A VARIATIONAL BAYESIAN APPROACH TO COMPRESSIVE SENSING BASED ON DOUBLE LOMAX PRIORS

Xiaojing Gu[⋆] *Henry Leung*[†] *Xingsheng Gu*[⋆]

[⋆] East China University of Science and Technology [†] University of Calgary
{xjing.gu,xsgu}@ecust.edu.cn leungh@ucalgary.ca

ABSTRACT

Automatic Relevance Determination (ARD) priors have been widely used to induce sparse reconstructions in Bayesian compressive sensing approaches. In this paper, we propose a new sparsity-promoting prior coined as Double Lomax prior. Its connection with the generalized inverse Gaussian distribution and Rayleigh distribution leads to a tractable full Variational Bayesian (VB) inference procedure here. It is shown that the proposed update procedure includes the canonical ARD update procedure as a special case, but provides a better global convergence performance and results in improved signal reconstructions.

Index Terms— Sparsity-promoting prior, Double Lomax distribution, automatic relevance determination (ARD), Variational Bayesian (VB), compressive sensing

1. INTRODUCTION

Compressive sensing (CS) theory [1–3] proposes new techniques to recover unknown sparse signals from underdetermined linear measurements and has become one of the main research topics in the signal processing area with various applications [4–10]. The CS system can be modeled as

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

where $\mathbf{y}$ is a $N \times 1$ measurement vector, $\mathbf{x}$ is a $M \times 1$ sparse vector, $\mathbf{e}$ is white Gaussian noise and $\mathbf{\Phi}$ is the $N \times M$ measurement matrix, with $N < M$. The associated reconstruction problem is given by

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

where $\|\mathbf{x}\|_0$ denotes the number of nonzero elements in the vector $\mathbf{x}$, ρ is the model parameter that controls the relative importance applied to error term and sparseness term. Since Eqn.(2) requires a combinatorial search which is NP-hard, linear programming methods with convex relaxation ℓ_1 norm [2, 11], reweighted norm algorithms [12, 13] and greedy methods [14, 15] have been wildly developed as alternatives.

A main issue of these methods is that the uncertainty of signal reconstructions is generally obscure. Recently, several algorithms have been investigated with in the Bayesian framework [16–19], with the advantage of providing probabilistic estimates that can guide adaptive measurement matrix design. Moreover, the Bayesian CS algorithms naturally model the unknown signal along with the model parameters, which result in fully automated algorithms estimating all required parameters.

In these Bayesian CS algorithms, the Automatic Relevance Determination (ARD) prior [20] is generally used, due to be non-log-concave for strong sparsity promotion and be Gaussian-integral-representable for efficient Bayesian analysis, simultaneously. It is defined by

$$ARD\left(x\right) = \int_0^\infty \mathcal{N}\left(x \left| 0, \frac{1}{\gamma}\right.\right) \mathcal{G}\left(\gamma \left| a, b\right.\right) d\gamma, \quad a \to 0, b \to 0, \tag{3}$$

where $\mathcal{G}\left(\gamma \left| a, b\right.\right)$ is the Gamma prior defined as $\mathcal{G}\left(\gamma \left| a, b\right.\right) = \frac{b^a}{\Gamma(a)}\gamma^{a-1}\exp\left(-b\gamma\right)$. Eqn.(3) invokes a two-level hierarchical prior model.

In this paper, we also formulate the CS reconstruction problem from a Bayesian perspective. Despite of the two-level ARD prior, 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 fully VB inference procedure is derived to solve for the CS problem using Double Lomax priors. It is shown that the proposed update procedure includes the canonical ARD update procedure as a special case, but results in a better global convergence performance due to the existence of extra latent variable.

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 [21]). That is,

$$\mathcal{DL}\left(x \left| 0, \eta, f\right.\right) = \frac{\eta}{2}\left(1 + \frac{\eta\left|x\right|}{f}\right)^{-(f+1)}, \ \eta > 0, f > 0. \tag{4}$$

It can be shown that $\mathcal{DL}\left(x \left| 0, \eta, f\right.\right) > 0$, $\int_{-\infty}^{+\infty} \mathcal{DL}\left(x \left| 0, \eta, f\right.\right) = 1$ and $\mathcal{DL}\left(-x \left| 0, \eta, f\right.\right) = \mathcal{DL}\left(x \left| 0, \eta, f\right.\right)$. Thus Eqn.(4) 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.

Theorem 1. *Double Lomax prior in Eqn.(4) is log-convex. When $f \to +0$, Double Lomax prior becomes proportional to Jeffreys prior; when $f \to +\infty$, Double Lomax prior becomes Laplace prior.*

Proof. The logarithm of density function Eqn.(4) is

$$\ln \mathcal{DL} = \ln\left(\frac{\eta}{2}\right) - (f+1)\ln\left(1+\frac{\eta\,|x|}{f}\right). \tag{5}$$

It is continuous on $(-\infty\ +\infty)$, and the second derivative is

$$\frac{d^2}{dx^2}\ln \mathcal{DL} = (f+1)\,\frac{\eta^2}{(f+\eta\,|x|)^2}\begin{cases} >0 & f \nrightarrow +\infty \\ =0 & f \to +\infty \end{cases}. \tag{6}$$

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

$$\lim_{f\to+0} \mathcal{DL}\,(x\,|0,\eta,f\,) \propto \frac{1}{|x|}. \tag{7}$$

$$\lim_{f\to+\infty} \mathcal{DL}\,(x\,|0,\eta,f\,) = \frac{\eta}{2}\exp\left(-\eta\,|x|\right). \tag{8}$$

Thus when the shape parameter goes to zero, Double Lomax prior is proportional to the noninformative Jeffreys prior $\mathrm{p}(x) \propto 1/|x|$; when the shape parameter goes to infinite, Double Lomax prior asymptotically becomes the Laplace prior with inverse-scale η. □

Remark 1. *Theorem 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 $f \to +\infty$) to the strongly concave log-sum penalty $\sum \log(|x|)$ (when $f \to +0$).*

Theorem 2. *Double Lomax prior in Eqn.(4) can be represented as a Gaussian integral. Given Gaussian prior $\mathcal{N}(\cdot)$, Exponential prior $\mathcal{E}(\cdot)$ and Gamma prior $\mathcal{G}(\cdot)$, we have*

$$\mathcal{DL} = \int_0^\infty\int_0^\infty \mathcal{N}\,(x\,|0,\alpha\,)\,\mathcal{E}\left(\alpha\,\middle|\,\frac{\eta^2\upsilon^2}{2}\right)\mathcal{G}\,(\upsilon|\,f,f)\,d\alpha d\upsilon. \tag{9}$$

Proof. Eqn.(4) can be expressed as,

$$\begin{aligned}
\mathcal{DL} =& \frac{\eta}{2} f^{f+1}(f+\eta\,|x|)^{-(f+1)} \\
=& \frac{\Gamma(f+1)}{\Gamma(f)}\frac{\eta}{2} f^f (f+\eta\,|x|)^{-(f+1)} \\
=& \frac{\eta f^f}{2\Gamma(f)}\int_0^\infty \upsilon^f \exp\{-\upsilon\,(f+\eta\,|x|)\}d\upsilon \\
=& \int_0^\infty \frac{\eta\upsilon}{2}\exp\{-\eta\upsilon\,|x|\}\,\frac{f^f}{\Gamma(f)}\upsilon^{f-1}\exp\{-f\upsilon\}\,d\upsilon \\
=& \int_0^\infty \mathcal{L}\,(x\,|0,\eta\upsilon)\mathcal{G}\,(\upsilon\,|f,f\,)\,d\upsilon.
\end{aligned} \tag{10}$$

Meanwhile, note that a Laplace prior has hierarchical representation

$$\mathcal{L}\,(x\,|0,\lambda\,) = \int_0^\infty \mathcal{N}\,(x\,|0,\alpha\,)\,\mathcal{E}\left(\alpha\,\middle|\,\frac{\lambda^2}{2}\right)d\alpha = \frac{\lambda}{2}\exp\left(-\lambda\,|x|\right). \tag{11}$$

Substituting Eqn.(11) into Eqn.(10) gives Eqn.(9). □

Remark 2. *Eqn.(9) corresponds to a three-level hierarchical Bayes model with two latent variables α and υ. 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 are generated from exponential priors with independent inverse-scales, which are formed by $\frac{1}{2}\eta^2$ multiplying a set of independent variables υ^2. On the third level, independent υ are generated from a Gamma prior with two parameters of the same value.*

3. BAYESIAN MODELING

In Bayesian modeling, each unknown variable is described by a probability distribution. We assume independent zero-mean Gaussian over noise $\mathbf{e}$,

$$\mathrm{p}\,(\mathbf{y}\,|\mathbf{x},\beta\,) = \prod_{n=1}^{N}\mathcal{N}\left(y_n\,|\boldsymbol{\phi}_n\mathbf{x},\beta^{-1}\right) = \mathcal{N}\left(\mathbf{y}\,|\boldsymbol{\Phi}\mathbf{x},\beta^{-1}\right), \tag{12}$$

where $\boldsymbol{\phi}_n$ is the nth row of $\boldsymbol{\Phi}$ and β is precision. We specify a Gamma prior for β:

$$\mathrm{p}\,(\beta) = \mathcal{G}\,(\beta\,|a_\beta,b_\beta\,). \tag{13}$$

We assume independent Double Lomax priors over the signal $\mathbf{x}$. Without loss of generality, the scale parameter η is assumed to be unity. Using the hierarchical representation in Eqn.(9), we have

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

$$\mathrm{p}\,(\alpha_m\,|\upsilon_m\,) = \mathcal{E}\left(\alpha_m\,\middle|\,\frac{\upsilon_m^2}{2}\right), \tag{15}$$

$$\mathrm{p}\,(\upsilon_m\,|f\,) = \mathcal{G}\,(\upsilon_m\,|f,f\,), \tag{16}$$

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

Therefore, the joint distribution of the measurements and all the unknown parameters can be factorize as

$$p(\mathbf{y},\beta,\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\upsilon}) = p(\mathbf{y}\,|\mathbf{x},\beta\,)p(\mathbf{x}\,|\boldsymbol{\alpha}\,)p(\boldsymbol{\alpha}\,|\boldsymbol{\upsilon}\,)p(\boldsymbol{\upsilon}\,|f\,)p(\beta). \tag{17}$$

4. INFERENCE PROCEDURE

The Bayesian inference is based on the posterior

$$\mathrm{p}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\upsilon}, \beta \,|\mathbf{y}\,) = \frac{\mathrm{p}(\mathbf{y}, \beta, \mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\upsilon})}{\mathrm{p}(\mathbf{y})}. \tag{18}$$

However the Bayesian integral $\mathrm{p}(\mathbf{y})$ is analytically intractable. Therefore, approximation methods are required. In this work, we incorporate a Variational Bayesian (VB) approach [22–24] for the inference, which approximates the posterior by introducing a factorable distribution, found by minimizing the Kullback-Leibler divergence between the posterior and its approximation. In this work, we factorize the approximation distribution $\mathrm{q}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\upsilon}, \beta)$ as $\mathrm{q}(\mathbf{x})\mathrm{q}(\boldsymbol{\alpha})\mathrm{q}(\boldsymbol{\upsilon})\mathrm{q}(\beta)$.

To compute $\mathrm{q}(\mathbf{x})$, we only need to consider those terms that are functionally dependent on $\mathbf{x}$. Because $\mathrm{p}(\mathbf{x}\,|\boldsymbol{\alpha}\,)$ is the conjugate prior for $\mathbf{x}$, $\hat{\mathrm{q}}(\mathbf{x})$ is still a Gaussian distribution

$$\begin{aligned}\hat{\mathrm{q}}\,(\mathbf{x}) &\propto \exp\left\langle \ln \mathrm{p}(\mathbf{y}\,|\mathbf{x}, \beta\,) + \ln \mathrm{p}(\mathbf{x}\,|\boldsymbol{\alpha}\,)\right\rangle_{\beta,\boldsymbol{\alpha}} \\ &= \mathcal{N}\,(\,\mathbf{x}|\,\boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}}),\end{aligned} \tag{19}$$

with

$$\boldsymbol{\mu}_{\mathbf{x}} = \langle\beta\rangle\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{y}. \tag{20}$$

$$\boldsymbol{\Sigma}_{\mathbf{x}} = \left(\langle\beta\rangle\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \langle\boldsymbol{\Lambda}\rangle\right)^{-1}. \tag{21}$$

where $\langle\boldsymbol{\Lambda}\rangle = diag(\langle\alpha_m^{-1}\rangle)$.

From Eqn.(14) and Eqn.(15) we observe that the terms involving $\boldsymbol{\alpha}$ are composed of M-order multiplicative terms over α_m, x_m and υ_m, a further factorization of this approximate posterior density becomes:

$$\mathrm{q}(\boldsymbol{\alpha}) = \prod_{m=1}^{M} \mathrm{q}(\alpha_m). \tag{22}$$

And

$$\begin{aligned}\hat{\mathrm{q}}\,(\alpha_m) &\propto \exp\langle \ln \mathrm{p}\,(x_m\,|\alpha_m\,) + \ln \mathrm{p}\,(\alpha_m\,|\upsilon_m\,)\rangle_{x_m,\upsilon_m} \\ &\propto \mathcal{N}\left(\sqrt{\langle x_m^2\rangle}\middle|\, 0, \alpha_m\right)\mathcal{E}\left(\alpha_m\middle|\frac{\langle\upsilon_m^2\rangle}{2}\right),\end{aligned} \tag{23}$$

where

$$\left\langle x_m^2\right\rangle = \mu_{\mathbf{x}(m)}^2 + \boldsymbol{\Sigma}_{\mathbf{x}(mm)}. \tag{24}$$

$\cdot_{(m)}$ and $\cdot_{(mm)}$ denote the m^{th} element of a vector and the m^{th} diagonal element of a matrix, respectively. By further inspecting its normalization, $\hat{\mathrm{q}}(\alpha_m)$ turns out to be a Generalized Inverse Gaussian Distribution ($\mathcal{GIG}$) with the probability function $\mathcal{GIG}(z\,|\omega,\,\chi,\psi) = \frac{(\psi/\chi)^{\omega/2}}{2K_\omega(\sqrt{\chi\psi})}z^{(\omega-1)}\exp\left(-\frac{1}{2}(\frac{\chi}{z}+\psi z)\right)$, where $K_\omega(\cdot)$ is the modified Bessel function of the third kind. Hence, here $\hat{\mathrm{q}}(\alpha_m)$ is $\mathcal{GIG}\left(\alpha_m\,\middle|\frac{1}{2},\,\langle x_m^2\rangle,\,\langle\upsilon_m^2\rangle\right)$. With moment analysis, we have

$$\left\langle\alpha_m^{-1}\right\rangle = \sqrt{\frac{\langle\upsilon_m^2\rangle}{\langle x_m^2\rangle}}, \tag{25}$$

$$\langle\alpha_m\rangle = \left\langle\alpha_m^{-1}\right\rangle^{-1} + \left\langle\upsilon_m^2\right\rangle^{-1}. \tag{26}$$

The approximate posterior density of $\boldsymbol{\upsilon}$ can be factorized as

$$\mathrm{q}(\boldsymbol{\upsilon}) = \prod_{m=1}^{M} \mathrm{q}(\upsilon_m). \tag{27}$$

And

$$\begin{aligned}\hat{\mathrm{q}}(\upsilon_m) &\propto \exp\left\langle \ln \mathrm{p}\,(\alpha_m\,|\upsilon_m\,) + \ln \mathrm{p}\,(\upsilon_m\,|f\,)\right\rangle_{\alpha_m} \\ &\propto \exp\left((f+1)\ln\upsilon_m - \frac{\langle\alpha_m\rangle}{2}\upsilon_m^2 - f\upsilon_m\right).\end{aligned} \tag{28}$$

Eqn.(28) contains logarithmic, square and linear terms, thus is not in correspondence with any standard distribution. However, when noninformative hyperprior is assumed, f is setted to zero, $\hat{\mathrm{q}}(\upsilon_m)$ becomes the Rayleigh distribution with the probability

$$\mathcal{R}(z\,|\lambda\,) = \frac{z}{\lambda^2}\exp\left(-\frac{z^2}{2\lambda^2}\right). \tag{29}$$

Here $\hat{\mathrm{q}}(\upsilon_m)$ is $\mathcal{R}(\upsilon_m\left|\frac{1}{\sqrt{\langle x_m^2\rangle}}\right.)$. Therefore,

$$\left\langle\upsilon_m^2\right\rangle = \frac{2}{\langle\alpha_m\rangle} = \frac{2}{\left\langle\alpha_m^{-1}\right\rangle^{-1} + \left\langle\upsilon_m^2\right\rangle^{-1}}. \tag{30}$$

Since Gamma distribution is conjugate to the precision of Gaussian function, we still obtain a Gamma distribution for β:

$$\hat{\mathrm{q}}(\beta) = \mathcal{G}(\beta\,|A_\beta, B_\beta\,), \tag{31}$$

$$\langle\beta\rangle = A_\beta/B_\beta, \tag{32}$$

where

$$A_\beta = \frac{N}{2} + a_\beta, \tag{33}$$

$$B_\beta = \frac{1}{2}\mathrm{trace}\left(\left\langle(\mathbf{y}-\boldsymbol{\Phi}\mathbf{x})(\mathbf{y}-\boldsymbol{\Phi}\mathbf{x})^{\mathrm{T}}\right\rangle\right) + b_\beta, \tag{34}$$

and

$$\left\langle(\mathbf{y}-\boldsymbol{\Phi}\mathbf{x})(\mathbf{y}-\boldsymbol{\Phi}\mathbf{x})^{\mathrm{T}}\right\rangle = \mathbf{y}\mathbf{y}^{\mathrm{T}} - 2\boldsymbol{\Phi}\boldsymbol{\mu}_{\mathbf{x}}\mathbf{y}^{\mathrm{T}} + \boldsymbol{\Phi}(\boldsymbol{\mu}_{\mathbf{x}}\boldsymbol{\mu}_{\mathbf{x}}{}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\mathbf{x}})\boldsymbol{\Phi}^{\mathrm{T}}. \tag{35}$$

Now let us inspect the relations between the proposed update procedure and the ARD update procedure. As noted previously, when the noninformative hyperprior of the third level is assumed ($f \to +0$), the Double Lomax prior becomes equivalent to the ARD prior, both are proportional to the noninformative Jeffreys prior. Consider Eqn.(25) and Eqn.(30) that induced by the Double Lomax priors, it can be proved that if just these two equations iterate to convergence, then the result is equivalent to directly setting

$$\left\langle\alpha_m^{-1}\right\rangle = \frac{1}{\langle x_m^2\rangle}, \tag{36}$$

which is the exact ARD update [25]. However, if the procedure is applied until all update equations are iterated to convergence, the difference 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 on $\langle x_m^2 \rangle_{(l)}$ and $\langle v_m^2 \rangle_{(l)}$, while $\langle v_m^2 \rangle_{(l)}$ is computed from all the former updates of $\langle \alpha_m^{-1} \rangle$, including $\langle \alpha_m^{-1} \rangle_{(l-1)}, \langle \alpha_m^{-1} \rangle_{(l-2)}, \ldots, \langle \alpha_m^{-1} \rangle_{(0)}$. Thus the Double Lomax procedure leads to a smoother update trace. In other words, if $\langle v_m^2 \rangle_{(l)}$ is considered as a smooth regularization, then the regularization effect can be propagated to other update variables by the alternating-update strategy of VB inference, leading to other's smoother update traces. Consequently, the smooth regularization effect may prevent the algorithm from a large deviation of the right direction when there are some disturbances or unreliable updates during the process of convergence.

5. EXPERIMENTS

In this section, the proposed VB-Double Lomax formulation (denoted with DL) is compared with the equivalent formulation with ARD prior (denoted with ARD). Furthermore, a hybrid formulation (denoted by Hybrid), which with first K iterations employing DL and the rest employing ARD until convergence is also considered here for a trade-off between performance and speed. This hybrid approach is based on the observation that ARD tends to converge to a local minima at early iterations.

In the experiments reported below, the hybrid number $K = 20$. Following default CS setup is used. $\mathbf{\Phi}_{CS}$ is induced by sampling columns i.i.d. 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 uniform ± 1 random spikes. We fix $M = 256$, $T = 80$, and vary the number of measurements N. 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. Fig.1 reports the number of measurements versus reconstruction error with error ranges (one standard deviations). It is shown that DL provides improved overall reconstruction performance over ARD for a reasonable number of measurements, and Hybrid provides a favorable performance close to DL. Fig.2 reports the number of iterative steps required for convergence versus the number of measurements. DL shows a slower convergence rate than ARD, but Hybrid has a comparable convergence rate as ARD.

6. 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 CS reconstruction problem using Double Lomax priors. Compare to the canonical ARD update, the proposed update procedure has one more latent variable which has the smoothness effect resulting in an improved performance. A hybrid formulation of Double Lomax and ARD is also considered as a trade-off between performance and computational speed. Experiments show favors of the proposed approach.

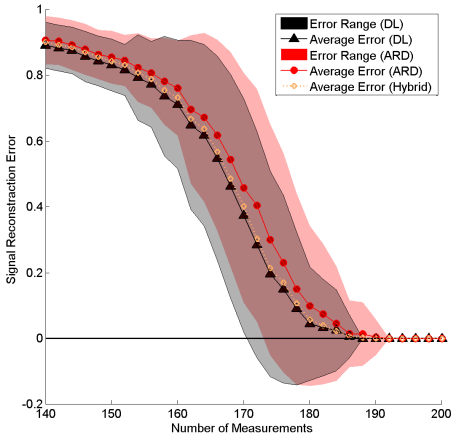

Fig. 1: Number of measurements versus reconstruction error with error ranges (one standard deviations).

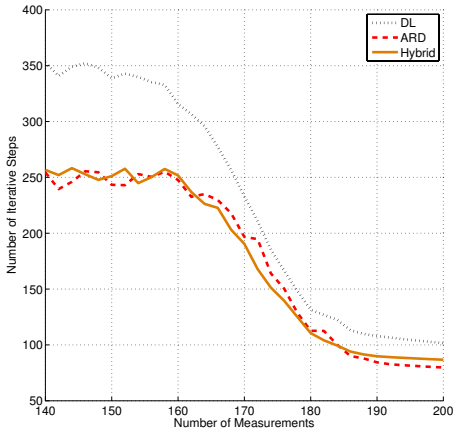

Fig. 2: The number of iterative step required for convergence.

7. ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China (No.61205017), China Postdoctoral Science Foundation funded project (No.2012M511058), and Shanghai Postdoctoral Sustentation Fund funded project (No.12R21412500).

8. REFERENCES

[1] D. L. Donoho, "Compressed sensing," *IEEE Trans. Information Theory*, vol. 52, no. 4, pp. 1289–1306, apr. 2006.

[2] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Information Theory*, vol. 52, no. 2, pp. 489 – 509, feb. 2006.

[3] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Information Theory*, vol. 51, no. 12, pp. 4203 – 4215, dec. 2005.

[4] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, Ting Sun, K. F. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 83–91, mar. 2008.

[5] A. Wagadarikar, R. John, R. Willett, and D. Brady, "Single disperser design for coded aperture snapshot spectral imaging," *Appl. Opt.*, vol. 47, no. 10, pp. B44–B51, apr. 2008.

[6] M. Rostami, O. Michailovich, and Z. Wang, "Image deblurring using derivative compressed sensing for optical imaging application," *IEEE Trans. Image Processing*, vol. 21, no. 7, pp. 3139 –3149, jul. 2012.

[7] A.S. Khwaja and J. Ma, "Applications of compressed sensing for sar moving-target velocity estimation and image compression," *IEEE Trans. Instrumentation and Measurement*, vol. 60, no. 8, pp. 2848 –2860, aug. 2011.

[8] J. Ma and F. X. Le Dimet, "Deblurring from highly incomplete measurements for remote sensing," *IEEE Trans. Geoscience and Remote Sensing*, vol. 47, no. 3, pp. 792 –802, mar. 2009.

[9] G. H. Chen, J. Tang, and S. Leng, "Prior image constrained compressed sensing (PICCS): A method to accurately reconstruct dynamic CT images from highly undersampled projection data sets," *Medical Physics*, vol. 35, no. 2, pp. 660–663, 2008.

[10] J. Ma, "Compressed sensing for surface characterization and metrology," *IEEE Trans. Instrumentation and Measurement*, vol. 59, no. 6, pp. 1600 –1615, jun. 2010.

[11] 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, 1999.

[12] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using focuss: a re-weighted minimum norm algorithm," *IEEE Trans. Signal Processing*, vol. 45, no. 3, pp. 600 –616, mar. 1997.

[13] E. J. Candes, M. B. Wakin, and S. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, dec. 2008.

[14] J.A. Tropp and A.C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Information Theory*, vol. 53, no. 12, pp. 4655 –4666, dec. 2007.

[15] D. L. Donoho, I. Drori, Y. Tsaig, and J. L. Starck, "Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit," Tech. Rep., 2006.

[16] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Processing*, vol. 52, no. 8, pp. 2153–2164, aug. 2004.

[17] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Processing*, vol. 56, no. 6, pp. 2346–2356, jun. 2008.

[18] S. Ji, D. Dunson, and L. Carin, "Multitask compressive sensing," *IEEE Trans. Signal Processing*, vol. 57, no. 1, pp. 92 –106, jan. 2009.

[19] F. Caron and A. Doucet, "Sparse bayesian nonparametric regression," in *International Conference on Machine Learning*, 2008.

[20] D. J. C. Mackay and C. Laboratory, "Bayesian nonlinear modeling for the prediction competition," in *ASHRAE Transactions*, 1994, vol. 100, pt. 2, pp. 1053–1062.

[21] C. Kleiber and S. Kotz, *Statistical size distributions in economics and actuarial sciences*, Wiley, Hoboken, NJ, 2003.

[22] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 1st edition, 2007.

[23] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, nov. 1999.

[24] D. J. C. MacKay, "Ensemble learning and evidence maximization," in *Advances in Neural Information Processing Systems*, 1995.

[25] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, jun. 2001.