Annotated Bibliography

Dav King

Datasets

Psychology Papers

Classical ML Papers

Bayesian Papers

(Lu & Lou, 2022): Bayesian approaches to variable selection

This paper discusses several Bayesian approaches to variable selection, especially focusing on applications in R. It focuses largely on four categories: Bayesian model selection, spike-and-slab priors, shrinkage priors, and the hybrid of both. These categories can be defined:

Bayesian Model Selection

Define a vector of binary latent indicator variables $\gamma = (\gamma_1, \dots, \gamma_p)^T$ that characterize a sub-model $\mathcal{M}_\gamma$:

$$y = \beta_0 + X_\gamma \beta_\gamma + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_n)$$

where $\gamma_j = 1$ indicates that the jth variable is included in the model and $\gamma_j = 0$ otherwise. Now each of the 2^p possible models can be represented by a specific instance of γ. Often, the independent Bernoulli distribution is used. There are many different criteria that can be used for comparing two models under γ, depending on whether conjugate priors are used.

Another common approach is Bayesian model averaging, where β is estimated based on a weighted average over the competing models. This often finds better performance.

Spike-and-Slab Prior

A spike-and-slab prior is a two-point mixture distribution which selects a subset of important predictors by forcing the β_j of non-selected variables to 0, allowing for jointly selecting the variables and estimating their regression coefficients. This prior has a generic form

$$\beta_j|\gamma_j \sim \gamma_j\phi_1\left(\beta_j\right) + (1-\gamma_j)\phi_0\left(\beta_j\right), \quad \gamma \sim \pi(\gamma)$$

where $\phi_1(\beta_j)$ is a diffuse "slab distribution" for modeling large effects and $\phi_0(\beta_j)$ is a concentrated "spike distribution" for modeling negligibly small effects, with γ_j being a binary latent indicator for the 2^p possible models. The spike distribution is either a degenerate distribution at zero or a continuous distribution centered at zero with small variance. Two common versions of spike-and-slab are stochastic search variable selection (SSVS), which uses normal distributions for both $\phi_1(\beta_j)$ and $\phi_0(\beta_j)$, and normal mixture of inverse gamma (NMIG), which imposes a spike-and-slab prior on the variances, indirectly inducing a spike-and-slab prior for regression coefficients (SSVS does this directly). NMIG is less sensitive to tuning parameters than NMIG.

Shrinkage Prior

A shrinkage prior is an absolutely continuous prior distribution which aims to restrict weak effects while retaining strong effects with minimal penalization. With a measure of fit function $l(\theta)$ and a penalty function $\text{pen}_\lambda(\beta)$ (where λ is a tuning parameter) corresponding to the negative logarithms of likelihood and prior distribution, respectively, this becomes an optimization problem:

$$\min_{\beta\in\mathbb{R}^p}\left\{l(y|\beta) + \text{pen}_\lambda(\beta)\right\}$$

This leads to a Bayesian hierarchical model:

$$p(y|\beta) \propto \exp\left\{-l(y|\beta)\right\}, \quad p_\lambda(\beta) \propto \exp\left\{-\text{pen}_\lambda(\beta)\right\}$$

The solution corresponds to the posterior mode of this hierarchical model. Bayesian approaches allow λ to be estimated as part of MCMC sampling, avoiding the need to use cross-validation for hyperparameter tuning. This approach offers the benefit of being able to select variables without needing to search all of $\mathcal{M}$, offering significant computational advantages over both Bayesian model selection and spike-and-slab priors.

Some examples of priors include Lasso-type priors (forcing some variables to 0 using the ℓ_1 norm), normal-gamma priors (offering more adaptive performance than the Bayesian Lasso), horseshoe priors (using a global hyperparameter to shrink all coefficients towards zero and a

local hyperparameter to adjust the scale of shrinkage at the local level), and Dirichlet-Laplace priors (a different form of global-local shrinkage prior).

Performance

This paper describes multiple experiments using a number of different priors, and it talks about R packages that are already set up to perform these calculations. The discussion section has several important findings that I will not copy here. There are particular concerns about two-point mixtures being sensitive to multi-collinearity among predictors, which will absolutely be at play in this project.

(Hoeting et al., 1999): Bayesian model averaging

Let Δ be the quantity of interest - e.g., an effect size, a course of action, etc. Its posterior distribution given data D is

$$\mathbb{P}(\Delta|D) = \sum_{\gamma=1}^{\Gamma} \mathbb{P}(\Delta|\mathcal{M}_\gamma, D)\mathbb{P}(M_\gamma|D)$$

This is an average of the posterior distributions under each of the models considered, weighted by their posterior model probability, which is defined

$$\mathbb{P}(\mathcal{M}_\gamma|D) = \frac{\mathbb{P}(D|\mathcal{M}_\gamma)\mathbb{P}(\mathcal{M}_\gamma)}{\sum_{j=1}^{\Gamma} \mathbb{P}(D|\mathcal{M}_j)\mathbb{P}(\mathcal{M}_j)}$$

The posterior mean and variance of Δ are as follows:

$$\mathbb{E}[\Delta|D] = \sum_{\gamma=0}^{\Gamma} \hat{\Delta}_\gamma \mathbb{P}(\mathcal{M}_\gamma|D)$$

$$\mathbb{V}[\Delta|D] = \sum_{\gamma=0}^{\Gamma} \left(\mathbb{V}[\Delta|D, \mathcal{M}_\gamma + \hat{\Delta}_\gamma^2\right) \mathbb{P}(\mathcal{M}_\gamma|D) - \mathbb{E}[\Delta|D]^2$$

where $\hat{\Delta}_k = \mathbb{E}[\Delta|D, \mathcal{M}_k]$.

There are some approaches that can make model selection easier. This includes the latent variable γ approach, as well as the idea that, if I compare two models $\mathcal{M}_1$ and $\mathcal{M}_2$ so that $\mathcal{M}_1$ is simpler, if I have definitive evidence that $\mathcal{M}_1$ performs better, I reject $\mathcal{M}_2$ and all models that are more complex from the average.

(George & McCulloch, 1993): Stochastic Search Variable Selection Prior

This paper develops the procedure for SSVS to select "promising" subsets of the data for further consideration. SSVS works by embedding the entire regression setup in a hierarchical Bayes normal mixture model using latent variables to identify subset choices, identifying "promising" subsets as those with higher posterior probability, and using Gibbs sampling to indirectly sample from the multinomial posterior distribution on the set of possible subset choices, allowing us to identify subsets with higher probability by their more frequent appearance in the Gibbs sample. This helps avoid the computational demands of calculating the posterior probabilities of all 2^p subsets. This paper differentiates SSVS from the spike-and-slab mixture prior because SSVS does not put a probability mass on $\beta_i = 0$.

This paper suggests that, if you do not know what the final effects of each β_i may be in the final model, you can take a semiautomatic approach to setting a cutoff size for β_i by considering the intersections of the marginal densities $\widehat{\beta}|\sigma_{\beta_i}, \gamma_i = 0 \sim N(0, \sigma^2_{\beta_i} + \tau_i^2)$ and $\widehat{\beta}_i|\sigma_{\beta_i}, \gamma_i = 1 \sim N(0, \sigma^2_{\beta_i} + c_i^2\tau_i^2)$. Then $t_i\sigma_{\beta_i}$ denotes the intersection point, and t_i may be thought of as the threshold at which the t statistic corresponds to an increased marginal probability that X_i should be included in the model.

SSVS uses the Gibbs sampler to generate a sequence $\gamma^{(1)}, \ldots, \gamma^{(m)}$, which in many cases converges rapidly in distribution to $\gamma \sim f(\gamma|Y)$. This will, with high probability in many cases, contain exactly the information relevant to variable selection because the γ with highest probability will appear most frequently. It is important to remember that you cannot consider the marginal frequency $\gamma_i = 1$ as evidence for or against the inclusion of X_i, unless there is minimal to no correlation among the X_i.

(Fahrmeir et al., 2010): Bayesian regularization in structured additive regression (STAR)

Bayesian regularization priors are very comparable to frequentist methods. In general for the approaches outlined here, all priors are conditionally Gaussian, while hyperpriors encourage shrinkage, smoothness, and selection.

Ridge Prior

For ridge regression with smoothing parameter λ, the frequentist approach requires some external mechanisms for selecting λ, often via cross-validation. In the Bayesian formulation, we can avoid this by estimating λ jointly with the regression coefficients. We assign a hyperprior to τ_j^2 (the regularization variance), generally using the inverse gamma prior because it is conjugate to the Gaussian prior for β_j. This yields a hierarchical prior formulation

$$\beta_j|\tau_j^2 \sim N(0,\tau_j^2), \quad \tau_j^2 \sim \text{Gamma}^{-1}(a,b)$$

in which case β_j follows a scaled t-distribution with $2a$ degrees of freedom and scale parameter $\sqrt{\frac{a}{b}}$. The heavier tails of the t-distribution lead to weaker penalization of large coefficients in this setting compared to the frequentist setting, which is desirable.

Lasso Prior

In this paper, the Bayesian analog to the lasso penalty is given by the Laplace distribution, which can be represented as a scale mixture of normals with exponential mixing distribution. This is the approach covered in (Park & Casella, 2008), so I won't touch on it too deeply here.

General L_p Priors

Also from (Park & Casella, 2008), more general L_p penalties ("bridge regression" penalties) with power exponential prior $p(\beta_j|\lambda) \propto \exp\left\{-\lambda|\beta_j|^p\right\}$ and $0 < p < 2$ can be expressed as scale mixtures of normals since

$$\exp\left\{-|\beta_j|^p\right\} \propto \int_0^\infty \exp\left\{-\frac{\beta_j^2}{2\tau_j^2}\right\} \frac{1}{\tau_j^6} s_{p/2}\left(\frac{1}{2\tau_j^2}\right) d\tau_j^2$$

where $s_p(\cdot)$ is the density of the positive stable distribution with index p. In this case, the mixing distribution is complicated and unwieldy, though it seems some progress has been made since 2008 (see (Kowal, 2022)).

Spike-and-Slab Priors

The general idea is that we model each component β_j of β as having come from either a distribution with most (or all) of its mass concentrated around zero ("spike") or a distribution that is diffuse with mass spread out over a large range of values ("slab"). Introducing binary latent indicators $\gamma_j \in \{0,1\}$, this is formally expressed as a mixture

$$\mathbb{P}\left(\beta_j|\gamma_j\right) = (1-\gamma_j)\phi\left(\beta_j; 0, v_{0j}^2\right) + \gamma_j\phi\left(\beta_j; 0, v_{1j}^2\right)$$

of Gaussian densities with variance $v_{0j} << v_{1j}$ and $\mathbb{P}(\gamma_j = 1|w_j) = 1 - \mathbb{P}(\gamma_j = 0|w_j) = w_j$. The idea is if $\gamma_j = 0$ and v_{0j} is close to 0, then β_j is likely to be close to 0. It is often enticing to set $v_{0j} = 0$, which means that $\gamma_j = 0$ implies β_j is not in the predictor. However, this prevents the Markov chain from converging, so special considerations are needed.

Normal Mixture of Inverse Gamma (NMIG)

Instead of placing spike-and-slab priors on the regression coefficients directly, consider putting them on the prior variances τ_j^2. Now the hierarchical prior $\beta_j|\tau_j^2$ is given by

$$\beta_j|\tau_j^2 \overset{\text{ind.}}{\sim} N\left(0, \tau_j^2\right), \quad j = 1, \ldots, q$$

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

with hyperparameters $0 < \nu_0 << \nu_1$. Using the Bernoulli prior, this works out to

$$\tau_j^2|w \sim (1-w)\text{Gamma}^{-1}(a_\tau, \nu_0 b_\tau) + w\text{Gamma}^{-1}(a_\tau, \nu_1 b_\tau)$$

If we integrate out τ_j^2, then the marginal distribution of β_j is a mixture of two scaled t-distributions.

Function Selection

Function selection boils down to two major things: first, whether or not $f_j(x_j)$ should be included in the model at all; second, whether $f_j(x_j)$ is linear or not. For the latter, we can decompose the function into linear and nonlinear terms:

$$f_j(x_j) = \beta_j x_j + \tilde{f}_j(x_j)$$

Importantly, β_j can be thought of as corresponding to all polynomials (in the general case), not just linear effects; then functions $\tilde{f}_j(x_j)$ need to be orthogonal to the polynomial function space.

(Rockova et al., 2012): Hierarchical Bayesian formulations for selecting variables in regression models

This paper mostly focuses on techniques that have already been discussed in this annotated bibliography. The paper mentions the elastic net prior, which is a compromise between LASSO and ridge - it has the variable selection property of the lasso, since it is not differentiable at zero, while also encouraging grouping like the ridge prior, allowing groups of strongly correlated variables to remain in the model with similar estimated coefficients. While this property may not be ideal in this setting, it is an intriguing consideration. The frequentist penalty ℓ_{net} is formulated $\ell_{\text{net}}(\beta) = a_1 \sum_{k=1}^q |\beta_k| + a_2 \sum_{k=1}^q \beta_k^2$, which corresponds to prior hierarchy

$$\beta_k|\tau_k \sim N\left(0, \left[\frac{a_2}{\sigma^2}\frac{\tau_k}{\tau_k - 1}\right]^{-1}\right)$$

$$\tau_k \sim \text{Gamma}\left[\frac{1}{2}, \frac{8a_2\sigma^2}{a_1^2}, (1, \infty)\right]$$

where Gamma$[a, b, (c, d)]$ refers to the truncated gamma distribution with shape a, scale b, and a support restricted to the interval (c, d).

This paper goes on to give some examples of various approaches to hierarchical Bayesian regression.

(Heuclin et al., 2024): Bayesian group fused priors

Most Bayesian approaches to variable selection do not take into account a potential group structure within covariates, though some have begun to be developed in recent years (see page 4). This paper proposes the horseshoe normal half-Cauchy (HS-NhC) fused prior, and compares it to several other proposed approaches.

References

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

Heuclin, B., Mortier, F., Tisne, S., Gibaud, J., Trottier, C., & Denis, M. (2024). *Bayesian group fused priors*. https://hal.science/hal-04486172v1

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., Lesaffre, E., Luime, J., & Löwenberg, B. (2012). Hierarchical Bayesian formulations for selecting variables in regression models. *Statistics in Medicine*, *31*, 1221–1237. https://doi.org/10.1002/sim.4439