Linear regression with horseshoe prior sampling: The nitty gritty details

Måns Magnusson

June 29, 2015

1 The MCMC sampler

Below are derivations for a bayesian linear regression model with the horseshoe prior in details. I would like to use the horseshoe prior in one in my own models and had a hard time finding how the MCMC sampler was derived in details. Hopefully these derivation can help others.

The linear model with the standard horseshoe prior can be expressed as follows.

$$\begin{aligned}
\mathbf{y}|\beta,\sigma^2,\mathbf{X} &\sim \mathrm{MVN}\left(\mathbf{X}\beta,\sigma^2\mathbf{I}_n\right)\\
\beta_i|\tau,\lambda_i,\sigma &\sim \mathrm{N}(0,\lambda_i^2\tau^2\sigma^2)\\
\sigma^2 &\sim \mathrm{IG}\left(a_n,b_n\right)\\
\tau &\sim C^+(0,1)\\
\lambda_i &\sim C^+(0,1)
\end{aligned}$$

where IG is the inverse-Gamma distribution, C^+ is the positive truncated Cauchy distribution and MVN is the multivariate normal distribution.

The joint posterior can be expressed as

$$\begin{aligned}
p(\beta,\sigma^2,\lambda_1,...,\lambda_p,\tau|\mathbf{y},\mathbf{X}) &\propto p(\mathbf{y}|\mathbf{X},\beta,\sigma^2)\cdot p(\beta,\sigma^2,\lambda_1,...,\lambda_p,\tau)\\
&= p(\mathbf{y}|\mathbf{X},\beta,\sigma^2)\cdot p(\beta|\sigma^2,\lambda_1,...,\lambda_p,\tau)\cdot p(\sigma^2)\cdot p(\tau)\cdot\prod_i^p p(\lambda_i)
\end{aligned}$$

1.1 Sampling the regression coefficients

The prior for β is

$$\Sigma_\beta=\sigma^2\Lambda_0^{-1}=\sigma^2\tau^2\begin{pmatrix}
\lambda_1^2 & 0 & 0 & 0 & 0\\
0 & \ddots & 0 & 0 & 0\\
0 & 0 & \lambda_i^2 & 0 & 0\\
0 & 0 & 0 & \ddots & 0\\
0 & 0 & 0 & 0 & \lambda_p^2
\end{pmatrix}$$

and based on this we use the updates for the ordinary bayesian linear regression.

$$
\begin{aligned}
p(\beta|\sigma^2,\lambda_1,...,\lambda_p,\tau,\mathbf{y},\mathbf{X}) &\propto p(\mathbf{y}|\mathbf{X},\beta,\sigma^2)\cdot p(\beta|\sigma^2,\lambda_1,...,\lambda_p,\tau) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)\right)\cdot\exp\left(-\frac{1}{2}\beta^T\Sigma_\beta^{-1}\beta\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\left[(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)+\sigma^2\beta^T\Sigma_\beta^{-1}\beta\right]\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\left[\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\mathbf{X}\beta-\beta^T\mathbf{X}^T\mathbf{y}+\beta^T\mathbf{X}^T\mathbf{X}\beta+\beta^T\mathit{\Lambda}_0\beta\right]\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\left[\beta^T\mathit{\Lambda}_n\beta-\mathbf{y}^T\mathbf{X}\beta-\beta^T\mathbf{X}^T\mathbf{y}\right]\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\left[\beta^T\mathit{\Lambda}_n\beta-\mu_n^T\mathit{\Lambda}_n\beta-\beta^T\mathit{\Lambda}_n\mu_n\right]\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\left[\beta^T\mathit{\Lambda}_n\beta-\mu_n^T\mathit{\Lambda}_n\beta-\beta^T\mathit{\Lambda}_n\mu_n+\mu_n^T\mathit{\Lambda}_n\mu_n\right]\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\left[(\beta-\mu_n)^T\mathit{\Lambda}_n(\beta-\mu_n)\right]\right)
\end{aligned}
$$

where

$$\mu_n = \Lambda_n^{-1}\mathbf{X}^T\mathbf{y}$$

and

$$\Lambda_n = (\mathbf{X}^T\mathbf{X}+\Lambda_0)$$

And hence

$$\beta \sim MVN(\mu_n, \sigma^2\Lambda_n^{-1})$$

1.2 Sampling of σ

Samling of sigma with an invers gamma prior follows the standard approach in bayesian linear regression. The posterior distribution of σ^2 is

$$\sigma^2 \sim \mathrm{IG}\left(a_n, b_n\right)$$

where

$$
\begin{aligned}
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2}\left(\mathbf{y}^T\mathbf{y} - \mu_n^T\Lambda_n\mu_n\right)
\end{aligned}
$$

1.3 Sampling of τ

$$
\begin{aligned}
p(\tau|\lambda,\sigma,\beta) &\propto p(\beta|\lambda,\tau,\sigma)\cdot p(\tau) \\
&= \frac{1}{\sqrt{(2\pi)^p|\Sigma_\beta|}}\exp\left(-\frac{1}{2}(\beta^{\mathrm{T}}\Sigma_\beta^{-1}\beta)\right)\cdot\frac{2}{\pi}\cdot\frac{1}{1+\tau^2} \\
& \quad \frac{1}{\sqrt{(2\pi)^p\sigma^{2p}\tau^{2p}\lambda_1^2\cdots\lambda_i^2\cdots\lambda_p^2}}\exp\left(-\frac{1}{2}(\beta^{\mathrm{T}}\Sigma_\beta^{-1}\beta)\right)\cdot\frac{2}{\pi}\cdot\frac{1}{1+\tau^2} \\
&\propto \frac{1}{\tau^p}\exp\left(-\frac{1}{2}(\beta^{\mathrm{T}}\Sigma_\beta^{-1}\beta)\right)\cdot\frac{1}{1+\tau^2} \\
&= \frac{1}{\tau^p}\exp\left(-\frac{1}{2\sigma^2\tau^2}\left(\sum_i\frac{\beta_i^2}{\lambda_i^2}\right)\right)\cdot\frac{1}{1+\tau^2}
\end{aligned}
$$

We then set $\gamma=\frac{1}{\tau^2}$ implies $\tau=\gamma^{-\frac{1}{2}}$ with

$$
\begin{aligned}
p(\gamma_i) &\propto \gamma^{\frac{p}{2}}\exp\left(-\frac{1}{2\sigma^2}\left(\sum_i\frac{\beta_i^2}{\lambda_i^2}\right)\gamma\right)\cdot\frac{1}{1+\gamma^{-1}}\left|\frac{d}{d\gamma}\gamma_i^{-\frac{1}{2}}\right| \\
&= \exp\left(-\frac{1}{2\sigma^2}\left(\sum_i\frac{\beta_i^2}{\lambda_i^2}\right)\gamma\right)\cdot\frac{\gamma^{\frac{p}{2}}}{\frac{\gamma+1}{\gamma}}\left|-\frac{1}{2}\gamma^{-\frac{3}{2}}\right| \\
&= \exp\left(-\frac{1}{2\sigma^2}\left(\sum_i\frac{\beta_i^2}{\lambda_i^2}\right)\gamma\right)\cdot\frac{\gamma^{\frac{p+2}{2}}}{\gamma+1}\frac{1}{2}\gamma^{-\frac{3}{2}} \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\left(\sum_i\frac{\beta_i^2}{\lambda_i^2}\right)\gamma\right)\cdot\frac{1}{\gamma+1}\gamma^{\frac{p-1}{2}}
\end{aligned}
$$

In Scott (2010, p. 6f.) they sample from this (and τ^2) with slice sampling by defining $\gamma=\frac{1}{\tau^2}$ (called η_i in Scott (2009, p. 6f.)) and $\hat{\mu}^2=\sum_p^P\left(\frac{\beta_p}{\lambda_p}\right)^2/\sigma^2=\sum_p^P\left(\frac{\beta_p}{\lambda_p\sigma}\right)^2$ with

$$
p(\gamma|\lambda_i,\hat{\mu})\propto\exp\left(-\frac{1}{2}\hat{\mu}^2\gamma\right)\gamma^{\frac{p-1}{2}}\frac{1}{1+\gamma}
$$

To sample τ we use the algorithm of using the same slice sampling procedure as in Damlen et al. (1999, section 3.2). Using this approach we get

$$
\begin{aligned}
l(\gamma) &= \frac{1}{1+\gamma} \\
\pi(\gamma) &= \exp\left(-\frac{1}{2}\hat{\mu}^2\gamma\right)\gamma^{\frac{p-1}{2}}
\end{aligned}
$$

so π is a gamma distribution with $\alpha = (p+1)/2$ and $\beta = \hat{\mu}^2$ we sample

$$\begin{aligned} u &\sim U(0,(1+\gamma)^{-1}) \\ \gamma &\sim \mathrm{G}\left(\alpha = \frac{1}{2}(p+1), \beta = \frac{1}{2}\hat{\mu}^2\right) I(\gamma < (1-u)/u) \end{aligned}$$

where $I()$ indicates the truncation region.

After sampling we transform back to $\tau$by $\tau = \gamma^{-\frac{1}{2}}$.

1.4 Sampling of λ_i

$$\begin{aligned} p(\beta|\lambda,\tau,\sigma)\cdot p(\lambda_i) &= \frac{1}{\sqrt{(2\pi)^p|\Sigma_\beta|}}\exp\left(-\frac{1}{2}(\beta^{\mathrm{T}}\Sigma_\beta^{-1}\beta\right)\cdot\frac{2}{\pi}\cdot\frac{1}{1+\lambda_i^2} \\ & \frac{1}{\sqrt{(2\pi)^p\sigma^{2p}\tau^{2p}\lambda_1^2\cdots\lambda_i^2\cdots\lambda_p^2}}\exp\left(-\frac{1}{2}(\beta^{\mathrm{T}}\Sigma_\beta^{-1}\beta\right)\cdot\frac{2}{\pi}\cdot\frac{1}{1+\lambda_i^2} \\ &\propto \frac{1}{\lambda_i}\exp\left(-\frac{1}{2\tau^2\sigma^2}\left(\sum_i\frac{\beta_i^2}{\lambda_i^2}\right)\right)\cdot\frac{1}{1+\lambda_i^2} \\ &\propto \frac{1}{\lambda_i}\exp\left(-\frac{\beta_i^2}{2\sigma^2\tau^2\lambda_i^2}\right)\cdot\frac{1}{1+\lambda_i^2} \end{aligned}$$

We then set $\gamma_i = \frac{1}{\lambda_i^2}$ with $\lambda_i = \gamma_i^{-\frac{1}{2}}$ with

$$\begin{aligned} p(\gamma_i) &\propto \gamma_i^{\frac{1}{2}}\exp\left(-\frac{\beta_i^2}{2\sigma^2\tau^2\gamma_i^{-1}}\right)\cdot\frac{1}{1+\gamma_i^{-1}}\left|\frac{d}{d\gamma}\gamma_i^{-\frac{1}{2}}\right| \\ &= \exp\left(-\frac{\beta_i^2}{2\sigma^2\tau^2}\gamma_i\right)\cdot\frac{\gamma_i^{\frac{1}{2}}}{\frac{\gamma_i+1}{\gamma_i}}\left|-\frac{1}{2}\gamma_i^{-\frac{3}{2}}\right| \\ &= \exp\left(-\frac{\beta_i^2}{2\sigma^2\tau^2}\gamma_i\right)\cdot\frac{\gamma_i^{\frac{3}{2}}}{\gamma_i+1}\frac{1}{2}\gamma_i^{-\frac{3}{2}} \\ &\propto \exp\left(-\frac{\beta_i^2}{2\sigma^2\tau^2}\gamma_i\right)\cdot\frac{1}{\gamma_i+1} \end{aligned}$$

In Scott (2009, p. 9f.) they sample from this (and τ^2) with slice sampling by defining $\gamma_i = \frac{1}{\lambda_i^2}$ (called η_i in Scott (2009, p. 9f.)) and $\hat{\mu_i} = \frac{\beta_i}{\tau\sigma}$ with

$$p(\gamma_i|\tau,\hat{\mu_i}) \propto \exp\left(-\frac{1}{2}\hat{\mu_i}^2\gamma_i\right)\frac{1}{1+\gamma_i}$$

To sample λ_i we use the algorithm of using the same slice sampling procedure as in Damlen et al. (1999, section 3.2). Using this approach we get

$$\begin{aligned} l(\gamma) &= \frac{1}{1+\gamma_i} \\ \pi(\gamma) &\propto \exp\left(-\frac{1}{2}\hat{\mu_i}^2\gamma_i\right) \end{aligned}$$

so we sample

$$
\begin{aligned}
u &\sim U(0, (1+\gamma_i)^{-1}) \\
\gamma &\sim \mathrm{Exp}\left(\frac{1}{2}\hat{\mu_i}^2\right) I(\gamma < (1-u)/u)
\end{aligned}
$$

where $I()$ indicates the truncation region. After sampling γ_i we convert back to λ_i with $\lambda_i = \gamma_i^{-\frac{1}{2}}$.

References

Damlen, P., Wakefield, J., Walker, S., 1999. Gibbs sampling for bayesian non-conjugate and hierarchical models by using auxiliary variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61 (2), 331–344.

Scott, J. G., 2009. Flexible learning on the sphere via adaptive needlet shrinkage and selection. Tech. rep.

Scott, J. G., 2010. Parameter expansion in local-shrinkage models. arXiv preprint arXiv:1010.5265.