used $u_n^{(k)} = -f^{'}(t_n^{(k)})$ to simplify notation.

We recall that for the observation model given by (1), when the signal is modeled i.i.d. Gaussian with zero mean and known variance σ^2, the MAP estimate of x_n is a Wiener estimate given by

$$x_n = \frac{\sigma^2}{\sigma^2 + \sigma_e^2}y_n \quad (25)$$

Therefore, we regard the proposed algorithm as a generalized Wiener estimate, because the variable u_n in equation (23) is a scaling factor that accounts for the heavy-tailed characteristic of the distribution, and the algorithm is an iterative algorithm.

3.2. Image denoising

Direct application of the proposed 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. We modify the algorithm such that localized information is used

Specifically, we replace the global scaling parameter s with a localized scaling parameter s_n which is given by

$$s_n^{(k+1)} = \frac{1}{2M+1}\sum_{m=-M}^{M} u_{n-m}^{(k)}\left[x_{n-m}^{(k+1)}\right]^2. \quad (26)$$

We notice from Table 2 that $u_n^{(k)}$ is a function of $\left[x_n^{(k)}\right]^2/s^{(k)}$ for the student-t and slash distribution. We replace it with $z_n^{(k)}/s_n^{(k)}$, where

$$z_n^{(k)} = \frac{1}{2M+1}\sum_{m=-M}^{M}\left[x_{n-m}^{(k)}\right]^2. \quad (27)$$

The estimate of the signal is then given by

$$x_n^{(k+1)} = \frac{s_n^{(k)}}{s_n^{(k)} + u_n^{(k)}\sigma_e^2}y_n \quad (28)$$

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

$$\sigma_e = \text{median}(|\mathbf{y}|)/0.6745 \quad (29)$$

3.3. Experimental results

The proposed denoising algorithm is called the iterative generalized Wiener estimate (IGWE) algorithm. Since the generalized Gaussian and a number of SMG distributions have been studied [8, 9, 17], we focus on the student-t and slash distributions which have

	IGWE-S	IGWE-T	[20]	[8]	[19]
Lena					
$\sigma_e = 10$	35.30	35.33	34.96	35.61	35.34
$\sigma_e = 15$	33.47	33.51	33.05	33.90	33.67
$\sigma_e = 20$	32.13	32.18	31.72	32.66	32.40
$\sigma_e = 25$	31.06	31.13	30.64	31.69	31.40
$\sigma_e = 30$	30.18	30.25	-	-	30.54
Barbara					
$\sigma_e = 10$	33.69	33.70	33.35	34.03	33.35
$\sigma_e = 15$	31.52	31.54	31.10	31.86	31.31
$\sigma_e = 20$	29.97	29.99	29.44	30.32	29.80
$\sigma_e = 25$	28.78	28.80	28.23	29.13	28.61
$\sigma_e = 30$	27.83	27.85	-	-	27.65

Table 3. A comparison of denoising results (dB) using different algorithms. Two test images are used under different noise levels.

not been widely applied to denoising problem. We use IGWE-T and IGWE-S to indicate the student-t and slash the distribution is being used, respectively. Experimental results show that best results are obtained for 3 to 4 iterations for the IGWE-T ($\nu = 3$) and IGWE-S ($\nu = 15$) algorithms[2]. 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 test the proposed algorithm using the complex wavelet representation [18,19]. In our experiments, we use the proposed algorithms (IGWE-T and IGWE-S) to process each individual image subband of the complex wavelet representation. We use exactly the same complex wavelet transform as that used in [19]. Results are shown in Table 3.

We compare results from three image denoising algorithms which use different overcomplete wavelet representations and different statistical models [8, 19, 20]. We can see from Table 3 that the performance of the proposed algorithms are comparable with that of the three state-of-the-art algorithms.

4. CONCLUSIONS

In this paper, we have developed MM algorithms for the MAP signal estimation for a linear Gaussian model. We study a family of heavy-tailed prior distributions which can be expressed as a scale mixture of Gaussian. By exploiting the log-convexity of the scale mixture of Gaussian, we show that an iterative algorithm can be developed. We then apply the proposed algorithm to image denoising. We show that the proposed algorithm can be regarded as a generalization of the classical Wiener estimate algorithm. We test the proposed algorithm using complex wavelet representations. Experimental results show that the performance of the proposed algorithm is competitive with three state-of-the-art algorithms.

5. REFERENCES

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

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

[3] J. Q. Fan and R. Z. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, pp. 939–955, Sept. 2001.

[4] R. Tibshirani, "Regression shrinkage and selection via lasso," *Journal of the Royal Statistical Society, Ser B*, vol. 57, pp. 257–288, Sept. 1995.

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

[6] 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.

[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] 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.

[9] J. Dias, "Bayesian wavelet-based image deconvolution: A GEM algorithm exploiting a class of heavy-tailed priors," IEEE Trans. Image Processing, 2005, to appear.

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

[11] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *American Statistician*, vol. 58, pp. 30–37, 2004.

[12] D. R. Hunter and R. Li, "Variable selection using MM algorithms," Annals of Statistics, 2005, to appear.

[13] F. Girosi, "Models of noise and robust estimates," MIT A.I. Memo No. 1287, 1991.

[14] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge University Press, 2004.

[15] J. Keilson, *Markov Chain Models - Rarity and Exponentiality*. New York: Springer-Verlag, 1979.

[16] G. Deng, "Generalized wiener estimation algorithm based on a family of heavy-tail distributions," in *Proc. IEEE International Conference on Image Processing*, Genoa, Italy, Oct. 2005.

[17] S. G. Chang, B. Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE Trans. Image Processing*, vol. 9, pp. 1532–1546, Sept. 2000.

[18] N. Kingsbury, "Image processing with complex wavelets," *Phil. Trans. R. Soc. Lond. Ser A*, vol. 357, pp. 2543–2560, 1999.

[19] 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.

[20] X. Li and M. T. Orchard, "Spatially adaptive image denoising under over-complete expansion," in *Proc. IEEE International Conference on Image Processing*, vol. 2, Sep 2000, pp. 311–314.

[2]The MAP estimate for ν is generally very computationally complicated [7] and is beyond the scope of this paper.