A Multiscale Bayesian Framework for Linear Inverse Problems and Its Application to Image Restoration *

*Yi Wan** and *Robert D. Nowak*

Department of Electrical and Computer Engineering, Rice University
6100 South Main St., Houston, TX 77005
Email: {wanyi, nowak}@rice.edu, Tel: (713)348-3776, Fax: (713)348-6196

EDICS Numbers: 2-WAVP, 2-REST

Submitted to the IEEE Transactions on Image Processing, January, 2001

Abstract

In this paper we develop a wavelet-based statistical method for solving linear inverse problems. The Bayesian framework developed here is general enough to treat a wide class of linear inverse problems involving (white or colored) Gaussian observation noise. In this approach, a signal prior is developed by modeling the signal/image wavelet coefficients as independent Gaussian mixture random variables. We first specify a uniform (non-informative) distribution on the mixing parameters, which leads to a simple and efficient iterative algorithm for MAP estimation. This algorithm is similar to the EM algorithm in that it alternates between a state estimation step and a maximization step. Moreover, we show that our algorithm converges monotonically to a local maximum of the posterior distribution. We next generalize the result to non-uniform priors and develop an efficient integer programming algorithm that enables a similar alternating optimization procedure. Experimental results show that this new method outperforms recent results, including multiscale Kalman filtering and wavelet-vaguelette type methods based on linear inverse filtering followed by wavelet coefficient denoising.

*This work was supported by the National Science Foundation, grant no. MIP–9701692, the Army Research Office, grant no. DAAD19-99-1-0290, and the Office of Naval Research, grant no. N00014-00-1-0390.

1 Introduction

Linear inverse problems are common in areas such as medical diagnostics, radar and sonar target estimation, seismology, radio astronomy, microscopy, and image restoration [1–3]. Such problems are mathematically described as

$$\mathbf{y} = \mathcal{L}\mathbf{x} + \mathbf{n} \tag{1}$$

where $\mathbf{x}$ is the original signal or image of interest, $\mathcal{L}$ is a linear operator, $\mathbf{n}$ is the additive noise term, and $\mathbf{y}$ is the observed data. The goal of the inverse problem is to recover $\mathbf{x}$ from $\mathbf{y}$; *i.e.,* to find an operator $\mathcal{H}$ (not necessarily linear) such that the estimate

$$\widehat{\mathbf{x}} = \mathcal{H}\mathbf{y} \tag{2}$$

is as close to $\mathbf{x}$ as possible according to some prescribed measure.

The model described above is plausible for a large variety of practical problems. First of all, many problems are linear in nature. For example, the differential operator, the Abel transform used in calculating the radial mass distribution of galaxies, the Radon transform in computed tomography (CT), and convolution operators are all linear operators. In addition, for any linear system with impulse response $\mathbf{h}(\mathbf{t}, \mathbf{t}')$, the output

$$\mathbf{y}(\mathbf{t}) = \int \mathbf{h}(\mathbf{t}, \mathbf{t}')\mathbf{x}(\mathbf{t}')d\mathbf{t}'$$

also falls into the formulation in (1). Even in non-linear situations such as the effect of non-linear sensor responses, it is often times still reasonable to linearize the operator at certain operating point and tackle the linearized (local) inverse problem.

In view of (1) and (2), one might be tempted to choose $\mathcal{H} = \mathcal{L}^{-1}$, but many times $\mathcal{L}^{-1}$ may not exist. Even when it does exist, it is often near singular and the solution may still be too sensitive to noise in the data. A good example in the case of discrete sampling of a continuous linear system is provided in [1]. In fact, such linear inverse problems have long been recognized as ill-posed problems and tremendous effort has been undertaken to produce satisfactory solutions [1, 3, 4]. Linear inverse problems can arise from a mathematical reformulation of the original problem. For example, stationary, correlated noise can be modeled as the output of white Gaussian noise passing through a known linear shift-invariant filter. Applying the corresponding whitening filter to the observed data, one is left with a (white noise) linear inverse problem.

To illustrate, suppose we have

$$\mathbf{y} = \mathcal{L}\mathbf{x} + \tilde{\mathbf{n}} \tag{3}$$

where in this case the noise term $\tilde{\mathbf{n}}$ is without loss of generality zero-mean noise with autocorrelation matrix Σ. If Σ is nonsingular, then it can be factored into the form $\Sigma = AA^T$ for some invertible matrix A. Hence if $\mathbf{n}$ is iid zero-mean white Gaussian noise of variance 1, then $\tilde{\mathbf{n}}$ and $A\mathbf{n}$ will have the same correlation statistics, namely, Σ. Thus we may view (3) as

$$\mathbf{y} = \mathcal{L}\mathbf{x} + A\mathbf{n}. \tag{4}$$

Multiplying A^{-1} on both sides of (4), we get

$$\mathbf{y}_1 = \mathcal{L}_1\mathbf{x} + \mathbf{n} \tag{5}$$

where $\mathbf{y}_1 = A^{-1}\mathbf{y}$ and $\mathcal{L}_1 = A^{-1}\mathcal{L}$. And we have the same linear inverse problem formulation as expressed in (1). For maximum *a posteriori* (MAP) estimation, we see from the following that the two formulations for denoising correlated noise are equivalent.

$$p(\mathbf{x}|\mathbf{y}) = p(\mathbf{x}|AA^{-1}\mathbf{y}) = p(\mathbf{x}|A\mathbf{y}_1). \tag{6}$$

There exist many classical and modern methods for solving linear inverse problems [1, 3, 5–7]. The classical methods usually are deterministic in nature. The singular value decomposition (SVD) paradigm has been used and recently it has been shown to be minimax best over certain classes of homogeneous functions [8, 9]. However, as pointed out in [7, 10], within such paradigm the bases are completely determined by the operator $\mathcal{L}$ and the underlying signal structure is totally ignored, which is very undesirable.

An often more successful treatment of this type of inverse problem is based on the method of regularization, in which an extra term is added in the minimization process of certain loss function

$L(\mathbf{x})$ to suppress irregular solutions,

$$L(\mathbf{x}) = - \parallel \mathbf{y} - \mathcal{L}\mathbf{x} \parallel^2 + \phi(\mathbf{x}; \alpha) \tag{7}$$

where α is a free parameter reflecting the degree of favoritism toward faithful inverse filtering versus noise suppression. There are many different ways of choosing the regularization function ϕ. Some typical choices of the function ϕ are discussed in, for example, [1,3,11]. However, the most common choice of ϕ, a quadratic functional, is not able to adapt to local inhomogeneity in the signals.

Recently several multiscale and wavelet-based approaches have been proposed; for examples, see [2,6,7,11–15]. The direction of this effort is clear. It is well known that wavelet transform of a large class of natural signals and images produces parsimonious representations [16,17]. Moreover, wavelets have spatially localized support, enabling wavelet-based denoising and restoration schemes to adapt to the local structure and behavior of the signal under study. Thus, wavelet-based methods avoid the pitfalls of classical quadratic regularization functions. In Figure 1 we show the histograms of the natural fruit image in the spatial domain and wavelet domain. We see that after the wavelet transform, the data has a super Gaussian distribution (with positive Kurtosis) centered at 0, showing sparsity, while in the original spatial domain there is no such property.

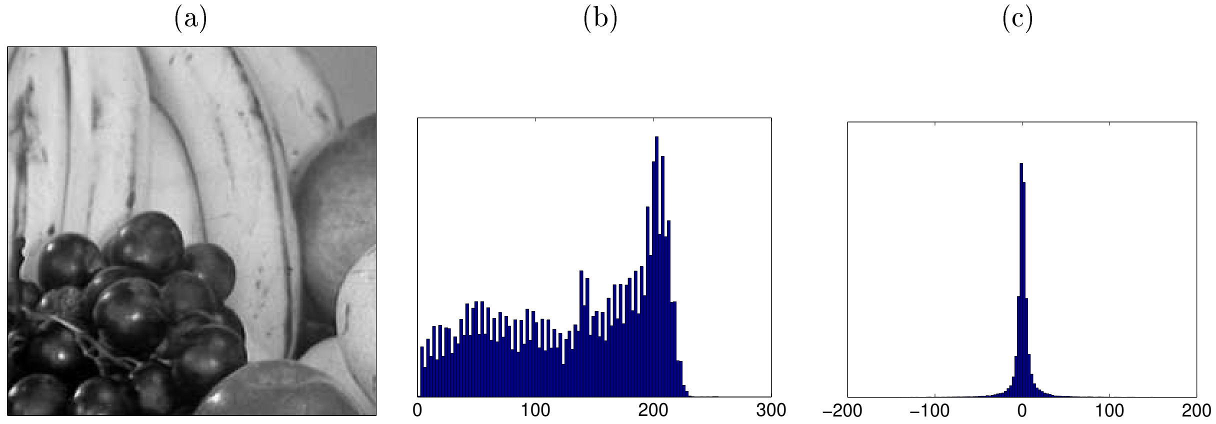

Figure 1: *Sparse representation through wavelet transform. (a) "Fruit" test image. (b) Histogram of pixel intensity values. (c) Histogram of the Haar wavelet coefficients.*

While multiscale/wavelet-based methods can avoid some of the drawbacks of traditional restoration methods, the existing methods do have some limitations. First of all, some of the multiscale methods are analogous to classical Wiener filtering, reinterpreted in the wavelet-domain. These methods may have computational advantages, but do not overcome some of the pitfalls (*e.g.,* over-smoothing) associated with linear filtering methods. However, Banham and Katsaggelos [13] developed a novel approach that circumvents this problem by switching the linear filter spatially based on an edge detection test, enabling the overall restoration procedure to automatically preserve details near detected edges. Other methods combine linear inverse filtering with wavelet coefficient thresholding/shrinkage operations. The iterative method of Starck *et al.* [2] introduces a thresholding step in an iterative restoration algorithm. Wavelet-vaguelette decomposition (WVD) methods (linear inverse filtering followed by wavelet coefficient thresholding), like the methods of Donoho [7] and Johnstone and Silverman [18], can be justified for certain classes of distortions (*e.g.,* scale-homogeneous operators) and noises (*e.g.,* stationary or $1/f$ type processes), but are generally not applicable to more general classes of noises and linear operators. Vaguelette-wavelet decomposition (VWD) [6], a natural alternative to WVD, has similar limitations. One limitation of WVD methods and related thresholding schemes is that they treat wavelet coefficients (or vaguelette coefficients) independently, ignoring possible (noise) correlations between coefficients that are introduced through linear inverse filtering. Notice that applying a linear inverse filter to the observed data $\mathbf{y}$ containing white Gaussian noise results in a correlated noise contamination that is difficult to remove by simple coefficient-wise thresholding. Recently a new method was developed by Neelamani *et al.* [15] that combines classical Fourier domain regularization methods with wavelet domain thresholding/shrinkage. The advantage of such a formulation is the suppression of noise in singular or near singular operator situations. Yet a mathematically sound statistical formulation is more desirable which ideally will render both inversion and noise suppression in one step.

We present an alternative wavelet-based Bayesian approach to linear inverse problems. Our approach was first described in [19], and here we investigate its performance and theoretical properties in greater detail. The method makes use of the wavelet domain statistical model proposed in [17, 19], and we develop a joint MAP estimator through a clean, direct and mathematically sound development. The suitable choice of Gaussian mixture model for the wavelet coefficients and an alternating maximization method leads to a very simple and efficient iterative algorithm that overcomes many of the difficulties encountered in previous methods. This method has some key features. Firstly, a unified Bayesian framework in which operator inversion, noise suppression, plus singularity (jump/edge) detection, are fully integrated in a joint MAP optimization criterion (as compared to other techniques which involve two-step approaches). In addition, the MAP optimization is carried out with a simple iterative algorithm which is guaranteed to converge monotonically to a local maximum of the *a posteriori* density function, irrespective of whether the linear operator $\mathcal{L}$ is singular or not. Moreover, this procedure is highly adaptive to the local inhomogeneity of the signal to be estimated. And finally, wavelet coefficient estimation is carried out jointly (in contrast to coefficient-wise thresholding procedures), insuring that statistical noise correlations introduced by the inverse operator are correctly modeled. Recently, another wavelet domain image restoration method was proposed by Belge *et al* [20], which is close in spirit to our approach in this paper. However, because their method is based on a generalized Gaussian (GG) density, various approximations to the model and parameter determination are necessary to devise a practical algorithm.

In Section 2 we develop the new method, investigate its convergence characteristics, and discuss some of its properties. In section 3, we apply the new algorithm to the special case of convolution operators for 1-D signals and 2-D images and comparisons with some other approaches are given. Finally, conclusions are drawn in Section 4.

2 Wavelet-domain Bayesian Model Formulation

2.1 Basic Formulation

It is widely accepted that the wavelet coefficients of natural images tend to be decorrelated. In addition, as demonstrated in Figure 1, the distribution of these coefficients usually has a peaky and heavy-tailed symmetric shape, centered at the origin [17]. This motivates the use of prior densities based on independent Gaussian mixture models (IGMMs) for the wavelet coefficients. In this paper, we emphasize two-component GMMs. Let

$$\boldsymbol{\theta} = \mathcal{W}\mathbf{x} \tag{8}$$

where $\mathcal{W}$ is the wavelet transform operator, then each wavelet coefficient θ_i is modeled by the distribution[1]

$$p(\theta_i) = (1 - q_i)\mathcal{N}(0, \sigma_0^2) + q_i\mathcal{N}(0, \sigma_1^2), \quad \sigma_0^2 \ll \sigma_1^2 \tag{9}$$

where $\mathcal{N}(\mu, \sigma^2)$ represents a Gaussian distribution with mean μ and variance σ^2. Associated with each wavelet coefficient are three parameters: the mixing probability q_i, and low and high variance parameters σ_0^2 and σ_1^2. Here σ_0^2 and σ_1^2 are fixed as constants over all wavelet coefficients. Then (1), formulated in terms of wavelet coefficients $\boldsymbol{\theta}$, becomes

$$\mathbf{y} = \mathcal{L}\mathcal{W}^{-1}\boldsymbol{\theta} + \mathbf{n} \tag{10}$$

where $\mathcal{W}^{-1}$ is the inverse wavelet transform operator. Let $\widetilde{\mathcal{L}} = \mathcal{L}\mathcal{W}^{-1}$, then the linear inverse problem formulated in the wavelet domain becomes

$$\mathbf{y} = \widetilde{\mathcal{L}}\boldsymbol{\theta} + \mathbf{n}. \tag{11}$$

Let $\mathbf{q} = \{q_i\}$, then the joint posterior distribution $p(\boldsymbol{\theta}, \mathbf{q}|\mathbf{y})$ is given by

$$\begin{aligned} p(\boldsymbol{\theta}, \mathbf{q}|\mathbf{y}) &= \frac{p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{q})p(\boldsymbol{\theta}, \mathbf{q})}{p(\mathbf{y})} \\ &\propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{q})p(\mathbf{q}) \end{aligned} \tag{12}$$

and the (joint) maximum *a posteriori* (MAP) estimator of $\boldsymbol{\theta}$ *and* $\mathbf{q}$ is

$$(\widehat{\theta}, \ \widehat{\mathbf{q}}) = \underset{\boldsymbol{\theta}, \ \mathbf{q}}{\operatorname{argmax}} \ \ p(\boldsymbol{\theta}, \mathbf{q} \mid \mathbf{y}). \tag{13}$$

The following assumptions are made to facilitate the development of the solution to (13):

1. $p(\boldsymbol{\theta}|\mathbf{q}) = \prod_{i=1}^{M} p(\theta_i|q_i)$, the wavelet coefficients are assumed to be statistically independent.

[1] This formulation is slightly different from another equivalent form which involves a hidden Bernoulli random variable [17]. This explicit expression involving the mixing probability will later on lead to an EM algorithm type iterative procedure, which is even more efficient than the EM algorithm used in [17] and has strong conceptual intuition.

2. Noise $\mathbf{n}$ is a zero-mean jointly Gaussian process: $\mathbf{n} \sim \mathcal{N}(0, \Sigma)$

Also the operator $\mathcal{L}$ is treated as an $M \times M$ matrix and all involved signals/images are represented as column vectors of length M.

Now in (12), treating the mixing probability $\mathbf{q}$ as being unknown *a priori*, we assign the non-informative uniform density on $[0, 1]$ to it, i.e.,

$$p(\mathbf{q}) = p(q_i) \equiv 1.$$

Taking logarithm of (12), we get

$$\begin{aligned}
L(\boldsymbol{\theta}, \mathbf{q}) &= \log p(\boldsymbol{\theta}, \mathbf{q}|\mathbf{y}) \\
&= \log p(\mathbf{y}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}|\mathbf{q}) + C_1 \\
&= \log p(\mathbf{y}|\boldsymbol{\theta}) + \sum_{i=1}^{M} \log p(\theta_i|q_i) + C_1 \\
&= -\frac{1}{2}(\mathbf{y} - \widetilde{\mathcal{L}}\boldsymbol{\theta})^T \Sigma^{-1} (\mathbf{y} - \widetilde{\mathcal{L}}\boldsymbol{\theta}) + \sum_{i=1}^{M} \log \left[\frac{1 - q_i}{\sigma_0} e^{-\frac{\theta_i^2}{2\sigma_0^2}} + \frac{q_i}{\sigma_1} e^{-\frac{\theta_i^2}{2\sigma_1^2}} \right] + C_2 \qquad (14)
\end{aligned}$$

where C_1 and C_2 are constants.

Usually the maximization of L is difficult because of the presence of mixture densities. However, the following lemma provides a necessary condition on any local maximal solution and will later on lead to an iterative algorithm that monotonically converges to a local maximum.

Lemma 1 *Let $(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{q}})$ be any local maximum of $L(\boldsymbol{\theta}, \mathbf{q})$, i.e.,*

$$L(\boldsymbol{\theta}, \mathbf{q}) \leq L(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{q}})$$

for all $(\boldsymbol{\theta}, \mathbf{q})$ in a sufficiently small neighborhood of $(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{q}})$. Then

1. *Each element $\widehat{q}_i$ of $\widehat{\mathbf{q}}$ is equal to either 0 or 1, and is given by*

$$\widehat{q}_i = \begin{cases} 0 & \text{if } \widehat{\theta}_i^2 \leq \frac{\log(\sigma_1^2/\sigma_0^2)}{\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}} \\ 1 & \text{if } \widehat{\theta}_i^2 > \frac{\log(\sigma_1^2/\sigma_0^2)}{\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}} \end{cases} . \qquad (15)$$

2. $\widehat{\boldsymbol{\theta}}$ *is uniquely expressed through* $\widehat{\mathbf{q}}$ *as*

$$\widehat{\boldsymbol{\theta}} = (\widetilde{\mathcal{L}}^T \Sigma^{-1} \widetilde{\mathcal{L}} + K^{-1})^{-1} \widetilde{\mathcal{L}}^T \Sigma^{-1} \mathbf{y} \tag{16}$$

where K *is a diagonal matrix with diagonal elements* $K_{ii} = \sigma^2_{\widehat{q}_i}$*.*

Proof. First note that

$$\begin{aligned}
\widehat{\mathbf{q}} &= \operatorname*{argmax}_{\mathbf{q}} L(\widehat{\boldsymbol{\theta}}, \mathbf{q}) \\
&= \operatorname*{argmax}_{\mathbf{q}} \sum_{i=1}^{M} \log p(\widehat{\theta}_i | q_i) \\
&= \operatorname*{argmax}_{\mathbf{q}} \sum_{i=1}^{M} \log \left[\frac{1 - q_i}{\sigma_0} e^{-\frac{\widehat{\theta}_i^2}{2\sigma_0^2}} + \frac{q_i}{\sigma_1} e^{-\frac{\widehat{\theta}_i^2}{2\sigma_1^2}} \right].
\end{aligned}$$

Since the last expression is in the decoupled format, we can maximize each of the summands separately. So

$$\widehat{q}_i = \begin{cases} 0 & \text{if } \frac{1}{\sigma_0} e^{-\frac{\widehat{\theta}_i^2}{2\sigma_0^2}} \geq \frac{1}{\sigma_1} e^{-\frac{\widehat{\theta}_i^2}{2\sigma_1^2}} \\ \\ 1 & \text{if } \frac{1}{\sigma_0} e^{-\frac{\widehat{\theta}_i^2}{2\sigma_0^2}} < \frac{1}{\sigma_1} e^{-\frac{\widehat{\theta}_i^2}{2\sigma_1^2}} \end{cases}$$

or

$$\widehat{q}_i = \begin{cases} 0 & \text{if } \widehat{\theta}_i^2 \leq \frac{\log(\sigma_1^2/\sigma_0^2)}{\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}} \\ 1 & \text{if } \widehat{\theta}_i^2 > \frac{\log(\sigma_1^2/\sigma_0^2)}{\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}} \end{cases}$$

for all $1 \leq i \leq M$.

Now based on this result, each $p(\theta_i | q_i)$ is just a simple Gaussian density with variance $\sigma^2_{\widehat{q}_i}$. Hence $p(\boldsymbol{\theta} | \widehat{\mathbf{q}}) \sim \mathcal{N}(0, K)$ where K is diagonal with diagonal elements $K_{ii} = \sigma^2_{\widehat{q}_i}$. Then from (14),

$$L(\boldsymbol{\theta}, \widehat{\mathbf{q}}) = -\frac{1}{2}(\mathbf{y} - \widetilde{\mathcal{L}}\boldsymbol{\theta})^T \Sigma^{-1} (\mathbf{y} - \widetilde{\mathcal{L}}\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T K^{-1} \boldsymbol{\theta} + C_3 \tag{17}$$

where C_3 is a constant independent of $\boldsymbol{\theta}$. Hence

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \widehat{\mathbf{q}})$$

$$= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} -\frac{1}{2}(\boldsymbol{\theta}-\mathbf{a})^T(\widetilde{\mathcal{L}}^T\Sigma^{-1}\widetilde{\mathcal{L}}+K^{-1})(\boldsymbol{\theta}-\mathbf{a})+C$$

where $\mathbf{a} = (\widetilde{\mathcal{L}}^T\Sigma^{-1}\widetilde{\mathcal{L}}+K^{-1})^{-1}\widetilde{\mathcal{L}}^T\Sigma^{-1}\mathbf{y}$ and C is a constant independent of $\boldsymbol{\theta}$.

Since $\widetilde{\mathcal{L}}^T\Sigma^{-1}\widetilde{\mathcal{L}} \geq 0$ and $K^{-1} > 0$, $\widetilde{\mathcal{L}}^T\Sigma^{-1}\widetilde{\mathcal{L}}+K^{-1}$ is strictly positive definite. Hence, given any such K, L is maximized uniquely when $\boldsymbol{\theta} = \mathbf{a}$, i.e.,

$$\widehat{\boldsymbol{\theta}} = (\widetilde{\mathcal{L}}^T\Sigma^{-1}\widetilde{\mathcal{L}}+K^{-1})^{-1}\widetilde{\mathcal{L}}^T\Sigma^{-1}\mathbf{y}.$$

□

Here we see that even though $\widetilde{\mathcal{L}}$ may be singular, $\widetilde{\mathcal{L}}^T\Sigma^{-1}\widetilde{\mathcal{L}}+K^{-1}$ is always invertible, showing that the solution can always be uniquely computed. This is reminiscent of the Moore-Penrose generalized inverse which is unique for any given matrix. However, in this case, even though the matrix is invertible, its condition number may be very large because some diagonal elements in K are of the value $1/\sigma_1^2 \approx 0$ since usually $\sigma_1^2 \gg 1$,

It is interesting to consider some special cases of (16). First, if the noise is white ($\Sigma = \sigma^2 I$), then $\widehat{\boldsymbol{\theta}}$ can be simplified to

$$\widehat{\boldsymbol{\theta}} = (\widetilde{\mathcal{L}}^T\widetilde{\mathcal{L}}+\sigma^2K^{-1})^{-1}\widetilde{\mathcal{L}}^T\mathbf{y}.$$

If in addition there is no "blurring" effect ($\mathcal{L} = I$), then (16) can be further simplified to

$$\theta_i = \frac{K_{ii}}{\sigma^2+K_{ii}}\left(\mathcal{W}y\right)_i, \tag{18}$$

which is exactly the same result as given by the MAP estimator in the wavelet-domain independent mixture (IM) denoising method [17]. Notice here that despite this similarity with wavelet shrinkage/thresholding methods, our approach is fundamentally different in situations where the linear operator is non-trivial or when the noise is correlated. In these cases, the wavelet coefficients are statistically correlated (due to correlated noise and/or inverse filtering the observed data) and hence wavelet coefficient estimation must be carried out *jointly*, in contrast to *diagonal* wavelet-

Step 0 Initialization: $\boldsymbol{\theta}^{(0)} = 0,\ i = 0$

Step 1 Estimate $\mathbf{q}^{(i+1)}$ using $\boldsymbol{\theta}^{(i)}$ as

$$q_j^{(i+1)} = \begin{cases} 0 & \text{if } \left(\theta_j^{(i)}\right)^2 \le \frac{\log\left(\sigma_1^2/\sigma_0^2\right)}{\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}} \\ 1 & \text{if } \left(\theta_j^{(i)}\right)^2 > \frac{\log\left(\sigma_1^2/\sigma_0^2\right)}{\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}} \end{cases} \qquad j = 1, 2, \cdots, M \tag{19}$$

Step 2 Using estimated $\mathbf{q}^{(i+1)}$ in step 1 to form the diagonal matrix $K^{(i+1)}$ with $K_{jj}^{(i+1)} = \sigma^2_{q_j^{(i+1)}}$, then compute $\boldsymbol{\theta}^{(i+1)}$ as

$$\boldsymbol{\theta}^{(i+1)} = \left[\widetilde{\mathcal{L}}^T \Sigma^{-1} \widetilde{\mathcal{L}} + \left(K^{(i+1)}\right)^{-1}\right]^{-1} \widetilde{\mathcal{L}}^T \Sigma^{-1} \mathbf{y} \tag{20}$$

Step 3 If $\mathbf{q}^{(i)} \neq \mathbf{q}^{(i+1)}$, then $i = i + 1$ and repeat steps 1 and 2; otherwise exit.

Table 1: Bayesian IGMM model restoration algorithm.

based estimators based on coefficient-wise shrinkage and thresholding schemes. The recent work of Johnstone and Silverman [18] deals with correlated noise removal, but focuses on some special classes of noises which allow individual wavelet coefficient shrinkage. The method presented here is general enough to treat any correlated noise as in (16). Comparisons will be presented later in the paper to illustrate this point.

Lemma 1 shows the coupling dependency between $\boldsymbol{\theta}$ and $\mathbf{q}$ at the optimal solution as expressed in (15) and (16). This also suggests an iterative method to search for a local maximum point. Namely, given a set of estimated values of $\boldsymbol{\theta}$ and $\mathbf{q}$, we can iteratively update $\mathbf{q}$ using (15) and $\boldsymbol{\theta}$ using (16). This kind of alternate optimization procedure has been used in similar problems, e.g., [21]. In the following we formulate such an iterative algorithm and then study its convergence behavior. Table 1 outlines our iterative algorithm.

In the procedure, interchanging Steps 1 and 2 will not affect the result. That is, we can also start from $\mathbf{q}^{(0)} = 0$ and then iteratively compute $\boldsymbol{\theta}^{(i+1)}$ followed by computing $\mathbf{q}^{(i+1)}$. This is because $\boldsymbol{\theta} = 0$ implies $\mathbf{q} = 0$ according to (15). In this algorithm, steps 1 and 2 always increase the value of L until a local maximum is reached, in which case the *state* of each wavelet coefficient remains the same. In general, for this type of maximization there are many local extreme points and the (local) maximum which the algorithm converges to depends on the choice of the initial estimate. Empirically the choice of some uniform constant value for $\boldsymbol{\theta}^{(0)}$ works well, hence in step 0 we set all wavelet coefficients to 0. This also agrees with the super Gaussian distribution of wavelet coefficients in general, which shows that most of the wavelet coefficients are near 0. Corresponding to this choice of initialization for $\boldsymbol{\theta}$, the states $\mathbf{q}$ are set to 0.

The following theorem summarizes the convergence behavior of the algorithm.

Theorem 1 *Let L be as in (14). Given any initial estimate $\boldsymbol{\theta}^{(0)}$ and $\mathbf{q}^{(0)}$, the iterative algorithm described in Table 1 will stop after S iterations with $S \leq 2^M$.*

Let $L^{(i)} = L(\boldsymbol{\theta}^{(i)}, \mathbf{q}^{(i)})$, then

$$L^{(0)} \leq L^{(1)} \leq \cdots \leq L^{(S)} = L^{(S+1)} = \cdots.$$

Choose $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(S)}$ and $\widehat{\mathbf{q}} = \mathbf{q}^{(S)}$, then $(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{q}})$ satisfies the coupling conditions (15) and (16) in Lemma 1. If in addition for all $\widehat{\theta}_i$, $i = 1, 2, \cdots, M$,

$$\theta_i^2 \neq \frac{\log\left(\sigma_1^2/\sigma_0^2\right)}{\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}},$$

then $(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{q}})$ is a local maximum point for L.

Proof. From the proof of Lemma 1, we see that for any fixed $\boldsymbol{\theta}$, L is uniquely maximized in $\mathbf{q}$. Also for any $\mathbf{q} \in \{0,1\}^M$, L is uniquely maximized in $\boldsymbol{\theta}$. Then for any $i \geq 0$,

$$L(\boldsymbol{\theta}^{(i)}, \mathbf{q}^{(i)}) \leq L(\boldsymbol{\theta}^{(i)}, \mathbf{q}^{(i+1)}) \leq L(\boldsymbol{\theta}^{(i+1)}, \mathbf{q}^{(i+1)}),$$

showing the monotonicity of the algorithm, i.e.,

$$L^{(0)} \leq L^{(1)} \leq L^{(2)} \leq \cdots.$$

Since in the algorithm $\mathbf{q}$ only takes values from the set $\{0,1\}^M$ of 2^M elements, there exist $i < j \leq 2^M$ such that $\mathbf{q}^{(i)} = \mathbf{q}^{(j)}$. Then we deduce that $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(j)}$ from the uniqueness of (16).

Hence $L^{(i)} = L^{(j)}$. It follows that

$$\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i+1)} = \cdots = \boldsymbol{\theta}^{(j)}$$

and

$$\mathbf{q}^{(i)} = \mathbf{q}^{(i+1)} = \cdots = \mathbf{q}^{(j)}.$$

Now let S be the number of total iterations of the algorithm. If $S = 2^M$, then $(\boldsymbol{\theta}^{(S)}, \mathbf{q}^{(S)})$ is clearly the maximal solution. In fact, it is not difficult to see that it's the global maximal solution over $\mathbf{R}^M \times \{0,1\}^M$. If $S < 2^M$, then we have $L^{(1)} < L^{(2)} < \cdots < L^{(S)} = L^{(S+1)}$, $\boldsymbol{\theta}^{(S)} = \boldsymbol{\theta}^{(S+1)}$ and $\mathbf{q}^{(S)} = \mathbf{q}^{(S+1)}$.

Let

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(S)}, \quad \widehat{\mathbf{q}} = \mathbf{q}^{(S)}.$$

We then have that $\widehat{\boldsymbol{\theta}}$ and $\widehat{\mathbf{q}}$ satisfy the local maximum point coupling condition as in (15) and (16).

Suppose no $\widehat{\theta}_i$ satisfies

$$\widehat{\theta}_i^2 = \frac{\log\left(\sigma_1^2/\sigma_0^2\right)}{\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}}$$

let $\delta = \frac{\log\left(\sigma_1^2/\sigma_0^2\right)}{\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}}$ and choose

$$\epsilon = \min_{1 \le i \le N} \left| \left| \widehat{\theta}_i \right| - \delta^{\frac{1}{2}} \right| > 0.$$

Then for all $\boldsymbol{\theta} \in \mathbf{R}^M$ such that $\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\| < \epsilon$ and all $\mathbf{q} \in [0,1]^M$ such that $\|\mathbf{q} - \widehat{\mathbf{q}}\| < \epsilon$, where $\|\cdot\|$ represents the Euclidean norm,

$$\widehat{\mathbf{q}} = \max_{\mathbf{q}} L(\widehat{\boldsymbol{\theta}}, \mathbf{q}) = \max_{\mathbf{q}} L(\boldsymbol{\theta}, \mathbf{q}),$$

since

$$|\theta_i - \widehat{\theta}_i| \le \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\| < \epsilon$$

implies that both θ_i^2 and $\widehat{\theta}_i^2$ will be on the same side of δ and hence (15) will give the same $\mathbf{q}$ value for all such $\boldsymbol{\theta}$.

It follows that

$$
\begin{aligned}
L(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{q}}) - L(\boldsymbol{\theta}, \mathbf{q}) &= [L(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{q}}) - L(\boldsymbol{\theta}, \widehat{\mathbf{q}})] + [L(\boldsymbol{\theta}, \widehat{\mathbf{q}}) - L(\boldsymbol{\theta}, \mathbf{q})] \\
&> 0.
\end{aligned}
$$

Thus $(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{q}})$ is a local maximum point and the proof is complete.

$\square$

Several points need to be addressed regarding the theorem. First, even though the total number of possible iterations to reach a local maximum can be as high as 2^M, in practice it usually converges to a fixed point in less than 10 iterations. Next, this iterative algorithm is highly nonlinear and signal dependent. Wavelet coefficients of large magnitude will tend to evolve into the high variance state, while wavelet coefficients of small magnitude will tend to be suppressed. These operations are carried out in a global way so that the difficulties in treating the problem as two separate stages (inverse filtering and denoising) as discussed in Section 1 do not arise here. Finally, the theorem shows that the algorithm converges to a local maximum point under very mild condition on $\widehat{\boldsymbol{\theta}}$. Even this condition can be circumvented through a slight perturbation of the threshold as described in the following corollary.

Corollary 1 *Let $(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{q}})$ be the solution produced from the algorithm. There exists a constant $c > 0$ such that for all $0 < \epsilon < c$, if*

$$
\delta = \frac{\log\left(\sigma_1^2/\sigma_0^2\right)}{\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}},
$$

no $\theta_j^{(i)}$ satisfies $\left(\theta_j^{(i)}\right)^2 = \delta + \epsilon$, and if (19) is replaced by

$$
q_j^{(i+1)} = \begin{cases} 0 & \text{if } \left(\theta_j^{(i)}\right)^2 < \delta + \epsilon \\ 1 & \text{if } \left(\theta_j^{(i)}\right)^2 > \delta + \epsilon \end{cases} \qquad j = 1, 2, \cdots, M \tag{21}
$$

Then the algorithm will converge to $(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{q}})$, which is a local maximum point.

Proof. Let

$$c = \min\{\left(\theta_j^{(i)}\right)^2 - \delta : \left(\theta_j^{(i)}\right)^2 > \delta,\ 1 \le j \le M,\ 1 \le i \le S\},$$

and for any $0 < \epsilon < c$, let $\delta' = \delta + \epsilon$. If we replace (19) by (21), then it's easy to see that $\left(\theta_j^{(i)}\right)^2 \neq \delta'$ for all $i,\ j$. And the maximum solution will be the same as before, namely, $(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{q}})$. By Theorem 1, it is a local maximum point. □

Thus we have shown that there exists a sufficiently small interval on the real line, which depends on the observed data $\mathbf{y}$ and can be as close to the theoretical threshold δ as needed, such that for any value in this interval to be taken as the new threshold, our algorithm converges to the same solution as the one using theoretical threshold δ and it is a local maximum. (Note that the maximal solution in general depends on the threshold level.) This kind of stability is desirable in applications, especially when in general the parameters cannot be accurately determined.

It is important to notice the difference between the algorithm developed in this paper and the popular *expectation-maximization* (EM) algorithm. Even though both are alternating procedures toward a (local) maximum likelihood value, they differ in both the formulation of the problem and its actual steps. In our algorithm the joint posterior distribution in the Bayesian framework is set up as the objective function. In this formulation there is no need to introduce the EM algorithm because the problem can be easily solved numerically and the algorithm comes out naturally. By comparison, in the usual formulation of the EM algorithm for GMM-based estimation problems the *states* are "hidden" and considered part of the so-called *complete data*. Beyond that, the same *states* in our algorithm are either "high" or "low" , while in the EM algorithm those states are usually somewhere in between, corresponding to the fact that the probability of the mixing variables are not single point masses of Dirac type. The computational efficiency is also different. In our algorithm at each "E" step only one comparison with a pre-computed threshold is needed. As one possible extension, this same procedure can be derived in a similar fashion to deal with m-mixture models.

2.2 Case of Non-uniform Prior

In the basic formulation of the previous subsection we set the prior $p(\mathbf{q})$ to be uniform. Now even though q_i for any fixed i can be conveniently viewed as uniformly distributed, we do know that wavelet coefficients for natural images tend to be sparse, i.e., relatively few wavelet coefficients are in the high-variance state. If we let $T = \sum_{i=1}^{M} q_i$, where $q_i \in \{0,\ 1\}$, then T/M will be closer to 0 than to 1.

If we express this prior knowledge into $p(\mathbf{q})$, the generalized form of (14) from (12) becomes

$$L(\boldsymbol{\theta}, \mathbf{q}) = -\frac{1}{2}(\mathbf{y} - \widetilde{\mathcal{L}}\boldsymbol{\theta})^T \Sigma^{-1} (\mathbf{y} - \widetilde{\mathcal{L}}\boldsymbol{\theta}) + \sum_{i=1}^{M} \log \left[\frac{1-q_i}{\sigma_0} e^{-\frac{\theta_i^2}{2\sigma_0^2}} + \frac{q_i}{\sigma_1} e^{-\frac{\theta_i^2}{2\sigma_1^2}} \right] + \log p(\mathbf{q}) + C_2 \quad (22)$$

In order to arrive at a similar alternate maximization procedure as before, we restrict $\mathbf{q}$ to be in the set $\{0,\ 1\}^M$. Next, we develop a similar alternate algorithm as the one in the previous basic formulation to maximize $L(\boldsymbol{\theta}, \mathbf{q})$ in (22). For any fixed $\mathbf{q} \in \{0,\ 1\}^M$, the expression for $\boldsymbol{\theta}$ to maximize L is the same as in (16). On the other hand, when fixing any $\boldsymbol{\theta}$ value, we need to maximize

$$H = \sum_{i=1}^{M} \log \left[\frac{1-q_i}{\sigma_0} e^{-\frac{\theta_i^2}{2\sigma_0^2}} + \frac{q_i}{\sigma_1} e^{-\frac{\theta_i^2}{2\sigma_1^2}} \right] + \log p(\mathbf{q}) \quad (23)$$

We assume that $\log p(\mathbf{q}) = \phi(T)$. The following theorem provides an efficient integer programming procedure to maximize H when ϕ is concave on $[0, M]$.

Theorem 2 *Let*

$$\begin{aligned}
a_i &= \log \left[\frac{1}{\sigma_0} e^{-\frac{\theta_i^2}{2\sigma_0^2}} \right], \\
b_i &= \log \left[\frac{1}{\sigma_1} e^{-\frac{\theta_i^2}{2\sigma_1^2}} \right], \\
d_i &= b_i - a_i, \\
\triangle_i &= \phi(i) - \phi(i-1), \\
h_i &= d_i + \triangle_i
\end{aligned}$$

for $1 \leq i \leq M$.

Without loss of generality, assume $d_1 \geq d_2 \geq \cdots \geq d_M$.

If $\phi(t)$ *is concave on* $[0, M]$, *i.e.,*

$$\phi((1-\alpha)t_1 + \alpha t_2)) \geq (1-\alpha)\phi(t_1) + \alpha\phi(t_2)$$

for any $\alpha \in [0,1]$ *and* $t_1, t_2 \in [0, M]$, *then*

1. $h_1 \geq h_2 \geq \cdots \geq h_M$.

2. *A* $\mathbf{q}$ *that globally maximizes* H *is given by the following formula*

$$q_i = \left\lfloor \frac{sgn(h_i) + 1}{2} \right\rfloor, \quad 1 \leq i \leq M, \tag{24}$$

   where the sign function $sgn(t)$ *is defined as*

$$sgn(t) = \begin{cases} 1 & t > 0 \\ 0 & t = 0 \\ -1 & t < 0 \end{cases}.$$

Proof.

1. By choosing $\alpha = \frac{1}{2}$, the concavity of ϕ shows that

$$\triangle_1 \geq \triangle_2 \geq \cdots \geq \triangle_M.$$

   Hence $h_1 \geq h_2 \geq \cdots \geq h_M$.

2. Since $\mathbf{q} \in \{0,1\}^M$, we have

$$H = \sum_{i=1}^{M} \log \left[\frac{1-q_i}{\sigma_0} e^{-\frac{\theta_i^2}{2\sigma_0^2}} + \frac{q_i}{\sigma_1} e^{-\frac{\theta_i^2}{2\sigma_1^2}} \right] + \log p(\mathbf{q})$$

$$
\begin{aligned}
&= \left[\sum_{i=1}^{M}(1-q_i)\log\left(\frac{1}{\sigma_0}e^{-\frac{\theta_i^2}{2\sigma_0^2}}\right)+q_i\log\left(\frac{1}{\sigma_1}e^{-\frac{\theta_i^2}{2\sigma_1^2}}\right)\right]+\phi(T)\\
&= \left[\sum_{i=1}^{M}(1-q_i)a_i+q_ib_i\right]+\phi(T)\\
&= \left[\sum_{i=1}^{M}q_id_i\right]+\phi(T)+C_4
\end{aligned}
$$

where $C_4=\sum_{i=1}^{M}a_i$ is a constant independent of $\mathbf{q}$.

Notice that for any fixed $T\in\{0,1,\cdots,M\}$, since $\{d_i\}$ is a decreasing sequence by assumption, the $\mathbf{q}$ that maximizes H is obtained by choosing

$$q_1=q_2=\cdots=q_T=1$$

and

$$q_{T+1}=q_{T+2}=\cdots=q_M=0,$$

for an appropriate value of T.

Hence the problem is reduced to determining T that maximizes H. In this case we can write H as

$$
\begin{aligned}
H &= \sum_{i=1}^{M}q_id_i+\sum_{i=1}^{M}q_i\triangle_i+C_5\\
&= \sum_{i=1}^{M}q_ih_i+C_5
\end{aligned}
$$

where $C_5=C_4+\phi(0)$ is a constant.

Then by the monotonicity of the sequence $\{h_i\}$, the $\mathbf{q}$ that maximize H are given by

$$q_i=\left\lfloor\frac{sgn(h_i)+1}{2}\right\rfloor,\quad 1\le i\le M.$$

$\square$

According to this theorem, the generalized form of the algorithm is outlined in Table 2.

Step 0 Initialization: $\boldsymbol{\theta}^{(0)} = 0,\ i = 0$

Step 1 Compute the sequence $\{d_i\}$, then sort it in decreasing order. Compute $\{h_i\}$ and then get $\mathbf{q}$ according to (24).

Step 2 Using estimated $\mathbf{q}^{(i+1)}$ in step 1 to form the diagonal matrix $K^{(i+1)}$ with $K_{jj}^{(i+1)} = \sigma^2_{q_j^{(i+1)}}$, then compute $\boldsymbol{\theta}^{(i+1)}$ as

$$\boldsymbol{\theta}^{(i+1)} = \left[\widetilde{\mathcal{L}}^T \Sigma^{-1} \widetilde{\mathcal{L}} + \left(K^{(i+1)}\right)^{-1}\right]^{-1} \widetilde{\mathcal{L}}^T \Sigma^{-1} \mathbf{y} \tag{25}$$

Step 3 If $\mathbf{q}^{(i)} \neq \mathbf{q}^{(i+1)}$, then $i = i + 1$ and repeat steps 1 and 2; otherwise exit.

Table 2: Bayesian IGMM model with non-uniform prior restoration algorithm.

In this generalized algorithm the major additional time-consuming part is the sorting of the sequence $\{d_i\}$, which is not a significant problem with various fast sorting algorithms available. Also two reasonable examples of $p(\mathbf{q})$ are the exponential form $p(\mathbf{q}) \propto exp\{-kT\}$, a complexity-type penalty [11], and the beta distribution $p(\mathbf{q}) = Beta_{\alpha,\beta}(T/M)$ with parameters α and β.

2.3 Nature of Adaptive Regularization

Even though the solution derived before is from a pure statistical formulation, it has an inherent connection with adaptive regularization. Recall that the threshold in (15) serves to categorize the wavelet coefficients into either the low or high variance state. When in the low variance state, a wavelet coefficient tends to be attenuated or shrunk, while it tends to be preserved in the high variance state. This is evident from the expression in (18). We restate it here in a slightly different form as

$$L = -\frac{1}{2} \parallel \Sigma^{-\frac{1}{2}}(y - \widetilde{\mathcal{L}}\boldsymbol{\theta}) \parallel^2 - \frac{1}{2} \parallel K^{-\frac{1}{2}}\boldsymbol{\theta} \parallel^2 \tag{26}$$

where $\| \cdot \|$ represents the Euclidean norm.

At first sight (26) looks like the method of regularization. But the key difference and advantage here is that the diagonal operator $K^{-\frac{1}{2}}$ adapts to the data, and its value at each position represents a trade-off between the two terms on the right side of (26) – either suppress or maintain the wavelet coefficients to varying degrees in a joint fashion.

(a) (b)

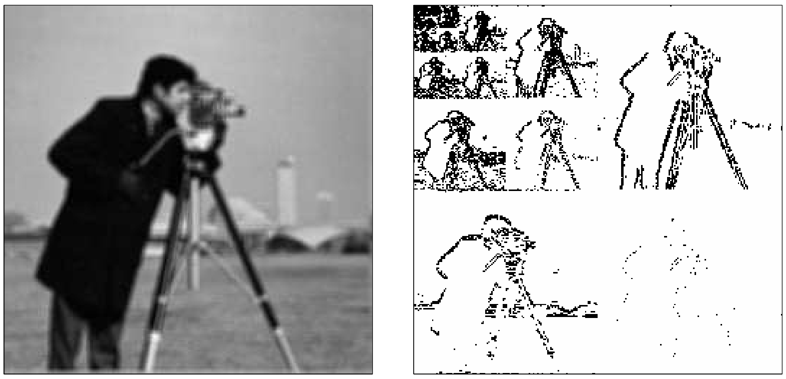

Figure 2: *Adaptive regularization nature. (a) A noisy blurred image. (b) Output* $\mathbf{q}$ *from the adaptive algorithm. Black dots correspond* $q_i = 1$. *White dots for* $q_i = 0$.

Figure 2 demonstrates the data adaptive behavior. We see that starting from all-low-variance states the algorithm converges to produce the $\mathbf{q}$ value that adjusts to the edges in the image.

3 Experiments

In this section we apply the developed algorithm to signal/image deconvolution. All wavelets used are Haar wavelets. We first give a few examples in 1-D to better illustrate some key points in the developed method. Then we experiment on 2-D deconvolution problems.

For the determination of σ_0^2 and σ_1^2, there are two approaches. One is to simply set $\sigma_0^2 = \sigma^2$, the noise power, and choose σ_1^2 to be several magnitudes higher, as is generally the case in signal denoising [17]. The other approach is based on estimating the variances from an initial restoration based on the settings of σ_0^2 and σ_1^2 above. Specifically, after an initial restoration with $\sigma_0^2 = \sigma^2$ and $\sigma_1^2 = 1000\sigma^2$, for example, we use an expectation-maximization (EM) algorithm to estimate improved values for the two variances. The EM algorithm fits a two-component Gaussian mixture to each scale of wavelet coefficients of the initial restoration. We then re-apply our algorithm with these "scale-adapted" variance settings. We will compare both approaches in the following experiments.

3.1 Application to 1-D Deconvolution

The first example is shown in Figure 4. Here the "Blocks" test signal is distorted by a 7-point uniform averaging (blurring) operator and further corrupted by an additive white Gaussian noise of small variance. The method we propose is compared with a pseudo-WVD type method[2] in which the inverse operator is first applied and then wavelet shrinkage is carried out on each wavelet coefficient corresponding to the noise variance associated with it (the noise variances associated with the wavelet coefficients of the filtered data can be easily computed from the inverse operator and the original noise variance). The soft thresholding function of Donoho [22] is used and, at any given scale, the threshold level for the i-th wavelet coefficient is given as $\tau_i = \alpha\sigma_i$, where σ_i is the noise standard deviation of the coefficient. To give the WVD method an advantage, at each scale we select α to minimize the total square error in that particular scale, using the true signal wavelet coefficients as a reference. This clairvoyant choice is not possible in practice, but here it allows us to evaluate the very best possible performance of the WVD method. From Figure 4, we see that the performance of the WVD soft thresholding scheme with the optimal (clairvoyant) threshold selection is much worse than our new method. This is because the WVD method based on coefficient thresholding ignores the (noise) correlations among the coefficients introduced by the inverse operator.

Notice also that there is performance improvement with estimated parameters based on the initially recovered signal. We could continue this process of re-estimating the parameters repeatedly. However, we find empirically that the restoration performance does not improve significantly with more iterations.

In Figure 5 we demonstrate the application of our method to signal restoration involving correlated noise. In this example, a white Gaussian noise is passed through a 5-point averaging low pass filter to produce a correlated noise of equal power. This noise is then added to the blurred signal used in Figure 4. We see that by taking into account the noise correlation there is an additional gain in SNR of over 5dB. The restored signals were both obtained by first obtaining an initial restoration, then estimating the proper variances at each scale and re-applying our procedure, as described above.

[2] The "true" WVD approach is not applicable to a boxcar blurring function [7].

	Bridge	Mandrill	Fruit	Camera man	Building
σ_0^2	0.0033	0.0136	0.0005	0.0002	0.0008
σ_1^2	0.4930	0.4911	0.6330	0.2841	0.7774

Table 3: σ_0^2 and σ_1^2 trained from EM algorithm. All image pixel intensities are linearly scaled to the range [0,1].

3.2 Application to 2-D Image Restoration

In this section we apply the method developed in this paper to image restoration. We first examine the wavelet coefficient distributions of a set of representative, natural images. In Table 3, we list the σ_0^2 and σ_1^2 values trained from some common test images. We see that σ_1^2 stays in a fairly stable range, while σ_0^2 has a large variation. σ_0^2 tends to be large for images with more texture content such as the "bridge" and "mandrill" test images. While it tends to be significantly smaller for less textural images such as the remaining 3 test images. So we can first roughly set $\sigma_0^2 = 0.01$ for textural images and $\sigma_0^2 = 0.0001$ for images such as "Fruit".

As in the 1-d case, we can use the initially restored image to reestimate the parameters and then go through the process again to get a more accurate result. It is known that wavelet coefficients have exponential decay property across scales, so we can also estimate such property based on the roughly estimated image at hand. In Figure 3 we compare the parameters trained from the original image and those trained from the initially restored image. We see that the exponential decay of σ_1^2 is well preserved, while σ_0^2 has a large change even though it remains roughly constant.

Hence in the following experiments we only reestimate σ_1^2 across scales while keeping σ_0^2 unchanged.

One drawback of wavelet transform that limits its application is its translation dependence, which is also the cause of "blocky" artifacts in, for example, wavelet-based denoising. In [23] a translation invariant approach is taken to yield better performance. The idea is simple, for each circularly translated version of the observed data, apply the wavelet-based processing algorithm, then inverse translate the results back and take the average of them. This is equivalent to using different wavelet bases, all of which are translated versions of each other, to process the data and then take the average result. Here if we let $\mathcal{W}_v$ represent the wavelet transform using the standard wavelet basis translated by vector v, then we have

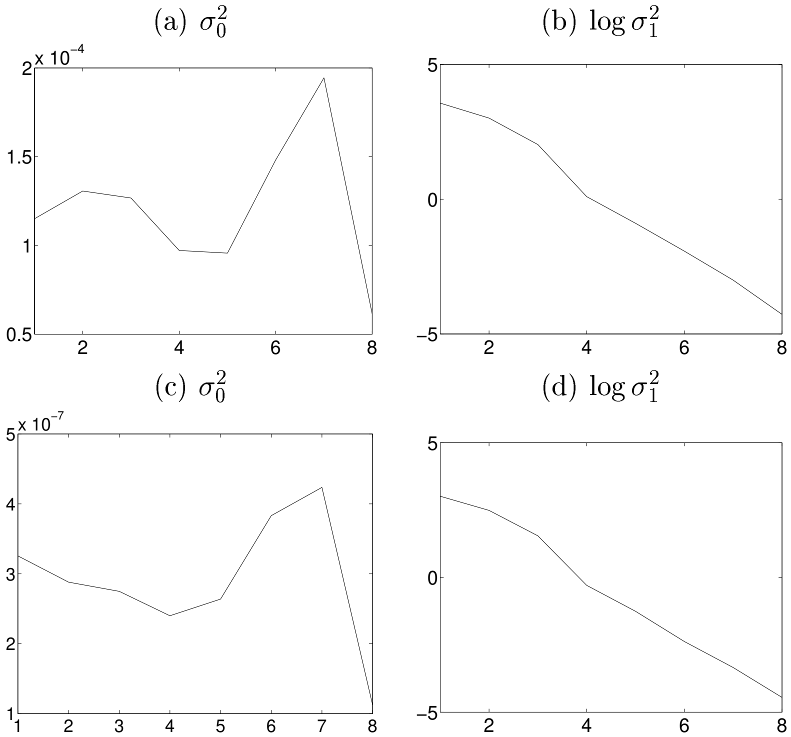

Figure 3: *Reestimation of σ_0^2 and σ_1^2 across scales for the "camera man" test image. (a), (b), from clean image. (c), (d), from restored image.*

$$y = \mathcal{L}\mathcal{W}_v^{-1}\boldsymbol{\theta}_v + \mathbf{n}$$

Notice that translation is also a linear operator. For each v, we apply the restoration process. In the end we take the average, i.e.,

$$\widehat{\mathbf{x}} = \frac{1}{|V|}\sum_{v\in V}\mathcal{W}_v^{-1}\widehat{\boldsymbol{\theta}}_v.$$

where V is the set of possible translations.

Since circular convolution operator tends to produce bad effect near the image boundary, we do a symmetric extention of the original image to double both the length and width of the original image, which is placed in the center of the new enlarged image. This is to facilitate the use of wavelet transform. Then the circular convolution is applied as usual and in then end we crop the resulting image to get back to the recovered image of the same size as that of the original.

In Figure 6 we apply our method to the restoration of cameraman image. We first use a rough set of parameters to do an initial recovery. The artifacts can be clearly seen. Then we use this recovery to estimate σ_1^2 scale by scale to get a better estimate of the parameters. After that we apply the algorithm for a collection of 25 translations (a 5×5 window of circular translations) and then take the average. The final improved SNR is 7.0dB, which is higher than the 6.7dB reported in [13]. The SNR of the final restored image is also comparable with that reported in [15]. Also we see that the parameter reestimation and translation invariance together provide an additional gain of over 3dB.

In Figure 7 we carry out the same procedure on two texture rich images. A higher σ_0^2 value is chosen according to what we discussed previously. From the final results, we see that the texture regions are relatively well recovered, especially in difficult areas such as the facial region of the mandrill.

4 Conclusions

In this paper, we develop a new wavelet-based method for solving linear inverse problem and apply it to image restoration from noisy and distorted observations. This Bayesian method employs a wavelet-domain independent GMM signal prior to produce a computationally efficient restoration algorithm that has produced good results. This approach has several advantages over existing multiscale/wavelet-based restoration methods. First, it is based on a unified Bayesian framework in which operator inversion and noise suppression are fully integrated in a joint MAP criterion. Moreover, MAP optimization is carried out with a simple iterative algorithm which is guaranteed to converge monotonically to a local maximum of the *a posteriori* density function and is highly signal adaptive. This new method estimates the image wavelet coefficients in a coordinated fashion, accounting for statistical correlations between coefficients due to correlated observation noise and/or inverse filtered observation noise. We extend the result to a more general case to handle non-uniform but concave prior function of the number of high variance states of the wavelet coefficients by developing an efficient integer programming algorithm. The extensions of the method developed in this paper are many. One is the direct application to problems involving any multi-mixture model, or use more realistic exponential decay across scale property on wavelet coefficients. Recently, graphical models, such as the wavelet-domain hidden Markov models (HMM) proposed by Crouse *et al.* [17, 24], have been applied in image processing with very good results. In this paper, we only present the simplest independent wavelet coefficient prior to illustrate the key ideas and basic results, but the framework could be extended to handle more sophisticated HMM type priors.

References

[1] H. C. Andrews and B. R. Hunt, *Digital Image Restoration.* Englewood Cliffs, New Jersey: Prentice Hall, 1977.

[2] F. M. J.-L Starck and A. Bijaoui, *Image Processing and Data Analysis.* Cambridge Univeristy Press, 1998.

[3] M. Bertero, "Linear inverse and ill-posed problems," *Advances in Electronics and Electron Physics*, vol. 75, no. 1, pp. 1–120, 1989.

[4] F. O'Sullivan, "A statistical perspective on ill-posed inverse problems," *Statistical Science.*, vol. 1, no. 4, pp. 502–527, 1986.

[5] K. Castleman, *Digital Image Processing.* Englewood Cliffs, New Jersey: Prentice-Hall, 1996.

[6] F. Abramovich and B. Silverman, "Wavelet decomposition approaches to statistical inverse problems," *Biometrika*, vol. 85, pp. 115–129, March 1998.

[7] D. L. Donoho, "Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition," *App. and Comp. Harmonic Analysis*, vol. 2, pp. 101–126, 1995.

[8] I. M. Johnstone and B. W. Silverman, "Speed of estimation in positron emission tomography and related inverse problems," *Annals of Statistics*, vol. 18, pp. 251–280, 1990.

[9] I. M. Johnstone and B. W. Silverman, "Discretization effects in statistical inverse problems," *Journal of Complexity*, vol. 7, pp. 1–34, 1991.

[10] F. Abramovich, T. Sapatinas, and B. W. Silverman, "Wavelet thresholding via a Bayesian approach," *J. Roy. Statist. Soc. Ser. B., 60, 725-749*, vol. 60, pp. 725–749, 1998.

[11] J. Liu and P. Moulin, "Complexity-regularized image restoration," *Proc. IEEE Int. Conf. on Image Proc. (ICIP'98)*, vol. 1, pp. 555–559, 1998.

[12] V. Koivunen, "A robust nonlinear filter for image restoration," *IEEE Trans. Image Processing*, vol. 4, no. 5, pp. 569–578, 1995.

[13] M. R. Banham and A. K. Katsaggelos, "Spatially adaptive wavlet-based multiscale image restoration," *IEEE Trans. Image Processing*, vol. 5, no. 4, pp. 619–634, 1996.

[14] J. Z. G. Wang and G.-W. Pan, "Solution of inverse problem in image processing by wavelet expansion," *IEEE Trans. Image Processing*, vol. 4, no. 5, pp. 579–593, 1995.

[15] R. Neelamani, H. Choi, and R. G. Baraniuk, "Wavelet-based deconvolution for ill-conditioned systems," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP'99*, vol. 1, 1999.

[16] D. L. Donoho, "Unconditional bases are optimal bases for data compression and for statistical estimation," *App. and Comp. Harmonic Analysis*, vol. 1, pp. 100–115, Dec. 1993.

[17] M. Crouse, R. Nowak, and R. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Processing*, vol. 46, pp. 886–902, 1998.

[18] I. Johnstone and B. Silverman, "Wavelet threshold estimators for data with correlated noise," *J. Roy. Statist. Soc. B*, vol. 59, pp. 319–351, 1997.

[19] Y. Wan and R. Nowak, "A bayesian multiscale approach to joint image restoration and edge detection," in *Proceedings of SPIE Conf. 3813 — Wavelet Application in Signal and Image Processing VII*, Denver, CO, July, 1999.

[20] M. Belge, M. E. Kilmer, and E. L. Miller, "Wavelet domain image restoration with adaptive edge-peserving regularity," *IEEE Trans. Image Processing*, vol. 9, no. 4, pp. 597–608, 2000.

[21] M. Figueiredo and J. Leitao, "Unsupervised image restoration and edge location using compound gauss-markov random fields and the mdl principle," *IEEE Transactions on Image Processing*, vol. 6, no. 8, pp. 1089–1102, August 1997.

[22] D. Donoho and I. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Statist. Assoc.*, vol. 90, pp. 1200–1224, Dec. 1995.

[23] R. Coifman and D. Donoho, "Translation invariant de-noising," in *Lecture Notes in Statistics: Wavelets and Statistics*, vol. New York: Springer-Verlag, pp. 125–150, 1995.

[24] R. Nowak, "Multiscale hidden Markov models for Bayesian image analysis," in *Bayesian Inference in Wavelet Based Models*, Springer-Verlag, 1999. Editors B. Vidakovic and P. Müller.

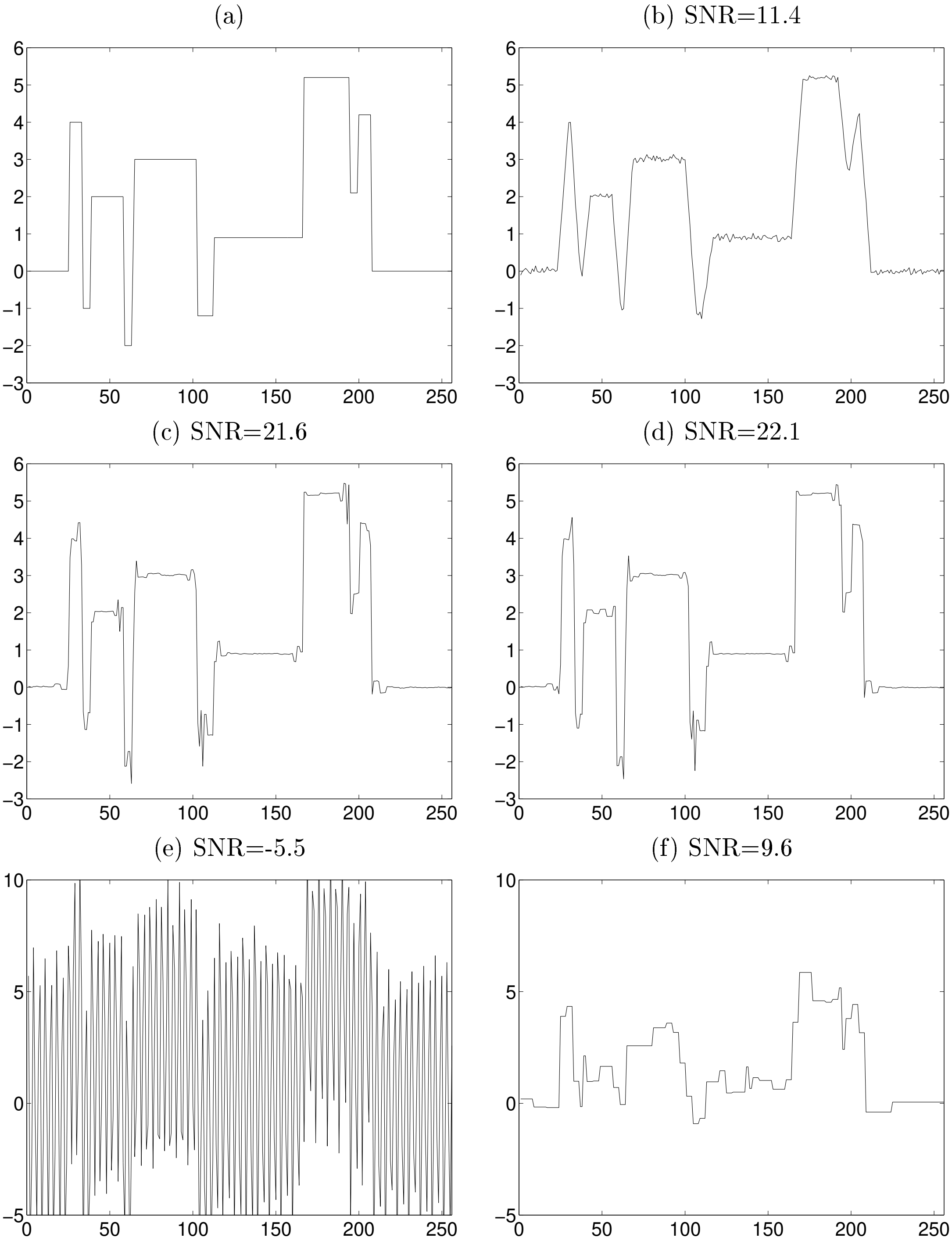

Figure 4: *1-D deconvolution performance comparison (a) Clean "Blocks" signal. (b) Corrupted signal after convolution and with added white Gaussian noise, $\sigma = 0.05$. (c) Bayesian IGMM restoration with $\sigma_0^2 = \sigma^2$, $\sigma_1^2 = 1000\sigma^2$. (d) Restoration using estimated parameters $\sigma_0^2 = 0.0019 \approx \sigma^2$, $\sigma_1^2 = 26.32 \approx 10^4 \cdot \sigma^2$. (e) Direct inversion. (f) Restoration by WVD type wavelet soft thresholding of the signal in (e) with optimal (clairvoyant) multiscale scaling factors.*

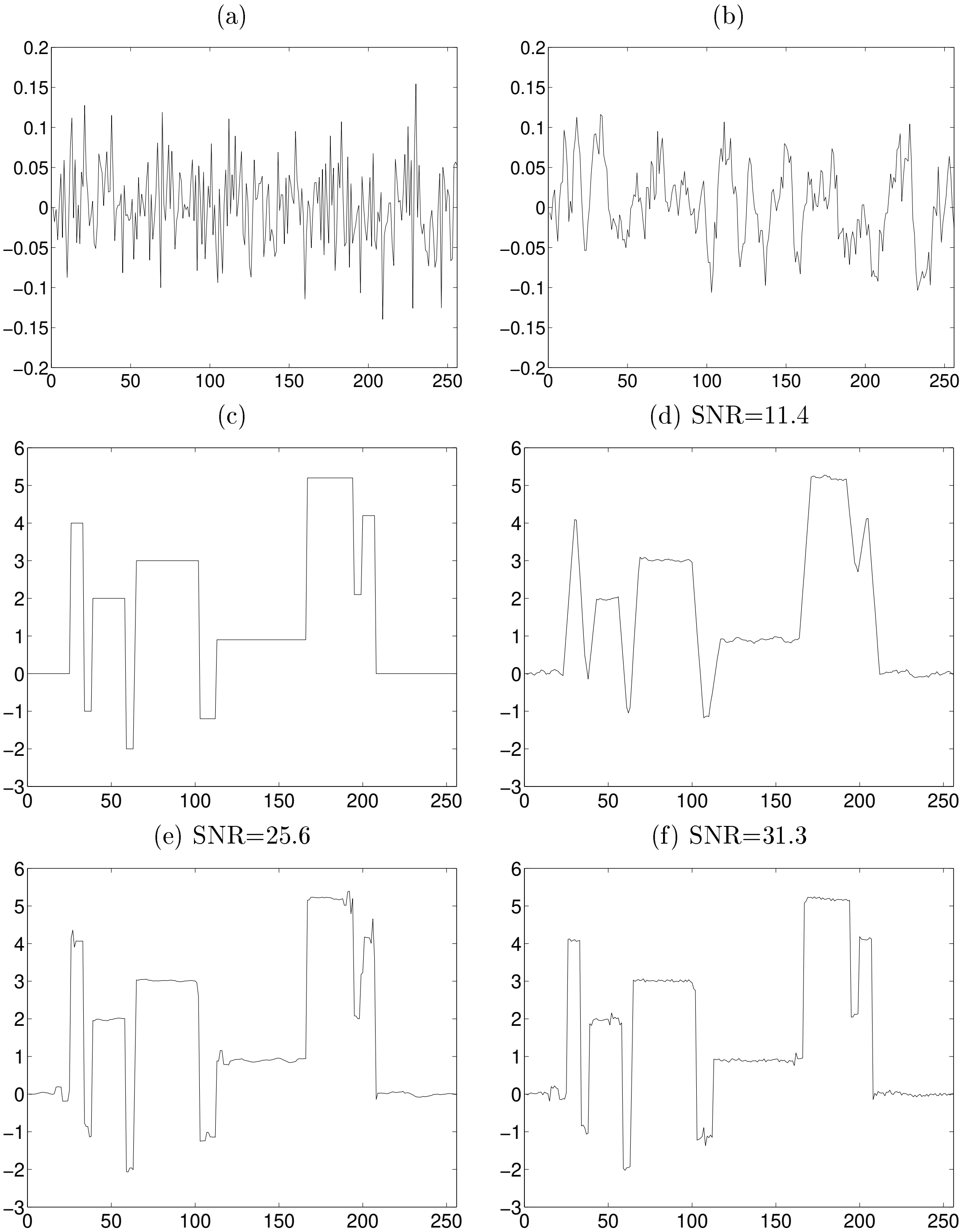

Figure 5: *(a) iid white Gaussian noise ($\sigma = 0.05$). (b) Correlated noise obtained by passing that in (a) through a low pass filter. (c) Original signal. (d) Blurred noisy signal. (e) Restoration by treating the noise as if it were white. (f) Restoration obtained by taking into account the matrix filter A in (4).*

(a) (b) SNR=15.3

(c) SNR=19.1 (d) SNR=22.3

Figure 6: *2-D image restoration. (a) Clean test image. (b) Observed image after blurring by* 9×9 *box-car matrix with added white Gaussian noise (40dB BSNR). (c), Initially recovered image using* $\sigma_0^2 = 10^{-4}$ *and* $\sigma_1^2 = 10^{-1}$*. (d) Improved result after reestimation of* σ_1^2 *with 25 translations.*

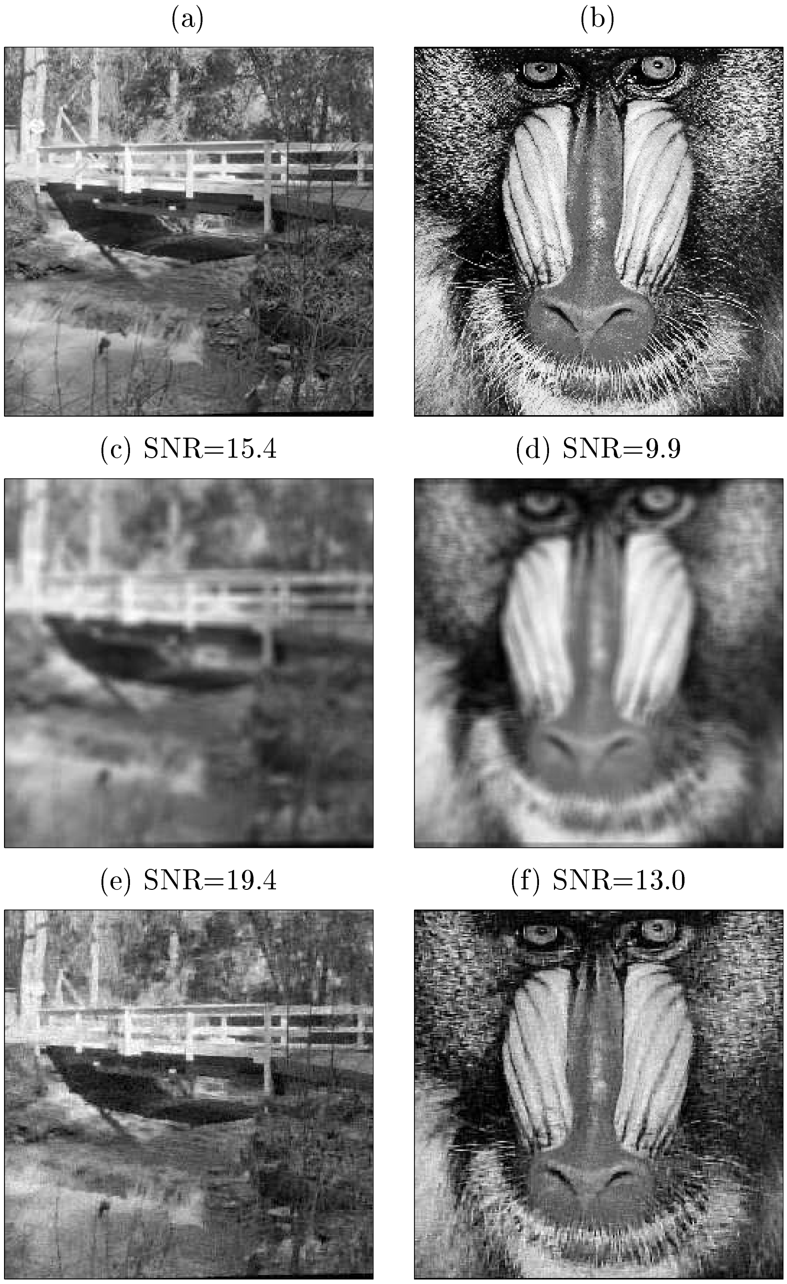

Figure 7: *Experiments on "bridge" and "mandrill" test images (a) (b), Clean images. (c) (d) Observed images after blurring by* 7×7 *box-car matrix with added white Gaussian noise (BSNR=30dB). (e) (f), Recovered images with initially chosen* $\sigma_0^2 = 10^{-2}$ *and* $\sigma_1^2 = 10^{-1}$.