UCLA Biostatistics 285: Homework 1

Instructor: Dr. Michele Guindani

Soumyakanti Pan, e-mail: span18@ucla.edu

April 28, 2023

1 Problem 1

We will carry out the analysis on the dataset `dataHW1` downloaded from class website. The dataset has one dependent variable `y` and 50 predictor variables given by `X`. Dr. Guindani is very kind to provide us a standardized dataset so we can drop the intercept from the model.

1.1 Spike and slab vs Horseshoe Regression

We can consider a hierarchical model as follows.

$$
\begin{aligned}
\mathbf{Y} \mid \boldsymbol{\beta}, \sigma^2 &\sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}) \\
\beta_j \mid \gamma_j, \sigma^2 &\sim (1-\gamma_j)\,\delta_0 + \gamma_j\,\mathcal{N}(0,\tau^2), j = 1,\ldots,p \\
\gamma_j \mid \pi_0 &\sim \mathrm{Ber}(\pi_0), j = 1,\ldots,p \\
\pi_0 &\sim \mathrm{Beta}(a,b) \\
\sigma^2 &\sim \mathrm{InvGamma}(\nu,\lambda)
\end{aligned}
$$

1.1.1 SSVS Gibbs sampler

In order to draw samples from the joint posterior, we cannot form an ergodic Markov chain if we carry out a full Gibbs sampler as done in George and McCulloch (1993) for normal mixture spike and slab priors. In order to make the chain irreducible, we need to draw γ from its marginal posterior

$$[\gamma \mid \mathbf{Y}] \propto [\mathbf{Y} \mid \gamma] \times [\gamma]$$

where we have integrated out the regression coefficients $\boldsymbol{\beta}$ and the error variance σ^2 (Malsiner-Walli and Wagner 2018). As shown in class, the marginal posterior can be given in closed form by

$$[\mathbf{Y} \mid \gamma] \propto \left|\mathbf{X}_\gamma^\top \mathbf{X}_\gamma + \frac{1}{\tau^2}\mathbf{I}\right|^{-1/2} \left(\lambda + \frac{1}{2}S_\gamma^2\right)^{-\nu-\frac{n}{2}-1}$$

where, $S_\gamma^2 = \mathbf{Y}^\top\mathbf{Y} - \mathbf{Y}^\top\mathbf{X}_\gamma\left(\mathbf{X}_\gamma^\top\mathbf{X}_\gamma + \frac{1}{\tau^2}\mathbf{I}\right)^{-1}\mathbf{X}_\gamma^\top\mathbf{Y}$. So, now we can write a SSVS Gibbs sampler as following in algorithm 1.

Following the Gibbs sampler described in algorithm 1, the functions `gamma_logmarginal` and `alpha_j` computes the logarithm of marginal posterior density of γ and the ratio $\alpha_j^{(t)}$ respectively. We have considered $\tau^2 = 100, \nu = 2, \lambda = 1, a = 2$ and $b = 2$. I have implemented the sampler using `R` and the output is given below. The output displays the hyperpriors used in the model and the corresponding value of the hyperparameter inputs and the number of samples and burn-in. The output also contains the time elapsed for the chain to complete every 1000 iterations.

We set the number of samples to be generated by `n.iter` with a burn-in of `n.burnin` samples.

Algorithm 1: Stochastic search variable selection (SSVS) via Gibbs sampler

```
Result: posterior samples $(\gamma^{(t)}, \sigma^{2(t)}, \pi_0^{(t)}, \beta^{(t)})_{t=0}^T$
Initialize: Start with $\gamma^{(0)} = (1, \ldots, 1)$ and $\pi_0^{(0)} = 0.5$.
while $t = 0, 1, \ldots, T$ do
    1. Sample $\gamma_j$ sequentially in a random permutation order given the rest of $\gamma$ vector, denoted by $\gamma_{-j}$ and $\mathbf{Y}$, by Bernoulli trials with success probability
       $$\Pr(\gamma_j^{(t+1)} = 1 \mid \gamma_{-j}^{(t)}, \mathbf{Y}) = \frac{1}{1 + \frac{1-\pi_0^t}{\pi_0^{(t)}} \alpha_j^{(t)}}, \text{ where } \alpha_j^{(t)} = \frac{p(\mathbf{Y} \mid \gamma_j = 0, \gamma_{-j}^{(t+1/2)})}{p(\mathbf{Y} \mid \gamma_j = 1, \gamma_{-j}^{(t+1/2)})}$$
       where, $\gamma_{-j}^{(t+1/2)}$ denotes the most updated $\gamma_{-j}$ vector between $\gamma^{(t)}$ and $\gamma^{(t+1)}$.
    2. Sample $\sigma^{2(t+1)}$ from InvGamma $\left(\nu + \frac{n}{2}, \lambda + \frac{1}{2} S^2_{\gamma^{(t+1)}}\right)$.
    3. Sample $\pi_0^{(t+1)}$ from Beta$(a + p_1^{(t+1)}, b + p - p_1^{(t+1)})$ where $p_1^{(t+1)} = \sum \gamma_j^{(t+1)}$.
    4. if $\gamma_j^{(t+1)} = 0$ then
           Set $\beta_j^{(t+1)} = 0$.
       else
           Sample $\boldsymbol{\beta}_\gamma$ from $\mathcal{N}\left(A_\gamma^{(t+1)} \mathbf{X}_{\gamma^{(t)}}^\top \mathbf{Y}, \sigma^{2(t+1)} A_\gamma^{(t+1)}\right)$ where $A_\gamma^{(t)} = \left(\mathbf{X}_{\gamma^{(t)}}^\top \mathbf{X}_{\gamma^{(t)}} + \frac{1}{\tau^2}\mathbf{I}\right)^{-1}$.
       end
end
```

```
Running SSVS Gibbs sampler with options:
 n.iter =  1000 , Burn-in =  100
 Hyperpriors:   π ~ Beta ( 2 , 2 )
                σ² ~ InvGamma ( 2 , 1 )
 ========================================
Progress:

Elapsed  100  iterations. ( 0.74 secs )
Elapsed  200  iterations. ( 1.42 secs )
Elapsed  300  iterations. ( 2.11 secs )
Elapsed  400  iterations. ( 2.79 secs )
Elapsed  500  iterations. ( 3.47 secs )
Elapsed  600  iterations. ( 4.15 secs )
Elapsed  700  iterations. ( 4.82 secs )
Elapsed  800  iterations. ( 5.46 secs )
Elapsed  900  iterations. ( 6.15 secs )
Elapsed  1000  iterations. ( 6.81 secs )
```

1.1.2 Mixing and convergence

To briefly discuss about mixing times for the chain, it is observed that the chain converges quite smoothly. The autocorrelation plots (figure 1) of all the parameters $\boldsymbol{\beta}, \sigma^2$ and π_0 show very low serial correlation indicating almost independent samples from the posterior distributions.

In addition, the traceplots also show visual diagnostics of the chains attaining stationarity as seen in figures 2 and 3.

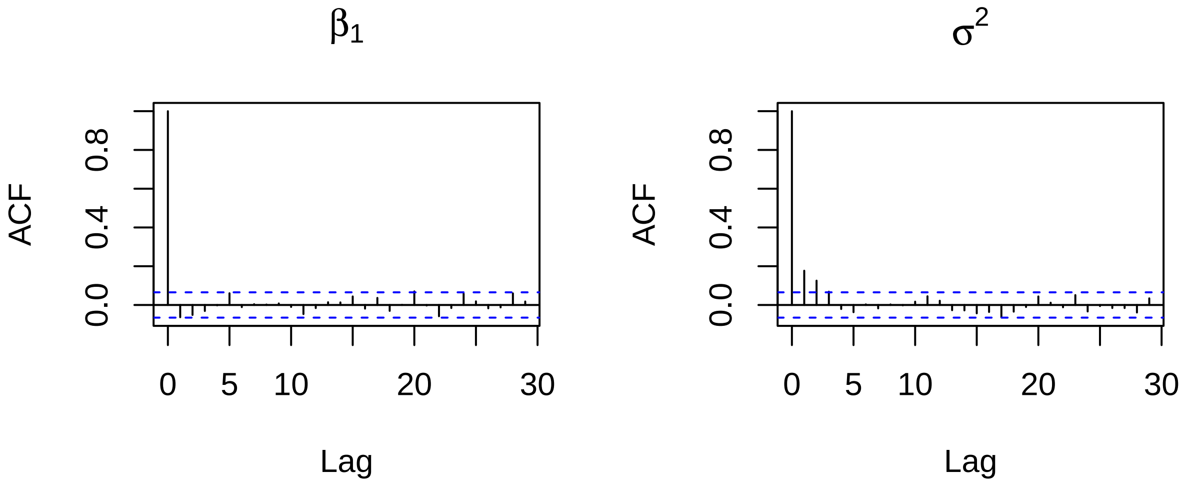

Figure 1: ACF plots of posterior samples

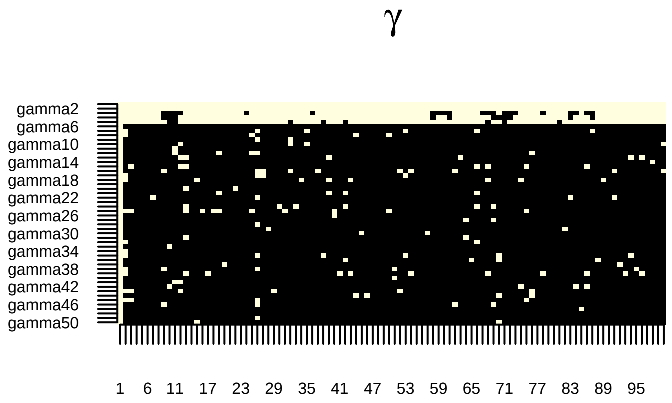

Figure 2: Summary of first 100 samples - Black: 0, Lightyellow: 1

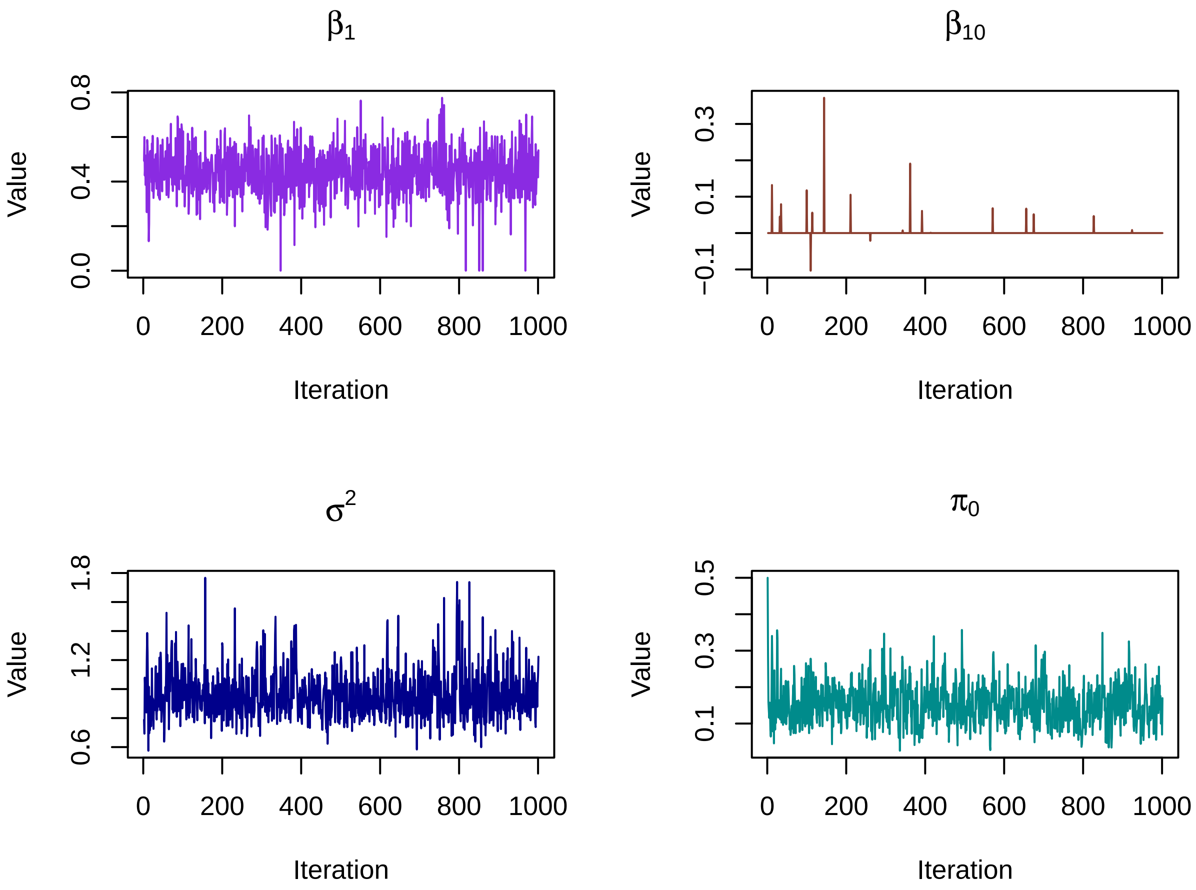

Figure 3: Traceplots of relevant parameters

Table 1: Posterior summary of first 10 effect sizes

	Mean	SD	2.5%	50%	97.5%
beta1	0.44	0.11	0.23	0.44	0.64
beta2	0.59	0.10	0.40	0.59	0.78
beta3	0.31	0.17	0.00	0.34	0.56
beta4	0.40	0.16	0.00	0.43	0.64
beta5	0.40	0.13	0.00	0.41	0.61
beta6	0.00	0.02	0.00	0.00	0.00
beta7	0.00	0.02	0.00	0.00	0.00
beta8	0.00	0.02	0.00	0.00	0.00
beta9	0.00	0.02	0.00	0.00	0.00
beta10	0.00	0.02	0.00	0.00	0.00

1.1.3 Posterior credible intervals and probability of inclusions (PPI)

Post cleaning the chain by removing the burn-in samples and subsequent thinning (not essential in this case), we can find the posterior credible intervals of the parameters.

The following plots (figure 4) shows a multiple box plot containing the posterior credible intervals of $\boldsymbol{\beta}$ and ordered posterior probability of inclusions (PPIs). The posterior probability of inclusion of the j-th predictor is calculated by the empirical relative frequencies as follows.

$$\Pr(\gamma_j = 1 \,|\, \mathbf{Y}) \approx \frac{1}{T} \sum_{t=1}^{T} \mathbf{1} \left(\gamma_j^{(t)} = 1 \right)$$

where, T is the total number of posterior samples and $\gamma_j^{(t)}$ is the t-th posterior sample of γ_j in the cleaned chain.

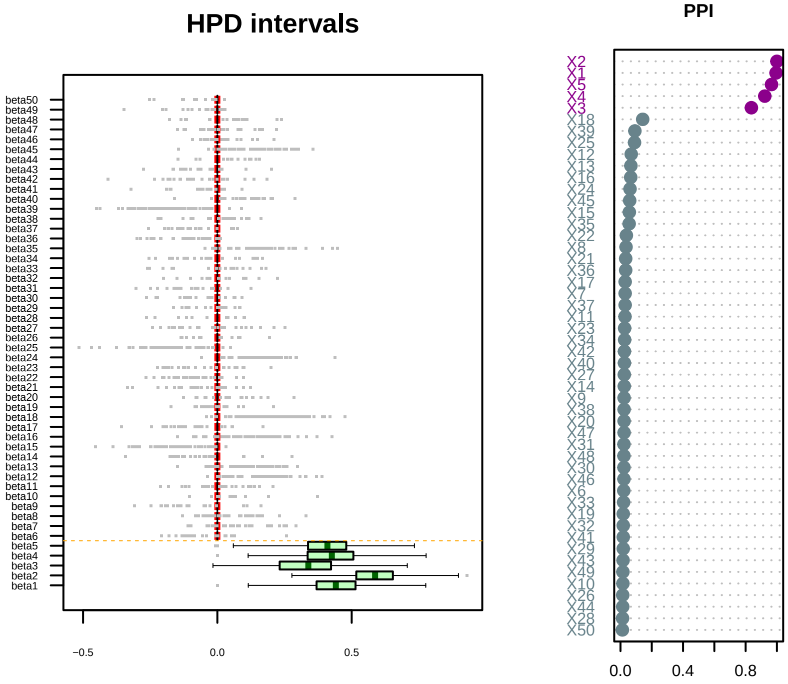

Figure 4: Posterior credible intervals of effect sizes and PPIs

1.1.4 Bayesian FDR

We shall find the optimal threshold for the PPIs as discussed in class. If we consider a loss function as a weighted sum of FP and FN counts across the hypotheses (Newton et al. 2004), then the posterior risk is minimized by thresholding the posterior probability of inclusion

$$\nu_j = \Pr(\gamma_j = 1 \,|\, \mathbf{Y}) < \frac{1}{1+\lambda}$$

where λ can be chosen to control the Bayesian FDR

$$\text{BFDR} = \frac{\sum(1-\nu_j)\delta_j}{\sum \delta_j}$$

at level α where δ_j denotes the indicator for rejecting the j-th comparison. In our case, we choose $\alpha = 0.05$ and will plot BFDR across a grid of possible values of λ and thereby choose its optimal value to ensure the FDR controlled at 0.05.

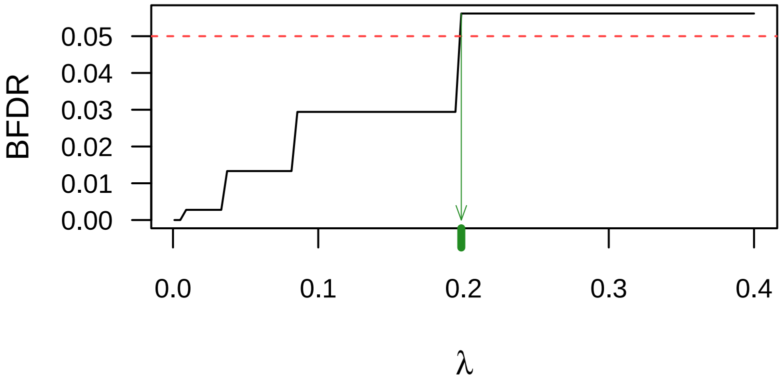

Figure 5: Choosing optimal cutoff in order to control bayesian FDR

Once we have found the optimal value of λ (0.1985), we can formulate the rejection rule as follows - we reject the null hypothesis $H_{0j} : \beta_j = 0$ if the posterior probability of inclusion of the j-th predictor, $\nu_j > 0.834$ for $j = 1, \ldots, p$.

1.2 Horseshoe prior

Here, we consider the horseshoe prior by Carvalho, Polson, and Scott (2010). The `horseshoe` package in `R` provides the `horseshoe` function which implements the following full Bayesian hierarchical model for variable selection.

$$\begin{aligned} \mathbf{Y} \,|\, \boldsymbol{\beta}, \sigma^2 &\sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n) \\ \beta_j \,|\, \sigma^2, \lambda_j, \tau^2 &\sim \mathcal{N}(0, \sigma^2 \lambda_j^2 \tau^2) \\ \lambda_j &\sim \mathrm{C}^+(0,1);\ \tau \sim \mathrm{C}^+(0,1) \\ p(\sigma^2) &\propto \frac{1}{\sigma^2} \end{aligned}$$

where C^+ denotes the standard half-Cauchy distribution on the positive reals with scale parameter 1.

1.2.1 Results from ‘horseshoe’ package

We see that both the methods identify the same effects to be significant and with almost similar estimates. In addition, we see slighly lower standard errors with SSVS Gibbs sampler than that of with horseshoe prior. Given the posterior credible intervals of β, using the `horseshoe` package, we can select variables by either

Table 2: Posterior summary of first 10 effect sizes using horseshoe prior

index	Mean	SD	X2.5.	X50.	X97.5.
beta1	0.39	0.13	0.07	0.39	0.63
beta2	0.54	0.12	0.29	0.54	0.77
beta3	0.19	0.15	-0.01	0.19	0.48
beta4	0.29	0.16	0.00	0.31	0.59
beta5	0.30	0.15	0.00	0.31	0.56
beta6	-0.01	0.05	-0.12	0.00	0.09
beta7	0.01	0.05	-0.08	0.00	0.15
beta8	0.01	0.05	-0.08	0.00	0.14
beta9	0.00	0.05	-0.11	0.00	0.09
beta10	0.01	0.05	-0.08	0.00	0.13

the `interval` method where we reject the null hypothesis $H_{0j} : \beta_j = 0$ if the posterior credible interval of β_j does not contain 0. The other method is `threshold` where we threshold the shrinkage coefficients, but that is only applicable to the sparse normal mean problem (Carvalho, Polson, and Scott 2010) which is essentially regression with identity design matrix, clear not our case.

```
hs.beta = HS.var.select(hs.object, Y, method = "interval")
cat(paste("beta", which(hs.beta == 1), sep = ""), "\n")
```

```
beta1 beta2
```

We find only the above 2 effects to be significant. However, one can also use the shrinkage coefficients to select variables. Note that under the above model, we can write the conditional posterior of $\boldsymbol{\beta}$ as

$$\tilde{\boldsymbol{\beta}} = \mathbb{E}\left[\boldsymbol{\beta} \,|\, \boldsymbol{\Lambda}, \tau, \sigma^2, \mathbf{Y}\right] = \tau^2 \boldsymbol{\Lambda}\left(\tau^2 \boldsymbol{\Lambda} + (\mathbf{X}^\top \mathbf{X})^{-1}\right)^{-1} \hat{\boldsymbol{\beta}}$$

where, $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate. Under the assumption that the predictors are uncorrelated with unit variance and zero mean, then $\mathbf{X}^\top \mathbf{X} \approx n\mathbf{I}$, and we can approximate

$$\tilde{\beta}_j \approx (1 - \kappa_j)\hat{\beta}_j$$

where,

$$\kappa_j = \frac{1}{1 + n\tau^2\lambda_j^2}$$

is the shrinkage factor for β_j (Piironen and Vehtari 2017). One can use the shrinkage coefficient to give a thresholding rule as in we will reject $H_{0j} : \beta_j = 0$ id $(1 - \hat{\kappa}_j) > c$, for some constant c. This c will depend on the loss function of which we are trying to minimize the posterior risk. If the loss function is a symmetric 0-1 loss function giving equal weights to false positives and false negatives, $c = 0.5$.

1.2.2 Prior for the global shrinkage parameter

Following from the expression of the shrinkage coefficient κ_j, when $\tau^2 = 1$, the prior distribution of the shrinkage coefficient $p(\kappa_j \,|\, \tau)$ will become Beta$(1/2, 1/2)$, our usual horseshoe prior. This means a priori, we give most of the mass to both the possibilities - relevant effect ($\kappa = 0$, no shrinkage) and irrelevant effects ($\kappa = 1$, shrinkage to 0). The interesting bit is that τ can be calibrated to push mass at either boundary - 0 or 1 (Piironen and Vehtari 2017). For example if τ is chosen such that $n\tau^2 = 0.1$, then majority of the prior mass of κ will be at 1 favoring complete shrinkage and as a result we should expect more variables to be shrunk to 0.

The usual options for the prior on the global shrinkage prior τ in this set-up is $\mathrm{C}^+(0, 1)$, standard half-Cauchy or $\mathrm{C}^+(0, a^2)$, standard half-Cauchy with scale parameter a. It has been pointed out that arbitrary choice of a

can give unnecessary mass to implausible values of τ. Piironen and Vehtari (2017) has focused on the a priori distribution of m_{eff}, the effective number of nonzero coefficients induced by the hyperpriors on τ and the scale pixture distribution on λ. The prior distribution on m_{eff} written as

$$m_{\text{eff}} = \sum_{j=1}^{p} (1 - \kappa_j)$$

can be calculated easily by composition sampling in the order

$$\begin{bmatrix} \tau \sim p(\tau) \\ \{\lambda_j\}_{j=1}^{p} \sim f_\lambda \end{bmatrix} \rightarrow \begin{bmatrix} \kappa_j \sim p(\kappa \,|\, \lambda_j, \tau), \\ j = 1, \ldots, p \end{bmatrix} \rightarrow p(m_{\text{eff}} \,|\, \tau, \lambda).$$

We can also easily calculate the a priori mean and variance of m_{eff}, which can be given as follows.

$$\mathbb{E}\left[m_{\text{eff}} \,|\, \tau\right] = \frac{p\tau\sqrt{n}}{1 + \tau\sqrt{n}}$$

$$\text{Var}\left[m_{\text{eff}} \,|\, \tau\right] = \frac{p\tau\sqrt{n}}{2(1 + \tau\sqrt{n})^2}$$

Hence, we will need τ to scale $1/\sqrt{n}$, failure to which will cause favoring models of varying sizes depending on the sample size n. Now, if we have a prior guess on the number of relevant variables as p_0, we can equate $\mathbb{E}[m_{\text{eff}} \,|\, \tau] = p_0$ and consider a prior which is concentrated around

$$\tau_0 = \frac{p_0}{p - p_0} \frac{1}{\sqrt{n}}$$

which is definitely not near 1, the choice the function `horseshoe` is using. Piironen and Vehtari (2017) has also suggested using τ_0 as different types of prior on τ such as point mass at τ_0, half-normal and half-Cauchy priors with scale parameter τ_0. This suggestion for the optimal value of τ also asymptotically aligns with the optimal value in terms of mean squared error and posterior contraction rates with respect to the true effects $\boldsymbol{\beta}^*$ for the sparse normal means problem as shown in Pas, Kleijn, and Vaart (2014).

References

Carvalho, Carlos M., Nicholas G. Polson, and James G. Scott. 2010. "The horseshoe estimator for sparse signals." *Biometrika* 97 (2): 465–80. https://doi.org/10.1093/biomet/asq017.

George, Edward I., and Robert E. McCulloch. 1993. "Variable Selection via Gibbs Sampling." *Journal of the American Statistical Association* 88 (423): 881–89. https://doi.org/10.1080/01621459.1993.10476353.

Malsiner-Walli, Gertraud, and Helga Wagner. 2018. "Comparing Spike and Slab Priors for Bayesian Variable Selection." https://arxiv.org/abs/1812.07259.

Newton, Michael A., Amine Noueiry, Deepayan Sarkar, and Paul Ahlquist. 2004. "Detecting differential gene expression with a semiparametric hierarchical mixture method." *Biostatistics* 5 (2): 155–76. https://doi.org/10.1093/biostatistics/5.2.155.

Pas, S. L. van der, B. J. K. Kleijn, and A. W. van der Vaart. 2014. "The horseshoe estimator: Posterior concentration around nearly black vectors." *Electronic Journal of Statistics* 8 (2): 2585–2618. https://doi.org/10.1214/14-EJS962.

Piironen, Juho, and Aki Vehtari. 2017. "On the Hyperprior Choice for the Global Shrinkage Parameter in the Horseshoe Prior." In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, edited by Aarti Singh and Jerry Zhu, 54:905–13. Proceedings of Machine Learning Research. PMLR. https://proceedings.mlr.press/v54/piironen17a.html.