Theory Notes

Dav King

```
library(VGAM)
```

```
Warning: package 'VGAM' was built under R version 4.4.2
```

```
Loading required package: stats4
```

```
Loading required package: splines
```

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
```

```
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
library(viridis)
```

```
Loading required package: viridisLite
```

```
library(patchwork)
```

Overview/Background

Variable selection is important in high-dimensional settings: We often expect that only a small handful of the predictors are actually associated with the outcome, but when we include many in the model, we have issues with computational complexity, sparse solutions, and potentially issues with finding variables significant which are not actually meaningful predictors.

Why is a Bayesian approach potentially better? Bayesian methods allow us to introduce prior information on the betas, which can help us introduce known structure into the variable selection setting and also stabilize inferences in high-dimensional settings (Lu & Lou, 2022).

I note that the introduction to (Rockova, 2013) has a lot of good background on this topic, and would be useful in explaining the setup for this presentation.

Priors

Spike-and-Slab Priors

The spike-and-slab prior is a two-point mixture on the β_j, which forces some of the β_j to zero and estimates the coefficients of the others. The generic form of the spike-and-slab prior is

$$\beta_j|\gamma_j \sim \gamma_j\phi_1(\beta_j) + (1-\gamma_j)\phi_0(\beta_j), \qquad \sim \pi(\ )$$

In this case, $\phi_1(\beta_j)$ is a diffuse "slab distribution" so that the β_j can reach their true coefficients, and $\phi_0(\beta_j)$ is a concentrated "spike distribution" pulling effects to 0, and γ_j is a binary latent indicator representing the 2^p possible models (Lu & Lou, 2022).

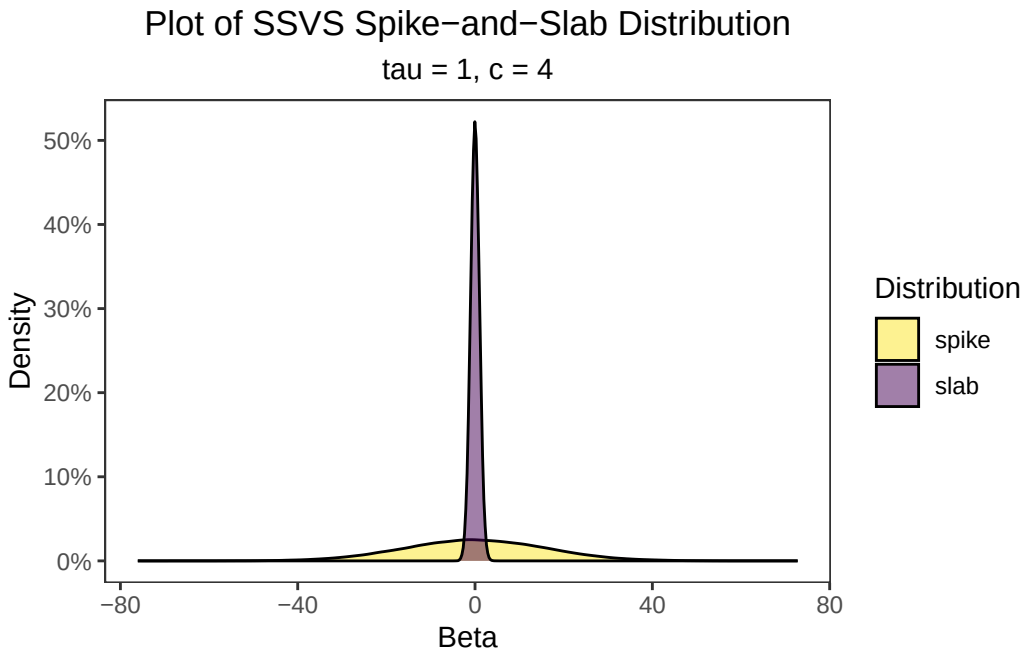
Stochastic Search Variable Selection (SSVS)

SSVS embeds the entire regression setup in a hierarchical Bayes normal mixture model, using latent variables to identify subset choices. Promising subsets of predictors have higher posterior probability, and Gibbs sampling can indirectly sample from the multinomial posterior distribution on the set of possible subset choices. Subsets with higher probability are identified by their more frequent appearance in the Gibbs sample, avoiding the problem of calculating the posterior probabilities for all 2^p subsets. Frequently, this converges quickly to near-optimal solutions (George & McCulloch, 1993).

According to George (1993), this can be represented using latent variable $\gamma_j \in \{0, 1\}$ with

$$\beta_j|\gamma_j \sim (1-\gamma_j)N(0,\tau_j^2) + \gamma_j N(0, c_j^2\tau_j^2)$$

```r
tau <- 1
c <- 4

slab <- rnorm(100000, 0, tau^2)
spike <- rnorm(100000, 0, c^2 * tau^2)

data.frame(slab, spike) %>%
  pivot_longer(everything()) %>%
  mutate(name = factor(name, levels = c("spike", "slab"))) %>%
  ggplot(aes(x = value, fill = name)) +
  geom_density(alpha = 0.5) +
  theme_bw() +
  labs(x = "Beta", y = "Density", fill = "Distribution",
       title = "Plot of SSVS Spike-and-Slab Distribution",
       subtitle = "tau = 1, c = 4") +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        panel.grid = element_blank()) +
  scale_fill_viridis(discrete = TRUE, direction = -1) +
  scale_y_continuous(labels = scales::percent_format())
```



Normal Mixture of Inverse Gamma (NMIG)

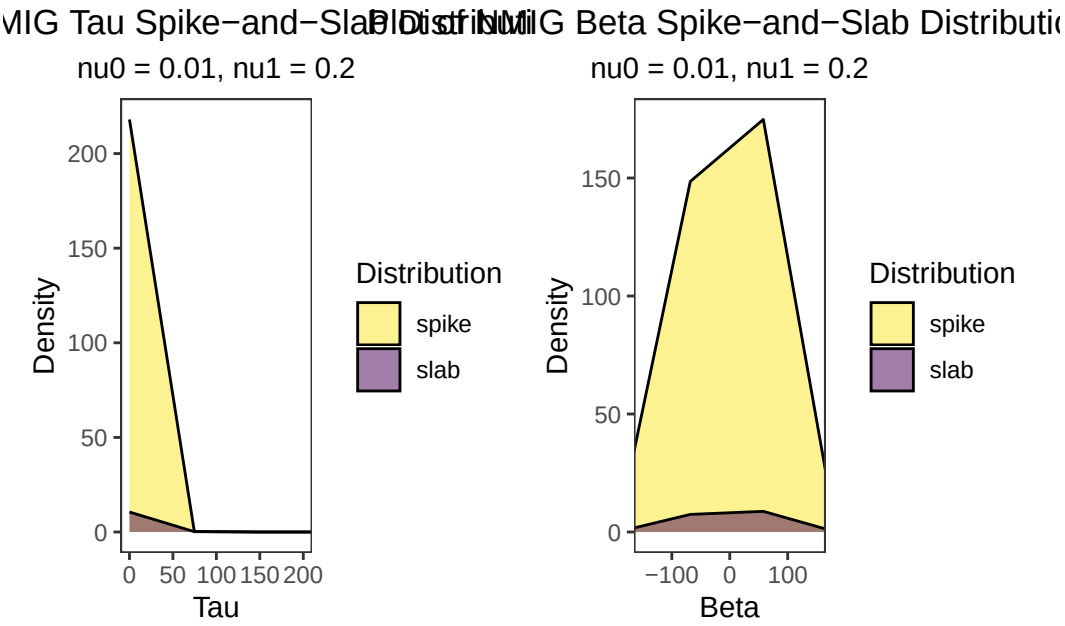
Given in (Fahrmeir et al., 2010) and Ishwaran and Rao (2003, 2005, find citations), The hierarchical prior for $\beta_j|\tau_j^2$ is

$$\beta_j|\tau_j^2 \sim N(0, \tau_j^2)\tau_j|\gamma_j \sim (1-\gamma_j)\text{IG}(a_\tau, \nu_0 b_\tau) + \gamma_j \text{IG}(a_\tau, \nu_1 b_\tau)$$

By placing the spike and slab priors on the variances instead of the coefficients themselves, we can have some robustness against tuning parameters (Lu & Lou, 2022).

```
nu_0 <- 1e-2
nu_1 <- 0.2

a_tau <- 1
b_tau <- 1

N_samples <- 100000

spike_tau <- 1 / rgamma(N_samples, a_tau, nu_0 * b_tau)
slab_tau <- 1 / rgamma(N_samples, a_tau, nu_1 * b_tau)

spike_beta <- rnorm(N_samples, 0, spike_tau)
slab_beta <- rnorm(N_samples, 0, slab_tau)

#data.frame(spike_tau, slab_tau, spike_beta, slab_beta) %>%
  #pivot_longer(everything()) %>%
  #separate(name, into = c("distribution", "level"), sep = "_") %>%
  #ggplot(aes(x = value, fill = distribution)) +
  #geom_density(alpha = 0.3) +
  #facet_wrap(~level, scales = "free") +
  #coord_cartesian(xlim = c(-100, 100))


nmig_tau <- data.frame(spike_tau, slab_tau) %>%
  pivot_longer(everything()) %>%
  mutate(name = substring(name, 1, nchar(name) - 4)) %>%
  mutate(name = factor(name, levels = c("spike", "slab"))) %>%
  ggplot(aes(x = value, fill = name)) +
  geom_density(alpha = 0.5) +
  coord_cartesian(xlim = c(0, 200)) +
  theme_bw() +
  labs(x = "Tau", y = "Density", fill = "Distribution",
```

```
       title = "Plot of NMIG Tau Spike-and-Slab Distribution",
       subtitle = "nu0 = 0.01, nu1 = 0.2") +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        panel.grid = element_blank()) +
  scale_fill_viridis(discrete = TRUE, direction = -1)


nmig_beta <- data.frame(spike_beta, slab_beta) %>%
  pivot_longer(everything()) %>%
  mutate(name = substring(name, 1, nchar(name) - 5)) %>%
  mutate(name = factor(name, levels = c("spike", "slab"))) %>%
  ggplot(aes(x = value, fill = name)) +
  geom_density(alpha = 0.5) +
  coord_cartesian(xlim = c(-150, 150)) +
  theme_bw() +
  labs(x = "Beta", y = "Density", fill = "Distribution",
       title = "Plot of NMIG Beta Spike-and-Slab Distribution",
       subtitle = "nu0 = 0.01, nu1 = 0.2") +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        panel.grid = element_blank()) +
  scale_fill_viridis(discrete = TRUE, direction = -1)

nmig_tau + nmig_beta
```



Shrinkage Priors

Shrinkage priors focus on simply pulling some of the coefficients towards zero, while retaining the stron geffects with minimum penalization. Shrinkage priors are continuous, as opposed to the spike-and-slab prior. As noted in (Bhadra et al., 2017), there are many settings (such as genomics) where many effects are negligible but not zero; this creates an argument for one-group global-local shrinkage priors, rather than two-group spike-and-slab priors.

LASSO Prior

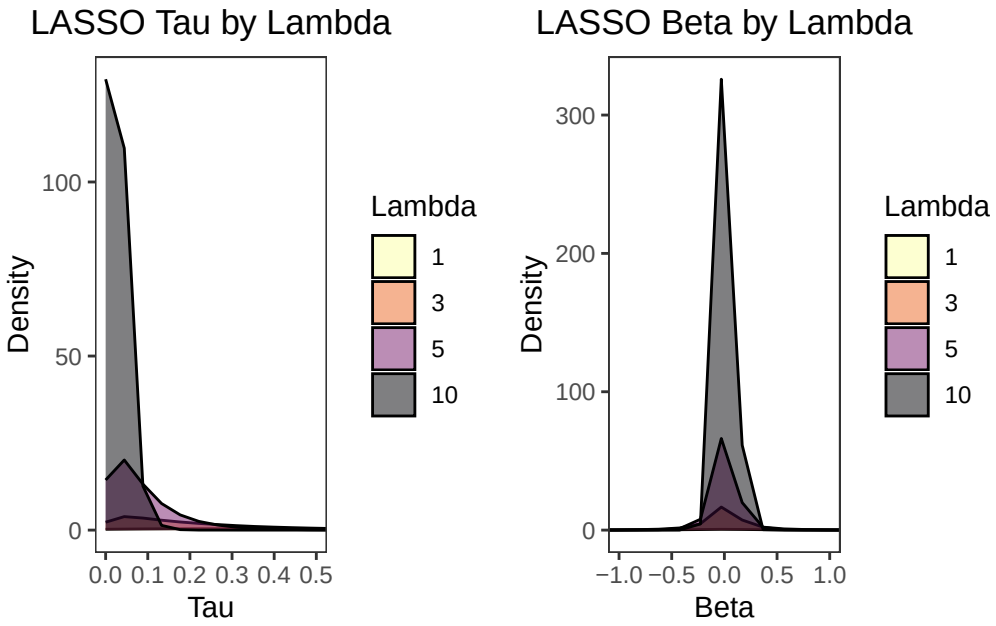
Originally written about in (Park & Casella, 2008), we can write the Laplace prior as

$$\beta_j|\tau_j \sim N(0, \sigma^2\tau_j^2)\tau_j^2|\lambda \sim \exp(\lambda^2/2)$$

We can Gibbs sample the λ parameter, instead of having to do cross-validation.

```
lambda <- c(1, 3, 5, 10)
sigma_2 <- 1

N_samples <- 100000

tau_values <- matrix(NA, nrow = N_samples, ncol = length(lambda))
beta_values <- matrix(NA, nrow = N_samples, ncol = length(lambda))
```

```r
for(l in 1:length(lambda)){
  tau_values[,l] <- rexp(N_samples, lambda[l]^2 / 2)
}

for(l in 1:length(lambda)){
  beta_values[,l] <- rnorm(N_samples, 0, sigma_2 * tau_values[,l])
}

lasso_tau <- data.frame(tau_values) %>%
  pivot_longer(everything(), names_to = "lambda_number") %>%
  mutate(lambda_number = as.integer(substring(lambda_number, 2))) %>%
  mutate(lambda = lambda[lambda_number]) %>%
  mutate(lambda = factor(lambda)) %>%
  ggplot(aes(x = value, fill = lambda)) +
  geom_density(alpha = 0.5) +
  coord_cartesian(xlim = c(0, .5)) +
  theme_bw() +
  labs(x = "Tau", y = "Density", fill = "Lambda",
       title = "LASSO Tau by Lambda") +
  theme(plot.title = element_text(hjust = 0.5),
        panel.grid = element_blank()) +
  scale_fill_viridis(discrete = TRUE, option = "B", direction = -1)


lasso_beta <- data.frame(beta_values) %>%
  pivot_longer(everything(), names_to = "lambda_number") %>%
  mutate(lambda_number = as.integer(substring(lambda_number, 2))) %>%
  mutate(lambda = lambda[lambda_number]) %>%
  mutate(lambda = factor(lambda)) %>%
  ggplot(aes(x = value, fill = lambda)) +
  geom_density(alpha = 0.5) +
  coord_cartesian(xlim = c(-1, 1)) +
  theme_bw() +
  labs(x = "Beta", y = "Density", fill = "Lambda",
       title = "LASSO Beta by Lambda") +
  theme(plot.title = element_text(hjust = 0.5),
        panel.grid = element_blank()) +
  scale_fill_viridis(discrete = TRUE, option = "B", direction = -1)

lasso_tau + lasso_beta
```



Note to myself: This might actually be why our sampler isn't working. Let's incorporate this into the actual thesis.

Normal-Gamma Prior

This is similar to the LASSO regression, but solves some of the problems. Bayesian LASSO is suboptimal: the shrinkage effect is too weak for small coefficients. The Normal-Gamma assumes the mxing distribution in the scale normal of mixture has a gamma distribution, given by:

$$\beta_j|\tau_j \sim N(0, \tau_j^2)\tau_j^2 \sim \text{gamma}(\lambda, 1/(2\xi^2))$$

In this case, both hyperparameters are important. Per (Griffin & Brown, 2010), this results in a distribution that places a lot of mass close to 0, but also has heavy tails, particularly as λ decreases. In many ways, this can be considered similar to a spike-and-slab prior with the right formulation.

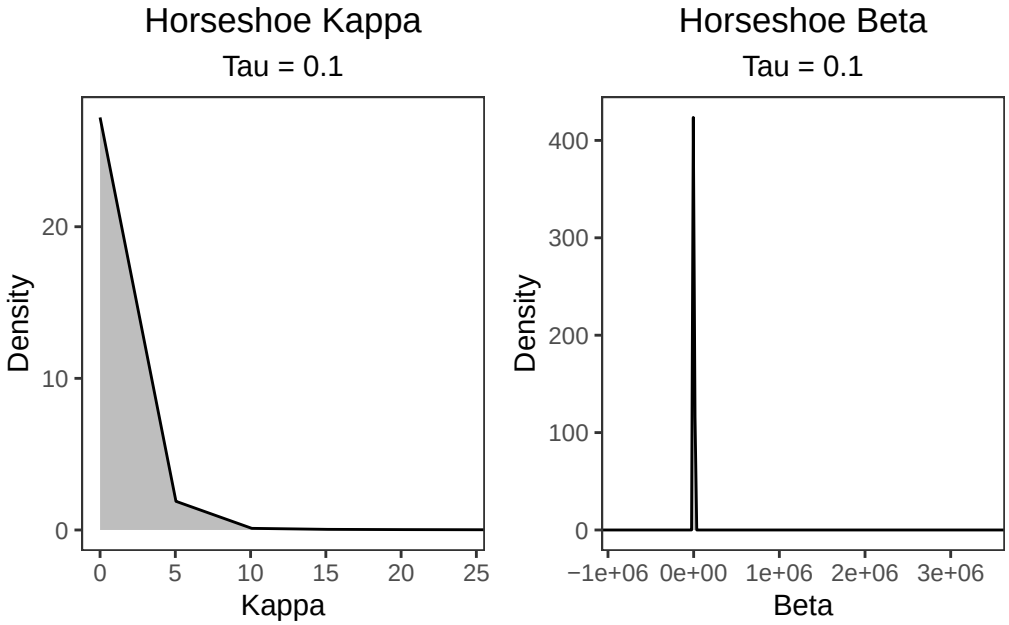
Horseshoe Prior

The horseshoe class of priors uses a global hyperparameter to shrink all coefficients towards zero, with a local hyperparameter to allow some coefficients to adjust the scale of shrinkage at the local level. These are known as **global-local** shrinkage priors. It can be represented as a scale mixture of normals,

$$\beta_j|\kappa_j \sim N(0, \kappa_j^2)\kappa_j|\tau \sim C^+(0, \tau)\tau|\sigma \sim C^+(0, \sigma)$$

In this case, $C^+(0, \tau)$ is a half-Cauchy distribution for the standard deviation κ_j. The κ_j is referred to as the local shrinkage parameter, while τ is gthe global shrinkage parameter.

```
tau <- 0.1

N_samples <- 100000

kappa <- rcauchy(N_samples * 2, 0, tau)
kappa <- kappa[kappa > 0]

beta <- rnorm(length(kappa), 0, kappa^2)

horseshoe_kappa <- ggplot(mapping = aes(x = kappa)) +
  geom_density(fill = "grey") +
  coord_cartesian(xlim = c(0, quantile(kappa, 0.9975))) +
  theme_bw() +
  labs(x = "Kappa", y = "Density", title = "Horseshoe Kappa",
       subtitle = "Tau = 0.1") +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        panel.grid = element_blank())

horseshoe_beta <- ggplot(mapping = aes(x = beta)) +
  geom_density(fill = "grey") +
  coord_cartesian(xlim = quantile(beta, c(.00001, .99999))) +
  theme_bw() +
  labs(x = "Beta", y = "Density", title = "Horseshoe Beta",
       subtitle = "Tau = 0.1") +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        panel.grid = element_blank())

horseshoe_kappa + horseshoe_beta
```



Going to need to set a seed for this one, it has absurdly different outcomes based on the run.

Per (Carvalho et al., 2010), this model has benefits in not needing to select any hyperparameters, since all unknowns have fully specified priors. This is different from most similar approaches in shrinkage models. The prior remains robust and highly adaptive even in the absence of these hyperparameters. The other benefit, compared to other models, is that it converges to the correct answer in sparse situations extremely quickly, while still demonstrating strong robustness against noise for obvious signals, which it leaves unshrunk.

Horseshoe+ Prior

An extension of the horseshoe prior is the horseshoe+ prior,

$$\beta_j|\kappa_j \sim N(0, \kappa_j^2)\kappa_j|\tau, \eta_j \sim C^+(0, \tau\eta_j)\eta_j \sim C^+(0, 1)$$

This extra latent variable provides an extra layer of local shrinkage, which is (in practice) often better in performance in terms of both MSE and computation time when dealing with ultra-sparse signals.

In terms of signals, the regression setting for noisy signals differentiates signals ($\kappa_i = 0$) from noise ($\kappa_i = 1$). The horseshoe+ prior has a "horseshoe" shape pushing posterior mass to either $\kappa_i = 0$ or $\kappa_i = 1$, which the horseshoe prior does not.

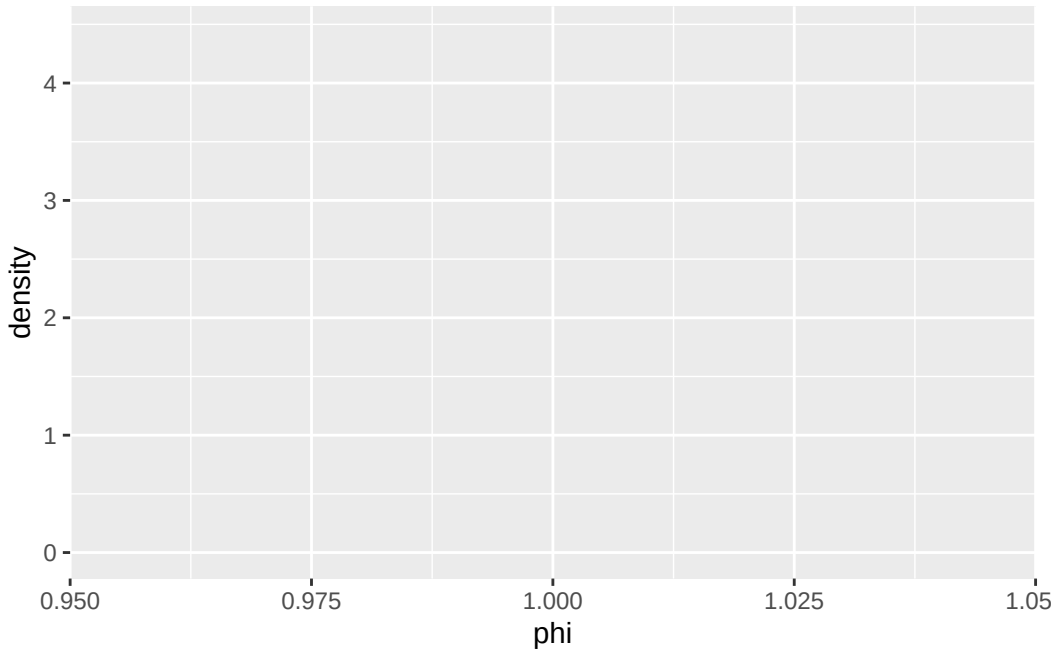
Dirichlet-Laplace Prior

This is another global-local shrinkage prior:

$$\beta_j|\sigma^2, \phi_j, \psi_j, \tau \sim N(0, \sigma^2\phi_j^2\psi_j\tau^2)\psi_j \sim \exp(1/2)(\phi_1, \dots, \phi_p) \sim \text{Dir}(a, \dots, a)\tau \sim \text{gamma}(pa, 1/2)$$

In this case, ϕ_j is the local shrinkage parameter, whereas τ is the global shrinkage parameter. $\text{Dir}(a, \dots, a)$ is the Dirichlet distribution with concentration vector $(a, \dots, a)$. This can be a hyperprior, but we can also choose: $a = 1/n$ if $p > n$ or there is a strong correlation between covariates; $a = 1/2$ when p is small and there is only moderate correlation between covariates.

```
p <- 10
a <- 1/2
sigma_2 <- 1

N_samples <- 100000

psi <- rexp(N_samples, 1/2)
phi <- rdiric(N_samples, a)
tau <- rgamma(N_samples, p * a, 1/2)
beta <- rnorm(N_samples, 0, sigma_2 * psi * phi * tau^2)

dl_psi <- ggplot(mapping = aes(x = psi)) +
  geom_density(fill = "grey") +
  theme_bw() +
  labs(x = "Psi", y = "Density", title = "Psi",
       subtitle = "Dirichlet-Laplace Prior") +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        panel.grid = element_blank())

ggplot(mapping = aes(x = phi)) +
  geom_density(fill = "grey")
```



This might not be super possible to plot, due to the nature of a Dirichlet distribution in high dimensions. We will see. There is a good paper (Bhattacharya et al., 2014), but I am not going to explain the nonsense happening with this distribution unless it somehow becomes demonstrably the best performing test.

Hybrid Priors

SSLASSO

The SSLASSO process considers a mixture of two Laplace distributions, where $\phi_1(\beta_j) = \frac{\lambda_1}{2}\exp\{-\lambda_1|\beta_j|\}$ with small λ_1 and $\lambda_0(\beta_j) = \frac{\lambda_0}{2}\exp\{-\lambda_0|\beta_j|\}$ with large λ_0. Using a binomial prior on γ,

$$\pi(\ |\ ) = \prod_{i=1}^{p} \left[\gamma_j\phi_1(\beta_j) + (1-\gamma_j)\phi_0(\beta_j)\right]$$

If $\phi_1(\beta_j) = \phi_0(\beta_j)$, we have the standard Lasso penalty.

Other Methods

Bayesian Model Averaging

In BMA, you essentially have a vector of latent variables that says whether or not each beta should be included in the model, and you sample this along with your other variables in the MCMC sampler. Remarkably, averaging over the outputs from these has shockingly good performance, and often provides a good estimate for the true model parameters without doing the full 2^p search. Many descriptions of this and similar approaches can be found in (Hoeting et al., 1999).

Best Subset Selection

This is talked about extensively in (Kowal, 2022). This approach de-empasizes *best* subset selection, which is often unstable under permutations of the data, and instead emphasizes *acceptable* or *good enough* subset selection, which is more stable and also simplifies the computation time. They use a modified Branch-and-Bound Algorithm (BBA) to implement this in the Bayesian setting. This approach focuses on the ℓ_0 norm, rather than ℓ_1 or ℓ_2.

References

Bhadra, A., Datta, J., Polson, N. G., & Willard, B. (2017). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis*, *12*(4), 1105–1131. https://doi.org/10.1214/16-BA1028

Bhattacharya, A., Pati, D., Pillai, N. S., & Dunson, D. B. (2014). Dirichlet-Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, *110*(512), 1479–1490. https://doi.org/10.1080/01621459.2014.960967

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*(2), 465–480. https://doi.org/10.1093/biomet/asq017

Fahrmeir, L., Kneib, T., & Konrath, S. (2010). Bayesian regularisation in structured additive regression: A unifying perspective on shrinkage, smoothing, and predictor selection. *Statistical Computing*, *20*, 203–219. https://doi.org/10.1007/s11222-009-9158-3

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881–889. https://doi.org/10.1080/01621459.1993.10476353

Griffin, J. E., & Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, *5*(1), 171–188. https://doi.org/10.1214/10-BA507

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*(4), 382–417. https://doi.org/10.1214/ss/1009212519

Kowal, D. R. (2022). Bayesian subset selection and variable importance for interpretable prediction and classification. *Journal of Machine Learning Research*, *23*, 1–38. https://doi.org/10.48550/arXiv.2104.10150

Lu, Z., & Lou, W. (2022). Bayesian approaches to variable selection: A comparative study from practical perspectives. *The International Journal of Biostatistics*, *18*(1), 83–108. https://doi.org/10.1515/ijb-2020-0130

Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the America Statistical Association*, *103*(482), 681–686. https://doi.org/10.1198/016214508000000337

Rockova, V. (2013). *Bayesian variable selection in high-dimensional applications.*