Research Article

Bayesian Sparse Estimation Using Double Lomax Priors

Xiaojing Gu,¹ Henry Leung,² and Xingsheng Gu¹

¹ East China University of Science and Technology, Meilong Road 130, Shanghai 200237, China

² University of Calgary, 2500 University Drive N.W., Calgary, AB, Canada T2N 1N4

Correspondence should be addressed to Xiaojing Gu; xjing.gu@gmail.com

Received 29 March 2013; Revised 6 July 2013; Accepted 8 July 2013

Academic Editor: Fenghui Yao

Copyright © 2013 Xiaojing Gu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Sparsity-promoting prior along with Bayesian inference is an effective approach in solving sparse linear models (SLMs). In this paper, we first introduce a new sparsity-promoting prior coined as Double Lomax prior, which corresponds to a three-level hierarchical model, and then we derive a full variational Bayesian (VB) inference procedure. When noninformative hyperprior is assumed, we further show that the proposed method has one more latent variable than the canonical automatic relevance determination (ARD). This variable has a smoothing effect on the solution trajectories, thus providing improved convergence performance. The effectiveness of the proposed method is demonstrated by numerical simulations including autoregressive (AR) model identification and compressive sensing (CS) problems.

1. Introduction

Consider the following linear model:

$$\mathbf{y} = \Phi \mathbf{x} + \mathbf{e}, \quad (1)$$

where $\Phi \in \mathbb{R}^{N \times M}$ is the designed matrix and $\mathbf{y} = [y_1, \dots, y_N]^T$ is the measurement vector. $\mathbf{e}$ is the measurement noise, and $\mathbf{x} = [x_1, \dots, x_M]^T$ is the unknown coefficient vector. This linear model is called underdetermined if $M \leq N$ and overdetermined otherwise. In the underdetermined case, there are in general numerous solutions, and additional assumptions need to be assumed on $\mathbf{x}$ for getting the desired one. In the applications we consider, the sparsity of $\mathbf{x}$ is an important assumption: most elements of $\mathbf{x}$ should be tiny or equal to zero when they are not required to tell the data, but few should be allowed to be large if necessary. A linear model with such an assumption on $\mathbf{x}$ is referred to the sparse linear model (SLM). SLMs can represent various types of signals in different transform domains [1–5]. Obtaining such a representation is equivalent to solving for a linear inverse problem with sparsity-promoting regularization. It is a fundamental issue of model (basis) selection [5, 6], compressive sensing [7, 8], source localization [9, 10], and data fusion [11, 12].

Under the sparsity assumption, one seeks for a representation of $\mathbf{y}$ with a minimal number of basis vectors in Φ . This

is equivalent to seeking for the sparsest solution of $\mathbf{x}$, which has a maximal number of elements equal to zero. The sparsest solution can be induced by the improper ℓ_0 -norm of $\mathbf{x}$ [13]: $\|\mathbf{x}\|_0 \triangleq \sum_m 1_{\{x_m \neq 0\}}$. The associated *maximum a posteriori* (MAP) estimation (assuming that $\mathbf{e}$ is white Gaussian noise) is a penalized least squares problem:

$$\min_{\mathbf{x}} \|\mathbf{y} - \Phi \mathbf{x}\|_2^2 + \rho \|\mathbf{x}\|_0, \quad (2)$$

where ρ is the model parameter that controls the relative importance applied to error term and sparseness term. The challenge in solving for (2) is that it requires a combinatorial search which is NP-hard in general. A natural approach is to replace $\|\cdot\|_0$ by a tractable approximation. Among all the tractable approximations, ℓ_1 -norm is the tightest convex relaxation of ℓ_0 -norm; it leads to the successful Lasso algorithm [1, 3, 4, 14–16]. However, a nonconvex relaxation is able to recover sparsity in a more efficient way. It is proved that a nonconvex relaxation that closely resembles the ℓ_0 -norm is able to find the correct solution with fewer measurements [17–20].

One difficulty with nonconvex relaxation is the existence of many local minima. Point-estimation methods such as MAP may easily get stuck in poor local solutions [21]. A posterior-estimation Bayesian approach can nicely alleviate this problem. Given the sparsity-promoting prior $p(\mathbf{x})$, which

is the probabilistic representation of a particular ℓ_0 -norm relaxation, the Bayesian inference can be expressed by

$$p(\mathbf{x} | \mathbf{y}) = \frac{p(\mathbf{y} | \mathbf{x}) p(\mathbf{x})}{\int p(\mathbf{y} | \mathbf{x}) p(\mathbf{x}) d\mathbf{x}}, \quad (3)$$

where $p(\mathbf{x} | \mathbf{y})$ is the posterior. The Bayesian approach computes the posterior mean $E[\mathbf{x} | \mathbf{y}] = \int \mathbf{x} p(\mathbf{x} | \mathbf{y}) d\mathbf{x}$ instead of its mode, and it integrates instead of maximizing over $p(\mathbf{x} | \mathbf{y})$. In case of many local minima, the mass of the probability density tends to concentrate at a strong minimum, which is more likely to be the global optimum. Although $E[\mathbf{x} | \mathbf{y}]$ is not necessarily the exact global optimum in general, it is proved that [22], by taking a “zero temperature limit,” $E[\mathbf{x} | \mathbf{y}]$ is able to converge to the global optimum. In other words, as $-\ln p(\mathbf{x})$ approaches the ℓ_0 -norm, $E[\mathbf{x} | \mathbf{y}]$ converges to the ℓ_0 solution.

In Bayesian formulation, the choice of prior $p(\mathbf{x})$ is important for both solution sparsity and computational tractability. The Student- t prior is a relatively standard prior that stands out by being non-log-concave (results in nonconvex penalty) and having a Gaussian-integral representation for efficient processing. With $\mathcal{S}(\cdot)$ denoting the Student- t prior, the distribution is defined as

$$\mathcal{S}(x | 0, \eta, f) = \frac{\Gamma((f+1)/2)}{\Gamma(f/2)} \left(\frac{\eta}{\pi f} \right)^{1/2} \left[1 + \frac{\eta}{f} x^2 \right]^{-(f+1)/2}, \quad \eta > 0, \quad f > 0, \quad (4)$$

where f controls the degree of sparsity enforced and is known as shape parameter and η is the scale parameter. Additionally, denoting the Gaussian prior with variance τ by $\mathcal{N}(x | 0, \tau)$ and the Gamma prior by $\mathcal{G}$, defined as $\mathcal{G}(z | a, b) = (b^a / \Gamma(a)) z^{a-1} \exp(-bz)$, the Student- t prior corresponds to a Gaussian integral as follows:

$$\mathcal{S}(x | 0, \eta, f) = \int_0^\infty \mathcal{N}(x | 0, (\eta\xi)^{-1}) \mathcal{G}\left(\xi | \frac{f}{2}, \frac{f}{2}\right) d\xi, \quad \eta > 0, \quad f > 0, \quad (5)$$

where ξ is the latent variable modifying the precision of the Gaussian mixture. The representation in (5) can be treated as an integral over scale parameter of the Gaussian density, thus is referred to as Gaussian scale mixtures (GSM) [23], which naturally invokes a two-level hierarchical Bayes model.

Specifically, setting $f \rightarrow +0$ leads to the noninformative-hyperpriors case of Student- t prior and gives the well known automatic relevance determination (ARD) prior (the scale parameter η is also assumed to be unity) [24] as follows:

$$\text{ARD}(x) = \mathcal{S}(x | 0, 1, f), \quad f \rightarrow 0^+. \quad (6)$$

As the limited case, the ARD prior provides the strongest sparsity-promoting potential among all Student- t priors with different values of f (see the remark of Proposition 1), and it has drawn tremendous attention in many problems [5, 8, 25].

Several methods have been proposed to solve for SLMs incorporating ARD priors, including variational Bayes (VB) inference [25], type II maximum likelihood (the “evidence framework”) [26], and a series of reweighted ℓ_1 problems [27].

In this paper, we formulate the SLM problem from a Bayesian perspective. Despite of the two-level Student- t prior, we propose the use of a new non-log-concave sparsity prior referred to as Double Lomax prior, which corresponds to a three-level hierarchical Bayes model. Additionally, a full VB inference procedure is derived to solve for the SLM using Double Lomax priors. As shown here, when noninformative hyperpriors are assumed ($f \rightarrow +0$), the updating procedure of proposed method includes the canonical ARD updating procedure as a special case, but it has a better global convergence performance that can result in improved sparse estimations.

The organization of this paper is as follows. In Section 2, we present the proposed Double Lomax prior and demonstrate its properties. In Section 3, we formulate the Bayesian model of SLM using Double Lomax priors based on a generalized noise model. In Section 4, a full VB inference procedure is developed. In Section 5, we discuss the relationship between the proposed method and the ARD approach. Experimental results are presented in Section 6. Two SLM applications with different Φ matrix properties are considered in this section. One application indicates order determination and coefficients estimation for autoregressive (AR) models, where the rows of Φ matrix are strongly correlated. The other application is sparse signal recovery from the standard compressive sensing (CS) setup, where the rows of Φ matrix are uncorrelated. Conclusions are given in Section 7.

2. Double Lomax Prior and Its Properties

The proposed Double Lomax prior can be derived from two back-to-back spliced Lomax distributions (also known as Pareto II distribution [28]). That is,

$$\mathcal{DL}(x | 0, \eta, f) = \frac{\eta}{2} \left(1 + \frac{\eta|x|}{f} \right)^{-(f+1)}, \quad \eta > 0, \quad f > 0. \quad (7)$$

It can be shown that $\mathcal{DL}(x | 0, \eta, f) > 0$, $\int_{-\infty}^{+\infty} \mathcal{DL}(x | 0, \eta, f) = 1$, and $\mathcal{DL}(-x | 0, \eta, f) = \mathcal{DL}(x | 0, \eta, f)$. Thus (7) is a probability distribution function of x defined on $(-\infty, +\infty)$, symmetrical about zero. η is the scale parameter, f is the shape parameter, and 0 refers to zero mean.

Moreover, note that

$$\begin{aligned} \lim_{f \rightarrow +\infty} \mathcal{DL}(x | 0, \eta, f) &= \frac{\eta}{2} \exp(-\eta|x|), \\ \lim_{f \rightarrow +0} \mathcal{DL}(x | 0, \eta, f) &= \begin{cases} 0, & x \neq 0, \\ \frac{\eta}{2}, & x = 0. \end{cases} \end{aligned} \quad (8)$$

Thus when the shape parameter goes to infinite, Double Lomax prior asymptotically becomes the Laplace prior with

inverse scale η ; when the shape parameter goes to zero, Double Lomax prior tends to strongly concentrate its mass on the zero value, distributing with the same summit but much longer and thinner tails. This behavior can be observed in Figure 1, which shows the Double Lomax priors with different values of f .

Proposition 1. Double Lomax prior in (7) is log-convex.

Proof. The logarithm of density function equation (7) is

$$\ln \mathcal{DL}(x | 0, \eta, f) = \ln\left(\frac{\eta}{2}\right) - (f+1) \ln\left(1 + \frac{\eta|x|}{f}\right). \quad (9)$$

It is continuous on $(-\infty, +\infty)$, and the derivatives are

$$\begin{aligned} \frac{d}{dx} \ln \mathcal{DL}(x | 0, \eta, f) &= -(f+1) \frac{\eta}{f + \eta|x|}, \quad x \neq 0, \\ \frac{d^2}{dx^2} \ln \mathcal{DL}(x | 0, \eta, f) &= (f+1) \frac{\eta^2}{(f + \eta|x|)^2} \begin{cases} > 0, & f \rightarrow +\infty, \quad x \neq 0, \\ = 0, & f \rightarrow +0, \quad x \neq 0. \end{cases} \end{aligned} \quad (10)$$

Thus, Double Lomax prior is log-convex for any $f > 0$ and is strictly log-convex for $f \rightarrow +\infty$, even though it does not have first and second derivative at the point $x = 0$. $\square$

Remark 2. Proposition 1 shows that the implicit penalty induced by Double Lomax prior, $-\ln \mathcal{DL}(x | 0, \eta, f)$, is (strictly) concave, spanning from the convex ℓ_1 -norm penalty (when $f \rightarrow +\infty$) to the strongly concave ℓ_0 -norm-like penalty (when $f \rightarrow +0$). Figure 2 is plotted $-\ln \mathcal{DL}(x | 0, \eta, f)$ for different values of f .

For Student- t prior in (4), $(d^2/dx^2) \ln \mathcal{S}(x | 0, \eta, f) = \eta(f+1)(\eta x^2 - f)/(f + \eta x^2)^2$, thus, the implicit penalty induced by Student- t prior, $-\ln \mathcal{S}(x | 0, \eta, f)$, has inconsistent concavity, that is, convex on the central interval $[-\sqrt{f/\eta}, \sqrt{f/\eta}]$ and concave on each of the outer intervals, $(-\infty, -\sqrt{f/\eta})$ and $(\sqrt{f/\eta}, +\infty)$. When $f \rightarrow +\infty$, the outer concave intervals are vanished, and the Student- t prior becomes the Gaussian prior with precision η and leads to the ℓ_2 -norm penalty that does not induce sparsity at all; when $f \rightarrow +0$, the central convex interval is vanished, and the Student- t prior is now referred to as ARD prior, which becomes proportional to the noninformative Jeffreys prior $p(x) \propto 1/|x|$ and leads to the log-sum penalty $\sum \log(|x|)$, providing similar sparsity-promoting effect as the Double Lomax prior does.

Proposition 3. Double Lomax prior in (7) can be represented as a GSM. Given Gaussian prior $\mathcal{N}(\cdot)$, exponential prior $\mathcal{E}(\cdot)$, and Gamma prior $\mathcal{G}(\cdot)$, one has

$$\begin{aligned} \mathcal{DL}(x | 0, \eta, f) &= \iint_0^\infty \mathcal{N}(x | 0, \alpha) \mathcal{E}\left(\alpha \mid \frac{\eta^2 v^2}{2}\right) \mathcal{G}(v | f, f) d\alpha dv. \end{aligned} \quad (11)$$

FIGURE 1: Double Lomax prior at zero position, with $\eta = 1$ for various values of f .

Proof. Equation (7) can be expressed as

$$\begin{aligned} \mathcal{DL}(x | 0, \eta, f) &= \frac{\eta}{2} f^{f+1} (f + \eta|x|)^{-(f+1)} \\ &= \frac{\Gamma(f+1) \eta}{\Gamma(f)} \frac{f^f}{2} (f + \eta|x|)^{-(f+1)} \\ &= \frac{\eta f^f}{2\Gamma(f)} \int_0^\infty v^f \exp\{-v(f + \eta|x|)\} dv \\ &= \int_0^\infty \frac{\eta v}{2} \exp\{-\eta v|x|\} \frac{f^f}{\Gamma(f)} v^{f-1} \exp\{-fv\} dv \\ &= \int_0^\infty \mathcal{L}(x | 0, \eta v) \mathcal{G}(v | f, f) dv. \end{aligned} \quad (12)$$

Meanwhile, note that the Laplace prior has a hierarchical representation as follows:

$$\begin{aligned} \mathcal{L}(x | 0, \lambda) &= \int_0^\infty \mathcal{N}(x | 0, \alpha) \mathcal{E}\left(\alpha \mid \frac{\lambda^2}{2}\right) d\alpha \\ &= \frac{\lambda}{2} \exp(-\lambda|x|), \end{aligned} \quad (13)$$

where $\mathcal{E}(z | \tau) = \tau \exp(-\tau z)$ is an exponential prior. Substituting (13) into (12) gives (11). This proves Proposition 3. $\square$

Remark 4. Equation (11) corresponds to a three-level hierarchical Bayes model with two latent variables α and v . On the first level, random variables are generated from zero-mean Gaussian priors with independent variance α . On the second level, a set of α is generated from exponential priors with independent inverse scales, which are formed by $(1/2)\eta^2$ multiplying a set of independent variables v^2 . On the third level, the positive squares of independent parameters v are generated from a Gamma prior with two parameters of the same value.

Moreover, besides this GSM model, Double Lomax prior can be considered as a Laplacian scaled mixture as well, and the inverse scale v_m can be thought as the width-controller of the corresponding Laplace prior: if v_m is large, the

FIGURE 2: Examples of the negative logarithm of Double Lomax prior for different values of f , normalized by $\ln(1 + \eta/f)$. The ℓ_1 -norm and ℓ_0 -norm are included for comparison.

corresponding distribution is narrow and the regularizing effect towards zero is strong.

3. Sparse Linear Model Using Double Lomax Priors

We assume independent Double Lomax priors over the coefficients and use the model in (1). Without loss of generality, the scale parameter η is assumed to be unity. Using (11), the coefficient model of (1) can be written as

$$p(\mathbf{x} | \boldsymbol{\alpha}) = \prod_{m=1}^M \mathcal{N}(x_m | 0, \alpha_m) = \mathcal{N}(\mathbf{x} | \mathbf{0}, \boldsymbol{\Lambda}^{-1}), \quad (14)$$

$$p(\alpha_m | v_m) = \mathcal{E}\left(\alpha_m | \frac{v_m^2}{2}\right), \quad (15)$$

$$p(v_m | f) = \mathcal{E}(v_m | f, f), \quad (16)$$

where α_m is the latent variable modifying the variance of the GSM model, v_m is the latent variable modifying the inverse scale of exponential prior, and $\boldsymbol{\Lambda} = \text{diag}(\alpha_m^{-1})$ denotes the precision matrix.

Traditionally the noise of SLM is presumed to be Gaussian distributed, which means that the measurements are also Gaussian distributed. However, in many real applications, measurements are distributed with tails that decay more slowly than Gaussian, which can not be described well by Gaussian noise model. Here we allow for very slow decay of tails by modeling the noise model using Student- t distribution. As a generalized noise model, Student- t can model a wide range of noises which from Gaussian to very heavy tailed. It also makes a connection of the robust statistic and the sparsity priors, under the view of the capability of modeling extreme events: those variables who have small probabilities but a large value range.

Using a hierarchical Bayes model which is similar to (5), denoting the shape parameter by d (also called degrees of freedom), the scale parameter by β (also called precision parameter), and the latent variable by δ_n , and imposing two

FIGURE 3: Directed graph representing the Bayesian modeling.

additional Gamma priors on β and d , the hierarchical form of Student- t distributed data can be written as

$$p(\mathbf{y} | \mathbf{x}, \beta, \boldsymbol{\delta}) = \prod_{n=1}^N \mathcal{N}(y_n | \phi_n \mathbf{x}, (\beta \delta_n)^{-1}) \quad (17)$$

$$= \mathcal{N}(\mathbf{y} | \boldsymbol{\Phi} \mathbf{x}, \beta^{-1} \mathbf{L}^{-1}),$$

$$p(\delta_n | d) = \mathcal{E}\left(\delta_n | \frac{d}{2}, \frac{d}{2}\right), \quad (18)$$

$$p(\beta) = \mathcal{E}(\beta | a_\beta, b_\beta), \quad (19)$$

$$p(d) = \mathcal{E}(d | a_d, b_d), \quad (20)$$

where ϕ_n is the n th row of $\boldsymbol{\Phi}$ and $\mathbf{L} = \text{diag}(\delta_n)$ is the precision matrix.

Figure 3 shows the complete Bayesian model and the interdependencies among the model parameters. All hyperparameters in this graph are set to small values for getting flat hyperpriors (e.g., $a_\beta = b_\beta = a_d = b_d = f = 10^{-4}$). It can be observed from the graph that the joint distribution of the observations and parameters $\boldsymbol{\Theta} = \{\beta, \boldsymbol{\delta}, d, \mathbf{x}, \boldsymbol{\alpha}, \mathbf{v}\}$ can be expressed as

$$\begin{aligned} p(\mathbf{y}, \boldsymbol{\Theta}) &= p(\mathbf{y}, \beta, \boldsymbol{\delta}, d, \mathbf{x}, \boldsymbol{\alpha}, \mathbf{v}) \\ &= p(\mathbf{y} | \mathbf{x}, \beta, \boldsymbol{\delta}) p(\mathbf{x} | \boldsymbol{\alpha}) p(\boldsymbol{\alpha} | \mathbf{v}) \\ &\quad \times p(\mathbf{v} | f) p(\beta) p(\boldsymbol{\delta} | d) p(d), \end{aligned} \quad (21)$$

where the $p(\mathbf{y} | \mathbf{x}, \beta, \boldsymbol{\delta})$, $p(\mathbf{x} | \boldsymbol{\alpha})$, $p(\boldsymbol{\alpha} | \mathbf{v})$, $p(\mathbf{v} | f)$, $p(\beta)$, $p(\boldsymbol{\delta} | d)$ and $p(d)$ are given in (17), (14), (15), (16), (19), (18), and (20), respectively.

4. Variational Bayesian Sparse Regression

It is usually difficult to find the analytical results of the integrals in Bayesian estimation, hence approximation methods are required. There are two major classes of the approximations. One is Markov chain Monte Carlo (MCMC), which represents the posterior by samples; unbiased results can be obtained in infinite time. However, MCMC is relatively time consuming [29]. The other is variational relaxations [22], which relaxes the Bayesian estimation into a series of feasible optimization problems, providing a good tradeoff between

computation accuracy and efficiency. VB inference [30–32] is one approach of this class.

VB approximates the posterior density $p(\Theta | y)$ by introducing a factorable distribution $q(\Theta)$. Using Jensen's inequality, we get the lower bound of the log probability of $p(y)$:

$$\ln p(y) \geq \int q(\Theta) \ln \frac{p(y, \Theta)}{q(\Theta)} d\Theta. \quad (22)$$

In this work, we factorize the approximation distribution $q(\Theta)$ over groups of parameters Θ_i as follows:

$$q(\Theta) = \prod_i q(\Theta_i) = q(x) q(\alpha) q(v) q(\beta) q(\delta) q(d). \quad (23)$$

Maximizing the right-hand side of (22) with respect to $q(\Theta_i)$, we can get a general expression of the optimal solution $\hat{q}(\Theta_i)$,

$$\hat{q}(\Theta_i) = \frac{\exp \langle \ln p(y, \Theta) \rangle_{k \neq i}}{\int \exp \langle \ln p(y, \Theta) \rangle_{k \neq i} d\Theta_i}, \quad (24)$$

where $\langle \cdot \rangle_{k \neq i}$ denotes the expectation operator with respect to the distributions $q(\Theta_k)$ for all $k \neq i$. $\int \exp \langle \ln p(y, \Theta) \rangle_{k \neq i} d\Theta_i$ is the integration over Θ_i which is a normalization constant. If the conjugate prior exists and is chosen for each group, the approximate posterior has the same functional form as the prior [30]. Hence the normalization constant does not need to be considered.

Now we derive a full VB inference procedure for the above problem. We use the general optimal solution in (24) to obtain the factorized approximate posterior densities $q(\Theta_i)$. However, different from most VB updates in the literature, the challenge here is, for several random variables in the proposed model, that there are no conjugate priors for their likelihoods, and hence, in order to obtain an iterative algorithm with closed form, extra integration analysis has to be incorporated in our derivation.

4.1. Coefficients, x . To compute $q(x)$, we only need to consider those terms that are functionally dependent on x . Because $p(x | \alpha)$ is the conjugate prior for x , $\hat{q}$ would still be a Gaussian. It is convenient to ignore the normalization constant. The optimal solution is given by

$$\begin{aligned} \hat{q}(x) &\propto \exp \langle \ln p(y | x, \beta, \delta) + \ln p(x | \alpha) \rangle_{\beta, \delta, \alpha} \\ &\propto \exp \langle \ln \mathcal{N}(y | \Phi x, \beta^{-1} L^{-1}) + \ln \mathcal{N}(x | 0, \Lambda^{-1}) \rangle_{\beta, \delta, \alpha} \\ &\propto \mathcal{N}(y | \Phi x, \langle \beta \rangle^{-1} \langle L \rangle^{-1}) \mathcal{N}(x | 0, \langle \Lambda \rangle^{-1}) \\ &= \mathcal{N}(x | \mu_x, \Sigma_x), \end{aligned} \quad (25)$$

where $\langle L \rangle = \text{diag}(\langle \delta_n \rangle)$ and $\langle \Lambda \rangle = \text{diag}(\langle \alpha_m^{-1} \rangle)$. Comparing the square term with standard Gaussian function, we can identify the mean and variance of this Gaussian, and get

$$\begin{aligned} \mu_x &= \langle \beta \rangle \Sigma_x \Phi^T \langle L \rangle y, \\ \Sigma_x &= (\langle \beta \rangle \Phi^T \langle L \rangle \Phi + \langle \Lambda \rangle)^{-1}. \end{aligned} \quad (26)$$

4.2. Coefficient's Latent Variables, α . From (14) and (15) we observe that the terms involving α are composed of M -order multiplicative terms over α_m , x_m , and v_m , a further factorization of this approximate posterior density becomes

$$\begin{aligned} q(\alpha) &= \prod_{m=1}^M q(\alpha_m), \\ \hat{q}(\alpha_m) &\propto \exp \langle \ln p(x_m | \alpha_m) + \ln p(v_m | \alpha_m) \rangle_{x_m, v_m} \\ &\propto \exp \left\langle \ln \mathcal{N}(x_m | 0, \alpha_m) + \ln \mathcal{E} \left(\alpha_m | \frac{v_m^2}{2} \right) \right\rangle_{x_m, v_m} \\ &\propto \mathcal{N} \left(\sqrt{\langle x_m^2 \rangle} | 0, \alpha_m \right) \mathcal{E} \left(\alpha_m | \frac{\langle v_m^2 \rangle}{2} \right), \end{aligned} \quad (27)$$

where

$$\langle x_m^2 \rangle = \mu_{x(m)}^2 + \Sigma_{x(mm)}. \quad (28)$$

$\cdot_{(m)}$ and $\cdot_{(mm)}$ denote the m th element of a vector and the m th diagonal element of a matrix, respectively.

However, since exponential distribution is not the conjugate prior for the variance of Gaussian distribution, the normalization term must be reinstated. Integrating over α_m , the normalization constant of this posteriori becomes

$$\begin{aligned} \mathcal{E}_{\hat{q}(\alpha_m)} &= \int_0^\infty \mathcal{N} \left(\sqrt{\langle x_m^2 \rangle} | 0, \alpha_m \right) \mathcal{E} \left(\alpha_m | \frac{\langle v_m^2 \rangle}{2} \right) d\alpha_m \\ &= \mathcal{L} \left(\sqrt{\langle x_m^2 \rangle} | 0, \sqrt{\langle v_m^2 \rangle} \right). \end{aligned} \quad (29)$$

Therefore, the complete form of this posteriori is

$$\begin{aligned} \hat{q}(\alpha_m) &= \frac{\mathcal{N} \left(\sqrt{\langle x_m^2 \rangle} | 0, \alpha_m \right) \mathcal{E} \left(\alpha_m | \langle v_m^2 \rangle / 2 \right)}{\mathcal{L} \left(\sqrt{\langle x_m^2 \rangle} | 0, \sqrt{\langle v_m^2 \rangle} \right)} \\ &= \frac{\sqrt{\langle v_m^2 \rangle}}{\sqrt{2\pi} \exp \left(-\sqrt{\langle v_m^2 \rangle} \langle x_m^2 \rangle \right)} \alpha_m^{-1/2} \\ &\quad \times \exp \left(-\frac{1}{2} \left(\frac{\langle x_m^2 \rangle}{\alpha_m} + \langle v_m^2 \rangle \alpha_m \right) \right). \end{aligned} \quad (30)$$

It is noted that the posteriori density in (30) corresponds to a probability function known as generalized inverse Gaussian distribution ($\mathcal{GIG}$), which has the probability function [33]

$$\mathcal{GIG}(z | \omega, \chi, \psi) = \frac{(\psi/\chi)^{\omega/2}}{2K_\omega(\sqrt{\chi\psi})} z^{(\omega-1)} \exp \left(-\frac{1}{2} \left(\frac{\chi}{z} + \psi z \right) \right), \quad (31)$$

where $K_\omega(\cdot)$ is the modified Bessel function of the third kind with index ω , one of its integral representations is given by

$$K_\omega(\tau) = \frac{1}{2} \int_0^\infty \tau^{\omega-1} \exp \left(-\frac{1}{2} \omega \left(\frac{1}{\tau} + \tau \right) \right) d\omega. \quad (32)$$

From (32), we have

$$K_{1/2}(\tau) = \sqrt{\frac{\pi}{2}} \tau^{-1/2} \exp(-\tau). \quad (33)$$

Substituting (33) into (30) and comparing with (31), it can be found that

$$\hat{q}(\alpha_m) = \mathcal{G}\mathcal{F}\mathcal{G}\left(\alpha_m \mid \frac{1}{2}, \langle x_m^2 \rangle, \langle \lambda_x \rangle\right). \quad (34)$$

The r th order moment of a $\mathcal{G}\mathcal{F}\mathcal{G}$ distribution is given by [33]

$$\langle z^r \rangle = \frac{K_{\omega+r}(\sqrt{\chi\psi})}{K_{\omega}(\sqrt{\chi\psi})} \left(\frac{\chi}{\psi}\right)^{r/2}. \quad (35)$$

In addition, from (32), it can be derived that

$$\begin{aligned} K_{\omega}(\tau) &= K_{-\omega}(\tau), \\ K_{\omega+1}(\tau) &= \frac{2\omega}{\tau} K_{\omega}(\tau) + K_{\omega-1}(\tau). \end{aligned} \quad (36)$$

Therefore, for the specific $\mathcal{G}\mathcal{F}\mathcal{G}$ in (34), using (35) with (36), we have:

$$\begin{aligned} \langle \alpha_m^{-1} \rangle &= \sqrt{\frac{\langle v_m^2 \rangle}{\langle x_m^2 \rangle}}, \\ \langle \alpha_m \rangle &= \left(\frac{1}{\sqrt{\langle x_m^2 \rangle \langle v_m^2 \rangle}} + 1 \right) \sqrt{\frac{\langle x_m^2 \rangle}{\langle v_m^2 \rangle}} \\ &= \langle \alpha_m^{-1} \rangle^{-1} + \langle v_m^2 \rangle^{-1}. \end{aligned} \quad (37)$$

4.3. Coefficient's Latent Variables, $\mathbf{v}$. The approximate posterior density of $\mathbf{v}$ can be factorized as

$$q(\mathbf{v}) = \prod_{m=1}^M q(v_m), \quad (38)$$

$$\begin{aligned} \hat{q}(v_m) &\propto \exp(\ln p(\alpha_m \mid v_m) + \ln p(v_m \mid f))_{\alpha_m} \\ &\propto \exp\left\langle \ln \mathcal{G}\left(\alpha_m \mid \frac{v_m^2}{2}\right) + \ln \mathcal{G}(v_m \mid f, f) \right\rangle_{\alpha_m} \\ &\propto \exp\left((f+1) \ln v_m - \frac{\langle \alpha_m \rangle}{2} v_m^2 - f v_m\right). \end{aligned} \quad (39)$$

Equation (39) contains logarithmic, square, and linear terms, thus it is not in correspondence with any standard distribution. Using a definite integration result in [34], the normalization term of $\hat{q}(v_m)$ can be expressed as

$$\begin{aligned} \mathcal{G}_{\hat{q}(v_m)} &= \langle \alpha_m \rangle^{-(f+2)/2} \Gamma(f+2) \\ &\times \exp\left(\frac{f^2}{4\langle \alpha_m \rangle}\right) D_{-(f+2)}\left(\frac{f}{\sqrt{\langle \alpha_m \rangle}}\right), \end{aligned} \quad (40)$$

where $D_{\omega}(\cdot)$ is the parabolic cylinder function, whose computation can be implemented [35]. Thereby,

$$\begin{aligned} \langle v_m^2 \rangle &= \frac{1}{C_{\hat{q}(v_m)}} \int_0^{\infty} v_m^2 \cdot v_m^{f+1} \exp\left(-\frac{\langle \alpha_m \rangle}{2} v_m^2 - f v_m\right) dv_m \\ &= \frac{1}{\langle \alpha_m \rangle} \frac{\Gamma(f+4)}{\Gamma(f+2)} \frac{D_{-(f+4)}\left(f/\sqrt{\langle \alpha_m \rangle}\right)}{D_{-(f+2)}\left(f/\sqrt{\langle \alpha_m \rangle}\right)}. \end{aligned} \quad (41)$$

However, when noninformative hyperprior is assumed, f is set to zero and $\hat{q}(v_m)$ becomes the Rayleigh distribution who has the probability function

$$\mathcal{R}(z \mid \lambda) = \frac{z}{\lambda^2} \exp\left(-\frac{z^2}{2\lambda^2}\right). \quad (42)$$

Here $\hat{q}(v_m)$ is $\mathcal{R}(v_m \mid 1/\sqrt{\langle x_m^2 \rangle}) = \langle \alpha_m \rangle v_m \exp(-(\langle \alpha_m \rangle/2)v_m^2)$. Therefore,

$$\begin{aligned} \langle v_m^2 \rangle &= \int_0^{\infty} v_m^2 \cdot \langle \alpha_m \rangle v_m \exp\left(-\frac{\langle \alpha_m \rangle}{2} v_m^2\right) dv_m \\ &= \langle \alpha_m \rangle \int_0^{\infty} v_m^3 \exp\left(-\frac{\langle \alpha_m \rangle}{2} v_m^2\right) dv_m \\ &= \frac{2}{\langle \alpha_m \rangle}. \end{aligned} \quad (43)$$

Equation (43) indicates that the value of $\langle v_m^2 \rangle$ is simply inversely proportional to $\langle \alpha_m \rangle$ when $f \rightarrow +0$.

4.4. Noise Precision Parameter, β . Since Gamma distribution is conjugate to the precision of Gaussian function, we still obtain a Gamma distribution for β as follows:

$$\hat{q}(\beta) = \mathcal{G}(\beta \mid A_{\beta}, B_{\beta}), \quad (44)$$

where

$$\begin{aligned} A_{\beta} &= \frac{N}{2} + a_{\beta}, \\ B_{\beta} &= \frac{1}{2} \text{trace}(\langle \mathbf{L} \rangle \langle (\mathbf{y} - \Phi \mathbf{x})(\mathbf{y} - \Phi \mathbf{x})^T \rangle) + b_{\beta}, \\ \langle (\mathbf{y} - \Phi \mathbf{x})(\mathbf{y} - \Phi \mathbf{x})^T \rangle &= \mathbf{y} \mathbf{y}^T - 2\Phi \mu_{\mathbf{x}} \mathbf{y}^T \\ &\quad + \Phi (\mu_{\mathbf{x}} \mu_{\mathbf{x}}^T + \Sigma_{\mathbf{x}}) \Phi^T. \end{aligned} \quad (45)$$

4.5. Noise's Latent Variables, δ . It is observed that the approximate posterior density of δ can be further factorized as

$$q(\delta) = \prod_{n=1}^N q(\delta_n). \quad (46)$$

And $\hat{q}(\delta_n)$ is still a Gamma distribution as follows:

$$\hat{q}(\delta_n) = \mathcal{G}(\delta_n \mid A_{\delta}, B_{\delta_n}), \quad (47)$$

where

$$\begin{aligned} A_{\delta} &= \frac{1}{2} + \frac{\langle d \rangle}{2}, \\ B_{\delta_n} &= \frac{\langle \beta \rangle}{2} \langle (y_n - \phi_n \mathbf{x})^2 \rangle + \frac{\langle d \rangle}{2}, \\ \langle (y_n - \phi_n \mathbf{x})^2 \rangle &= y_n^2 - 2y_n \phi_n \boldsymbol{\mu}_x + \phi_n (\boldsymbol{\Sigma}_x + \boldsymbol{\mu}_x \boldsymbol{\mu}_x^T) \phi_n^T. \end{aligned} \quad (48)$$

4.6. Noise's Degrees of Freedom, d . It is noted that

$$\begin{aligned} \hat{q}(d) \propto \exp \left(-N \ln \Gamma \left(\frac{d}{2} \right) + \frac{N}{2} d \ln \frac{d}{2} + (a_d - 1) \ln d \right. \\ \left. + \frac{d}{2} \left(\sum_{n=1}^N \langle \ln \delta_n \rangle - \sum_{n=1}^N \langle \delta_n \rangle - 2b_d \right) \right), \end{aligned} \quad (49)$$

which is not in correspondence with any standard distribution. This problem has been solved by expanding $\ln \Gamma(d/2)$ using Sterling's approximation [6], $\ln \Gamma(d/2) \approx (d/2 - 1/2) \ln(d/2) - d/2$. Thus,

$$\begin{aligned} \hat{q}(d) \propto \exp \left(\left(\frac{N}{2} + a_d - 1 \right) \ln d \right. \\ \left. + \frac{1}{2} \left(N + \sum_{n=1}^N \langle \ln \delta_n \rangle - \sum_{n=1}^N \langle \delta_n \rangle - 2b_d \right) d \right) \\ = \mathcal{G}(d | A_d, B_d), \end{aligned} \quad (50)$$

where

$$\begin{aligned} A_d &= \frac{N}{2} + a_d, \\ B_d &= -\frac{1}{2} \left(N + \sum_{n=1}^N \langle \ln \delta_n \rangle - \sum_{n=1}^N \langle \delta_n \rangle \right) + b_d, \end{aligned} \quad (51)$$

where

$$\langle \ln \delta_n \rangle = \psi(A_{\delta}) - \ln(B_{\delta_n}), \quad (52)$$

$\psi(\cdot)$ is the digamma function, defined by $\psi(z) = (d/dz) \ln \Gamma(z)$.

We summarize the VB inference of Double Lomax formulation (noninformative hyperprior) as follows.

For coefficient model, consider

$$\hat{q}(\mathbf{x}) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x), \quad (53)$$

$$\boldsymbol{\mu}_x = \langle \beta \rangle \boldsymbol{\Sigma}_x \boldsymbol{\Phi}^T \langle \mathbf{L} \rangle \mathbf{y}, \quad (54)$$

$$\boldsymbol{\Sigma}_x = (\langle \beta \rangle \boldsymbol{\Phi}^T \langle \mathbf{L} \rangle \boldsymbol{\Phi} + \langle \boldsymbol{\Lambda} \rangle)^{-1}, \quad (55)$$

$$\hat{q}(\alpha_m) = \mathcal{G}\mathcal{F}\mathcal{G} \left(\alpha_m | \frac{1}{2}, \langle x_m^2 \rangle, \langle \lambda_x \rangle \right), \quad (56)$$

$$\langle \alpha_m^{-1} \rangle = \sqrt{\frac{\langle v_m^2 \rangle}{\langle x_m^2 \rangle}}, \quad (57)$$

$$\hat{q}(v_m) = \mathcal{R} \left(v_m | \frac{1}{\sqrt{\langle x_m^2 \rangle}} \right), \quad (58)$$

$$\langle v_m^2 \rangle = \frac{2}{\langle \alpha_m \rangle} = \frac{2}{\langle \alpha_m^{-1} \rangle^{-1} + \langle v_m^2 \rangle^{-1}}. \quad (59)$$

For noise (observation) model, consider

$$\hat{q}(\beta) = \mathcal{G}(\beta | A_{\beta}, B_{\beta}), \quad (60)$$

$$\langle \beta \rangle = \frac{A_{\beta}}{B_{\beta}}, \quad (61)$$

$$\hat{q}(\delta_n) = \mathcal{G}(\delta_n | A_{\delta}, B_{\delta_n}), \quad (62)$$

$$\langle \delta_n \rangle = \frac{A_{\delta}}{B_{\delta_n}}, \quad (63)$$

$$\hat{q}(d) = \mathcal{G}(d | A_d, B_d), \quad (64)$$

$$\langle d \rangle = \frac{A_d}{B_d}. \quad (65)$$

When initial expectations are assumed, the iterative updating can be carried out in closed form.

5. Relations to ARD Update Procedure

Now let us inspect the relations between the proposed Double Lomax updating procedure and the ARD updating procedure.

Considering (57) and (59) that are induced by the Double Lomax priors, it can be shown that if just (57) and (59) are iterated to convergence, then the final result is equivalent to directly setting

$$\langle \alpha_m^{-1} \rangle = \frac{1}{\langle x_m^2 \rangle}, \quad (66)$$

which is the exact ARD update [26]. To prove this, construct a function as $\mathcal{F}(\langle \alpha_m^{-1} \rangle, \langle v_m^2 \rangle) = \langle x_m^2 \rangle \langle \alpha_m^{-1} \rangle + \langle v_m^2 \rangle / \langle \alpha_m^{-1} \rangle - \ln(\langle v_m^2 \rangle)$. It can be shown that (57) minimizes $\mathcal{F}$ partially with respect to $\langle \alpha_m^{-1} \rangle$, and (59) minimizes $\mathcal{F}$ partially with respect to $\langle v_m^2 \rangle$. If we do the complete minimization over both $\langle \alpha_m^{-1} \rangle$ and $\langle v_m^2 \rangle$, then we directly get (66). Therefore, when noninformative hyperprior is assumed, the proposed Double Lomax updates only have one more latent variable $\langle v_m^2 \rangle$ than the canonical ARD updates; this latent variable can be further vanished if we only consider the updating equations of (57) and (59).

However, if we consider the full iteration procedure involving all the updating equations, (54), (55), (57), (59),

FIGURE 4: Regression coefficient estimation results using the Double Lomax formulation (DL) and the ARD formulation (ARD), reported with means and standard deviations.

(61), (63), and (65), then the difference is shown: the solution trajectories of Double Lomax are not necessarily consistent with those of ARD. To see this, we need to denote ARD's Equation (66), and Double Lomax's Equations (57) and (59) from an iteration perspective, so we have

$$\langle \alpha_m^{-1} \rangle_{(l+1)} = \frac{1}{\langle x_m^2 \rangle_{(l)}} \quad (67)$$

for ARD and

$$\langle \alpha_m^{-1} \rangle_{(l+1)} = \sqrt{\frac{\langle v_m^2 \rangle_{(l)}}{\langle x_m^2 \rangle_{(l)}}}, \quad (68)$$

$$\langle v_m^2 \rangle_{(l+1)} = \frac{2}{\langle \alpha_m^{-1} \rangle_{(l)}^{-1} + \langle v_m^2 \rangle_{(l)}^{-1}} \quad (69)$$

for Double Lomax, where l denotes the l th iteration.

It can be seen from (69) that the latent variable $\langle v_m^2 \rangle_{(l)}$ of Double Lomax can be expressed as a function of $\langle \alpha_m^{-1} \rangle$, which involves all the updates of $\langle \alpha_m^{-1} \rangle$ from former iterations. That is,

$$\langle v_m^2 \rangle_{(l)} = \Omega_{(l)} \left(\langle \alpha_m^{-1} \rangle_{(l-1)}, \langle \alpha_m^{-1} \rangle_{(l-2)}, \dots, \langle \alpha_m^{-1} \rangle_{(0)} \right). \quad (70)$$

Considering the above expression and comparing (67) with (68), the difference between ARD and Double Lomax ($f \rightarrow +0$) is shown: in contrast to the ARD's $\langle \alpha_m^{-1} \rangle_{(l+1)}$ that depends only on $\langle x_m^2 \rangle_{(l)}$, the Double Lomax's $\langle \alpha_m^{-1} \rangle_{(l+1)}$ depends not

only on $\langle x_m^2 \rangle_{(l)}$ but also on all the former updates of $\langle \alpha_m^{-1} \rangle$, including $\langle \alpha_m^{-1} \rangle_{(l-1)}, \langle \alpha_m^{-1} \rangle_{(l-2)}, \dots, \langle \alpha_m^{-1} \rangle_{(0)}$. Consequently, it makes the Double Lomax method being able to smooth the solution trajectory of $\langle v_m^2 \rangle_{(l)}$ and also the trajectories of all other variables, as the results of smoothness propagation effect caused by the alternating-update strategy of VB inference. With these smooth trajectories, the algorithm may be prevented from a large deviation of the right direction when there are some disturbances or unreliable updates. As will be shown experimentally, comparing to ARD, Double Lomax tends to move more carefully around local minima and provides a better convergence performance.

Following, we use linear regression as an example, the coefficient vector is $\mathbf{x} = [1, -1, 0, 0, 1, -1, 1, 0, 0, -1]^T$, the basis matrix Φ is a 12×10 random matrix whose entries are independently chosen from $\mathcal{N}(0, 1)$, and $\mathbf{e}$ is zero-mean Gaussian noise with a standard deviation of 0.01. The VB-Double Lomax ($f \rightarrow +0$) formulation established in the above section (denoted with DL) and an equivalent formulation with ARD prior (denoted with ARD) are used here for comparison. The maximum number of iterations are set to 50 for both of the algorithms. Figure 4 shows the estimation results of 50 trials, and the estimation mean (denoted with mean) and standard deviation (denoted with std) are also reported. Neither DL nor ARD can recover the exact coefficients due to the noise. In general, DL has shown more cases of convergence to global minima, while ARD has shown more cases of convergence to local minima, which has a relatively large deviation of global minima though with a small probability since that does not change the mean values much. Figure 5(a) reports the coefficient update traces of a typical case when ARD converges to a local minima but DL does not. Figure 5(b) reports the coefficient update traces of a typical case when both ARD and DL converge to the global minima. In both situations, the DL appears to have smoother coefficient update traces, which is in agreement with the above analysis. Furthermore, the update traces of latent variables $\langle \alpha_m^{-1} \rangle$ and noise parameter β when both ARD and DL converge to the global minima (the case in Figure 5(b)) are also reported in Figure 6. Comparing to the traces of ARD's $\langle \alpha_m^{-1} \rangle$, smoother traces of DL's are observed, even if they are plotted on a log scale. It is noted that the trace shapes of DL's $\langle \alpha_m^{-1} \rangle$ are similar to those of DL's $\langle v_m^2 \rangle$. But the differences can be seen in Figures 6(d) and 6(e) (nonzero coefficient's traces are shown in Figure 6(d) while zero coefficients are shown in Figure 6(e) for better visualization). If we force $\langle v_m^2 \rangle$ to equal to be $\langle \alpha_m^{-1} \rangle$, an instance of ARD update is obtained. It is worth noting that since an extra level of iterations is introduced by $\langle v_m^2 \rangle$, a slower convergence speed of DL can be expected in general.

6. Computer Simulations

In this section, the proposed VB-Double Lomax ($f \rightarrow +0$) formulation (denoted with DL) is compared with the equivalent formulation with ARD prior (denoted with ARD). Furthermore, a hybrid formulation (denoted by hybrid), with first K iterations employing DL and the rest employing ARD

(a) The update traces when ARD converges to a local minima but DL does not

(b) The update traces when both ARD and DL converge to the global minima

FIGURE 5: The update traces of coefficients from the regression coefficient estimation experiment shown in Figure 4. (A) The update traces by ARD. (B) The update traces by DL. (C) The update traces of the first three coefficients $[1, -1, 0]^T$.

FIGURE 6: The update traces of latent variables and noise parameter when both ARD and DL converge to the global minima (shown in Figure 5(b)).

until convergence, is also considered here for a tradeoff between performance and speed. This hybrid approach is based on the observation that ARD tends to converge to a local minima at early iterations. Two kinds of SLM applications with different Φ matrices are considered here. One is the AR model identification where the rows of Φ matrix

are strongly correlated, the other is the standard compressive sensing problem where the rows of Φ matrix are uncorrelated. In the experiments reported below, the hybrid number $K = 20$, and an experimental criterion, $\|\mathbf{x}(k) - \mathbf{x}(k-1)\| < 10^{-3}$ and lasting for more than 50 iterations, is used to terminate the iterative procedure.

FIGURE 7: AR model identification using DL. Left: coefficients and order estimation. Center: expected values of one-step-ahead predictions versus the observations. Right: excitation noise estimation.

Autoregressive model is widely used to model time series data, which attempts to predict the future values based on previous values. An AR model of order p is defined as

$$y_n = \sum_{i=1}^p x_i y_{n-i} + e_n, \quad (71)$$

where y_n is the n th observation (measurement) in the series, x_i is the AR coefficients, and e_n is the excitation noise. Using matrix notation, we have

$$\mathbf{y} = \Phi_{\text{AR}} \mathbf{x} + \mathbf{e}, \quad (72)$$

where Φ_{AR} is the $N \times p$ matrix whose n th row contains the lags for elements y_n , that is, $[y_{n-1}, \dots, y_{n-p}]$ and $\mathbf{x}$ is the AR coefficient vector to be estimated. If the model order p is unknown, the strategy used here is choosing an M where M is large enough to ensure $M > p$ then computing the coefficient vector of $M \times 1$, and the extra coefficients are supposed to have estimates of zero.

Figure 7 shows an example of AR identification using DL with $N = 300$, $p = 8$, and $M = 20$. The excitation noise follows a Student- t distribution with $d = 3$ and $\beta = 1$. It is shown in the left of the figure that for coefficients $M \leq p$, the algorithm gives relatively satisfactory estimates, and for the extra coefficients $M > p$, the estimates have been set closely to zero. The center is plotted for the expected values of one-step-ahead predictions compared with the observations, and the right gives the comparison of the estimated excitation noise distribution with that used to generate the observations.

To compare DL, ARD, and hybrid, we use three AR models with $p = 5, 10$, and 15 . In each case, $M = 20$ and the excitation noise is Student- t distributed with $d = 3$ and $\beta = 1$. The coefficient estimation error is calculated as $\|\hat{\mathbf{x}} - \mathbf{x}\|_2 / \|\mathbf{x}\|_2$, where $\hat{\mathbf{x}}$ and $\mathbf{x}$ are the estimated and true coefficient vectors. The prediction mean square errors (MSEs) are

calculated as $E[(\hat{y} - y)^2]$, where $\hat{y}$ and y are the one-step-ahead predictions and actual observations. Each experiment is carried out 200 times and the average results for AR(5), AR(10), and AR(15) are shown in Figures 8(a), 8(b), and 8(c), respectively. The figures on the left report the coefficient estimation error with error bars (one standard deviation). It is shown that DL results in smaller errors than ARD, and the improvement is more significant for AR(10) and AR(15). It is also observed that hybrid gives almost the same performance as DL. As the number of observations increases, the coefficient estimation errors of DL, ARD, and Hybrid decrease and their differences become less significant. Similar observations are found in the prediction MSEs, which are reported in the center of Figure 8. However, the prediction MSEs do not decrease monotonously with the number of observations. This may be due to the reason that the current SLM formulation is based on the overall coefficient estimation error but does not put the partial autocorrelation of the AR model in consideration. Finally, the right of the figure reports the number of iterative steps required for convergence versus the number of observations. It is shown that for the same number of observations, DL requires more iterations than ARD for convergence. However, hybrid can achieve a comparable convergence rate as ARD. In general, as the number of observations increases, the convergence speeds of DL, ARD, and Hybrid also increase.

CS is a technique for acquiring and recovering sparse signals with underdetermined linear systems that can be represented as

$$\mathbf{y} = \Phi_{\text{CS}} \mathbf{x} + \mathbf{e}, \quad (73)$$

where $\mathbf{y}$ is $N \times 1$ observations, Φ_{CS} is $N \times M$ measurement matrix, $N < M$, $\mathbf{e}$ represents the acquisition noise, and $\mathbf{x}$ is the $M \times 1$ unknown sparse signal, that is, most of its coefficients are zero.

FIGURE 8: Comparison of DL, ARD, and hybrid for AR(5), AR(10), and AR(15). Left: coefficient estimation error with error bars (one standard deviation). Center: prediction MSE. Right: the number of iterative step required for convergence.

FIGURE 9: Number of measurements versus reconstruction error with error ranges (one standard deviation) for the noise-free case.

We use the following default CS setup in our experiments. Φ_{CS} is induced by sampling i.i.d. columns from a unit sphere in R^N . T coefficients at random locations of the signal are drawn from four different probability distributions, and the rest $(M - T)$ of the coefficients are set to zero. The nonzero coefficients of the sparse signals are realizations of the following four distributions: (1) uniform ± 1 random spikes, (2) nonnegative unit spikes, (3) zero-mean unit variance Gaussian, and (4) zero-mean Student- t with unit

scale parameter and shape parameter equal 3. We fix $M = 256$ and $T = 80$, and we vary the number of measurements N . Moreover, both noiseless and noisy acquisitions are considered, where a zero-mean white Gaussian noise with a standard deviation of 0.005 is used in the noisy experiment. The reconstruction error is calculated as $\|\hat{\mathbf{x}} - \mathbf{x}\|_2 / \|\mathbf{x}\|_2$, where $\hat{\mathbf{x}}$ and $\mathbf{x}$ are the estimated and true signals. Each experiment is carried out 200 times and the average results are reported. Figure 9 reports the number of measurements

FIGURE 10: The number of iterative step required for convergence for the noise-free case.

versus reconstruction error with error ranges (one standard deviation) for the noise-free case. It is shown that DL provides improved overall reconstruction performance over ARD for a reasonable number of measurements, the improvement is more significant for signal types (1) and (2) and less significant for types (3) and (4) (but the differences can still be seen from the error ranges). And the hybrid can provide a performance close to DL. Figure 10 reports the number of

iterative steps required for convergence versus the number of measurements for the noise-free case. DL shows a slower convergence rate than ARD, and hybrid has a comparable rate as of the ARD. Figure 11 shows the number of estimated nonzero coefficients for the noise-free case, whose values are larger than 10^{-3} . Overall, DL gives the sparsest estimation results, then hybrid followed by ARD. Figure 12 reports the results for noisy acquisitions. Similar observations to the

FIGURE 11: The number of estimated nonzero coefficients for the noise-free case.

noise-free case (shown in Figure 9) are found in this figure, except that this time none of the algorithms can recover the coefficients exactly due to the noise.

7. Conclusion

In this paper, we propose a new non-log-concave sparsity prior, referred to as Double Lomax Prior, which corresponds to a three-level hierarchical Bayes model. A VB inference

procedure is introduced to solve for the SLM using Double Lomax priors. When noninformative hyperprior is assumed ($f \rightarrow +0$), the proposed update procedure includes the canonical ARD update procedure as a special case. However, comparing to the canonical ARD update, the proposed update procedure has one more latent variable which has the smoothness effect that can result in an improved performance. A hybrid formulation of Double Lomax and ARD is also considered as a tradeoff between performance and

FIGURE 12: Number of measurements versus reconstruction error with error ranges (one standard deviation) for noisy observations.

computational speed. Experiments on both correlated and uncorrelated SLM simulations with applications to AR model identification and compressive sensing are carried out to demonstrate the effectiveness of the proposed approach.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive suggestions. This work is supported by the National Natural Science Foundation of China

(no. 61205017), the China Postdoctoral Science Foundation funded project (no. 2012M511058), and the Shanghai Postdoctoral Sustentation Fund funded project (no. 12R21412500).

References

- [1] J.-J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Transactions on Information Theory*, vol. 50, no. 6, pp. 1341–1344, 2004.

- [2] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 5, pp. 2197–2202, 2003.
- [3] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society B*, vol. 58, no. 1, pp. 267–288, 1996.
- [4] B. D. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Transactions on Signal Processing*, vol. 47, no. 1, pp. 187–200, 1999.
- [5] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2153–2164, 2004.
- [6] J. Christmas and R. Everson, "Robust autoregression: student-t innovations using variational Bayes," *IEEE Transactions on Signal Processing*, vol. 59, no. 1, pp. 48–57, 2011.
- [7] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.
- [8] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Transactions on Signal Processing*, vol. 56, no. 6, pp. 2346–2356, 2008.
- [9] D. Wipf and S. Nagarajan, "A unified Bayesian framework for MEG/EEG source imaging," *NeuroImage*, vol. 44, no. 3, pp. 947–966, 2009.
- [10] I. F. Gorodnitsky, J. S. George, and B. D. Rao, "Neuromagnetic source imaging with FOCUSS: a recursive weighted minimum norm algorithm," *Electroencephalography and Clinical Neurophysiology*, vol. 95, no. 4, pp. 231–251, 1995.
- [11] B. Huang and H. Song, "Spatiotemporal reflectance fusion via sparse representation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 10, pp. 9707–9716, 2012.
- [12] B. Yang and S. Li, "Pixel-level image fusion with simultaneous orthogonal matching pursuit," *Information Fusion*, vol. 13, no. 1, pp. 10–19, 2012.
- [13] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Review*, vol. 51, no. 1, pp. 34–81, 2009.
- [14] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.
- [15] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.
- [16] B. D. Rao, K. Engan, S. F. Cotter, J. Palmer, and K. Kreutz-Delgado, "Subset selection in noise based on diversity measure minimization," *IEEE Transactions on Signal Processing*, vol. 51, no. 3, pp. 760–770, 2003.
- [17] R. Chartrand and V. Staneva, "Restricted isometry properties and nonconvex compressive sensing," *Inverse Problems*, vol. 24, no. 3, Article ID 035020, 14 pages, 2008.
- [18] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted l_1 minimization," *The Journal of Fourier Analysis and Applications*, vol. 14, no. 5–6, pp. 877–905, 2008.
- [19] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.
- [20] R. Chartrand, "Exact reconstruction of sparse signals via non-convex minimization," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 707–710, 2007.
- [21] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, New York, NY, USA, 2002.
- [22] M. J. Wainwright and M. I. Jordan, *Graphical Models, Exponential Families, and Variational Inference*, Now Publishers, 2008.
- [23] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Iteratively reweighted least squares for linear regression when errors are normal/independent distributed," in *Multivariate Analysis V*, P. R. Krishnaiah, Ed., pp. 35–57, North-Holland, 1980.
- [24] D. J. C. MacKay and C. Laboratory, "Bayesian nonlinear modeling for the prediction competition," in *Proceedings of the ASHRAE Annual Meeting*, vol. 100, pp. 1053–1062, June 1994.
- [25] C. M. Bishop and M. E. Tipping, "Variational relevance vector machines," in *Proceedings of the 16th Conference Uncertainty in Artificial Intelligence*, pp. 46–53, 2000.
- [26] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, no. 3, pp. 211–244, 2001.
- [27] D. Wipf and S. Nagarajan, "A new view of automatic relevance determination," in *Proceedings of the 21st Annual Conference on Neural Information Processing Systems (NIPS '07)*, December 2007.
- [28] C. Kleiber and S. Kotz, *Statistical Size Distributions in Economics and Actuarial Sciences*, Wiley-Interscience, Hoboken, NJ, USA, 2003.
- [29] M. Seeger and D. Wipf, "Variational bayesian inference techniques," *IEEE Signal Processing Magazine*, vol. 27, no. 6, pp. 81–91, 2010.
- [30] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, NY, USA, 1st edition, 2007.
- [31] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "Introduction to variational methods for graphical models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, 1999.
- [32] D. J. C. MacKay, "Ensemble learning and evidence maximization," in *Annual Conference on Neural Information Processing Systems (NIPS '95)*, 1995.
- [33] B. Jørgensen, *Statistical Properties of the Generalized Inverse Gaussian Distribution*, vol. 9, Springer, New York, NY, USA, 1982.
- [34] I. S. Gradshteyn, I. M. Ryzhik, A. Jeffrey, and D. Zwillinger, *Table of Integrals, Series, and Products*, Elsevier, Boston, Mass, USA, 7th edition, 2007.
- [35] E. Cojocar, "Parabolic cylinder functions implemented in Matlab," <http://arxiv.org/abs/0901.2220>.