

Bayesian Variable Selection

STATISTICAL COMPUTING WORKSHOP, ISI DELHI

DAY 2



Bayesian Variable Selection

- Instead of selecting an optimal model, the Bayesian
 - Estimates parameters for the largest model
 - Computes marginal posterior inclusion probability (PIP) for each β
 - Selects model post-hoc based on these PIPs
- Different Approaches exist
 - Indicator Model Selection
 - Stochastic Search Variable Selection
 - Continuous Shrinkage Priors

Indicator Model Selection

- Define $\theta_j = I_j \beta_j$
- I_j is an indicator parameter, β_j is a continuous valued parameter
- Then θ_j is either 0 or it is the effect of the j th covariate
- Methods differ on how to put priors on (I_j, β_j)
- Kuo and Mullick (1998) proposes priors such that $\pi(I_j, \beta_j) = \pi(I_j)\pi(\beta_j)$
- This usually does not work well as it almost never flips from $I_j = 0$ to $I_j = 1$

Gibbs Variable Selection (Dellaportas *et al.*, 2000)

- This still uses the Indicator Model selection paradigm, with
$$\pi(I_j, \beta_j) = \pi(\beta_j | I_j) \pi(I_j)$$
- Typical choice for $\pi(\beta_j | I_j) = (1 - I_j)N(\tilde{\mu}, S) + I_j N(0, \tau^2)$
- The parameter τ^2 is the prior variance of β
- This configuration ensures, that the posterior distribution of β_j is not affected by the prior parameters for β_j when $I_j = 0$
- Requires a lot of tuning and good proposal distribution

Stochastic Search Variable Selection (George and McCulloch, 1993)

- We have $\theta_j = \beta_j$ and $\pi(I_j, \beta_j) = \pi(\beta_j | I_j) \pi(I_j)$
- The choice for $\pi(\beta_j | I_j)$ is
$$\pi(\beta_j | I_j) = (1 - I_j)N(0, \tau^2) + I_j N(0, g\tau^2)$$
- Now, the posterior for β_j is affected by parameters when $I_j = 0$
- Some tuning is required here as well.
- One strategy is to fix g at, say, 100 and take τ^2 to be a parameter. This reduces tuning dependency, but does not eliminate it.

Continuous Shrinkage Priors

- Introducing a continuous shrinkage prior on β_j (without any indicator variables) pulls near-zero coefficients towards 0 and retains large significant ones.
- This implies that these methods introduce shrinkage, but not variable selection.
- There are no I_j variables, based on which you can compute a MPPI
- You need to introduce some sort of thresholding mechanism to do variable selection afterwards.
- Common methods include:
 - Constructing 95% CI's based on the samples and checking if 0 is in the CI
 - Creating a meaningful substitute for MPPI and threshold based on that

Bayesian LASSO (Park and Casella, 2008)

- In frequentist LASSO, a penalty term of $-\lambda|\beta_j|$ is attached to the **log**-likelihood.
- This is same as saying we are multiplying the likelihood by $\exp(-\lambda|\beta_j|)$
- This would then be similar to say that the posterior should look like

$$\pi(\beta|Y, X) \propto L(\beta; Y, X) \prod_j \exp(-\lambda|\beta_j|)$$

- If we take $\pi(\beta) = \prod_j \pi_j(\beta_j)$ with $\pi_j(\beta) \propto \exp(-\lambda|\beta|)$
- This is equivalent to saying

$$\beta_j \sim DE(\lambda)$$

Bayesian LASSO (Contd.)

- The set-up for Bayesian LASSO is therefore,

$$\begin{aligned} Y|X, \beta_0, \beta, \sigma^2 &\sim N(\beta_0 + X\beta, \sigma^2 I_n) \\ \beta_j|\lambda &\sim DE(\lambda) \\ \sigma^2 &\sim \pi(\sigma^2) \end{aligned}$$

- However, this set-up does not have the nice property of conjugacy for β_j 's
- Instead, we reparametrize the Double Exponential (or Laplace) prior as a scale mixture of Normal densities with exponential mixing density:

$$\frac{a}{2} \exp(-a|z|) = \int_0^\infty \frac{1}{\sqrt{2\pi s}} \exp\left(-\frac{z^2}{2s}\right) \frac{a^2}{2} \exp\left(-\frac{a^2 s}{2}\right) ds$$

Bayesian LASSO (Contd.)

- The final set-up for Bayesian LASSO is

$$Y|X, \beta_0, \beta, \sigma^2 \sim N(\beta_0 + X\beta, \sigma^2 I_n)$$

$$\beta_i|\lambda, \sigma^2, \tau_i^2 \sim N(0, \sigma^2 \tau_i^2)$$

$$\tau_i^2|\lambda \sim \text{Exp}\left(\frac{\lambda^2}{2}\right)$$

$$\sigma^2 \sim \pi(\sigma^2)$$

- λ remains a tuning parameter.
- Once can put a prior on λ
- A 95% CI can be constructed to determine if a variable is significant

Horseshoe Prior (Carvalho *et al.*, 2010)

- The construction for variable selection using Horseshoe prior is:

$$Y|X, \beta_0, \beta, \sigma^2 \sim N(\beta_0 + X\beta, \sigma^2 I_n)$$

$$\beta_i|\lambda_i \sim N(0, \lambda_i^2)$$

$$\lambda_i|\tau \sim C^+(0, \tau)$$

$$\tau|\sigma \sim C^+(0, \sigma)$$

$$\sigma^2 \sim \pi(\sigma^2)$$

- This produces a global (via τ) and local (via λ_i^2) shrinkage.
- Is unbounded at 0 and the density is otherwise bounded by $\log\left(1 + \frac{2}{x^2}\right)$ and $\log\left(1 + \frac{4}{x^2}\right)$ on the two sides.
- Significant variables selected by $E\left(\frac{1}{1+\lambda_i^2} \middle| Y\right)$ being small.

Horseshoe Prior vs Others

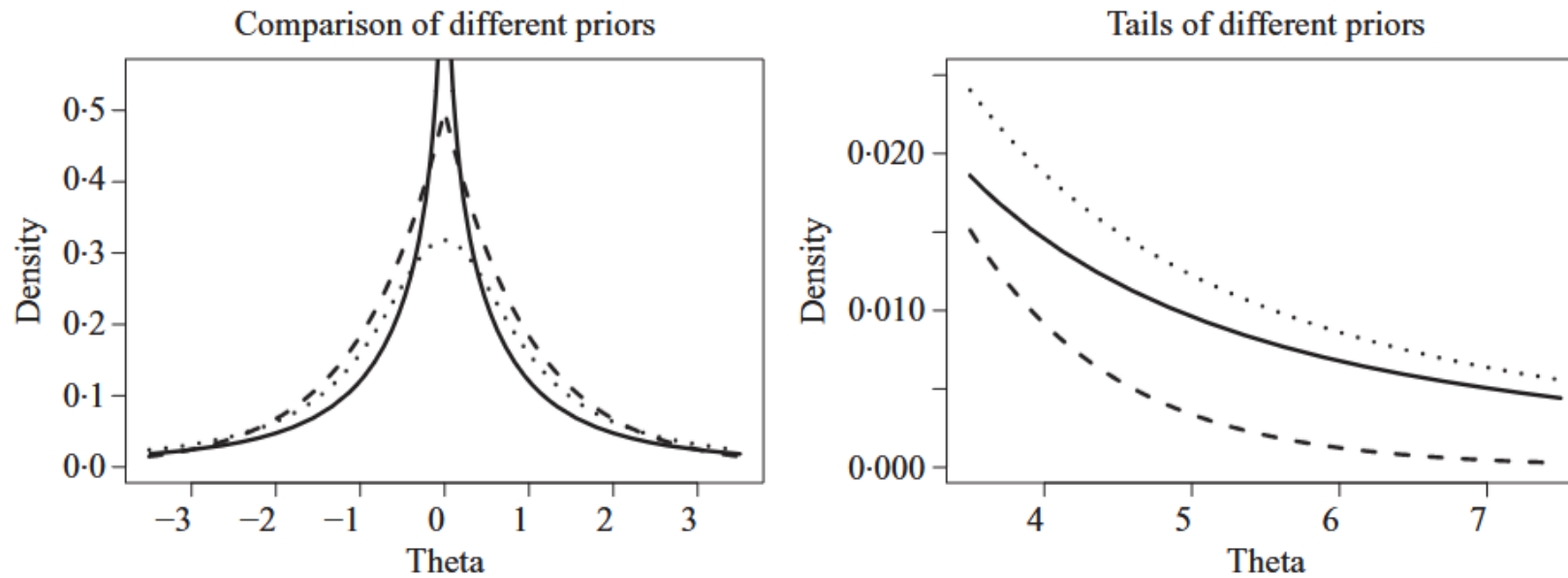


Fig. 1. Comparison of the horseshoe (solid), Cauchy (dotted) and double-exponential (dashed) densities.

Dirichlet-Laplace Prior (Bhattacharya *et al.*, 2015)

- For a Dirichlet-Laplace Prior based variable selection model:

$$Y|X, \beta_0, \beta, \sigma^2 \sim N(\beta_0 + X\beta, \sigma^2 I_n)$$

$$\beta_i|\tau, \sigma^2, \phi \sim DE(\phi_i \tau \sigma)$$

$$(\phi_1, \dots, \phi_p) \sim Dir(a, a, \dots, a)$$

$$\tau \sim Gamma\left(na, \frac{1}{2}\right)$$

$$\sigma^2 \sim \pi(\sigma^2)$$

- Has a closed-form prior: $\pi_a(\beta) = \frac{1}{2^{\frac{1+a}{2}} \Gamma(a)} |\beta|^{\frac{(a-1)}{2}} K_{1-a}\left(\sqrt{2|\beta|}\right), 0 < a < 1$

DL prior

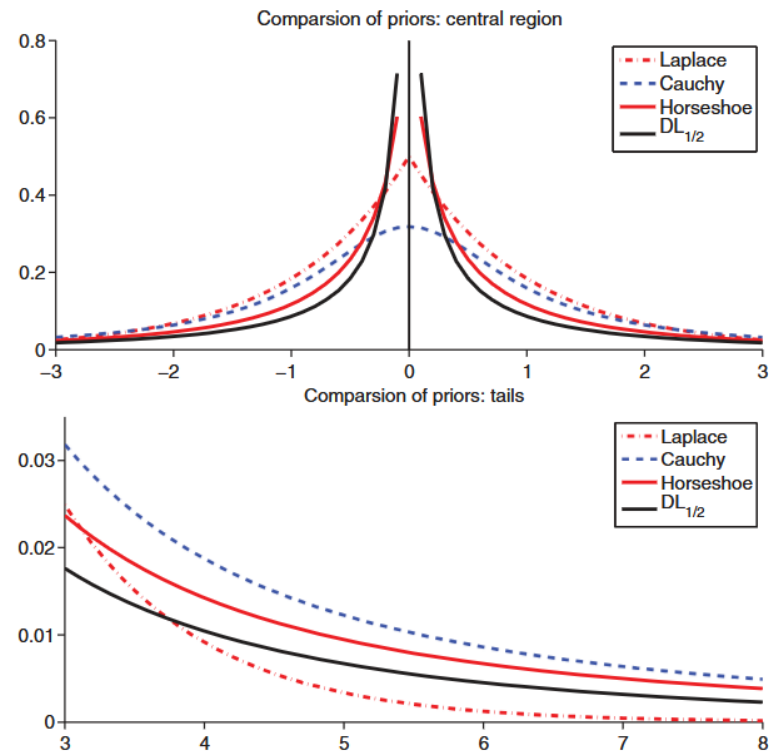


Figure 1. Marginal density of the DL prior with $a = 1/2$ in comparison to other shrinkage priors.

DL Prior (Contd.)

- For posterior sampling ease, we recall that Laplace prior can be alternatively written as exponential scale mixture of normals.

- Hence, the DL prior is often written as:

$$Y|X, \beta_0, \beta, \sigma^2 \sim N(\beta_0 + X\beta, \sigma^2 I_n)$$

$$\beta_i|\tau, \sigma^2, \phi \sim N(0, \psi_i \phi_i \tau \sigma)$$

$$\psi_i \sim \text{Exp}\left(\frac{1}{2}\right)$$

$$(\phi_1, \dots, \phi_p) \sim \text{Dir}(a, a, \dots, a)$$

$$\tau \sim \text{Gamma}\left(na, \frac{1}{2}\right)$$

$$\sigma^2 \sim \pi(\sigma^2)$$

References

1. Kuo, L., & Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, 65-81.
2. Dellaportas, P., Forster, J. J., & Ntzoufras, I. (2000). Bayesian variable selection using the Gibbs sampler. *Biostatistics-Basel-*, 5, 273-286.
3. George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423), 881-889.
4. Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the american statistical association*, 103(482), 681-686.
5. Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2), 465-480.
6. Bhattacharya, A., Pati, D., Pillai, N. S., & Dunson, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512), 1479-1490.