Automated Recovery of Compressedly Observed Sparse Signals From Smooth Background

Zhaofu Chen, Rafael Molina, Member, IEEE, and Aggelos K. Katsaggelos, Fellow, IEEE

Abstract—We propose a Bayesian based algorithm to recover sparse signals from compressed noisy measurements in the presence of a smooth background component. This problem is closely related to robust principal component analysis and compressive sensing, and is found in a number of practical areas. The proposed algorithm adopts a hierarchical Bayesian framework for modeling, and employs approximate inference to estimate the unknowns. Numerical examples demonstrate the effectiveness of the proposed algorithm and its advantage over the current state-of-the-art solutions.

Index Terms—Bayesian algorithm, robust principal component analysis, compressive sensing

EDICS Category: MLR-BAYL, SSP-HIER, SSP-SSEP

I. INTRODUCTION

Consider the measurement system expressed as
\[ Y = \Phi X + B + N, \]
where the signal of interest \( X \in \mathbb{R}^{M \times N} \) undergoes a transformation \( \Phi \in \mathbb{R}^{N \times M} \) and is corrupted by both noise \( N \) and a smooth background \( B \). The signal \( X \) is assumed to have sparse columns, i.e., \( \| x_i \|_0 \ll M \) for \( i = 1, \ldots, N \), where \( x_i \) denotes the \( i^{\text{th}} \) column of \( X \) and \( \| \cdot \|_0 \) is the \( \ell_0 \)-norm. The smooth background \( B \) is a low-rank matrix. The transformation \( \Phi \) in general has the effect of compression, i.e., \( L \leq M \).

The model in (1) is found in a number of applications. In network anomaly detection [1], \( X \) consists of the temporal snapshots of flow anomalies, \( \Phi \) represents the network routing operation, and \( B \) contains the smooth link measurements resulted from the normal traffic flows, respectively. As another application, in video surveillance from compressed measurements [2], \( X \) denotes the moving objects in the foreground, \( \Phi \) is a known measurement matrix, and \( B \) is the compressed version of the background.

The model in (1) is also closely related to Compressive Sensing (CS) [3]–[5] and Robust Principal Component Analysis (RPCA) [6]. For CS, where \( B \) is not present, algorithms generally employ regularized optimization (e.g., in [7]–[9]) or Bayesian approaches (e.g., in [10]–[12]). For RPCA, where \( \Phi = I \), regularized optimization problems are solved with proper convex relaxation to sparsity and rank [13]–[15]. Alternatively, Bayesian approaches model the sparse and low-rank components with appropriate prior structures and employ approximate inference techniques for estimation [16], [17].

Recently, the model in (1) has received much attention from the signal processing community. [1] provides the conditions for identifiability and recoverability. Algorithms based on convex optimization have been proposed in [1], [18], [19]. The optimization-based algorithms usually require proper selection of user parameters. In contrast, [20] takes a Bayesian perspective and automatically estimates all model parameters. A limitation of [20], though, is the memory and computational requirement incurred by the use of Hessian information.

In this paper, we incorporate the Laplace prior to model the sparse signal. As explained in [21], this prior has been shown to promote sparsity to a higher level than the Sparse Bayesian Learning (SBL) prior used in [20]. In addition, in order to reduce memory consumption, we develop a constructive approach based on the principles in [22] and [12] that essentially replaces the demanding matrix inversion with a sequence of efficient rank-one updates. The algorithm proposed herein is free of user parameters, making it amenable to be deployed for automatic operation.

This paper is organized as follows. In Section II we introduce the hierarchical Bayesian model. The inference procedure with the constructive algorithm is outlined in Section III. Numerical examples demonstrating the effectiveness of the proposed algorithm are provided in Section IV. Finally, we draw concluding remarks in Section V.

II. HIERARCHICAL BAYESIAN MODEL

A. Modeling Additive Noise

\[ N = \{ n_{ij} \}, \] in (1) contains uncorrelated noise. We employ an independent and identically distributed (i.i.d.) Gaussian distribution \( p(n_{ij} | \beta) = N(n_{ij} | 0, \beta^{-1}) \), with precision \( \beta \) modeled using the conjugate Gamma prior, i.e., \( p(\beta | a_0, b_0) = \mathcal{G}(\beta | a_0, b_0) \), where the hyperparameters \( a_0 \) and \( b_0 \) are fixed at small values to approximate the non-informative Jeffreys prior.

B. Modeling Smooth Background

Consider the factorization of the smooth background \( B = PQ^T \), where \( P \) and \( Q \) are \( L \times K \) and \( N \times K \) matrices, respectively, and \( K \) is a loose upper bound for the rank of...
B. Smoothness of $\mathbf{B}$ is resulted when most of the $K$ outer products in $\mathbf{PQ}^T$ are zeros. To achieve this, common sparsity promoting priors are simultaneously assigned to the columns of $\mathbf{P}$ and $\mathbf{Q}$, as in [20]
\[
\begin{align*}
    p(\mathbf{P}|\gamma) &= \prod_{i=1}^K N(\mathbf{p}_i|0, \gamma_i^{-1}I) \\
    p(\mathbf{Q}|\gamma) &= \prod_{j=1}^N N(\mathbf{q}_j|0, \gamma_j^{-1}I),
\end{align*}
\]  
where the precisions $\{\gamma_i\}$ are assigned Gamma priors, i.e.,
\[
p(\gamma_i|a_1, b_1) = G(\gamma_i|a_1, b_1).
\]  
The hyperparameters $a_1$ and $b_1$ are fixed at small values to yield broad distributions.

C. Modeling Sparse Signal

In least squares fitting problems $\ell_1$-norm is commonly used for regularization so that sparse solutions are preferred. This is equivalent to adopting a Laplace prior on the sparse signals
\[
p(x_j|\omega_j) = \lambda_j^{M/2}/2^M \exp(-\sqrt{\lambda_j}|x_j|_1).
\]  
The Laplace prior, though being sparsity promoting, is not conjugate to the Gaussian model. To overcome this limitation, a hierarchical model is established as in [23], where
\[
p(x_j|\omega_j) = \lambda_j/2 \exp(-\lambda_j|x_j|^2/2).
\]  
Finally, the $\lambda_j$ is modeled as $p(\lambda_j|a_2, b_2) = G(\lambda_j|a_2, b_2)$, where $a_2$ and $b_2$ are fixed at small values.

D. Complete System Model

By combining the observation and prior models, the joint distribution is expressed as
\[
p(\mathbf{Y}, \mathbf{X}, \mathbf{P}, \mathbf{Q}, \Omega, \lambda, \gamma, \beta|a_0, b_0, a_1, b_1, a_2, b_2) = p(\mathbf{Y}|\mathbf{X}, \mathbf{P}, \mathbf{Q}, \beta)p(\beta|a_0, b_0)p(\mathbf{P}|\gamma)p(\mathbf{Q}|\gamma) \times p(\gamma|a_1, b_1)p(\mathbf{X}\Omega)p(\Omega|\lambda)p(\lambda|a_2, b_2),
\]  
where $\Omega = [\omega_1, \ldots, \omega_N]$, $\lambda = [\lambda_1, \ldots, \lambda_N]$, and $\gamma = [\gamma_1, \ldots, \gamma_K]$.

III. APPROXIMATE BAYESIAN INFERENCE

Bayesian approaches seek the posterior distributions of the unknowns given the observations. Approximations are usually employed since the exact posterior distributions are analytically intractable. Common approximate approaches include (1) point estimation, such as Maximum Likelihood (ML) and Maximum A Posteriori (MAP) estimation, (2) sampling approximations, such as Gibbs Sampler (GS), and (3) Variational Bayes (VB), etc.

In this paper, we employ VB to approximate the posterior distributions of $\mathbf{P}$, $\mathbf{Q}$ and $\gamma$. For $\mathbf{X}$, $\Omega$ and $\lambda$, we introduce a constructive algorithm based on marginalization. Finally, the noise precision $\beta$ is estimated via VB.

A. Inference of Smooth Background

The inference of $\mathbf{P}$, $\mathbf{Q}$ and $\gamma$ follows from the invocation of the mean-field approximation and variational calculus. Omitting the details that can be found in [20], the posterior means of $\mathbf{P}$ and $\mathbf{Q}$ are given by
\[
\begin{align*}
    \bar{\mathbf{P}} &= \bar{\beta} (\mathbf{Y} - \Phi \mathbf{X}) \mathbf{Q} \Sigma_p, \\
    \bar{\mathbf{Q}} &= \bar{\beta} (\mathbf{Y} - \Phi \mathbf{X})^T \mathbf{P} \Sigma_q,
\end{align*}
\]  
where $\bar{\beta}$ denotes the posterior mean of $\beta$, and the covariance matrices $\Sigma_p$ and $\Sigma_q$ are inter-related as
\[
\begin{align*}
    \Sigma_p &= (\bar{\beta} \bar{\gamma}^2 \mathbf{Q} + N \bar{\beta} \Sigma_q + \Gamma)^{-1}, \\
    \Sigma_q &= (\bar{\beta} \bar{\gamma}^2 \bar{\mathbf{P}} + L \bar{\beta} \Sigma_p + \Gamma)^{-1}.
\end{align*}
\]  
The term $\bar{\Gamma}$ above is a diagonal matrix constructed from $\bar{\gamma}$. Similarly, the approximate posterior mean of $\gamma_i$ is found to be
\[
\bar{\gamma}_i = \frac{1}{\bar{\theta}_i^2 + \frac{1}{\bar{\theta}_j^2} + \frac{N \sigma_{ii}^2 + 2b_i}{L}}
\]  
where $\sigma_{ii}^2$ and $\sigma_{ji}^2$ are the $i$th elements on the diagonals of $\Sigma_p$ and $\Sigma_q$, respectively. In the iterations, the posterior means in (6) and (7) are used as the estimates of the corresponding variables.

B. Inference of Sparse Signal

If we continued applying VB inference we would have to find, within the iterative process, the distributions $q(\lambda)$, $q(\Omega)$, and $q(\mathbf{X})$ minimizing the Kullback-Leibler divergence
\[KL(q(\lambda)q(\Omega)q(\mathbf{X})p(\mathbf{G}|\mathbf{X}, \beta)p(\mathbf{X}|\Omega)p(\Omega|\lambda)p(\beta))\], where $\mathbf{G} = \mathbf{Y} - \mathbf{PQ}^T$, $p(\mathbf{G}|\mathbf{X}, \beta) = \prod jN(g_j|x_j, \beta^{-1}I)$, and $q(\lambda)$ and $q(\Omega)$ are degenerate distributions. Then we would have
\[
q(\mathbf{X}) = p(\mathbf{X}|\mathbf{G}, \beta, \Omega) = \prod_{j=1}^N p(x_j|g_j, \beta, \omega_j),
\]  
and with some algebra it follows that this conditional distribution is a multivariate Gaussian $N(x_j|x_j^0, \Sigma^x_{ji})$ with
\[
\begin{align*}
    x_j &= \beta \Sigma^x_{ji} \Phi^T g_j, \\
    \Sigma^x_{ji} &= (\Phi^T \Phi + \Omega_j^{-1})^{-1},
\end{align*}
\]  
where $\Omega_j = diag(\omega_{j1}, \ldots, \omega_{jM})$.

Two observations follow from (9). First, $\Sigma^x_{ji}$ involves the inversion of an $M \times M$ matrix, which is memory and computationally intensive. Second, if an $\omega_j$ is zero, then the corresponding $x_j$ must be zero (see (4)), and then the $\ell^2$ column and row in $\Sigma^x_{ji}$ are eliminated by removing $\omega_j$ from the diagonal of $\Omega_j^{-1}$ and the $\ell^i$ column of $\Phi$. Therefore, the dimension of $\Sigma^x_{ji}$ is determined by the number of nonzeros in $\omega_j$. Since $x_j$ is sparse, most of its entries are expected to be zero with zero variance.

Let us now find the nonzero components of $\omega_j$. We fix $q(\mathbf{X})$ to (8), where $\lambda$, $\Omega$, and $\beta$ are estimated from
\[
p(\mathbf{G}, \lambda, \Omega, \beta) = p(\mathbf{G}) \prod_{j=1}^N p(g_j, \omega_j, \lambda_j|\beta) = p(\beta) \prod_{j=1}^N N(g_j|0, \sigma_{jj})p(\omega_j|\lambda_j)p(\lambda_j),
\]  
where $C_j = \beta^{-1}I + \Phi \Phi^T$.

To determine the optimal $\omega_j$, we take the logarithm of (10), drop the terms independent of $\omega_j$ and maximize
\[
\begin{align*}
    \mathcal{L}(\omega_j) &= -\frac{1}{2} \log |C_j| - \frac{1}{2} \Phi^T C_j^{-1} \mathbf{e}_j - \frac{1}{2} \sum_{i=1}^M \omega_i, \\
    C_j &= \beta^{-1}I + \sum_{k \neq i} \omega_k \phi_k \phi_k^T + \omega_i \phi_i \phi_i^T
\end{align*}
\]  
where $-iC_j$ denotes the portion of $\Sigma_j$ with the contribution from $\omega_j$ excluded. Utilizing properties of determinant, we have
\[
\begin{align*}
    \log |C_j| &= \log |-iC_j| + \log (1 + \omega_i \phi_i^T (-iC_j)^{-1} \phi_i) \\
    &= \log |\tilde{\Sigma} | + \log (1 + \omega_i r_{ij}),
\end{align*}
\]  
where we define $r_{ij} = \phi_i^T (-iC_j)^{-1} \phi_j$ for notational clarity. Invoking matrix inversion lemma, we have
\[
C_j^{-1} = (-iC_j)^{-1} - \frac{\omega_i (-iC_j)^{-1} \phi_i \phi_i^T (-iC_j)^{-1}}{1 + \omega_i r_{ij}}
\]
and
\[ g_j^T C_j^{-1} g_j = g_j^T (-C_j)^{-1} g_j - \frac{\omega_{ij} (\phi_i^T (-C_j)^{-1} g_j)^2}{1 + \omega_{ij} r_{ij}} \]
\[ = g_j^T (-C_j)^{-1} g_j - \frac{\omega_{ij} s_{ij}^2}{1 + \omega_{ij} r_{ij}}, \]
where \( s_{ij} = \phi_i^T (-C_j)^{-1} g_j \) is defined for notational clarity.

Substituting (13) and (15) into (11), and retaining only terms dependent on \( \omega_{ij} \), it follows that
\[ l(\omega_{ij}) = \frac{1}{2} \left[ \log \frac{1}{1 + \omega_{ij} r_{ij}} + \frac{\omega_{ij} s_{ij}^2}{1 + \omega_{ij} r_{ij}} - \omega_{ij} \right]. \]
(16)
It is clear that the \( \omega_{ij} \) maximizing (16) also maximizes (11).

Note in (16) both \( \omega_{ij} \) and \( \omega_{ij} r_{ij} \) are independent of \( \omega_{ij} \). The univariate function \( l(\omega_{ij}) \) can be maximized by examining its derivative and taking into account the domain of \( l(\omega_{ij}) \). With some algebra, the optimal \( \omega_{ij} \) is found to be
\[ \omega_{ij}^* = \begin{cases} \frac{-r_{ij} + 2 \lambda_j}{2 \lambda_j r_{ij}} + \sqrt{\frac{r_{ij}^2 + 4 \lambda_j s_{ij}^2}{4 \lambda_j r_{ij}^2}}, & s_{ij} > r_{ij} > \lambda_j \\ 0, & \text{otherwise} \end{cases} \]
(17)
As discussed above, a zero-valued \( \omega_{ij} \) indicates that the corresponding row and column in \( \Sigma_j^x \) are all zeros. Therefore, (17) provides a guideline as how the model should be adjusted. In the iterative procedure, depending on the previous estimate of \( \omega_{ij} \) and the condition \( s_{ij}^2 > r_{ij} \leq \lambda_j \), there are three possible adjustments to the model, namely
\[ \begin{cases} \text{Add } \omega_{ij}, & \text{if previously excluded and } s_{ij}^2 > r_{ij} > \lambda_j \\ \text{Remove } \omega_{ij}, & \text{if previously included and } s_{ij}^2 > r_{ij} \leq \lambda_j \\ \text{Update } \omega_{ij}, & \text{if previously included and } s_{ij}^2 > r_{ij} > \lambda_j \end{cases} \]
(18)
Note that, for a fixed \( j \), all these updates only involve efficient rank-one modifications to \( \Sigma_j^x \) and \( x_j \), rather than requiring matrix inversions (see [22] for details). Also note that the effective dimension of \( \Sigma_j^x \) and \( x_j \) is much smaller than \( M \), as only a small subset of \( \{\omega_{ij}\} \) are included at any time in the iterative process. In this way the constructive approach alleviates the memory and computational requirement. The selection of an \( \omega_{ij} \) for update can be done either randomly, or such that the largest increase in the log-likelihood is obtained.

For a fixed \( j \), whenever the model is adjusted, \( \{r_{ij}\}_i \) and \( \{s_{ij}\}_i \) need to be updated efficiently. Invocation of matrix inversion lemma yields
\[ C_j^{-1} = \beta I - \beta^2 \Phi \Sigma_j^x \Phi^T, \]
where \( \beta = \frac{\lambda_j}{(M - 1 + \omega_{ij}^2) r_{ij}} \).
(19)
Finally, the update of \( \lambda_j \) is done by maximizing the logarithm of (10) with respect to \( \lambda_j \) to obtain
\[ \lambda_j^* = \frac{(M - 1 + \omega_{ij})}{\sum_i \omega_{ij}/2 + b_2}. \]
(25)
The updates in (18) and (21) to (25) are iterated until convergence. In each iteration an \( \omega_{ij} \) is selected for update. One criterion for convergence is that the relative change in log-likelihood (11) falls below a pre-defined threshold. In practice, we find that a small number of iterations (e.g., 20) is usually sufficient for good performance. When the iterations converge, the approximate posterior mean \( x_j \) is used as an estimate of \( x_j \).

C. Inference of noise power

Via mean-field approximation, the approximate posterior distribution of the noise precision \( \beta \) is found to be Gamma, with mean
\[ \beta = \frac{LN + a_0}{\|Y - \Phi X - PQ^\top\|^2 + b_0}, \]
where
\[ \|Y - \Phi X - PQ^\top\|^2 = \|Y - \Phi X - \hat{P}Q^\top\|^2 + N \times \text{trace}(P^\top \Sigma X Q) + L \times \text{trace}(Q^\top \Sigma P) \]
(27)
The iterative procedure described above continues until convergence is reached. One criterion for convergence is that the relative change in the variables, e.g., \( X \) or \( B \), falls below a pre-defined threshold.

IV. Numerical Results

A. Simulation

1) Recovery Accuracy: We first demonstrate the performance of the proposed algorithm on simulated data. The data generation is described as follows. We consider problems of varying scales \( M \in \{100, 200, 400, 600\} \). For each scale, \( L = M/2 \) and \( N = M \). The ground truth sparsity is fixed at \( ||X||_0 = 5\% MN \), and nonzeros of \( X \) are independently drawn from uniform \( U(-10, 10) \) distribution. The ground truth rank of \( B \) is fixed at \( r = 0.05L \). The smooth component \( B \) is generated as the product of an \( L \times r \) matrix and an \( r \times N \) matrix, whose elements are drawn from i.i.d. \( N(0, 100/L) \) and \( N(0, 100/N) \) distributions, respectively. Additive noise with standard deviation \( \sigma_n = 0.05 \) is added to the measurement. The transformation \( \Phi \) is generated similarly to that in [1], which consists of random orthonormal rows.

The proposed algorithm is applied to recover the sparse signal and smooth background, with estimates denoted as \( \hat{X} \) and \( \hat{B} \), respectively. All hyperparameters in the model are set to a small value (e.g., \( 10^{-5} \)), making the algorithm fully automated. As a comparison, we show the performance on the same test data of two alternative algorithms, namely the Variational Bayes Sparse Estimator (VBSE) proposed in [20], and the state-of-the-art Alternating Directions Method of Multipliers (ADMM) proposed in [1]. For the problem considered herein, VBSE is the only available Bayesian approach to our knowledge and ADMM is reported to have high recovery accuracy in the literature.

In Table I four metrics are used to evaluate the performance: rank of \( B \), number of nonzeros in \( X \), reconstruction error...
The ADMM algorithm is most sensitive to noise due to its empirical smallest $\epsilon^\times$ and tracks the rank of $\hat{B}$, and ADMM-2 that gives almost correct $r(\hat{B})$, and ADMM-3 that gives correct $r(\hat{B})$, and ADMM-1 that yields the lowest $\epsilon^\times$, ADMM-2 that gives correct $r(\hat{B})$, and ADMM-3 that gives almost correct $||B||_0$.

It is clear from Figure 1 that the proposed algorithm is robust to additive noise, thanks to its constructive nature, has lower memory requirement than VBSE, making it scale better with problem dimensions. Last but not least, although the numbers seem to imply that ADMM is close to an order of magnitude more efficient than the proposed algorithm, we would like to point out that the numbers shown here do not include the overhead of parameter tuning. Parameter tuning may take significantly longer than the running time shown in the table.

2) Robustness to Noise: To examine robustness of the proposed algorithm to noise, we fix $L = 100$, $M = N = 200$, $r(\hat{B}) = 5$, $||X||_0 = 200$, and vary $\sigma_n$ from 0.01 to 1. Three parameter settings for ADMM are considered, namely ADMM-1 that yields the lowest $\epsilon^\times$, ADMM-2 that gives correct $r(\hat{B})$, and ADMM-3 that gives almost correct $||B||_0$.

It is clear from Figure 1 that the proposed algorithm is robust to additive noise, thanks to its estimation of noise level. In particular, it accurately recovers the nonzeros in $X$ and tracks the rank of $B$ even at very low Signal-to-Noise-Ratio (SNR) conditions (e.g., SNR < 10 for $\sigma_n = 1$). The ADMM algorithm is most sensitive to noise due to its inherent modeling assumption. Moreover, it requires careful tuning of user parameters to achieve reasonable results.

3) Ability to Identify Nonzero Signal Elements: To investigate the ability of the proposed algorithm in detecting nonzeros of $X$ under different conditions, we consider $||X||_0$ from 1% to 11% and $r(B) \in \{5, 10, 15\}$. $\sigma_n$ is fixed at 0.01. The data generation is similar to that above.

Figure 2 shows the F-scores at various conditions. The F-score, defined as the harmonic mean of precision and recall, is a balanced indicator of detection performance. From the figure it is seen the proposed algorithm gives reliable detections over a wider range of operational conditions than its alternatives.

V. Conclusions

In this paper we proposed a Bayesian based approach for recovery of sparse signals from compressed measurements when additive noise and a smooth background are present. A memory and computationally efficient constructive algorithm is developed under the framework of Bayesian inference. Advantages over the current state-of-the-art alternative solutions are demonstrated in numerical examples.
REFERENCES


