A HYBRID METHOD FOR DECONVOLUTION OF BERNOULLI-GAUSSIAN PROCESSES

Sinan Yıldırım[a], *A. Taylan Cemgil*[b], *Ayşın B. Ertüzün*[a]

[a] Department of Electrical and Electronics Engineering, Boğaziçi University, Bebek, İstanbul, Turkey
[b] Department of Computer Engineering, Boğaziçi University, Bebek, İstanbul, Turkey

ABSTRACT

We investigate a hybrid method which improves the quality of state inference and parameter estimation in blind deconvolution of a sparse source modeled by a Bernoulli-Gaussian process. In this problem, when both the signal and the filter are jointly estimated, the true posterior is typically highly multimodal. Therefore, when not properly initialized, standard stochastic inference methods, (MCEM, SEM or SAEM), tend to get stuck and suffer from poor convergence. In our approach, we first relax the Bernoulli-Gaussian prior model by a Student-*t* model. Our simulations suggest that deterministic inference in the relaxed model is not only efficient, but also provides a very good initialization for the Bernoulli-Gaussian model. We provide simulation studies that compare the results obtained with and without our initialization method for several combinations of state inference and parameter estimation methods used for the Bernoulli-Gaussian model.

Index Terms— Sparsity, Bernoulli-Gaussian, Student-*t*, Expectation-Maximization (EM), Markov Chain Monte Carlo (MCMC)

1. INTRODUCTION

In blind deconvolution, we usually encounter problems where a latent input signal $\mathbf{x}$ is passed through a *linear time-invariant* (LTI) system with unknown impulse response $\mathbf{g}$, and the resulting signal is corrupted by additive Gaussian noise $\mathbf{v}$, giving the observed signal $\mathbf{y}$. In general, this problem is ill-posed unless further assumptions about the signal and filter are made. A reasonable assumption that holds for many practical problems is that the latent input signal is *sparse*, where it is zero -or very close to zero- almost everywhere but has a few non-zero values, often referred to as spikes [1]. In such cases (seismology, audio processing), it is usually important to find the locations of these spikes, as well as $\mathbf{x}$ itself, and $\mathbf{g}$.

There are several methods for solving this problem. The approaches differ in the way how they model the unknown variables, or the statistical inference/optimization methods they use. The most common approach for modeling a sparse $\mathbf{x}$ is to assume that it is a *Bernoulli-Gaussian* process [2, 3, 1], where one Gaussian has a very small variance and the other's is far from zero. Bernoulli-Gaussian model is easy to handle, since it leads to well known statistical inference/optimization methods, as well as provides an intuitive semantic interpretation by indicating the locations of spikes [2]. However, experience suggests that its performance is highly dependent on initial values of estimated parameters, which is a serious problem when we do not have any prior information about the system parameters.

Modeling samples of a sparse signal with a Student-*t* distribution [4] is another approach. This can be achieved by assigning each sample of $\mathbf{x}$ a Gaussian random variable whose variance is an inverse-Gamma random variable. This prior assumption is used in source separation and blind deconvolution problems [5, 6]. In contrast to the Bernoulli-Gaussian case, the inference problems tend to be easier, since inverse-gamma assumption leads to more smooth distributions/objective functions.

In this paper, we aim to provide empirical evidence that the Student-*t* model can be used to provide suitable initial values for the unknown parameters for the Bernoulli-Gaussian model. We are motivated by the fact that the two models are qualitatively very similar but have significantly different behaviors for inference.

2. MODEL AND PROBLEM STATEMENT

We consider the following filtering process

$$y_k = \sum_{i=0}^{L-1} g_i x_{k-i} + \sigma_v v_k, \qquad k = 0, 1, ..., n \qquad (1)$$

where y_k's are the noisy observations. Here, the system impulse response is represented by a $1 \times L$ vector of filter coefficients, $\mathbf{g} \equiv g_{0:L-1}$, and $\sigma_v v_k$ is the observation noise with $\{v_k\}$ being i.i.d Gaussian with zero mean and unity variance, that is, $v_k \sim \mathbf{N}(0, 1)$. We assume that the samples of the input signal $\mathbf{x}$ are independently Gaussian, such as $x_k = \sigma_{w,k} w_k$, where $w_k \sim \mathbf{N}(0, 1)$, and the standard deviations $\sigma_{w,k}$ are random variables. Hence, conditionally $x_k | \sigma_{w,k} \sim \mathbf{N}(0, \sigma_{w,k}^2)$. As the variance is varying with time, the resulting process is non-stationary.

This work is supported by DPT project number DPT 2007K120610, and TUBITAK project number 107E050.

The key point here is the choice of the prior distribution of $\sigma_{w,k}$, so that $\mathbf{x}$ is a sparse process. In the sequel, we will introduce two simple choices, that are qualitatively similar but have significantly different behaviors for inference.

Model 1: $\sigma_{w,k} = \sigma_{q_k}$, where q_k is Bernoulli. It is a common approach in the literature to condition $\sigma_{w,k}$ on an *indicator* (Bernoulli random variable) $q_k \in \{0, 1\}$ along with success probability $\lambda = p(q_k = 1)$, such that $\sigma_{w,k} = \sigma_{q_k}$ [2, 3]. Taking $\sigma_1 \gg \sigma_0$ and assigning a small value for λ, one can obtain a sparse and impulsive trace. Here, the indicators q_k can be used to indicate the locations of spikes. In many applications, such as seismology [2], typically correct estimation of the spike locations is the primary concern, and this parametrization provides an intuitive semantic interpretation.

Model 2: $\sigma^2_{w,k}$ is inverse-Gamma. An alternative to the first approach is to assign an inverse-Gamma distribution for the variables $\sigma^2_{w,k}$. That is, $\sigma^2_{w,k} \sim \mathcal{IG}(\alpha, \beta)$. The reason for choosing an inverse-Gamma distribution for $\sigma^2_{w,k}$ is analytical convenience leading to a relatively easier objective function than *Model 1* and efficient inference algorithms [5].

Having introduced the models, our problem can be defined: Given the observation data $\mathbf{y}$, we wish to estimate the latent signal $\mathbf{x}$ and the locations of spikes in $\mathbf{x}$, as well as the filter coefficients $\mathbf{g}$, and possibly the model parameters Θ. For *Model 1*, the parameters are denoted by $\Theta_1 = \{\sigma^2_v, \sigma_0, \sigma_1, \lambda\}$ and for *Model 2*, $\Theta_2 = \{\sigma^2_v, \alpha, \beta\}$.

3. STATE INFERENCE AND PARAMETER ESTIMATION

3.1. *Model 1*

It can easily be seen that, *Model 1* is equivalent to a conditionally Gaussian linear state-space model (CGLSSM) [7], with indicators q_k. So we can write

$$\mathbf{x}_{k+1} = \mathbf{H}\mathbf{x}_k + \mathbf{U}_{q_{k+1}} w_{k+1} \tag{2}$$

$$y_k = \mathbf{g}\mathbf{x}_k + \sigma_v v_k \tag{3}$$

where

$$\mathbf{x}_k = \begin{bmatrix} x_k & x_{k-1} & \dots & x_{k-L+1} \end{bmatrix}^\top \tag{4}$$

and $\mathbf{H}$ is the one-tap delay matrix of size $L \times L$, and $\mathbf{U}_{q_k} = \begin{bmatrix} \sigma_{q_k} & 0 & \dots & 0 \end{bmatrix}^\top$.

3.1.1. *State inference*

The joint posterior distribution of $\mathbf{x}$ and $\mathbf{q} \equiv q_{0:n}$ is

$$p(\mathbf{x}, \mathbf{q}|\mathbf{y}, \mathbf{g}, \Theta) = p(\mathbf{q}|\mathbf{y}, \mathbf{g}, \Theta) p(\mathbf{x}|\mathbf{q}, \mathbf{y}, \mathbf{g}, \Theta). \tag{5}$$

The factors on the right hand side of (5) can be evaluated separately: The first factor can be written as

$$p(\mathbf{q}|\mathbf{y}, \mathbf{g}, \Theta) \propto p(\mathbf{y}|\mathbf{q}, \mathbf{g}, \Theta) p(\mathbf{q}) \tag{6}$$

where the conditional density $p(\mathbf{y}|\mathbf{q}, \mathbf{g}, \Theta)$, which is the *likelihood* of the data, can easily be found by running the Kalman filter [7].

Evaluation of the second factor in (5) is straightforward, too. Given the indicators $\mathbf{q}$ and $\mathbf{y}$, each $\mathbf{x}_k$ is Gaussian with mean $\hat{\mathbf{x}}_{k|n}$ and covariance $\hat{\mathbf{\Sigma}}_{k|n}$, and can be evaluated by using one of the smoothing algorithms for Gaussian linear state-space systems [7].

However, the main difficulty is due to the fact that there are $2^{(n+1)}$ distinct configurations of $\mathbf{q}$. Hence, we can not practically evaluate (5) for every configuration of $\mathbf{q}$, unless n is very small. Therefore the exact joint posterior $p(\mathbf{x}, \mathbf{q}|\mathbf{y}, \mathbf{g}, \Theta)$ is intractable. To overcome this problem, we use two different Markov Chain Monte Carlo (MCMC) sampling methods [8] to draw samples from the joint posterior distribution.

1. Sample indicators, perform exact inference for $\mathbf{x}$: In this type of MCMC sampling, we first estimate the best sequence of indicators $\hat{\mathbf{q}}$ using Gibbs sampling, then use this indicator sequence for exact inference of $\mathbf{x}$. The Gibbs sampler for indicators uses the useful property of CGLSSM's that, when the filtering, and smoothing moments as well as the innovation and innovation covariance at time k for $q_k = c$ are available, the likelihood of observations can be calculated up to a proportionality which does not depend on the value c. In this way, at iteration i and time k, we can sample $q_k^{(i)}$ from the marginal distribution $p(q_k|\mathbf{q}^{(i)}_{0:k-1}, \mathbf{q}^{(i-1)}_{k+1:n}, \mathbf{y})$ [7].

2. Sample both q_k and x_k jointly: In the second MCMC sampling method, we use the fact that, given all the state values other than x_k, denoted by, $\mathbf{x}_{-k}$ one can obtain and sample from the posterior distribution of (x_k, q_k) [9]. Notice that

$$\begin{aligned} p(x_k, q_k|\mathbf{x}_{-k}, \mathbf{y}, \mathbf{g}, \Theta) \\ = p(q_k|\mathbf{x}_{-k}, \mathbf{y}, \mathbf{g}, \Theta) p(x_k|q_k, \mathbf{x}_{-k}, \mathbf{y}, \mathbf{g}, \Theta) \end{aligned} \tag{7}$$

where the factors in (7) can easily be evaluated.

So, the Gibbs sampling algorithm for joint simulation of q_k and x_k for N iterations is as follows [2]:

For $i = 1, ..., N$: for $k = -L + 1, ..., n$:

- Sample $q_k^{(i)}$ from the distribution $p(q_k|\mathbf{x}^{(i-1)}_{-k}, \mathbf{y}, \mathbf{g}, \Theta)$.
- Sample $x_k^{(i)}$ from $u \sim p(x_k|q_k^{(i)}, \mathbf{x}_{-k}, \mathbf{y}, \mathbf{g}, \Theta))$.

The first algorithm is computationally more demanding as a subset of variables is integrated out analytically. In the literature, it is advised that analytical computation should be favored as the associated Markov Chain tends to have superior mixing properties making it worth to do the extra computation [10]. We provide simulation results to test this claim in the context of sparse deconvolution.

3.1.2. *Parameter Estimation*

For parameter estimation, a common approach is to use *Expectation Maximization* (EM) algorithm [7]. In our case, we find the ML estimates for $\mathbf{g}$ and Θ in the maximization step. In EM algorithm, this corresponds to computing the quantity

$$Q^{(i)} = \mathrm{E}\left[\log p(\mathbf{y}, \mathbf{x}, \mathbf{q}|\mathbf{g}; \Theta)|\mathbf{y}, \mathbf{g}^{(i-1)}; \Theta^{(i-1)}\right] \quad (8)$$

at the expectation stage of the i^{th} iteration. The *maximum likelihood* (ML) estimates for $\mathbf{g}$ and Θ are the ones those maximize (8)

$$(\mathbf{g}^{(i)}, \Theta^{(i)}) = \arg\max_{\mathbf{g},\Theta} Q^{(i)}. \quad (9)$$

We decompose the factor in the logarithm as

$$p(\mathbf{y}, \mathbf{x}, \mathbf{q}|\mathbf{g}; \Theta) = p(\mathbf{q})p(\mathbf{x}|\mathbf{q})p(\mathbf{y}|\mathbf{x}, \mathbf{q}, \mathbf{g}; \Theta). \quad (10)$$

Taking the expectation of the logarithm of (10), we obtain the ML solutions for $\hat{\mathbf{g}}^{(i)}$, $\hat{\sigma}_v^{2(i)}$, $\hat{\sigma}_q^{2(i)}$, and $\hat{\lambda}^{(i)}$ [2]. However, this requires sufficient statistics obtained by taking expectations under the posterior distribution $p(\mathbf{x}, \mathbf{q}|\mathbf{y}, \mathbf{g}^{(i-1)}, \Theta^{(i-1)})$, which is intractable. Therefore, we apply *Monte-Carlo* (MC) methods [7] which are based on averaging over simulations to approximate the values of concern.

When sufficient statistics are estimated using Monte Carlo, several EM techniques can be used to optimize over the parameters. Below, we will mention *Monte Carlo EM* (MCEM), *Stochastic EM* (SEM), and *Stochastic Approximation EM* (SAEM). The general expression of SEM and MCEM algorithms is as follows [7]:

Given an initial parameter set $(\mathbf{g}^{(0)}, \Theta^{(0)})$, do, for i = 1,2,...

- *Simulation:* Draw m_i samples for $(\mathbf{x}, \mathbf{q})$, $(\hat{\mathbf{x}}^{i,1}, \hat{\mathbf{q}}^{i,1}), ..., (\hat{\mathbf{x}}^{i,m_i}, \hat{\mathbf{q}}^{i,m_i})$ from $p(\mathbf{x}, \mathbf{q}|\mathbf{y}, \mathbf{g}^{(i-1)}, \Theta^{(i-1)})$.
- *Maximization:* Compute $\mathbf{g}^{(i)}$ and $\Theta^{(i)}$ that maximizes the function $\hat{Q}^{(i)}$, where

$$\hat{Q}^{(i)} = (1-\nu_i)\hat{Q}^{(i-1)} \quad (11)$$
$$+ \nu_i\{\frac{1}{m_i}\sum_{j=1}^{m_i} \log p(\mathbf{y}, \hat{\mathbf{x}}^{i,j}, \hat{\mathbf{q}}^{i,j}|\mathbf{g}^{(i-1)}; \Theta^{(i-1)})\}.$$

In (11), if $\{\nu_i\}_{i\geq 1} > 0$ and $\nu_1 \neq 1$, we have the SAEM algorithm. If $\nu_1 = 1$ and m_i increases with time , then the algorithm is called MCEM. If $\nu_1 = 1$ and m_i is constant, we have the SEM algorithm.

3.2. *Model 2*

Recall that, in *Model 2*, it is assumed that the variance of x_k, $\sigma_{w,k}^2$ is an inverse-Gamma random variable with $\sigma_{w,k}^2 \sim \mathcal{IG}(\alpha, \beta)$. It is important to note that given the parameters $\mathbf{g}$, σ_v^2, and $\sigma_{w,k}^2$ for $k = 0, ..., n$, we can make exact inference for $\mathbf{x}$. Moreover, given $\mathbf{x}$, we can find the *maximum a posteriori* (MAP) estimate of the variances $\sigma_{w,k}^2$. If we denote $\Theta = (\sigma_v^2, \alpha, \beta)$, the EM steps for the current model are

- E-step: Find the posterior distribution

$$p(\hat{\mathbf{x}}^i|\mathbf{y}, \mathbf{g}^{(i-1)}, \Theta^{(i-1)}, \{\sigma_{w,k}^{2(i-1)}\})$$

- M-step: Calculate the parameters $(\mathbf{g}^{(i)}, \Theta^{(i)})$, and $\{\sigma_{w,k}^{2(i)}\}$ that maximize the EM quantity $Q^{(i)}$

$$(\mathbf{g}^{(i)}, \Theta^{(i)}, \{\sigma_{w,k}^{2(i)}\}) = \arg\max_{\mathbf{g},\Theta} Q^{(i)} \quad (12)$$

where

$$Q^{(i)} = \mathrm{E}\left[\log p(\mathbf{y}, \mathbf{x}, \{\sigma_{w,k}^2\}|\mathbf{g}, \Theta)|\mathbf{y}, \mathbf{g}^{(i-1)}, \Theta^{(i-1)}\right]. \quad (13)$$

Similar to what we have done for the previous model, we can decompose the term in the expectation in (13) such as

$$p(\mathbf{y}, \mathbf{x}, \{\sigma_{w,k}^2\}|\mathbf{g}, \Theta) = p(\mathbf{y}|\mathbf{x}, \mathbf{g}, \Theta)p(\mathbf{x}|\{\sigma_{w,k}^2\})p(\{\sigma_{w,k}^2\})$$

where each factor can be written in a straightforward manner. It can be seen that the solutions for $\mathbf{g}$ and σ_v^2 are the same as the ones in *Model 1*. It can also be derived that updates for $\{\sigma_{w,k}^2\}$ can be performed separately for $k = 0, ..., n$, to obtain [5]

$$\hat{\sigma}_{w,k}^{2(i)} = (\mathrm{E}\left[x_k^2\right] + 2\beta)/(2\alpha + 3). \quad (14)$$

There is no analytical ML solution for the inverse-Gamma parameters α and β, given the estimated variances $\{\sigma_{w,k}^2\}$. However, since the likelihood surface is well-behaved, numerical methods for finding the maximum works sufficiently well.

4. RESULTS AND CONCLUSIONS

We investigate the performance of each state inference technique in section 3.1.1 combined with each of the methods for parameter estimation in section 3.1.2, using the results obtained by averaging over 20 simulations. For each simulation, we generated synthetic sparse input signals of length 200 under the assumptions of *Model 1*. We take $\lambda = 0.03, \sigma_v = 0.1, \sigma_1 = 0.01$, and $\sigma_2 = 1$. The length of $\mathbf{g}$ is taken $L = 10$, and $\mathbf{g}$ is generated randomly from $\mathbf{N}(0, 4I)$. For each simulation, we treat data as generated under the assumptions of *Model 2* and apply the EM algorithm in section 3.2. Resulting estimates for $\mathbf{g}$ is used as the initial estimates for the methods used in *Model 1*. We compared the results when this initialization approach is used with the ones when $\hat{\mathbf{g}}$ is initialized randomly. Table 1 shows for each case the mean squared error (MSE) values for estimates of $\mathbf{x}$ and $\mathbf{g}$, and detection rate of spikes as well as the average number of false alarms for spikes. It is important to note that, there is a scale invariance between $\mathbf{x}$ and $\mathbf{g}$ by the nature of our model formulations. Therefore for an estimate of $\mathbf{x}$, $\hat{\mathbf{x}}$, we calculate MSE for $k\hat{\mathbf{x}}$ and $\hat{\mathbf{g}}/k$ where $k = \sum_{i=0}^{n-1} x_i\hat{x}_i / \sum_{i=0}^{n-1} \hat{x}_i^2$. Figure 1 shows

Table 1. Performances of the methods with and without the assistance of the proposed initialization method. M1,M2: *first and second MCMC methods in 3.1.1*, MSEx/MSEg: *MSE of estimates for* $\mathbf{x}$/$\mathbf{g}$, RI/PI: *random/proposed initialization of* $\mathbf{g}$, D/F:*detection rate/average number of false detections for spikes*

Par.est.	St.inf	$\hat{\mathbf{g}}^{(0)}$	MSEg	MSEx	D	F
MCEM	M1	PI	0.2732	0.0053	0.76	2.40
		RI	3.4699	0.0259	0.69	23.8
	M2	PI	0.2706	0.0053	0.78	2.30
		RI	22.376	0.0261	0.69	23.5
SEM	M1	PI	0.2250	0.0052	0.78	2.40
		RI	0.3722	0.0273	0.68	24.0
	M2	PI	0.2347	0.0052	0.78	2.20
		RI	3.7820	0.0272	0.63	24.2
SAEM	M1	PI	0.2314	0.0052	0.78	2.4
		RI	6.8040	0.0267	0.70	27.4
	M2	PI	0.2368	0.0053	0.80	2.25
		RI	24.561	0.0277	0.60	26.0

for one simulation the true $\mathbf{x}$ and $\mathbf{g}$ as well as their estimates with and without using our initialization method.

As it should be clear from the results, performances of all the methods for *Model 1* strongly depend on the initial value $\hat{\mathbf{g}}^{(0)}$, and the use of *Model 2* to assist *Model 1* gives out significantly better results in every measure that we concerned. We also observe that the MCMC methods for state inference exhibit similar performance for sparse data, so there is not much need to use the first method, which has much more computational complexity. Furthermore, one can notice the scale invariance between $\mathbf{x}$ and $\mathbf{g}$ from the figure.

Concluding, we have a powerful method for real data applications on sparse deconvolution, since the effect of initial value problem is reduced. For example, in seismology, the earth response is found either by the help of big earthquakes, where you can measure the excitation signal [11], or by generating the excitation signal using a wavelet, such as in marine seismic works [2], where you have strong priors. We plan to apply our approach on daily seismological data where the excitation signal is small in power and mostly unknown. However, we still have the identifiability problem because of the scale invariance mentioned above. Therefore, a forward step may be using Bayesian inference techniques integrated with the hybrid method to avoid this problem.

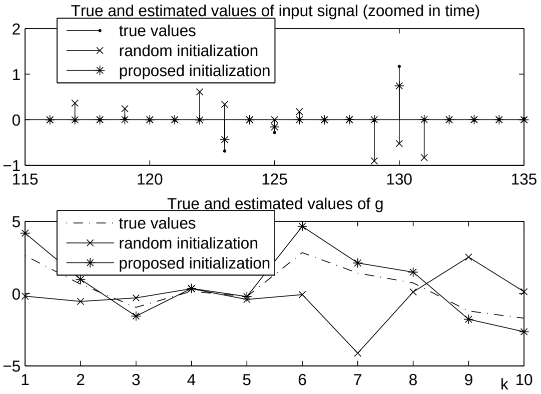

Fig. 1. True and estimated $\mathbf{x}$ and $\mathbf{g}$ with and without proposed initialization method (SEM algorithm with inference method M1 is used)

5. REFERENCES

[1] C. Labat and J. Idier, “Sparse blind deconvolution accounting for time-shift ambiguity,” in *2006 IEEE International Conference*, 2006.

[2] O. Rosec, J. M. Boucher, B. Nsiri, and T. Chonavel, “Blind marine seismic deconvolution using statistical MCMC methods,” *IEEE Oceanic Engineering*, vol. 28, pp. 502–512, 2003.

[3] O. Cappé, A. Doucet, M. Lavielle, and E. Moulines, “Simulation based methods for blind maximum likelihood filter identification,” *Signal Processing*, vol. 73, no. 1, pp. 3–25(23), 1999.

[4] A. Gelman, J.B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, ”Chapman & Hall”, 1995.

[5] A. T. Cemgil, C. Fevotte, and S. J. Godsill, “Variational and stochastic inference for Bayesian source separation,” *Digital Signal Processing*, vol. 17, no. 5, pp. 891–913, 2007.

[6] D. Tzikas, A. Likas, and N. Galatsanos, “Variational bayesian blind image deconvolution with student-t priors,” in *IEEE International Conference*. ICIP, 2007.

[7] O. Cappé, E. Moulines, and T. Rydén, *Inference in Hidden Markov Models*, Springer, 2005.

[8] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, Chapman & Hall/CRC, 1996.

[9] M. Lavielle, “A stochastic algorithm for parametric and non-parametric estimation in the case of incomplete data,” *Signal Processing*, vol. 42, pp. 3–17, 1995.

[10] J. S. Liu, *Monte Carlo strategies in scientific computing*, Springer, 2004.

[11] Y. Özakın, “Crustal structure of Southwestern Anatolia using p-receiver function analysis,” M.S. thesis, Boğaziçi University, 2008.