Generalized Beta Mixture of Gaussians

Artin Armagan, David B. Dunson and Merlise Clyde

Presented by Shaobo Han, Duke University
July 22, 2013
Two alternatives in Bayesian sparse learning:

- **Point mass mixture priors**
  \[
  \theta_j \sim (1 - \pi)\delta_0 + \pi g_{\theta}, \quad j = 1, \ldots, p
  \]  

- **Continuous shrinkage priors**
  - Inducing regularization penalties
  - Global-local shrinkage [N. G. Polson and J. G. Scott (2010)]
  \[
  \theta_j \sim \mathcal{N}(0, \psi_j \tau), \quad \psi_j \sim f, \tau \sim g
  \]  
e.g. the *normal-exponential-gamma*, the *horseshoe* and the *generalized double Pareto*.
  - Computationally attractive

---

[1] Shrinkage globally, act locally: sparse Bayesian regularization and prediction, Bayesian statistics, 2010
The horseshoe estimator

[Sparse normal-means problem] Suppose a \( p \)-dimensional vector \( y|\theta \sim \mathcal{N}(\theta, I) \) is observed.

\[
\theta_j|\tau_j \sim \mathcal{N}(0, \tau_j), \quad \tau_j^{1/2} \sim \mathcal{C}^+(0, \phi^{1/2}), \quad j = 1, \ldots, p
\]  

(3)

With an appropriate transformation \( \rho_j = 1/(1 + \tau_j) \),

\[
\theta_j|\rho_j \sim \mathcal{N}(0, 1/\rho_j - 1), \quad \pi(\rho_j|\phi) \propto \rho_j^{-1/2}(1 - \rho_j)^{-1/2} \frac{1}{1 + (\phi - 1)\rho_j}
\]  

(4)

For fixed values \( \phi = 1 \) [C. M. Carvalho, et. al. (2010)]\(^2\),

\[
\mathbb{E}(\theta_j|y) = \int_0^1 y_j(1 - \rho_j)\rho(\rho_j|y)d\rho_j = \{1 - \mathbb{E}(\rho_j|y)\}y_j
\]  

(5)

Two major issues: robustness to large signals and shrinkage of noise.

- Horseshoe prior: \( \phi = 1, \rho_j \sim \mathcal{B}(1/2, 1/2) \).
- Strawderman-Berger prior: \( \phi = 1, \rho_j \sim \mathcal{B}(1, 1/2) \).

\(^2\)The horseshoe estimator for sparse signals, Biometrika, 2010
TPB normal scale mixture prior

**Definition 1**

The three-parameter beta (TPB) distribution for a random variable $X$ is defined by the density function

$$f(x; a, b, \phi) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \phi^b x^{b-1}(1 - x)^{a-1} \{1 + (\phi - 1)x\}^{-(a+b)}$$

for $0 < x < 1$, $a > 0$, $b > 0$ and $\phi > 0$ and is denoted by $\mathcal{TPB}(a, b, \phi)$.

**Definition 2**

The TPB normal scale mixture representation for the distribution of random variable $\theta_j$ is given by

$$\theta_j | \rho_j \sim \mathcal{N}(0, 1/\rho_j - 1), \quad \rho_j \sim \mathcal{TPB}(a, b, \phi)$$

where $a > 0$, $b > 0$ and $\phi > 0$. The resulting marginal distribution on $\theta_j$ is denoted by $\mathcal{TPBN}(a, b, \phi)$. 
Shrinkage effects for different values of $a$, $b$ and $\phi$

\[
\theta_j | \rho_j \sim \mathcal{N}(0, 1/\rho_j - 1), \quad \rho_j \sim \mathcal{T}\mathcal{P}\mathcal{B}(a, b, \phi)
\]  \hspace{1cm} \text{(8)}

(a,b)=(1/2,1/2) \hspace{2cm} (a,b)=(1,1/2) \hspace{2cm} (a,b)=(1,1)

(a,b)=(1/2,2) \hspace{2cm} (a,b)=(2,2) \hspace{2cm} (a,b)=(5,2)

Figure 1:
\[\phi = \{1/10, 1/9, 1/8, 1/7, 1/6, 1/5, 1/4, 1/3, 1/2, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}\]
c onsidered for all pairs of $a$ and $b$. The line corresponding to the lowest value of $\phi$ is drawn with a dashed line.
Equivalence of hierarchies

Proposition 3

If $\theta_j \sim \mathcal{T}_{PBN}(a, b, \phi)$, then
1. $\theta_j \sim \mathcal{N}(0, \tau_j)$, $\tau_j \sim \mathcal{G}(a, \lambda_j)$ and $\lambda_j \sim \mathcal{G}(b, \phi)$.
2. $\theta_j \sim \mathcal{N}(0, \tau_j)$, $\pi(\tau_j) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \phi^{-a}\tau^{a-1}(1 + \tau_j/\phi)^{-(a+b)}$ which implies that $\tau_j\phi \sim \beta'(a, b)$, the inverted beta distribution with parameter $a$ and $b$.

Corollary 4

If $a = 1$ in Proposition 3.1, then $\mathcal{TPBN} \equiv \mathcal{NEG}$. If $(a, b, \phi) = (1, 1/2, 1)$ in Proposition 3.1, then $\mathcal{TPBN} \equiv \mathcal{SB} \equiv \mathcal{NEG}$

Corollary 5

If $\theta_j \sim \mathcal{N}(0, \tau_j)$, $\tau_j^{1/2} \sim \mathcal{C}^+(0, \phi^{1/2})$ and $\phi^{1/2} \sim \mathcal{C}^+(0, 1)$, then $\theta_j \sim \mathcal{T}_{PBN}(1/2, 1/2, \phi)$, $\phi \sim \mathcal{G}(1/2, \omega)$, and $\omega \sim \mathcal{G}(1/2, 1)$. 
Experiments: sparse linear regression

Figure 2: Relative ME at different \((a, b, \phi)\) values for (a) Case 1 and (b) Case 2. Both contains on average 10 nonzero elements.

Figure 3: Posterior mean by sampling (square) and by approximate inference (circle). \((a, b, \phi) = (1, 1/2, 10^{-4})\)