Appendix to Hierarchical Infinite Divisibility for Multiscale Shrinkage

Xin Yuan, Vinayak Rao, Shaobo Han and Lawrence Carin
Department of Electrical and Computer Engineering Duke University,
Durham, NC 27708-0291 USA

Appendix I: Inference of the Model

This appendix presents the MCMC sampling, heuristically mean-field variational Bayesian (VB) [1] and Expectation-Maximization (EM) for posteriori distribution inference. The Generalized Inverse Gaussian (GIG) distribution is denoted by:

$$\mathrm{GIG}(x; a, b, p) = \frac{(a/b)^{\frac{p}{2}}}{2K_p(\sqrt{ab})} x^{p-1} \exp\left(-\frac{1}{2}(ax + \frac{b}{x})\right), \tag{1}$$

where $K_p(\theta)$ is the modified Bessel function of the second kind

$$K_p(\theta) = \int_0^\infty \frac{1}{2}\theta^{-p} t^{p-1} \exp\left(-\frac{1}{2}(t + \frac{\theta^2}{t})\right) dt. \tag{2}$$

A. Shrinkage Prior without Tree Structure (Flat Model)

For the simplification of the inference, we use $\gamma_i \sim \mathrm{Ga}(1/n, 1),\ \forall i = 1, \dots, n$; and $\tilde{\gamma}_i = \gamma_i / \sum_{i'} \gamma_{i'}$. We denote ith column of $\mathbf{\Psi}$ with $\mathbf{\Psi}_i$. The posteriori distributions for MCMC sampling are:

$$
\begin{aligned}
p(x_i|-) &\propto \mathcal{N}(\mu, \sigma^2), && (3)\\
\sigma^2 &= (\tau\alpha_i\alpha_0 + \alpha_0 \boldsymbol{\Psi}_i^T \boldsymbol{\Psi}_i)^{-1}, && (4)\\
\mu &= \alpha_0 \sigma^2 \boldsymbol{\Psi}_i^T \boldsymbol{y}_{-i}, && (5)\\
\boldsymbol{y}_{-i} &= \boldsymbol{y} - \boldsymbol{\Psi}\boldsymbol{x} + \boldsymbol{\Psi}_i x_i; && (6)\\
p(\alpha_i|-) &\propto \mathrm{GIG}(x_i^2\tau\alpha_0, \frac{1}{\gamma_i}, -0.5); && (7)\\
p(\gamma_i|-) &\propto \mathrm{GIG}(2, \frac{1}{\alpha_i}, \frac{1}{n} - 1); && (8)\\
p(\tau|-) &\propto \mathrm{Ga}(c_0 + \frac{1}{2}n, d_0 + \frac{1}{2}\sum_i x_i^2 \alpha_i \alpha_0); && (9)\\
p(\alpha_0|-) &\propto \mathrm{Ga}(a_0 + \frac{1}{2}n + \frac{1}{2}m, b_0 + \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\Psi}\boldsymbol{x}\|_2^2 + \frac{1}{2}\sum_i x_i^2 \tau \alpha_i). && (10)
\end{aligned}
$$

The variational Bayesian update equations are:

$$\begin{aligned}
\langle x_i \rangle &= \mu_{x_i} = \sigma_{x_i}^2 \langle \alpha_0 \rangle \boldsymbol{\Psi}_i^T (\boldsymbol{y} - \sum_{l=1, l\neq i}^{N} \boldsymbol{\Psi}_l \langle x_l \rangle), && (11)\\
\sigma_{x_i}^2 &= \langle \alpha_0 \rangle^{-1} \left(\langle \tau \rangle \langle \alpha_i \rangle + \boldsymbol{\Psi}_i^T \boldsymbol{\Psi}_i \right)^{-1}, \\
\langle x_i^2 \rangle &= \mu_{x_i}^2 + \sigma_{x_i}^2; && (12)\\
\langle \alpha_i \rangle &= \frac{\sqrt{\langle \frac{1}{\gamma_i} \rangle} K_{0.5}(\sqrt{\langle x_i^2 \rangle \langle \tau \rangle \langle \alpha_0 \rangle \langle \frac{1}{\gamma_i} \rangle})}{\sqrt{\langle x_i^2 \rangle \langle \tau \rangle \langle \alpha_0 \rangle} K_{-0.5}(\sqrt{\langle x_i^2 \rangle \langle \tau \rangle \langle \alpha_0 \rangle \langle \frac{1}{\gamma_i} \rangle})}, && (13)\\
\langle \frac{1}{\alpha_i} \rangle &= \frac{\sqrt{\langle x_i^2 \rangle \langle \tau \rangle \langle \alpha_0 \rangle} K_{1.5}(\sqrt{\langle x_i^2 \rangle \langle \tau \rangle \langle \alpha_0 \rangle \langle \frac{1}{\gamma_i} \rangle})}{\sqrt{\langle \frac{1}{\gamma_i} \rangle} K_{0.5}(\sqrt{\langle x_i^2 \rangle \langle \tau \rangle \langle \alpha_0 \rangle \langle \frac{1}{\gamma_i} \rangle})}; && (14)\\
\langle \gamma_i \rangle &= \frac{\sqrt{\langle \frac{1}{\alpha_i} \rangle} K_{1/n}(\sqrt{2 \langle \frac{1}{\alpha_i} \rangle})}{\sqrt{2} K_{1/n-1}(\sqrt{2 \langle \frac{1}{\alpha_i} \rangle})}, && (15)\\
\langle \frac{1}{\gamma_i} \rangle &= \frac{\sqrt{2} K_{2-1/n}(\sqrt{2 \langle \frac{1}{\alpha_i} \rangle})}{\sqrt{\langle \frac{1}{\alpha_i} \rangle} K_{1-1/n}(\sqrt{2 \langle \frac{1}{\alpha_i} \rangle})}; && (16)\\
\langle \tau \rangle &= \frac{c_0 + 0.5n}{d_0 + \frac{1}{2} \sum_i \langle x_i^2 \rangle \langle \alpha_i \rangle \langle \alpha_0 \rangle}; && (17)\\
\langle \alpha_0 \rangle &= \frac{a_0 + 0.5n + 0.5m}{b_0 + \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{\Psi} \langle \boldsymbol{x} \rangle\|_2^2 + \frac{1}{2} \mathrm{trace} \left(\boldsymbol{\Psi}^T \boldsymbol{\Psi} [\langle \boldsymbol{x}\boldsymbol{x}^T \rangle - \langle \boldsymbol{x} \rangle \langle \boldsymbol{x}^T \rangle] \right) + \frac{1}{2} \sum_i \langle x_i^2 \rangle \langle \tau \rangle \langle \alpha_i \rangle} && (18)\\
&\approx \frac{a_0 + 0.5n}{b_0 + \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{\Psi} \langle \boldsymbol{x} \rangle\|_2^2 + \frac{1}{2} \sum_{i=1}^{n} \boldsymbol{\Psi}_i^T \boldsymbol{\Psi}_i [\langle x_i^2 \rangle - \langle x_i \rangle^2] + \frac{1}{2} \sum_i \langle x_i^2 \rangle \langle \tau \rangle \langle \alpha_i \rangle}, && (19)
\end{aligned}$$

where the approximation is used as the same as in [2], and for equations (14),(16), please refer to [3].

For the EM estimation, if we want a point estimate of γ_i, we can get it by the mode of GIG distribution:

$$\gamma_i = \frac{(\frac{1}{n} - 2) + \sqrt{(\frac{1}{n} - 2)^2 + \frac{2}{\alpha_i}}}{2}. \quad (20)$$

B. Tree Structure Model (s-HM Model)

Considering we have n_ℓ wavelet coefficients at level $\ell = 0, \ldots, L$, and the ith element at level ℓ of the wavelet coefficient is $x_{\ell,i}$. We denote the kth column of the matrix $\boldsymbol{\Psi}$ as $\boldsymbol{\Psi}_k$ corresponding to the ith element at level ℓ of the wavelet coefficient.

For MCMC sampling, the posteriori distributions are:

$$\begin{aligned} p(x_{0,i}|-) &\propto \mathcal{N}(x_{0,i}|\mu_{x_{0,i}}, \sigma^2_{x_{0,i}}), &(21)\\ \sigma^2_{x_{0,i}} &= \alpha_0^{-1}(\tau_0 + \boldsymbol{\Psi}_k^T \boldsymbol{\Psi}_k)^{-1}, &(22)\\ \mu_{x_{0,i}} &= \alpha_0 \sigma^2_{x_{0,i}} \boldsymbol{\Psi}_k^T(\boldsymbol{y} - \boldsymbol{\Psi}\boldsymbol{x} + \boldsymbol{\Psi}_k x_k); &(23)\\ p(\tau_0|-) &\propto \text{Ga}(c_0 + 0.5N_0, d_0 + 0.5\sum_{i=1}^{N_0} x^2_{0,i}\alpha_0); &(24)\\ p(x_{\ell,i}|-) &\propto \mathcal{N}(\mu_{x_{\ell,i}}, \sigma^2_{x_{\ell,i}}), &(25)\\ \sigma^2_{x_{\ell,i}} &= \alpha_0^{-1}(\tau_\ell \alpha_{\ell,i} + \boldsymbol{\Psi}_k^T \boldsymbol{\Psi}_k)^{-1}, &(26)\\ \mu_{x_{\ell,i}} &= \alpha_0 \sigma^2_{x_{\ell,i}} \boldsymbol{\Psi}_k^T(\boldsymbol{y} - \boldsymbol{\Psi}\boldsymbol{x} + \boldsymbol{\Psi}_k x_{\ell,i}),\ \forall \ell = 1, \ldots, L; &(27) \end{aligned}$$

$$\begin{aligned} p(\alpha_{\ell,i}|-) &\propto \text{GIG}\left(x^2_{\ell,i}\tau_\ell\alpha_0, \frac{1}{\gamma_{\ell,i}}, -0.5\right); &(28)\\ p(\tau_\ell|-) &\propto \text{Ga}\left(c_0 + 0.5n_\ell, d_0 + 0.5\sum_{i=1}^{n_\ell} x^2_{\ell,i}\alpha_{\ell,i}\alpha_0\right); &(29) \end{aligned}$$

In order to sample γ_ℓ, recall:

$$\begin{aligned} \gamma_{\ell,i} &\sim \text{Ga}(T^{(\ell-1)}_{pa(\ell,i)}/n_c, 1), &(30)\\ \tilde{\gamma}_{\ell,i} &= \frac{\gamma_{\ell,i}}{\sum_{i'} \gamma_{\ell,i'}}. &(31) \end{aligned}$$

Note $\tilde{\boldsymbol{\gamma}}_\ell$ is the normalization form of $\boldsymbol{\gamma}_\ell$ and $T^{(\ell-1)}_{pa(\ell,i)} = \tilde{\gamma}_{(\ell-1),pa(\ell,i)}$ is the normalization form of $\gamma_{\ell-1}$. If $\ell = 1$, $T^{(0)}_{pa(\ell,i)} = 1$ and $n_c = n_1$.

For $\ell = 1, \ldots, L-1$, from

$$p(\tilde{\gamma}_{\ell,i}|-) = \text{InvGa}(\alpha_{\ell,i}|1, (2\tilde{\gamma}_{\ell,i})^{-1})\text{Dir}(\tilde{\boldsymbol{\gamma}}_\ell|\tilde{\boldsymbol{\gamma}}_{\ell-1})\text{Dir}(\tilde{\boldsymbol{\gamma}}_{\ell+1}|\tilde{\boldsymbol{\gamma}}_\ell), \tag{32}$$

we have

$$\begin{aligned} p(\gamma_{\ell,i}|-) &= \text{InvGa}\left(\alpha_{\ell,i}|1, \frac{\sum_i \gamma_{\ell,i}}{2\gamma_i}\right) \text{Ga}\left(\gamma_\ell|\tilde{\gamma}_{(\ell-1),pa(\ell,i)}/n_c, 1\right) \text{Dir}\left(\tilde{\boldsymbol{\gamma}}_{\ell+1}|\tilde{\boldsymbol{\gamma}}_\ell\right)\\ &\propto \text{GIG}\left(2, \frac{\sum_{j\neq i}\gamma_{\ell,j}}{\alpha_{\ell,i}}, \frac{\tilde{\gamma}_{(\ell-1),pa(\ell,i)}}{n_c} - 1\right) \text{Dir}\left(\tilde{\boldsymbol{\gamma}}_{\ell+1}|\tilde{\boldsymbol{\gamma}}_\ell\right) \sum_j \gamma_{\ell,j}. &(33) \end{aligned}$$

Similarly, at layer L, $\gamma_{L,i}$ does not have children,

$$p(\gamma_{L,i}|-) \propto \text{GIG}\left(2, \frac{\sum_{j\neq i}\gamma_{L,j}}{\alpha_i}, \frac{\tilde{\gamma}_{(L-1),pa(L,i)}}{n_c} - 1\right) \sum_j \gamma_{L,j}. \tag{34}$$

We here use the Metropolis-Hastings (MH) algorithm [4] to sample $\gamma_{\ell,i}$. We propose to use the distribution Q of generalized-inverse-Gaussian (the first term in (33)-(34)) in a MH independence chain and accept $\boldsymbol{\gamma}_\ell^{*\ t+1} = \boldsymbol{\gamma}_\ell^{*\ \prime}$ with probability $\min\{p_\ell, 1\}$, where

$$p_\ell = \frac{\prod_{i=1}^{n_\ell} p(\gamma_{\ell,i}^{*\ \prime})}{\prod_{i=1}^{n_\ell} p(\gamma_{\ell,i}^{*\ t})} \frac{\prod_{i=1}^{n_\ell} Q(\gamma_{\ell,i}^{*\ t})}{\prod_{i=1}^{n_\ell} Q(\gamma_{\ell,i}^{*\ \prime})}. \tag{35}$$

In the experiments, we found the accept ratio of the proposed distribution is around 80%, and we update the $\boldsymbol{\gamma}$ level-by-level.

In the following variational Bayesian (VB) inference, we use the mean of this generalized-inverse-Gaussian to approximate the mean value of $\boldsymbol{\gamma}$. We can also use the Monte Carlo integration to sample $\gamma_{\ell,i}$ and then estimate the mean values. The VB update equations different from the flat model are:

$$\langle x_{0,i} \rangle = \sigma_{x_{0,i}}^2 \langle \alpha_0 \rangle \boldsymbol{\Psi}_k^T \left(\boldsymbol{y} - \sum_{l=1, l\neq k}^{n} \boldsymbol{\Psi}_l \langle x_l \rangle \right) = \mu_{x_{0,i}}, \tag{36}$$

$$\sigma_{x_{0,i}}^2 = \langle \alpha_0^{-1} \rangle \left(\langle \tau_0 \rangle + \boldsymbol{\Psi}_k^T \boldsymbol{\Psi}_k \right)^{-1}, \tag{37}$$

$$\langle x_{0,i}^2 \rangle = \mu_{x_{0,i}}^2 + \sigma_{x_{0,i}}^2; \tag{38}$$

$$\langle \tau_0 \rangle = \frac{c_0 + 0.5 n_0}{d_0 + 0.5 \sum_{i=1}^{n_0} \langle x_{0,i}^2 \rangle \langle \alpha_0 \rangle}; \tag{39}$$

$$\langle x_{\ell,i} \rangle = \langle \alpha_0 \rangle \sigma_{x_{\ell,i}}^2 \boldsymbol{\Psi}_k^T \left(\boldsymbol{y} - \sum_{l=1, l\neq k}^{n} \boldsymbol{\Psi}_l \langle x_l \rangle \right), \tag{40}$$

$$\sigma_{x_{\ell,i}}^2 = \langle \alpha_0 \rangle^{-1} (\langle \tau_\ell \rangle \langle \alpha_{\ell,i} \rangle + \boldsymbol{\Psi}_k^T \boldsymbol{\Psi}_k)^{-1}, \tag{41}$$

$$\langle x_{\ell,i}^2 \rangle = \langle x_{\ell,i} \rangle^2 + \sigma_{x_{\ell,i}}^2, \quad \forall \ell = 1, \ldots, L; \tag{42}$$

$$\langle \alpha_{\ell,i} \rangle = \frac{\sqrt{\langle \frac{1}{\gamma_{\ell,i}} \rangle} K_{0.5}(\sqrt{\langle x_{\ell,i}^2 \rangle \langle \tau_\ell \rangle \langle \alpha_0 \rangle \langle \frac{1}{\gamma_{\ell,i}} \rangle})}{\sqrt{\langle x_{\ell,i}^2 \rangle \langle \tau_\ell \rangle \langle \alpha_0 \rangle} K_{-0.5}(\sqrt{\langle x_{\ell,i}^2 \rangle \langle \tau_\ell \rangle \langle \alpha_0 \rangle \langle \frac{1}{\gamma_{\ell,i}} \rangle})}, \tag{43}$$

$$\langle \frac{1}{\alpha_{\ell,i}} \rangle = \frac{\sqrt{\langle x_{\ell,i}^2 \rangle \langle \tau_\ell \rangle \langle \alpha_0 \rangle} K_{1.5}(\sqrt{\langle x_{\ell,i}^2 \rangle \langle \tau_\ell \rangle \langle \alpha_0 \rangle \langle \frac{1}{\gamma_{\ell,i}} \rangle})}{\sqrt{\langle \frac{1}{\gamma_{\ell,i}} \rangle} K_{0.5}(\sqrt{\langle x_{\ell,i}^2 \rangle \langle \tau_\ell \rangle \langle \alpha_0 \rangle \langle \frac{1}{\gamma_{\ell,i}} \rangle})}; \tag{44}$$

$$\langle \gamma_{\ell,i} \rangle = \frac{\sqrt{\langle \frac{1}{\alpha_i} \rangle \sum_{j\neq i} \langle \gamma_{\ell,j} \rangle} K_{\tilde{\gamma}_{(\ell-1),pa(\ell,i)}/n_c}(\sqrt{2 \langle \frac{1}{\alpha_{\ell,i}} \rangle \sum_{j\neq i} \langle \gamma_{\ell,j} \rangle})}{\sqrt{2} K_{(\tilde{\gamma}_{(\ell-1),pa(\ell,i)}/n_c - 1)}(\sqrt{2 \langle \frac{1}{\alpha_{\ell,i}} \rangle \sum_{j\neq i} \langle \gamma_{\ell,j} \rangle})}, \tag{45}$$

$$\langle \frac{1}{\gamma_{\ell,i}} \rangle = \frac{\sqrt{2} K_{(2-\tilde{\gamma}_{(\ell-1),pa(\ell,i)}/n_c)}(\sqrt{2 \langle \frac{1}{\alpha_{\ell,i}} \rangle \sum_{j\neq i} \langle \gamma_{\ell,j} \rangle})}{\sqrt{\langle \frac{1}{\alpha_i} \rangle \sum_{j\neq i} \langle \gamma_{\ell,j} \rangle} K_{(1-\tilde{\gamma}_{(\ell-1),pa(\ell,i)}/n_c)}(\sqrt{2 \langle \frac{1}{\alpha_{\ell,i}} \rangle \sum_{j\neq i} \langle \gamma_{\ell,j} \rangle})}; \tag{46}$$

$$\langle \tau_\ell \rangle = \frac{c_0 + 0.5 n_\ell}{d_0 + 0.5 \sum_{i=1}^{n_\ell} \langle (x_i^{(\ell)})^2 \rangle \langle \alpha_{\ell,i} \rangle \langle \alpha_0 \rangle}, \forall \ell = 1, \ldots, L. \tag{47}$$

For the point estimation of $\gamma_{\ell,i}$ with EM, we use the mode of GIG:

$$\gamma_{\ell,i} = \frac{(\tilde{\gamma}_{(\ell-1),pa(\ell,i)}/n_c - 2) + \sqrt{(\tilde{\gamma}_{(\ell-1),pa(\ell,i)}/n_c - 2)^2 + \frac{2\sum_{j \neq i} \gamma_{\ell,j}}{\alpha_{\ell,i}}}}{2}. \tag{48}$$

REFERENCES

[1] M. J. Beal, "Variational algorithms for approximate bayesian inference," *Ph.D. dissertation, University College London*, 2003.

[2] L. He, H. Chen, and L. Carin, "Tree-structured compressive sensing with variational bayesian analysis," *IEEE Signal Processing Letters*, vol. 17, no. 3, pp. 233–236, 2010.

[3] A. Armagan, M. Clyde, and D. B. Dunson, "Generalized beta mixtures of gaussians," *Advances in Neural Information Processing Systems (NIPS)*, pp. 523–531, 2011.

[4] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," *Biometrika*, pp. 97–109, 1970.