Horseshoe Estimator for Constrained Normal Means

We consider the model $\boldsymbol{y}|\boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ where $\boldsymbol{\mu} \in \mathcal{K} = \{\boldsymbol{\mu} : \boldsymbol{A}\boldsymbol{\mu} \geq 0\}$. First we consider a straightforward generalization of the global-local prior to the constrained case when the prior on $\boldsymbol{\mu}$ is supported on $\mathcal{K}$. Extension of the two-component mixture prior to the convex cone restriction is more nuanced and is discussed second.

2.1 Horseshoe Extension

Specifically, let

$$\begin{aligned} \boldsymbol{\mu}|\boldsymbol{\lambda}, \tau &\sim TN(0, \tau^2 \boldsymbol{\Lambda}, \mathcal{K}) \\ (\lambda_1, \ldots, \lambda_n) &\sim \prod_{i=1}^{n} p_\lambda(\lambda_i) \\ \tau &\sim p_\tau(\tau) \end{aligned} \tag{1}$$

where $\boldsymbol{\Lambda} = diag\{\lambda_1^2, \ldots, \lambda_n^2\}$ and $TN(\psi, \boldsymbol{\Sigma}, \mathcal{K})$ denotes normal with the distribution of a multivariate normal with mean ψ, variance matrix $\boldsymbol{\Sigma}$ and truncated to the cone $\mathcal{K}$.

It is interesting is to investigate the effect of the conic geometry on the Bayes estimates. Of course, the truncated normal prior will be conjugate, yielding a truncated normal posterior. If the model is $\boldsymbol{y}|\boldsymbol{\mu} \sim N(\boldsymbol{0}, \boldsymbol{\Omega})$, and the prior density for $\boldsymbol{\mu}$ is $TN(\boldsymbol{0}, \boldsymbol{V}, \mathcal{K})$ and $\boldsymbol{Q} = (\boldsymbol{\Omega}^{-1} + \boldsymbol{V}^{-1})^{-1}$, then the posterior is $\boldsymbol{\mu}|\boldsymbol{y} \sim TN(\boldsymbol{Q}\boldsymbol{\Omega}^{-1}\boldsymbol{y}, \boldsymbol{Q}, \mathcal{K})$. Thus, the posterior mean could be directly computed for that of a truncated normal, albeit truncated to a general convex polyhedral cone. We derive the expression for the posterior mean using a slightly different argument which is instructive in the sense it provides explicit expressions for the marginal of $\boldsymbol{y}$ using hidden truncation argument.

Before we give our main result, we define some useful notation. Let $\Phi^{(r)}(\boldsymbol{z}; \boldsymbol{\xi}, \boldsymbol{W}) = P(\boldsymbol{Z} \leq \boldsymbol{z})$ for $\boldsymbol{Z} \sim N(\boldsymbol{\xi}, \boldsymbol{W})$ where Φ is the standard normal cdf. Also, for $\boldsymbol{x} = (x_1, \ldots, x_n)'$, let $\phi^{(n)}(\boldsymbol{x}) = \prod_{i=1}^{n} \phi(x_i)$ where ϕ is the standard normal pdf.

The following result provides some insight to how the half-plane restrictions, $\boldsymbol{A}$ appears in the expression for the posterior mean.

Theorem 1. *Let $\boldsymbol{y}|\boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \boldsymbol{\Omega})$ where it is known a priori that $\boldsymbol{\mu} \in \mathcal{K}$, a polyhedral convex cone defined by $\mathcal{K} = \{\boldsymbol{\mu} : \boldsymbol{A}\boldsymbol{\mu} \geq \boldsymbol{0}\}$ for some matrix $\boldsymbol{A}$ of dimension $k \times n$. Let the prior on $\boldsymbol{\mu}$ be $\boldsymbol{\mu} \sim N(0, \boldsymbol{V})_{\mathcal{K}}$. Let $\boldsymbol{F} = \boldsymbol{A}\boldsymbol{Q}\boldsymbol{A}' = \boldsymbol{D}\boldsymbol{R}\boldsymbol{D}'$, say, where $\boldsymbol{D}$ be a diagonal matrix with entries equal to the square root of the diagonal entries of $\boldsymbol{F}$ and $\boldsymbol{Q} = (\boldsymbol{\Omega}^{-1} + \boldsymbol{V}^{-1})^{-1}$. Also for $i = 1, \ldots, k$, let $\boldsymbol{R}_{-i}$ be $\boldsymbol{R}$ without the ith column and the ith row and let $\boldsymbol{r}_{-i}$ denote the ith column of $\boldsymbol{R}$ without the ith diagonal element. Let $\boldsymbol{B}_i = [\boldsymbol{I} : -\boldsymbol{r}_{-i} r_i^{-1}]$ where $\boldsymbol{I}$ is the identity matrix of dimension $(k-1)$ and let $\boldsymbol{u} = (u_1, \ldots, u_k)' = \boldsymbol{D}^{-1}\boldsymbol{A}\boldsymbol{Q}\boldsymbol{\Omega}^{-1}\boldsymbol{y}$. Assuming $\boldsymbol{\Omega}$ and $\boldsymbol{V}$ are fixed and given, we have*

$$E(\boldsymbol{\mu}|\boldsymbol{y}) = \boldsymbol{Q}[\boldsymbol{\Omega}^{-1}\boldsymbol{y} + \boldsymbol{A}'\boldsymbol{D}^{-1}\boldsymbol{v}]$$

where $\boldsymbol{v} = (v_1, \ldots, v_k)'$ and $v_i = \phi(-u_i)\Phi^{(k-1)}(\boldsymbol{B}_i\boldsymbol{u}; \boldsymbol{0}, \boldsymbol{R}_{-i} - r_i^{-1}\boldsymbol{r}_{-i}\boldsymbol{r}_{-i}^T)/\Phi^{(k)}(\boldsymbol{u}; \boldsymbol{0}, \boldsymbol{R})$.

Corresponding Author: Anindya Roy
Email: anindya@umbc.edu

Proof. The joint model for $\boldsymbol{y}$ and $\boldsymbol{\mu}$ is

$$\begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{\mu}\end{pmatrix} \sim N\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix} \quad \begin{pmatrix}\boldsymbol{\Omega}+\boldsymbol{V} & \boldsymbol{V}\\ \boldsymbol{V} & \boldsymbol{V}\end{pmatrix}\right)$$

Hence that of $\boldsymbol{y}$ and $\boldsymbol{A\mu}$ is

$$\begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{\mu}\end{pmatrix} \sim N\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix} \quad \begin{pmatrix}\boldsymbol{\Omega}+\boldsymbol{V} & \boldsymbol{VA}'\\ \boldsymbol{AV} & \boldsymbol{AVA}'\end{pmatrix}\right)$$

Then following Arnold (2009), the marginal density formula for $\boldsymbol{y}$ under the hidden truncation $\boldsymbol{A\mu} \geq \mathbf{0}$, is

$$p_y(\boldsymbol{y}) = |\boldsymbol{\Sigma}_{11}|^{-1/2}\phi^{(n)}(\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{y})\frac{\Phi^{(k)}(-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{y};\mathbf{0},\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})}{\Phi^{(k)}(\mathbf{0};\mathbf{0},\boldsymbol{\Sigma}_{22})}$$

where

$$\begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Omega}+\boldsymbol{V} & \boldsymbol{VA}'\\ \boldsymbol{AV} & \boldsymbol{AVA}'\end{pmatrix}.$$

Then by the multiparameter version of Tweedie's formula (Robbins, 1956), we have

$$\mathrm{E}(\boldsymbol{\mu}|\boldsymbol{y}) = \boldsymbol{y} + \boldsymbol{\Omega}\nabla_{\boldsymbol{y}}\log p_y(\boldsymbol{y})$$

The gradient of $\log p_y(\boldsymbol{y})$ has two parts, The first part is $\nabla_{\boldsymbol{y}}\log\phi^{(n)}(\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{y})$. By the chain rule of vector differentiation, we get

$$\nabla_{\boldsymbol{y}}\log\phi^{(n)}(\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{y}) = -\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{y} = -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{y} = -(\boldsymbol{\Omega}+\boldsymbol{V})^{-1}\boldsymbol{y}.$$

Therefore,

$$\begin{aligned}\mathrm{E}(\boldsymbol{\mu}|\boldsymbol{y}) &= \boldsymbol{y} - \boldsymbol{\Omega}(\boldsymbol{\Omega}+\boldsymbol{V})^{-1}\boldsymbol{y} + \boldsymbol{\Omega}\nabla_{\boldsymbol{y}}\log\Phi^{(k)}(-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{y};\mathbf{0},\boldsymbol{F}),\\ &= \boldsymbol{Q}\boldsymbol{\Omega}^{-1}\boldsymbol{y} + \boldsymbol{\Omega}\nabla_{\boldsymbol{y}}\log\Phi^{(k)}(-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{y};\mathbf{0},\boldsymbol{F}),\end{aligned}$$

We further note that

$$\begin{aligned}\nabla_{\boldsymbol{y}}\log\Phi^{(k)}(-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{y};\mathbf{0},\boldsymbol{F}) &= \nabla_{\boldsymbol{y}}\log\Phi^{(k)}(\boldsymbol{AQ\Omega}^{-1}\boldsymbol{y};\mathbf{0},\boldsymbol{F})\\ &= \nabla_{\boldsymbol{y}}\log\Phi^{(k)}(\boldsymbol{D}^{-1}\boldsymbol{AQ\Omega}^{-1}\boldsymbol{y};\mathbf{0},\boldsymbol{R})\\ &= \boldsymbol{\Omega}^{-1}\boldsymbol{QA}'\boldsymbol{D}^{-1}\ \nabla_{\boldsymbol{u}}\log\Phi^{(k)}(\boldsymbol{u};0,\boldsymbol{R})\\ &= \boldsymbol{\Omega}^{-1}\boldsymbol{QA}'\boldsymbol{D}^{-1}\ \frac{\nabla_{\boldsymbol{u}}\Phi^{(k)}(\boldsymbol{u};0,\boldsymbol{R})}{\Phi^{(k)}(\boldsymbol{u};0,\boldsymbol{R})}\end{aligned}$$

To compute the gradient we use the standard formula for partial derivatives of the multivariate cdf of a random vector $\boldsymbol{X} = (X_i,\ldots,X_k)$ given by $\frac{\partial}{\partial x_i}F(x_1,\ldots,x_k) = f_i(x_i)F_{-i|i}(x_{-i})$ where f_i is the marginal density of X_i, $F_{-i|i}$ is the conditional cdf of the rest of the components of $\boldsymbol{X}$ given X_i and $\boldsymbol{x}_{-i}$ is the vector $\boldsymbol{x} = (x_1,\ldots,x_k)$ without the ith component. Using the conditional distribution of multivariate normal and the fact that $\boldsymbol{B}_i\boldsymbol{u} = \boldsymbol{u}_{-i} - \boldsymbol{r}_{-i}r_i^{-1}u_i$, we have

$$\begin{aligned}\frac{\partial}{\partial u_i}\Phi^{(k)}(\boldsymbol{u};\mathbf{0},\boldsymbol{R}) &= \phi(u_i)\ \Phi^{(k-1)}(\boldsymbol{u}_{-i};\ \boldsymbol{r}_{-i}r_i^{-1}u_i,\ \boldsymbol{R}_{-i} - r_i^{-1}\boldsymbol{r}_{-i}\boldsymbol{r}_{-i}^T)\\ &= \phi(u_i)\ \Phi^{(k-1)}(\boldsymbol{B}_i\boldsymbol{u};\ 0,\ \boldsymbol{R}_{-i} - r_i^{-1}\boldsymbol{r}_{-i}\boldsymbol{r}_{-i}^T)\end{aligned}$$

Therefore,

$$\mathrm{E}(\boldsymbol{\mu}|\boldsymbol{y}) \quad = \quad \boldsymbol{Q}\boldsymbol{\Omega}^{-1}\boldsymbol{y} + \boldsymbol{Q}\boldsymbol{A}\boldsymbol{D}^{-1}\ \phi(u_i)\ \frac{\Phi^{(k-1)}(\boldsymbol{B}_i\boldsymbol{u};\ 0,\ \boldsymbol{R}_{-i} - r_i^{-1}\boldsymbol{r}_{-i}\boldsymbol{r}_{-i}^T)}{\Phi^{(k)}(\boldsymbol{u}; 0, \boldsymbol{R})}.$$

□

The expression for the posterior mean has two parts. The first part $\boldsymbol{Q}\boldsymbol{\Omega}^{-1}\boldsymbol{y}$ is the usual Bayes estimator normal-normal conjugacy which is the unbiased estimator $\boldsymbol{y}$ plus a Bayes correction. However, under the conic constraint the second term acts as a correction for the restriction to the convex cone.

Let the entries of the covariance matrices be functions of some lower dimensional parameters $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_d\}$. For example, for the usual Horseshoe prior formulation, $\boldsymbol{\theta} = \{\sigma^2, \lambda_1^2, \ldots, \lambda_n^2, \tau^2)$. Even though the expression for the posterior mean in Theorem 1 is derived with fixed $\boldsymbol{\Omega}$ and $\boldsymbol{V}$, it is instructive to write the posterior mean as $\mathrm{E}(\boldsymbol{\mu}|\boldsymbol{y}, \boldsymbol{\theta})$. If priors are specified on $\boldsymbol{\theta}$, then the posterior mean for $\boldsymbol{\mu}$ an be obtained as

$$\mathrm{E}(\boldsymbol{\mu}|\boldsymbol{y}) = \mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{y}}(\mathrm{E}(\boldsymbol{\mu}|\boldsymbol{y}, \boldsymbol{\theta})),$$

where the first expectation on the right hand side is taken over the marginal posterior of $\boldsymbol{\theta}$.

The marginal distribution of $\boldsymbol{y}$ given the truncated normal prior is $p_y(\boldsymbol{y})$ and it belonds to the *closed (fundamental) skew normal* family; see Gonzalez-Farias *et al.* (2004) (**?**), Arellano-Vallea and Genton, (2005) (**?**). The marginal distribution can be used for estimation of hyper-parameter to obtain the marginal posterior of $\boldsymbol{\theta}$. For example, one could use the fundamental skew normal likelihood directly.