1 Basic set up

Consider the linear regression model

$$y = X\beta + \epsilon$$

where $y = (y_1, \ldots, y_n)'$ is a vector of response and X is a $n \times p$ covariate matrix and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)'$ is a vector of errors. We assume for simplicity $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$. We deal with the case when p is possibly larger than n.

(β, σ^2) is our unknown parameter and the goal of Bayesian inference is to place appropriate prior distributions on (β, σ^2). The posterior distribution is given by

$$\pi(\beta, \sigma^2 \mid y, X) = \frac{\pi(\beta, \sigma^2)\mathrm{N}(y; X\beta, \sigma^2 I)}{\int \pi(\beta, \sigma^2)\mathrm{N}(y; X\beta, \sigma^2 I) d\beta d\sigma^2}$$

When $n \gg p$, $\pi(\beta, \sigma^2 \mid y, X) \approx \mathrm{N}(\beta; \hat{\beta}, I(\beta)^{-1})$, where $I(\beta)$ is the Fisher Information matrix. The above is called the Bernstein-von Mises theorem or the Bayesian central limit theorem. This essentially means that when $n \gg p$, prior does not have much role in determining the posterior. In fact, the likelihood swamps the prior and we essentially get equivalent results from frequentist and Bayesian inference. This rosy picture breaks down when p is large. Prior has profound effect for large p and it is essential to carefully design the prior.

Priors should be designed in such a way that the posterior of β concentrates around the "true" β_0. Also, its pretty much impossible to recover the true parameter β_0 unless the effective dimension of β_0 is smaller or equal to the sample size. The goal of this section is to design appropriate priors which takes into account such information. On the other hand, a prior should be devised in an objective way in the sense that we cannot incorporate too much knowledge about the true parameter β_0 other than assuming that its sparse.

Motivated by the idea of sparsity, one popular approach is to impose sparsity on β through prior distributions. One natural prior is to consider

$$\beta_j \sim \pi_0 \delta_0 + (1 - \pi_0) g \tag{1}$$

where $\pi_0 \in [0, 1]$ and g is a suitable density on $\mathbb{R}$. The density g is supposed to capture the non-zero coefficients and π_0 is supposed to mimic the true sparsity. But you do not know the true sparsity. How to circumvent the problem?

1.1 Choosing π_0

Define the variable inclusion indicator by $\gamma_j = I(\beta_j \neq 0)$. Therefore, $\gamma_1, \ldots, \gamma_p$ indicate which predictors are included in the model, $\gamma = (\gamma_1, \ldots, \gamma_p)$. Note that, depending on whether a variable is included or excluded, the total number of candidate models is 2^p. A candidate model is represented

by γ. The size of this model $p_\gamma = \sum_{j=1}^{p} \gamma_j$. Then $p_\gamma \sim \text{Binomial}(p, 1-\pi_0)$ and the expected model size is $p(1-\pi_0)$.

Note that the joint distribution of (1) can also be described in the following hierarchical manner.

$$\begin{aligned} p_\gamma &\sim \text{Bernoulli}(p, 1-\pi_0) \\ \gamma &\sim \text{Uniform}\{\gamma : \sum_{j=1}^{p} \gamma_j = p_\gamma\}, \quad \beta_j \mid \gamma_j = 1 \sim g, \quad \beta_j \mid (\gamma_j = 0) = 0. \end{aligned}$$

Clearly, if we fix π_0 and p is big, it gives a lot of prior information on the model size. Clearly the *innocuous* looking prior with $\pi_0 = 0.5$ is highly informative since it forces the expected model size to be $p/2$. Not only that, under this distribution $P(p_\gamma = s) = \binom{p}{s} 2^{-p} \asymp e^{s \log p/s} 2^{-p}$ which is extremely small for small values of s. One way to penalize large models is by placing a hyperprior on π_0. π_0 is an important parameter and generally assigned a beta prior. If $\pi_0 \sim U(0,1)$, then the marginal distribution of p_γ becomes a discrete uniform on $\{0, 1, \ldots, p\}$. Here also the expected value is $p/2$ but the mass is uniformly distribution on all possible values of p_γ rather than concentrated on $p/2$ as in setting $\pi_0 = 1/2$. More aggressive shrinkage can be achieved by having $\pi_0 \sim \text{Beta}(1, \kappa p)$ where κ is a tuning parameter.

The prior distribution on the space of models plays a crucial role in high-dimensional settings and is connected to multiplicity adjustment (Scott and Berger, 2010). $\pi_0 \sim U(0,1)$ corresponds to the Scott–Berger prior which assigns equal prior probability to all model sizes, and then distributes the probability evenly to all models of the same size. The resulting prior has the form

$$\pi(\gamma) = \frac{1}{p+1} \frac{1}{\binom{p}{|\gamma|}} \tag{2}$$

As discussed before, more aggressive penalty on the model size can be obtained by replacing the $U(0,1)$ prior on π_0 with a $\text{Beta}(1, \kappa\, p)$ prior on π_0 for some constant $\kappa > 0$ which leads to an *a priori* $O(1)$ expected model size. Such priors are special examples of complexity priors, for which the prior on the model size p_γ satisfies an exponential decay $\pi(p_\gamma) \propto e^{-a p_\gamma \log(bp/p_\gamma)}$. A similar prior considered by Yang et al 2015 assumes the form

$$\pi(\gamma) \propto \left(\frac{1}{p}\right)^{\kappa|\gamma|} I(|\gamma| \leq s_0), \tag{3}$$

where $s_0 \leq n$. This prior also implies a penalty of the order $e^{-p_\gamma \log p}$ for models of size p_γ akin to (2), and additionally assigns zero prior probability to any model with size larger than a specified cut-off s_0. A default choice of $s_0 = n$ precludes the model size from exceeding the sample size.

1.2 Spike and slab variable selection

Consider the prior

$$\pi(\beta) = \prod_{j=1}^{p} \{\delta_0(\beta_j)\pi_{0j} + (1-\pi_{0j}) N(\beta_j; 0, c_j^2)\}$$

where π_{0j} is the prior probability of excluding the j-th predictor by setting its coefficient to 0. Introducing the γ, this prior can also be written hierarchically as

$$\begin{aligned}\gamma \sim \pi(\gamma) &= \binom{p}{|\gamma|}\left\{\prod_{j=1}^{p}(1-\pi_{0j})^{\gamma_j}\pi_{0j}^{1-\gamma_j}\right\}\times\frac{1}{\binom{p}{|\gamma|}}\\ \pi(\beta_j \mid \gamma_j) &= \delta_0(\beta_j)(1-\gamma_j)+\gamma_j N(\beta_j;0,c_j^2).\end{aligned}$$

It is important to point out here is that $\gamma \mid \beta$ is fully deterministic and it does not make sense to do a Gibbs sampling of (β,γ) as $\beta \mid \gamma, y, X$ and $\gamma \mid \beta, y, X$. In the following, we develop a Gibbs sampler that converges to $\beta \mid y, X$ by cycling through the updates $\beta_j \mid \beta_{-j}, y, X$. Gibbs sampler proceeds by sampling from conditional posterior of β_j , for bioske $j = 1, \ldots, p$,

$$\pi(\beta_j \mid \beta_{-j}, y, X) = \hat{\pi}_j\delta_0(\beta_j) + (1-\hat{\pi}_j)\mathrm{N}(\beta_j; E_j, V_j)$$

where $V_j = (c_j^{-2} + X_j'X_j)^{-1}, E_j = V_jX_j'(y - X_{-j}\beta_{-j}), X_j = j$ th column of X, $X_{-j} = X$ with j th column excluded, $\beta_{-j} = \beta$ with j th element excluded, and

$$\hat{\pi}_j = \frac{\pi_{0j}}{\pi_{0j} + (1-\pi_{0j})\frac{N0;0,c_j^2)}{N(0;E_j,V_j)}}$$

is the conditional probability of $\beta_j = 0$. Clearly this algorithm is slow as it updates β_j one at a time and the Markov chain for β has high autocorrelation. This can be circumvented by the marginalizing out the β. Denote by $\beta_S = \{\beta_j : \gamma_j = 1\}$ where S denotes the $\{j : \gamma_j = 1\}$. Denote by π_S the joint prior on β_S. In this case $\pi_S = \prod_{j\in S}\mathrm{N}(\beta_j : 0, c_j^2)$. Then

$$\pi(\gamma \mid y, X) = \frac{\int p(y \mid X_S, \beta_S)\pi_S(\beta_S)d\beta_S}{\sum_{S\subset\{1,\ldots,p\}}\int p(y \mid X, \beta_S)\pi_S(\beta_S)d\beta_S}$$

An example of such a model is given by the g-prior as in Yang et al 2015.

$$\begin{aligned}Y \mid \beta_S, S &\sim \mathrm{N}(X_S\beta_S, 1/\phi I_n)\\ \beta_S \mid S &\sim \mathrm{N}(0, g\phi^{-1}(X_S'X_S)^{-1}), \pi(\phi) = 1/\phi\\ \pi(\gamma) &\propto \left(\frac{1}{p}\right)^{\kappa|\gamma|} I(|\gamma| \le s_0).\end{aligned}$$

One can develop a Metropolis Hastings algorithm to update γ. In general terms, a Metropolis-Hastings random walk is an iterative and local-move based procedure involving three steps:

1. Use the current state γ to define a neighborhood $N(\gamma)$ of proposal states.
2. Choose a proposal state γ' in $N(\gamma)$ according to some probability distribution $S(\gamma, \cdot)$ over the neighborhood, e.g. the uniform distribution.
3. Move to the new state γ' with probability $R(\gamma, \gamma')$, and stay in the original state γ with probability $1 - R(\gamma, \gamma')$, where the acceptance ratio is given by

$$R(\gamma, \gamma') = \min\left\{1, \frac{\pi(\gamma' \mid Y)S(\gamma', \gamma)}{\pi(\gamma \mid Y)S(\gamma, \gamma')}\right\}$$

The specific form of Metropolis-Hastings update analyzed in this section is obtained by randomly selecting one of the following two schemes to update γ, each with probability 0.5.

1. Single flip update: Choose an index $j \in \{1, \ldots, p\}$ uniformly at random, and form the new state γ' by setting $\gamma'_j = 1 - \gamma_j$
2. Double flip update: Define the subsets $S(\gamma) = \{j \in \{1, \ldots, p\} : \gamma_j = 1\}$. Choose an index pair $(k, l) \in S(\gamma) \times S^c(\gamma)$ uniformly at random, and form the new state γ' by flipping γ_k from 1 to 0 and γ_l from 0 to 1. (If the set $S(\gamma)$ is empty, then we do nothing.)

1.3 Speeding up computation: Block update of β_j

Due to the intractability of calculating the posterior probabilities exactly, stochastic search is often used. Stochastic Search Variable Selection (SSVS) moves between multiple models and comes back to models which are more representative of the data. SSVS (George & McCulloch, 1993, JASA) rely on MCMC to conduct this search.

$$\beta_j \sim \pi_{0j} N(0, \nu_{0j}) + (1 - \pi_{0j}) N(0, \nu_{1j}),$$

where ν_{0j} small, ν_{1j} "reasonably" big (away from 0). George & McCulloch suggested taking $\nu_{0j} = \tau_j^2$, $\nu_{1j} = g_j^2 \tau_j^2$, where g_j big, τ_j^2 is small. Introduce $\gamma = (\gamma_1, \ldots, \gamma_p)$ such that

$$\beta_j \mid \gamma_j \sim (1 - \gamma_j) N(0, \nu_{0j}) + \gamma_j N(0, \nu_{1j}),$$

Note here that having known γ does not specify β completely. SO its possible to derive a Gibbs sampler for (β, γ). Note that

$$\pi(\beta, \gamma, \sigma^2) = \left[\prod_{j=1}^{p} \pi(\beta_j \mid \sigma^2, \gamma_j) \right] \pi(\sigma^2)$$

and note that

$$\beta \mid \gamma \sim \mathrm{N}(0, D)$$

where $D = \mathrm{diag}(a_1 \tau_1^2, \ldots, a_p \tau_p^2)$ where $a_j = 1$ if $\gamma_j = 0$ and $a_j = g_j^2$ if $\gamma_j = 1$. Thus

$$\begin{aligned} \pi(\beta \mid -) &\propto N(y \mid X\beta, \sigma^2 I) N(\beta \mid 0, D) \\ P(\gamma_j = 1 \mid -) &= h_1/(h_1 + h_2), h_1 = \pi_{0j} \mathrm{N}(\beta_j \mid 0, g^2 \tau_j^2), h_2 = (1 - \pi_{0j}) \mathrm{N}(\beta_j \mid 0, \tau_j^2). \end{aligned}$$

2 Continuous shrinkage priors

Although point mass mixture priors are intuitively appealing and possess attractive theoretical properties, posterior sampling requires a stochastic search over an enormous space, leading to slow mixing and convergence. Computational issues and consideration that many of the θ_js may be small but not exactly zero has motivated a rich literature on continuous shrinkage priors. Polson

et al 2010 noted that essentially all such shrinkage priors can be represented as global-local (GL) mixtures of Gaussians,

$$\theta_j \sim \mathrm{N}(0, \psi_j \tau), \quad \psi_j \sim f, \quad \tau \sim g, \tag{4}$$

where τ controls global shrinkage towards the origin while the local scales $\{\psi_j\}$ allow deviations in the degree of shrinkage. If g puts sufficient mass near zero and f is appropriately chosen, GL priors in (4) can intuitively approximate point mass priors but through a continuous density concentrated near zero with heavy tails.

2.1 Robust estimation of sparse signals

Consider $y \sim N(\theta, 1)$. The prior for the mean is $\pi(\theta)$. The marginal density is given by $m(y) = \int p(y-\theta)\pi(\theta)d\theta$. For one sample of y,

$$E(\theta \mid y) = y + \frac{d}{dy} \log m(y)$$

The following result speaks to the horseshoe's robustness to large outlying signals.

THEOREM **2.1**. *Suppose $y \sim N(\theta, 1)$. Let $m(y)$ denote the predictive density under the horseshoe prior for known scale parameter $\tau < \infty$, i.e. where $(\theta \mid \lambda) \sim N(0, \tau^2\lambda^2)$ and $\lambda \sim Ca^+(0, 1)$. Let $E(\theta \mid y)$ denote the posterior mean. Then $\lim_{|y|\to\infty} d \log m(y)/dy = 0$.*

This is NOT true for Bayesian Lasso (Park and Casella, 2008) given by

$$\theta_j \mid \tau \sim DE(\tau) \Leftrightarrow \theta_j \mid \tau, \psi_j \sim N(0, \psi_j\tau^2), \psi_j \sim \mathrm{Exp}(1/2)$$

In sparse situations, posterior learning τ allows most noise observations to be shrunk very near zero. Yet this small value of τ will not inhibit the estimation of large signals. Under the double-exponential prior, for example, small values of τ can also lead to strong shrinkage near the origin. This shrinkage, however, can severely compromise performance in the tails. For DE, smaller value of τ may reduce the risk at the origin, But do so at the expense of increased risk in the tails $|E(\theta_i \mid y_i) - y_i| \approx \sqrt{2}/\tau$ for large y_i.

2.2 Dirichlet Kernel priors

Let ϕ_0 denote the standard normal density on $\mathbb{R}$. Also, let $\mathrm{DE}(\tau)$ denote a zero mean double-exponential or Laplace distribution with density $f(y) = (2\tau)^{-1}e^{-|y|/\tau}$ for $y \in \mathbb{R}$. Integrating out the local scales ψ_j's, (4) can be equivalently represented as a global scale mixture of a kernel $\mathcal{K}(\cdot)$,

$$\theta_j \overset{\text{i.i.d.}}{\sim} \mathcal{K}(\cdot\,, \tau), \quad \tau \sim g, \tag{5}$$

where $\mathcal{K}(x) = \int \psi^{-1/2}\phi_0(x/\sqrt{\psi})g(\psi)d\psi$ is a symmetric unimodal density on $\mathbb{R}$ and $\mathcal{K}(x, \tau) := \tau^{-1/2}\mathcal{K}(x/\sqrt{\tau})$. For example, $\psi_j \sim \mathrm{Exp}(1/2)$ corresponds to a double-exponential kernel $\mathcal{K} \equiv \mathrm{DE}(1)$, while $\psi_j \sim \mathrm{IG}(1/2, 1/2)$ results in a standard Cauchy kernel $\mathcal{K} \equiv \mathrm{Ca}(0, 1)$.

These choices lead to a kernel which is *bounded* in a neighborhood of zero. However, if one instead uses a half Cauchy prior $\psi_j^{1/2} \sim \mathrm{Ca}_+(0, 1)$, then the resulting horseshoe kernel is unbounded with

a singularity at zero. This phenomenon coupled with tail robustness properties leads to excellent empirical performance of the horseshoe. However, the joint distribution of θ under a horseshoe prior is understudied and further theoretical investigation is required to understand its operating characteristics. One can imagine that it concentrates more along sparse regions of the parameter space compared to common shrinkage priors since the singularity at zero potentially allows most of the entries to be concentrated around zero with the heavy tails ensuring concentration around the relatively small number of signals.

The above class of priors rely on obtaining a suitable kernel $\mathcal{K}$ through appropriate normal scale mixtures. In this article, we offer a fundamentally different class of shrinkage priors that alleviate the requirements on the kernel, while having attractive theoretical properties. In particular, our proposed class of Dirichlet-kernel (Dk) priors replaces the single global scale τ in (5) by a vector of scales $(\phi_1\tau, \ldots, \phi_n\tau)$, where $\phi = (\phi_1, \ldots, \phi_n)$ is constrained to lie in the $(n-1)$ dimensional simplex $\mathcal{S}^{n-1} = \{x = (x_1, \ldots, x_n)^{\mathrm{T}} : x_j \geq 0, \sum_{j=1}^n x_j = 1\}$ and is assigned a $\mathrm{Dir}(a, \ldots, a)$ prior:

$$\theta_j \mid \phi_j, \tau \sim \mathcal{K}(\cdot\,, \phi_j\tau), \quad \phi \sim \mathrm{Dir}(a, \ldots, a). \tag{6}$$

In (6), $\mathcal{K}$ is any symmetric (about zero) unimodal density with exponential or heavier tails; for computational purposes, we restrict attention to the class of kernels that can be represented as scale mixture of normals. While previous shrinkage priors obtain marginal behavior similar to the point mass mixture priors our construction aims at resembling the *joint distribution* of θ under a two-component mixture prior.

We focus on the Laplace kernel from now on for concreteness, noting that all the results stated below can be generalized to other choices. The corresponding hierarchical prior given τ,

$$\theta_j \mid \phi, \tau \sim \mathrm{DE}(\phi_j\tau), \quad \phi \sim \mathrm{Dir}(a, \ldots, a), \tag{7}$$

is referred to as a Dirichlet–Laplace prior, denoted $\theta \mid \tau \sim \mathrm{DL}_a(\tau)$.

To understand the role of ϕ, we undertake a study of the marginal properties of θ_j conditional on τ, integrating out ϕ_j. The results are summarized in Proposition 2.1 below.

PROPOSITION **2.1**. *If $\theta \mid \tau \sim \mathrm{DL}_a(\tau)$, then the marginal distribution of θ_j given τ is unbounded with a singularity at zero for any $a < 1$. Further, in the special case $a = 1/n$, the marginal distribution is a wrapped Gamma distribution $\mathrm{WG}(\tau^{-1}, 1/n)$, where $WG(\lambda, \alpha)$ has a density $f(x; \lambda, \alpha) \propto |x|^{\alpha-1} e^{-\lambda|x|}$ on $\mathbb{R}$.*

Thus, marginalizing over ϕ, we obtain an unbounded kernel $\mathcal{K}$, so that the marginal density of $\theta_j \mid \tau$ has a singularity at 0 while retaining exponential tails.

The parameter τ plays a critical role in determining the tails of the marginal distribution of θ_j's. We consider a fully Bayesian framework where τ is assigned a prior g on the positive real line and learnt from the data through the posterior. Specifically, we assume a $\mathrm{gamma}(\lambda, 1/2)$ prior on τ with $\lambda = na$. We continue to refer to the induced prior on θ implied by the hierarchical structure,

$$\theta_j \mid \phi, \tau \sim \mathrm{DE}(\phi_j\tau), \quad \phi \sim \mathrm{Dir}(a, \ldots, a), \quad \tau \sim \mathrm{gamma}(na, 1/2), \tag{8}$$

as a Dirichlet–Laplace prior, denoted $\theta \sim \mathrm{DL}_a$. Posterior computation can be done using two different ways:

Exact sampler:

The proposed class of DL priors leads to straightforward posterior computation via an efficient data augmented Gibbs sampler. The DL_a prior (8) can be equivalently represented as

$$\theta_j \sim \text{N}(0, \psi_j \phi_j^2 \tau^2),\ \psi_j \sim \text{Exp}(1/2),\ \phi \sim \text{Dir}(a, \ldots, a),\ \tau \sim \text{gamma}(na, 1/2).$$

We detail the steps in the normal means setting noting that the algorithm is trivially modified to accommodate normal linear regression, robust regression with heavy tailed residuals, probit models, logistic regression, factor models and other hierarchical Gaussian cases. To reduce autocorrelation, we rely on marginalization and blocking as much as possible. Our sampler cycles through (i) $\theta \mid \psi, \phi, \tau, y$, (ii) $\psi \mid \phi, \tau, \theta$, (iii) $\tau \mid \phi, \theta$ and (iv) $\phi \mid \theta$. We use the fact that the joint posterior of (ψ, ϕ, τ) is conditionally independent of y given θ. Steps (ii) - (iv) together give us a draw from the conditional distribution of $(\psi, \phi, \tau) \mid \theta$, since

$$[\psi, \phi, \tau \mid \theta] = [\psi \mid \phi, \tau, \theta][\tau \mid \phi, \theta][\phi \mid \theta].$$

Steps (i) – (iii) are standard and hence not derived. Step (iv) is non-trivial and we develop an efficient sampling algorithm for jointly sampling ϕ. Usual one at a time updates of a Dirichlet vector lead to tremendously slow mixing and convergence, and hence the joint update in Theorem 2.2 is an important feature of our proposed prior; a proof can be found in the Appendix. Consider the following parametrization for the three-parameter generalized inverse Gaussian (giG) distribution: $Y \sim \text{giG}(\lambda, \rho, \chi)$ if $f(y) \propto y^{\lambda-1} e^{-0.5(\rho y + \chi/y)}$ for $y > 0$. For $\lambda = -1/2$ we get the inverse Gaussian distribution (iG).

THEOREM **2.2**. *The joint posterior of $\phi \mid \theta$ has the same distribution as $(T_1/T, \ldots, T_n/T)$, where T_j are independently distributed according to a giG$(a-1, 1, 2|\theta_j|)$ distribution, and $T = \sum_{j=1}^n T_j$.*

Proof. Integrating out τ, the joint posterior of $\phi \mid \theta$ has the form

$$\pi(\phi_1, \ldots, \phi_{n-1} \mid \theta) \propto \prod_{j=1}^n \left[\phi_j^{a-1} \frac{1}{\phi_j}\right] \int_{\tau=0}^{\infty} e^{-\tau/2} \tau^{\lambda-n-1} e^{-\sum_{j=1}^n |\theta_j|/(\phi_j \tau)} d\tau. \tag{9}$$

for $\lambda = an$.

We now state a result from the theory of normalized random measures. Suppose $T_1, \ldots, T_n$ are independent random variables with T_j having a density f_j on $(0, \infty)$. Let $\phi_j = T_j/T$ with $T = \sum_{j=1}^n T_j$. Then, the joint density f of $(\phi_1, \ldots, \phi_{n-1})$ supported on the simplex $\mathcal{S}^{n-1}$ has the form

$$f(\phi_1, \ldots, \phi_{n-1}) = \int_{t=0}^{\infty} t^{n-1} \prod_{j=1}^n f_j(\phi_j t) dt, \tag{10}$$

where $\phi_n = 1 - \sum_{j=1}^{n-1} \phi_j$. Setting $f_j(x) \propto \frac{1}{x^\delta} e^{-|\theta_j|/x} e^{-x/2}$ in (10), we get

$$f(\phi_1, \ldots, \phi_{n-1}) = \left[\prod_{j=1}^n \frac{1}{\phi_j^\delta}\right] \int_{t=0}^{\infty} e^{-t/2} t^{n-1-n\delta} e^{-\sum_{j=1}^n |\theta_j|/(\phi_j t)} dt. \tag{11}$$

We aim to equate the expression in (11) with the expression in (9). Comparing the exponent of ϕ_j gives us $\delta = 2 - a$. The other requirement $n - 1 - n\delta = \lambda - n - 1$ is also satisfied, since $\lambda = na$. The proof is completed by observing that f_j corresponds to a giG$(a-1, 1, 2|\theta_j|)$ when $\delta = 2 - a$. $\diamondsuit$

Summaries of each step are provided below.

(i) Assuming a Gaussian sequence model $y \sim \mathrm{N}(\theta, I_n)$, to sample $\theta \mid \psi, \phi, \tau, y$, draw θ_j independently from a $\mathrm{N}(\mu_j, \sigma_j^2)$ distribution with

$$\sigma_j^2 = \{1 + 1/(\psi_j \phi_j^2 \tau^2)\}^{-1}, \quad \mu_j = \{1 + 1/(\psi_j \phi_j^2 \tau^2)\}^{-1} y.$$

(ii) The conditional posterior of $\psi \mid \phi, \tau, \theta$ can be sampled efficiently in a block by independently sampling $\tilde{\psi}_j \mid \phi, \theta$ from an inverse-Gaussian distribution $\mathrm{iG}(\mu_j, \lambda)$ with $\mu_j = \phi_j \tau / |\theta_j|, \lambda = 1$ and setting $\psi_j = 1/\tilde{\psi}_j$.

(iii) Sample the conditional posterior of $\tau \mid \phi, \theta$ from a $\mathrm{giG}(\lambda - n, 1, 2\sum_{j=1}^n |\theta_j|/\phi_j)$ distribution.

(iv) To sample $\phi \mid \theta$, draw $T_1, \ldots, T_n$ independently with $T_j \sim \mathrm{giG}(a-1, 1, 2|\theta_j|)$ and set $\phi_j = T_j/T$ with $T = \sum_{j=1}^n T_j$.

Collapsed sampler: Note that the distribution of $\psi = \phi \times \tau$ follows a distribution given by $\prod_{j=1}^p \mathrm{Ga}(a, 1/2)$. Then the model simply becomes

$$\theta_j \mid \psi \sim \mathrm{DE}(\psi_j), \psi_j \sim \mathrm{Ga}(a, 1/2).$$

which is equivalent to

$$\theta_j \mid \psi, q \sim N(0, q_j \psi_j^2), \quad q_j \sim \exp(1/2), \quad \psi_j \sim \mathrm{Ga}(a, 1/2).$$

So the full conditionals of q_j and ψ_j are Inverse Gaussian and generalized inverse Gaussian respectively.

3 Variational Inference with spike and slab priors

Going back to the linear regression model

$$Y = X\beta + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2)$$

with

$$\beta_j \mid \gamma_j \sim \delta_0(\beta_j)(1 - \gamma_j) + \gamma_j \mathrm{N}(\beta_j; 0, \sigma^2 \sigma_\beta^2)\}.$$

Here $P(\gamma_j = 1) = \pi$. Since results can be sensitive to the choice of hyperparameters $\theta = (\sigma^2, \sigma_\beta^2, \pi)$, we estimate θ from the data by introducing a prior on θ, and integrating over values of θ. We do not assume a specific form for the prior on θ one feature of our variational method is that it works with any prior on the hyperparameters.

A collection of choices of variational approximation for $(\beta, \gamma, \sigma^2)$ is given by

1. $q(\beta, \gamma, \sigma^2) = q(\beta, \gamma) q(\sigma^2)$
2. $q(\beta, \gamma, \sigma^2) = \prod_{j=1}^p q(\beta_j, \gamma_j) q(\sigma^2)$
3. $q(\beta, \gamma, \sigma^2) = q(\beta) q(\sigma^2) \prod_{j=1}^p q(\gamma_j)$

We begin by decomposing the posterior inclusion probabilities as

$$PIP(j) = p(\gamma_j = 1|X, y, \theta)p(\theta|X, y)d\theta.$$

There are two components to our our inference strategy. One component approximates posterior probabilities $p(\gamma_j = 1|X, y, \theta)$ by minimizing the Kullback-Leibler divergence between an approximating distribution on β, γ and the posterior of β, γ given θ. The second component estimates $p(\theta|X, y)$ by importance sampling, using the variational solution from the first component to compute the importance weights.

We shall work with the representation (2). In particular, we let

$$q(\beta_j, \gamma_j) = \alpha_j N(\beta_j; \mu_j, s_j^2) + (1 - \alpha_j)\delta_0(\beta_j).$$

The coordinate descent updates for this optimization problem can be obtained by taking partial derivatives of the ELBO and setting the partial derivatives to zero, and solving for the parameters α_j, μ_j and s_j^2. This yields coordinate updates

$$\begin{aligned} s_j^2 &= \frac{\sigma^2}{(X'X)_{jj} + 1/\sigma_\beta^2} \\ \mu_j &= \frac{s_j^2}{\sigma^2}[(X'y)_j - \sum_{k \neq j}(X'X)_{kj}\alpha_k\mu_k] \\ \frac{\alpha_j}{1 - \alpha_j} &= \frac{\pi}{1 - \pi} \times \frac{s_j}{\sigma_\beta \sigma} e^{\mu_j^2/2(s_j^2)} \end{aligned}$$