ECON5120 Bayesian Data Analysis

Week 9: Model Uncertainty and Variable Selection

Santiago Montoya-Blandón

University of Glasgow
Adam Smith Business School

March 11, 2026

Lecture outline

1. Motivation
2. Model Comparison
   - Pitfalls of using Bayes factors
   - Specification Uncertainty
3. Stochastic Search Variable Selection (SSVS) or Spike-and-Slab Priors
4. Bayesian Penalized Regression
   - Bayesian Shrinkage
   - Bayesian LASSO and Variants
5. Bayesian Model Averaging (BMA)
   - Model Averaging in Large Dimensions

Model Uncertainty

- We are about to take Bayesian to a meta level
- So far we have talked about *uncertainty* in a parameter θ given an *assumed known model* M when having access to data y
- In more generality, we usually assume M is one of a class of models $\mathcal{M}$

Questions for you

1. What if the model $M \in \mathcal{M}$ is uncertain?

Model Uncertainty

- We are about to take Bayesian to a meta level
- So far we have talked about *uncertainty* in a parameter θ given an *assumed known model* M when having access to data y
- In more generality, we usually assume M is one of a class of models $\mathcal{M}$

Questions for you

1. What if the model $M \in \mathcal{M}$ is uncertain?
2. Can we say anything about model M given data y?

Model Uncertainty

- Remember we are Bayesian: we hope to summarize uncertainty through a probability distribution
- Nothing stops us from using Bayesian analysis for this problem
- So now we just need a probability distribution over $\mathcal{M}$!

Model Uncertainty

- Remember we are Bayesian: we hope to summarize uncertainty through a probability distribution
- Nothing stops us from using Bayesian analysis for this problem
- So now we just need a probability distribution over $\mathcal{M}$!
- Specifically, we will need a *likelihood* and *prior* to use Bayes rule
- Prior: $p(M)$ over $M \in \mathcal{M}$
- Likelihood: $p(y \mid M)$ as a function of $M \in \mathcal{M}$
- Using Bayes rule (standard probability results) we obtain for any $M \in \mathcal{M}$:

$$p(M \mid y) = \frac{p(y \mid M)p(M)}{p(y)}$$

Lecture outline

1. Motivation
2. Model Comparison
   - Pitfalls of using Bayes factors
   - Specification Uncertainty
3. Stochastic Search Variable Selection (SSVS) or Spike-and-Slab Priors
4. Bayesian Penalized Regression
   - Bayesian Shrinkage
   - Bayesian LASSO and Variants
5. Bayesian Model Averaging (BMA)
   - Model Averaging in Large Dimensions

Model Uncertainty for Bayesians

- We have just used Bayes rule to understand model uncertainty!

$$p(M \mid y) = \frac{p(y \mid M)p(M)}{p(y)} \propto p(y \mid M)p(M)$$

- For any model $M \in \mathcal{M}$:
  - $p(M)$ is the *prior model probability*
  - $p(y \mid M)$ is the *marginal likelihood*
  - $p(M \mid y)$ is the *posterior model probability*
- As usual, we end up using the posterior to summarize our knowledge
- These are general and powerful results from simple probability rules

Marginal Likelihood and Model Evidence

- $p(y \mid M)$ is also called the *evidence* for model M in data y
- We have avoided computing quantities like $p(y \mid M)$ throughout the course
- This is a difficult high-dimensional integral over a (possibly model-specific) parameter space Θ_M
- Let model $M \in \mathcal{M}$ be characterized by model-specific parameters θ^M
- Prior $p(\theta^M \mid M)$ and likelihood $p(y \mid \theta^M, M)$

$$p(y \mid M) = \int_{\Theta_M} p(y \mid \theta^M, M) p(\theta^M \mid M) d\theta^M$$

- This also means solving this problem for different models in $\mathcal{M}$

Model comparison: Bayes factors

- To compare models, use *posterior odd ratios* between models $M_1, M_2 \in \mathcal{M}$

$$PO_{1,2} := \frac{p(M_1 \mid y)}{p(M_2 \mid y)} = \frac{p(y \mid M_1)}{p(y \mid M_2)} \cdot \frac{p(M_1)}{p(M_2)}$$

- Can be written as a product of the *Bayes factor* and the *prior odds ratio* between models M_1 and M_2

$$BF_{1,2} := \frac{p(y \mid M_1)}{p(y \mid M_2)} \quad \text{and} \quad \text{Prior Odds}_{1,2} := \frac{p(M_1)}{p(M_2)}$$

- Bayes factor is evidence in favour of model M_1 over M_2
- $BF_{1,2}$ is large when model M_1 is more likely to fit the data than model M_2
- Different authors define different scales for comparing models according to this evidence ratio

Bayes factors: Hypothesis testing

- Bayes factor can also be used to test hypotheses about models
- Let H_0 and H_1 represent a null and alternative hypothesis, respectively
- Let M_1 be the model that imposes H_0 and M_2 be the model that imposes H_1
- $BF_{1,2}$ compares the ratio of how much the data supports M_1 (the null H_0) vs M_2 (the alternative H_1)
- Using evidence scales, one can effectively perform a test of H_0 against H_1

Bayes factors: Hypothesis testing

- Interesting special case: $M_1 =$ Restricted model, $M_2 =$ Unrestricted model
- Let $\theta = (\theta_1, \theta_2)$. M_1 assumes $\theta_2 = \theta^*$, but model M_2 lets $\theta_2 \neq \theta^*$
- **Savage-Dickey ratio:** If the prior is such that $p(\theta_1 \mid \theta_2 = \theta^*, M_2) = p(\theta_1 \mid M_1)$, then

$$BF_{1,2} = \frac{p(\theta_2 = \theta^* \mid y, M_2)}{p(\theta_2 = \theta^* \mid M_2)}$$

Bayes factors: Hypothesis testing

- Interesting special case: $M_1 =$ Restricted model, $M_2 =$ Unrestricted model
- Let $\theta = (\theta_1, \theta_2)$. M_1 assumes $\theta_2 = \theta^*$, but model M_2 lets $\theta_2 \neq \theta^*$
- **Savage-Dickey ratio:** If the prior is such that $p(\theta_1 \mid \theta_2 = \theta^*, M_2) = p(\theta_1 \mid M_1)$, then

$$BF_{1,2} = \frac{p(\theta_2 = \theta^* \mid y, M_2)}{p(\theta_2 = \theta^* \mid M_2)}$$

- **Generalized Savage-Dickey ratio:** If no restriction is imposed in the prior $p(\theta_1 \mid M_1)$, then

$$BF_{1,2} = \frac{p(\theta_2 = \theta^* \mid y, M_2)}{p(\theta_2 = \theta^* \mid M_2)} \cdot \mathbb{E}_{p(\theta_1 \mid \theta_2 = \theta^*, y, M_2)} \left[\frac{p(\theta_1 \mid M_1)}{p(\theta_1 \mid \theta_2 = \theta^*, M_2)} \right]$$

Bayes factor computation

- In general, Bayes factors have no closed-form solution
- Note the Bayes factor is independent of $p(y)$, the marginal likelihood of the data across all models
- For a finite model class $\mathcal{M}$ ($|\mathcal{M}| := R < \infty$, where $|\cdot|$ is set cardinality)

$$p(y) = \sum_{r=1}^{R} p(y \mid M_r) p(M_r)$$

Bayes factor computation

- In general, Bayes factors have no closed-form solution
- Note the Bayes factor is independent of $p(y)$, the marginal likelihood of the data across all models
- For a finite model class $\mathcal{M}$ ($|\mathcal{M}| := R < \infty$, where $|\cdot|$ is set cardinality)

$$p(y) = \sum_{r=1}^{R} p(y \mid M_r) p(M_r)$$

- We can no longer avoid computing evidences for models M_1 and M_2
- This means at least computing $p(y \mid M_1)$ and $p(y \mid M_2)$

Bayes factor computation

A few methods to compute $p(y \mid M)$ exist in the literature:

- Analytical solutions (conjugate priors); e.g.,
  - Binomial likelihood with Beta prior
  - Normal likelihood with Normal-Inverse-Gamma priors
- Gelfand and Dey (1994)
- Chib (1995) from Gibbs sampling output
- Chib and Jeliazkov (2001) from Metropolis-Hastings sampling output
- Bridge sampling (Bennett, 1976; Meng & Wong, 1996)

Misleading Bayes factors

Bayes factors can be misleading in some circumstances:

- Using improper prior distributions
- Some non-informative but proper priors
- Continuous model classes

In light of some of the pitfalls associated to using Bayes factors, there are several alternatives in the literature

1. Stochastic search variable selection (also known as spike-and-slab priors)
2. LASSO, elastic nets and other variants of penalized regression
3. Bayesian model averaging (BMA)

Sources of Model Uncertainty

- So far, we have been vague about the nature of model uncertainty
- Imagine data is given as outcome y and covariates X
- Example model: $y = f(\theta, X) + u$
- Many sources of uncertainty:
  - Error distribution $p(u \mid X)$
  - Features to include in X
  - Functional form f
  - Priors on θ

Model Uncertainty: Linear Regression

- To ground discussion, we will focus on a particular source of model uncertainty: *variable selection*
- Focus on a normal linear regression model with homoskedastic errors:

$$y_i = x_i^\top \beta + u_i \text{ for } i = 1, \ldots, n \implies y = X\beta + u \,, \text{where } u \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$$

Model Uncertainty: Linear Regression

- To ground discussion, we will focus on a particular source of model uncertainty: *variable selection*
- Focus on a normal linear regression model with homoskedastic errors:

$$y_i = x_i^\top \beta + u_i \text{ for } i = 1, \ldots, n \implies y = X\beta + u \,, \text{where } u \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$$

- In practice and with large datasets, you have access to many variables X and need to select which of them to include in your regression
- Examples:
  - Predict housing prices based on house and market features
  - Explain economic growth based on country characteristics
- Assuming everything else about the model was correct, we simply need to worry about selecting variables in X

Variable Selection

- Let X represent p possible features with a sample size of n
- Model space is finite: $\mathcal{M} = \{M_1, \ldots, M_R\}$ with $R = 2^p$
- Intuitively, you can either include or exclude any of the p variables
- If p is small, can simply run Bayesian linear regression on all possible combinations of features in X

Variable Selection

- Let X represent p possible features with a sample size of n
- Model space is finite: $\mathcal{M} = \{M_1, \ldots, M_R\}$ with $R = 2^p$
- Intuitively, you can either include or exclude any of the p variables
- If p is small, can simply run Bayesian linear regression on all possible combinations of features in X
- With moderate p, say $p > 40$, this is computationally demanding
- What to do with large p? Or if $p > n$?
- Both situations are likely with big data

Variable Selection: Useful Equivalence

- Define $\gamma := (\gamma_1, \ldots, \gamma_p)$, where each element is such that

$$\gamma_j := \begin{cases} 1 & \text{if feature } X_j \text{ is included in the regression} \\ 0 & \text{otherwise} \end{cases}$$

- As there are p features, the dimensionality of β is also p
- In a linear regression, if a variable X_j has coefficient $\beta_j = 0$, then X_j is irrelevant to predict y
- This means we have the following useful equivalent representation for all $j = 1, \ldots, p$

$$\gamma_j = \begin{cases} 1 & \text{if } \beta_j \neq 0 \\ 0 & \text{if } \beta_j = 0 \end{cases}$$

Variable Selection: Useful Equivalence

- γ has support equal to $\{0,1\} \times \cdots \times \{0,1\} = \{0,1\}^p$
- That is, the support can take on 2^p values, just like our original model space!
- Consequence: Any model M that includes a given set of covariates in X can be equivalently represented with a value of γ

Variable Selection: Useful Equivalence

- γ has support equal to $\{0,1\} \times \cdots \times \{0,1\} = \{0,1\}^p$
- That is, the support can take on 2^p values, just like our original model space!
- Consequence: Any model M that includes a given set of covariates in X can be equivalently represented with a value of γ
- This might seem surprising, but it is easy to show
- Example: $p = 5$, such that $X = (X_1, X_2, X_3, X_4, X_5)$
- Model M includes X_1, X_3 and X_4 into the regression
- Simply set $\gamma = (1, 0, 1, 1, 0)$
- The same can be done for any $M \in \mathcal{M}$

Lecture outline

1. Motivation
2. Model Comparison
   - Pitfalls of using Bayes factors
   - Specification Uncertainty
3. Stochastic Search Variable Selection (SSVS) or Spike-and-Slab Priors
4. Bayesian Penalized Regression
   - Bayesian Shrinkage
   - Bayesian LASSO and Variants
5. Bayesian Model Averaging (BMA)
   - Model Averaging in Large Dimensions

Spike-and-slab Priors

- We can use our previous equivalence to provide a simple prior for variable selection
- As both the variables to include and their regression coefficients are uncertain, this means both γ and β are unknown
- As Bayesians, we assign a joint prior distribution $p(\beta, \gamma) = p(\beta \mid \gamma)p(\gamma)$
- One prior for the variable inclusion indicators is to have simple independent Bernoulli distributions with prior inclusion probability $\underline{p}_j$:

$$p(\gamma) = \prod_{j=1}^{p} p(\gamma_j) = \prod_{j=1}^{p} \gamma_j^{\underline{p}_j} \cdot (1 - \gamma_j)^{1-\underline{p}_j}$$

- A priori, we include every covariate independently with probability $\underline{p}_j$

Spike-and-slab Priors

- Conditional on the value of γ, the set of included covariates is fixed
- Recall: any value of γ corresponds to some model $M \in \mathcal{M}$
- If $\gamma_j = 0$, then $\beta_j = 0$ for any variable $j = 1, \ldots, p$
- If $\gamma_j = 1$, then β_j is completely unrestricted
- This means our prior will have two components:
  1. Spike: point mass at 0 associated to $\gamma_j = 0$
  2. Slab: unrestricted smooth density centered at 0 when $\gamma_j = 1$
- The point mass is discontinuous and can be difficult to deal with, so we can additionally smooth out the spike component

Spike-and-slab Priors

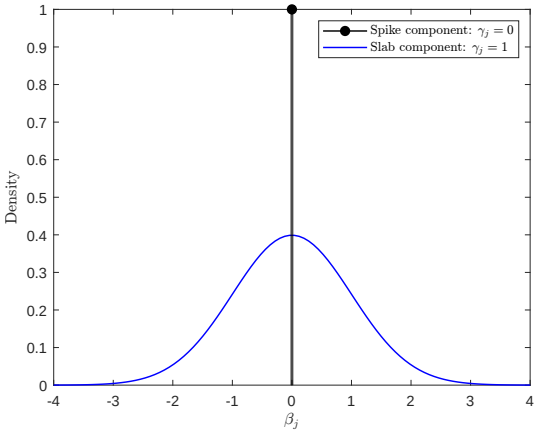

Figure: Ideal Spike-and-Slab Prior

Spike-and-slab Priors

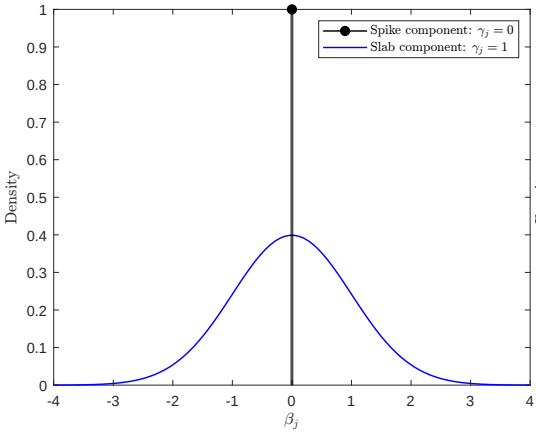

Figure: Ideal Spike-and-Slab Prior

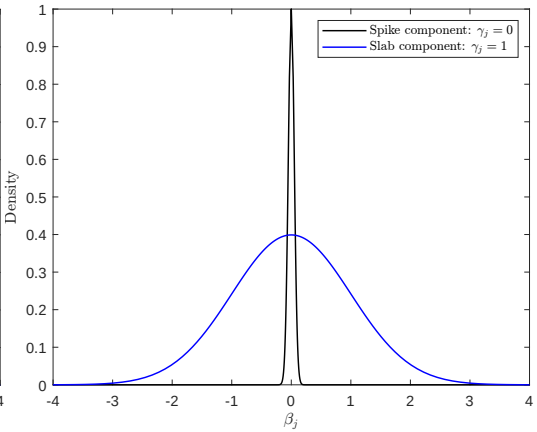

Figure: Smoothed Spike-and-Slab Prior

Spike-and-slab Priors

- We can represent this prior formally using a *mixture distribution*
- Spike: $\mathcal{N}(0, \tau^{(1)})$, Slab: $\mathcal{N}(0, \tau^{(2)})$
- If $\tau^{(1)} \ll \tau^{(2)}$ with $\tau^{(1)}$ close to 0, spike resembles point mass at 0
- Independent priors for β_j conditional on γ_j for each $j = 1, \ldots, p$

$$\beta_0 \sim \mathcal{N}(0, \underline{V}_0)$$
$$\beta_j \mid \gamma_j \sim (1 - \gamma_j) \cdot \mathcal{N}(0, \tau_j^{(1)}) + \gamma_j \cdot \mathcal{N}(0, \tau_j^{(2)})$$

- Intuition:
  - Constant is assigned a standard prior irrespective of γ
  - If β_j comes from the spike component ($\gamma_j = 0$), it has mean 0 and variance $\tau_j^{(1)} \approx 0$. It is essentially as if removing it from the model (setting $\beta_j \approx 0$)
  - If β_j comes from the slab component ($\gamma_j = 1$), even if mean is 0, $\tau_j^{(2)}$ is large enough so β_j is unrestricted

Stochastic Search Variable Selection (SSVS)

- The only remaining uncertain element is σ^2, use standard prior

$$\sigma^2 \sim \text{Inv-Gamma}(\underline{\alpha}/2, \underline{\delta}/2)$$

- Conditional posteriors available in closed form:

$$\begin{aligned}
\beta \mid \gamma, \sigma^2, y, X &\sim \mathcal{N}_{p+1}(\bar{\beta}, \bar{B}) \\
\sigma^2 \mid \gamma, \beta, y, X &\sim \text{Inv-Gamma}(\bar{\alpha}/2, \bar{\delta}/2) \\
\gamma_j \mid \beta_j, \sigma^2, y, X &\sim \text{Bernoulli}(\bar{p}_j),\ j = 1, \ldots, p
\end{aligned}$$

- With these conditional posteriors, one can set up a Gibbs sampler
- Drawing γ from its posterior means drawing a new candidate model
- Intuition: just add the variable selection indicators as another set of parameters and compute their posterior!

Stochastic Search Variable Selection (SSVS)

- Posterior hyperparameters are given as

$$\bar{B} = \left[\underline{V}(\gamma)^{-1} + \frac{X^\top X}{\sigma^2}\right]^{-1}, \quad \bar{\beta} = \bar{B}\frac{X^\top y}{\sigma^2}$$

$$\bar{\alpha} = \underline{\alpha} + n, \quad \bar{\delta} = \underline{\delta} + (y - X\beta)^\top (y - X\beta)$$

$$\bar{p}_j = \frac{\underline{p}_j \cdot \mathcal{N}(\beta_j \mid 0, \tau_j^{(2)})}{(1 - \underline{p}_j) \cdot \mathcal{N}\left(\beta_j \mid 0, \tau_j^{(1)}\right) + \underline{p}_j \cdot \mathcal{N}\left(\beta_j \mid 0, \tau_j^{(2)}\right)}$$

$$\underline{V}_{1,1}(\gamma) = \underline{V}_0, \quad \underline{V}_{j,\ell}(\gamma) = 0 \text{ for } j \neq \ell$$

$$\underline{V}_{j+1,j+1}(\gamma) = (1 - \gamma_j) \cdot \tau_j^{(1)} + \gamma_j \cdot \tau_j^{(2)}; \; j = 1, \ldots, p$$

- Posterior again is a combination of prior and data quantities

Back to Model Uncertainty

- Recall we introduce SSVS to handle model uncertainty through variable selection
- In addition to summarizing the posteriors for β and σ^2, we can also provide summaries for the posterior of γ
- Each draw of γ corresponds to a model $M \in \mathcal{M}$ supported by the data
- This means we can summarize posterior probabilities over models!
  - Proportion of times $\gamma_j = 1$ across draws is the same as the proportion of times X_j is selected as a relevant variable in the regression
  - Can study most likely model size (number of ones in γ)
  - Can study selection accuracy
- Tuning is essential for good performance: $\tau_j^{(1)}$, $\tau_j^{(2)}$ and $\underline{p}_j$, $j = 1, \ldots, p$

Lecture outline

1. Motivation
2. Model Comparison
   - Pitfalls of using Bayes factors
   - Specification Uncertainty
3. Stochastic Search Variable Selection (SSVS) or Spike-and-Slab Priors
4. Bayesian Penalized Regression
   - Bayesian Shrinkage
   - Bayesian LASSO and Variants
5. Bayesian Model Averaging (BMA)
   - Model Averaging in Large Dimensions

Penalized Regression

- Before introducing the specific form of the LASSO, let us introduce the idea of *penalization*
- This is not a new concept: Bayesian shrinkage is a form of penalization
- General framework in a linear regression context:

$$\min_{\beta \in \mathbb{R}^p} (y - X\beta)^\top (y - X\beta) + \lambda \cdot \rho(\beta)$$

- λ is a *tuning parameter* controlling the strength of penalization
- $\rho(\beta)$ is a penalty function with respect to slope coefficients β
- As $\lambda \to \infty$, then $\beta \to 0$
- As $\lambda \to 0$, then $\beta \to \hat{\beta}_{\text{OLS}} = (X^\top X)^{-1} X^\top y$
- Usually assume data is *standardized*; i.e., mean 0 and variance 1

Penalty Functions

- Different penalties amount to different ways of regularizing β
- We will examine three important cases that can be closely connected to the Bayesian framework

$$\rho_{\text{LASSO}}(\beta) = \sum_{j=1}^{p} |\beta_j| =: \|\beta\|_1$$

$$\rho_{\text{Ridge}}(\beta) = \sum_{j=1}^{p} \beta_j^2 =: \|\beta\|_2^2$$

$$\rho_{\text{Elastic Net},\alpha}(\beta) = \alpha \cdot \rho_{\text{LASSO}}(\beta) + (1-\alpha) \cdot \rho_{\text{Ridge}}(\beta)$$

Penalty Functions

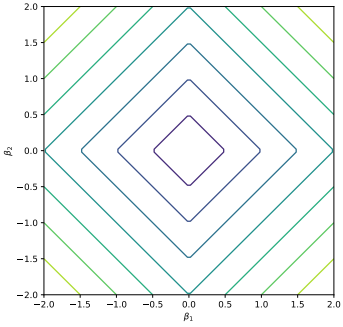

Figure: LASSO Penalty

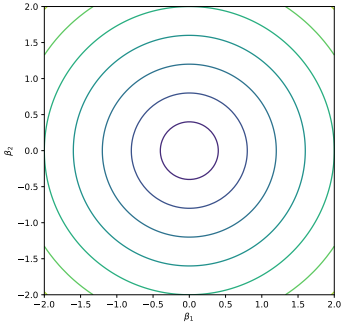

Figure: Ridge Penalty

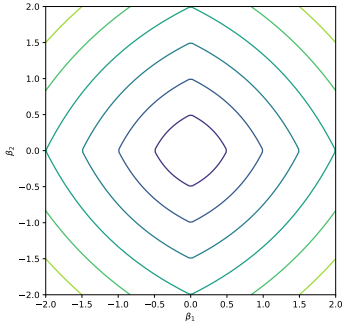

Figure: Elastic Net Penalty

Sparse Solutions

- LASSO stands for Least Absolute *Shrinkage* and *Selection* Operator
- Shrinkage is induced by all penalized regression methods, including Bayesian
- The discontinuities at the corners of the LASSO and elastic net penalties mean both methods can also produce *sparse* solutions
- This means some coefficients β_j are set *exactly to* 0 even for finite λ
- The same is not true for ridge regression: the solutions can be set close to 0, but will never be exactly 0 for any finite λ (only true at the limit)

Sparse Solutions

- Sparsity is what allows us to deal with cases with many covariates p or $p > n$
- Assume only a subset $p_0 < p$ of the β coefficients are non-zero, with $p_0 < n$
- Techniques such as LASSO aim to set as many coefficients to 0 as possible while keeping only relevant predictors
- Advances in computation allow us to fit penalized methods over a grid of λ values in seconds, even for large datasets

Sparse Solutions

- Sparsity is what allows us to deal with cases with many covariates p or $p > n$
- Assume only a subset $p_0 < p$ of the β coefficients are non-zero, with $p_0 < n$
- Techniques such as LASSO aim to set as many coefficients to 0 as possible while keeping only relevant predictors
- Advances in computation allow us to fit penalized methods over a grid of λ values in seconds, even for large datasets
- Bayesian techniques usually require additional post-processing to impose sparsity numerically
- Bayesian solutions capture uncertainty naturally, even for the excluded (zero) coefficients
- Capturing this uncertainty is difficult in the frequentist perspective

Shrinkage and Selection

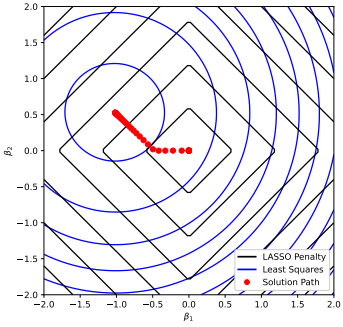

Figure: LASSO Solution

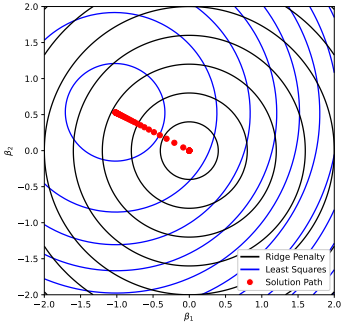

Figure: Ridge Solution

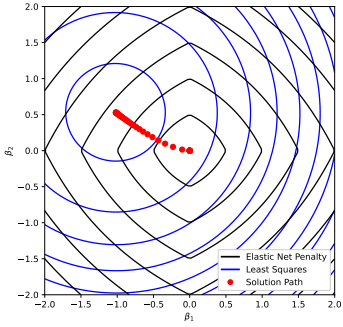

Figure: Elastic Net Solution

Shrinkage and Selection

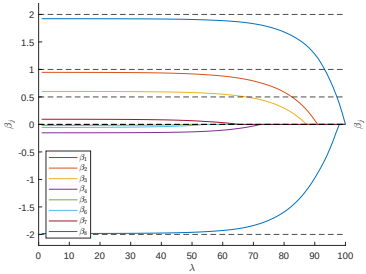

Figure: LASSO Path

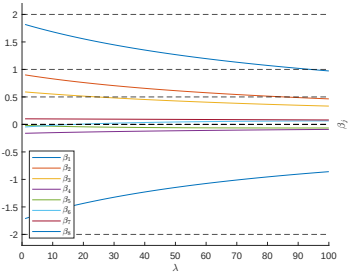

Figure: Ridge Path

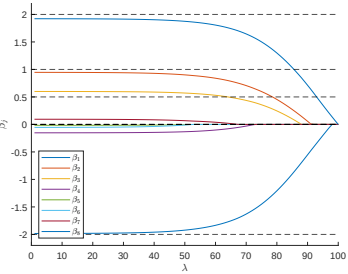

Figure: Elastic Net Path

Bayesian Shrinkage

- I have mentioned several times over the course that Bayesian is also a method of penalization or shrinkage
- *Bayesian induces shrinkage of solutions towards the prior*

Bayesian Shrinkage

- I have mentioned several times over the course that Bayesian is also a method of penalization or shrinkage
- *Bayesian induces shrinkage of solutions towards the prior*
- To make the connection clearer, let us show a Bayesian interpretation for both LASSO and Ridge penalties
- Start with Ridge regression, as we need to introduce no new notation
- Recall the objective of Ridge regression:

$$\hat{\beta}_{\text{Ridge}} := \arg\min_{\beta \in \mathbb{R}^p} (y - X\beta)^\top (y - X\beta) + \lambda \cdot \sum_{j=1}^{p} \beta_j^2$$

Bayesian Interpretation of Ridge

- We will show the Ridge solution corresponds to the posterior mode of a normal likelihood with a normal conjugate prior centered at 0

$$y \mid X, \beta, \sigma^2, \lambda \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$$
$$\beta \mid \sigma^2, \lambda \sim \mathcal{N}_p\left(0_p, \frac{\sigma^2}{\lambda} I_p\right)$$

- Prior centered at 0: shrinkage of coefficients towards 0
- Notice that λ controls the precision of the prior
- Prior for σ^2 unrestricted for now

Bayesian Interpretation of Ridge

- From the likelihood and prior, we have

$$p(y \mid X; \beta, \sigma^2, \lambda) \propto \left(\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)^\top (y - X\beta)\right]$$

$$p(\beta \mid \sigma^2, \lambda) \propto \exp\left[-\frac{\lambda}{2\sigma^2} \cdot \sum_{j=1}^{p} \beta_j^2\right] \qquad \left(\text{using } \beta^\top I_p \beta = \sum_{j=1}^{p} \beta_j^2\right)$$

- Using Bayes' rule: Posterior proportional to likelihood times prior

$$p(\beta \mid y, X; \sigma^2, \lambda) \propto \exp\left\{-\frac{1}{2\sigma^2}\left[(y - X\beta)^\top (y - X\beta) + \lambda \cdot \sum_{j=1}^{p} \beta_j^2\right]\right\}$$

Bayesian Interpretation of Ridge

- Recall for d-dimensional X with a continuous distribution $f(\cdot)$, the mode is defined as

$$\text{mode}(X) \coloneqq \arg\max_{x \in \mathbb{R}^d} f(x)$$

- Generalizes intuitive notion of mode as "most likely value"
- Using this definition of mode, we find the posterior mode for β

$$\text{mode}(\beta \mid y, X; \sigma^2, \lambda) \coloneqq \arg\max_{\beta \in \mathbb{R}^p} p(\beta \mid y, X; \sigma^2, \lambda) \qquad (\max -f = \min f)$$

$$= \arg\min_{\beta \in \mathbb{R}^p} (y - X\beta)^\top (y - X\beta) + \lambda \cdot \sum_{j=1}^{p} \beta_j^2$$

Bayesian Interpretation of Ridge

- Recall for d-dimensional X with a continuous distribution $f(\cdot)$, the mode is defined as

$$\text{mode}(X) \coloneqq \arg\max_{x \in \mathbb{R}^d} f(x)$$

- Generalizes intuitive notion of mode as "most likely value"
- Using this definition of mode, we find the posterior mode for β

$$\text{mode}(\beta \mid y, X; \sigma^2, \lambda) \coloneqq \arg\max_{\beta \in \mathbb{R}^p} p(\beta \mid y, X; \sigma^2, \lambda) \qquad (\max -f = \min f)$$

$$= \arg\min_{\beta \in \mathbb{R}^p} (y - X\beta)^\top (y - X\beta) + \lambda \cdot \sum_{j=1}^{p} \beta_j^2$$

- This is the same as the Ridge solution! $\text{mode}(\beta \mid y, X; \sigma^2, \lambda) = \hat{\beta}_{\text{Ridge}}$

Bayesian Interpretation of Ridge: Final Details

- Recall setting is normal linear regression with normal prior centered at 0
- Previous proof shows Ridge regression can be interpreted as the mode of the *conditional* (on σ^2) posterior of β
- We previously showed mode does not change after integrating out σ^2, so this holds for the *marginal* posterior of β, $p(\beta \mid y, X; \lambda)$, as well

Important Result

In a normal model for the data, the amount of shrinkage implied by Bayesian with a normal prior is the same as that implied by Ridge regression using penalty $\rho_{\text{Ridge}}(\beta)$.

Bayesian Interpretation of Ridge: Final Details

- How to select tuning parameter λ to improve predictive performance?
- Advances in computation mean one can fit penalized regression methods across a whole grid of λ values in seconds, even for large datasets
- Can use cross-validation (CV) or other techniques to select the $\lambda \in \{\lambda_1, \ldots, \lambda_L\}$ that maximizes a given performance measures
- Can also maximize the marginal likelihood as this is available in closed-form for the normal model with normal prior

Bayesian Interpretation of Ridge: Final Details

- How to select tuning parameter λ to improve predictive performance?
- Advances in computation mean one can fit penalized regression methods across a whole grid of λ values in seconds, even for large datasets
- Can use cross-validation (CV) or other techniques to select the $\lambda \in \{\lambda_1, \ldots, \lambda_L\}$ that maximizes a given performance measures
- Can also maximize the marginal likelihood as this is available in closed-form for the normal model with normal prior
- In Bayesian, can use optimal value from marginal likelihood or CV
- Can also go full Bayesian and provide a prior for it: $p(\lambda)$
- We will see examples of this second approach in the specific case of LASSO

Bayesian Interpretation of LASSO and Variants

- Note the Ridge shrinkage structure was imposed purely through the prior
- Likelihood always remains a normal with homoskedastic error
- This means we can impose different shrinkage properties by using different prior structures
- **Bayesian interpretation of LASSO and company:** posterior mode of normal model with priors that match the penalty structure

Bayesian Interpretation of LASSO and Variants

- Note the Ridge shrinkage structure was imposed purely through the prior
- Likelihood always remains a normal with homoskedastic error
- This means we can impose different shrinkage properties by using different prior structures
- **Bayesian interpretation of LASSO and company:** posterior mode of normal model with priors that match the penalty structure
- Different penalized estimators correspond to different *shrinkage priors*
- Technicality: Require conditioning on σ^2 for unimodality of posterior

Bayesian Interpretation of LASSO

- Let us now focus on the *Bayesian LASSO*
- Given λ and σ^2, let us use a Laplace or double-exponential prior distribution

$$p(\beta \mid \sigma^2, \lambda) \propto \exp\left[-\frac{\lambda}{2\sigma^2}\sum_{j=1}^{p}|\beta_j|\right]$$

- As before, λ controls the precision of the prior
- Prior imposes sharper selection properties due to shape of density
- Same steps as before produce the conditional posterior for β

$$p(\beta \mid y, X; \sigma^2, \lambda) \propto \exp\left\{-\frac{1}{2\sigma^2}\left[(y - X\beta)^\top(y - X\beta) + \lambda \cdot \sum_{j=1}^{p}|\beta_j|\right]\right\}$$

Shrinkage Priors

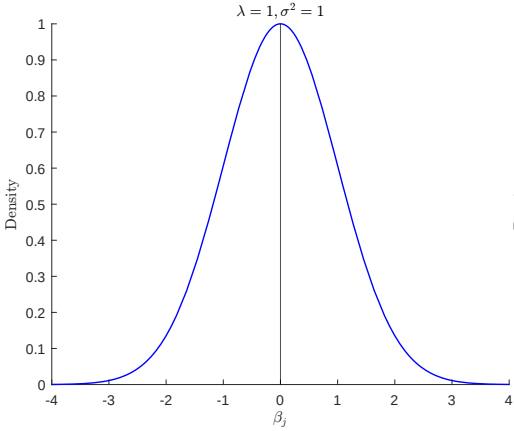

Figure: Normal prior corresponding to Ridge regression

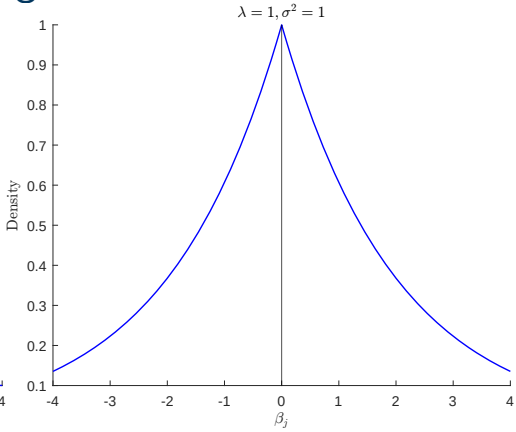

Figure: Laplace (double-exponential) prior corresponding to LASSO

Bayesian Interpretation of LASSO

- LASSO is the same solution as the posterior mode using the Laplace prior!
- Tibshirani (1996) already presented this result in the original LASSO paper
- This sparked a body of literature that gave computationally feasible Bayesian solutions to many penalized regression problems
- Includes variations of LASSO with additional structure and complex priors
  - Group LASSO for grouped covariates (e.g., polynomials, interactions)
  - Fused LASSO for dependence structures in the observations (e.g., time series, spatial)
  - Adaptive LASSO: individual λ_j parameters

Bayesian LASSO: Implementation

- Bayesian LASSO implementation from Park and Casella (2008)
- Their idea: Replace Laplace prior with mixture of normals with exponential mixing density
- Practically, this means using the following hierarchical priors:

$$\begin{aligned}
\beta_j \mid \sigma^2, \tau_j^2 &\overset{iid}{\sim} \mathcal{N}(0, \sigma^2 \cdot \tau_j), \text{ over } j = 1, \ldots, p \\
\tau_j \mid \lambda &\overset{iid}{\sim} \text{Exp}(\lambda^2/2) \quad \text{with } \tau_j^2 > 0, \text{ over } j = 1, \ldots, p \\
p(\sigma^2) &\propto \frac{1}{\sigma^2}
\end{aligned}$$

- Can additionally include a Gamma$(\underline{r}, \underline{\eta})$ prior on λ^2 for full Bayesian inference

Bayesian LASSO: Implementation

- Conditional posteriors available for Gibbs sampling

$$\beta \mid y, X; \sigma^2, \tau_1^2, \ldots, \tau_p^2, \lambda \sim \mathcal{N}(\bar{\beta}, \sigma^2 \bar{B})$$

$$\frac{1}{\tau_j^2} \mid y, X; \beta, \sigma^2, \lambda \overset{iid}{\sim} \text{Inv-Gaussian}\left(\sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}}, \lambda^2\right), \text{ with } \tau_j^2 > 0$$

$$\text{independently across } j = 1, \ldots, p,$$

$$\sigma^2 \mid y, X; \beta, \tau_1^2, \ldots, \tau_p^2, \lambda \sim \text{Inv-Gamma}(\bar{\alpha}/2, \bar{\delta}/2)$$

- Conditional posterior for λ^2 if included is Gamma$(\bar{r}, \bar{\eta})$
- If not included, selection can be done using an optimal value of λ

Bayesian LASSO: Implementation

- Posterior hyperparameters for the conditionals in previous slide

$$\bar{B} = \left(D^{-1} + X^\top X\right)^{-1}, \quad \bar{\beta} = \bar{B}X^\top y, \quad D = \text{diag}\{\tau_1^2, \ldots, \tau_p^2\}$$
$$\bar{\alpha} = n - 1 + p, \quad \bar{\delta} = (y - X\beta)^\top (y - X\beta) + \beta^\top D^{-1}\beta$$
$$\bar{r} = \underline{r} + p, \quad \text{and} \quad \bar{\eta} = \underline{\eta} + \frac{1}{2}\sum_{j=1}^{p} \tau_j^2$$

- Recall we require to process the draws we get from Gibbs sampler to impose sparsity (exactly zero coefficients) numerically
- Model selection as by-product: focus on non-zero coefficients
- As Bayesians, we have full uncertainty quantification over coefficients, even those set to 0 numerically

Advanced Shrinkage Priors

- We have discussed the LASSO priors at length
- Other shrinkage priors exist with different, sometimes superior properties
- These priors can be extended to more complex data structures, just like LASSO variants
- Other priors are more clever in the way they shrink coefficients
  - t distribution
  - Spike-and-slab
  - Horseshoe prior: $\tau_j \overset{iid}{\sim} \text{Cauchy}^+(0,1); j = 1, \ldots, p$

Advanced Shrinkage Priors

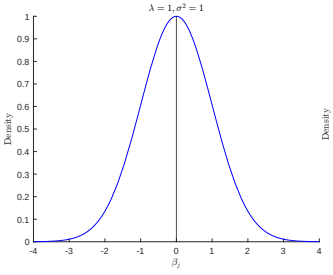

Figure: Normal prior

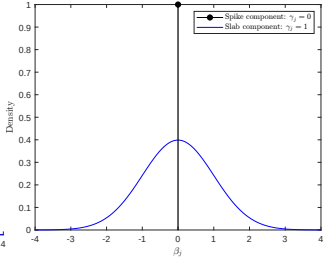

Figure: Spike-and-Slab

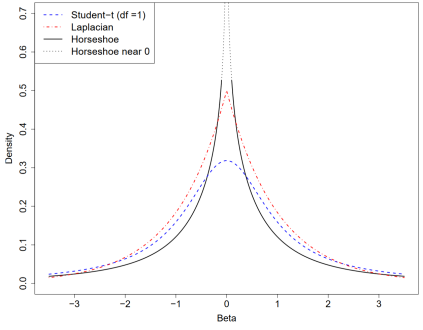

Figure: Horseshoe prior (Carvalho et al., 2009)

Lecture outline

Model Uncertainty

- Model selection is a way to acknowledge uncertainty about correct model
- Aim to select an optimal model amongst a pool of candidate models

Model Uncertainty

- Model selection is a way to acknowledge uncertainty about correct model
- Aim to select an optimal model amongst a pool of candidate models
- What if instead of selecting an optimal model, we combine all models?
- We could aggregate by taking averages across all candidate models
  - Many models likely to perform poorly
  - Only few models likely to have good performance
  - Weigh models according to their probability of being correct!
- Wisdom of the masses: large uninformed crowds perform better than single experts in many tasks
- In our context: aggregating across models usually performs better than a single optimal model, even if each model performs poorly on average

Bayesian Model Averaging

- Bayesian allows us to average across models using simple rules of probability
- Finite set of R candidate models: $\mathcal{M} = \{M_1, \ldots, M_R\}$
- Let ϕ be a parameter that has a common interpretation across all models
- Recall $p(M \mid y)$ are posterior model probabilities of model $M \in \mathcal{M}$
- Posterior of ϕ across the space of models is a mixture:

$$p(\phi \mid y) = \sum_{M \in \mathcal{M}} p(\phi \mid y, M) p(M \mid y) = \sum_{r=1}^{R} p(\phi \mid y, M_r) p(M_r \mid y)$$

- For any function $h(\phi)$, the rules of conditional expectation imply

$$\mathbb{E}[h(\phi) \mid y] = \sum_{M \in \mathcal{M}} \mathbb{E}[h(\phi) \mid y, M] p\left(M \mid y\right)$$

Bayesian Model Averaging

- Note that $p(\phi \mid y, M)$ is the posterior of ϕ for each specific model $M \in \mathcal{M}$
- Posterior of ϕ across all models is a mixture of the model-specific posteriors using posterior model probabilities $p(M \mid y)$ as mixture weights
- Posterior integrates out model uncertainty using posterior model probabilities
- Intuition: average across models using $p(M \mid y)$ as weights
- Same intuition holds for expectations over the posterior of ϕ

Bayesian Model Averaging

- Note that $p(\phi \mid y, M)$ is the posterior of ϕ for each specific model $M \in \mathcal{M}$
- Posterior of ϕ across all models is a mixture of the model-specific posteriors using posterior model probabilities $p(M \mid y)$ as mixture weights
- Posterior integrates out model uncertainty using posterior model probabilities
- Intuition: average across models using $p(M \mid y)$ as weights
- Same intuition holds for expectations over the posterior of ϕ
- Common parameter interpretation is crucial: if models estimate different objects, average across models is not meaningful
- Example: ϕ in M_1 represents mean, but ϕ in M_2 represents variance
- Average across M_1 and M_2 will be like comparing apples to oranges

BMA in Linear Regression

- Linear regression is a great case study, as regression coefficients β always have the same interpretation if data is standardized
- Let us formally study *Bayesian Model Averaging* (BMA) for linear regression
- Focus again on *variable selection* amongst p covariates to simplify discussion
- Model space: $\mathcal{M} = \{M_1, \ldots, M_R\}$ with $R = 2^p$
- Assume for now p is small enough to fit model with all variable combinations

Priors for BMA: g-prior

- Need to fit 2^p models. How to calibrate 2^p priors?
- Model $M \in \mathcal{M}$ has p_M covariates: need to tune prior mean (p_M-dimensional vector) and variance ($p_M \times p_M$ matrix) for each model
- Let X_M denote the $n \times p_M$ matrix of regressors in model $M \in \mathcal{M}$
- The g-prior is one such way to set priors in this setting:

$$p(\beta_0) \propto 1 \quad \text{and} \quad p(\sigma^2) \propto \frac{1}{\sigma^2}$$
$$\beta^{(M)} | \sigma^2 \sim \mathcal{N}(\underline{\beta}^{(M)}, \sigma^2 \underline{V}_M)$$

- g-prior standard choice: $\underline{\beta}^{(M)} = 0_{p_M}$ and $\underline{V}_M = (g_M X_M^\top X_M)^{-1}$
- g-prior reduces prior choice to a single scalar g_M per model $M \in \mathcal{M}$

Posteriors for BMA

- For any given model $M \in \mathcal{M}$, this is just a standard Bayesian linear regression with conjugate priors!
- Analytical solutions for any given model $M \in \mathcal{M}$

$$\beta^{(M)}|y, M \sim \mathrm{t}_{p_M}(n, \bar{\beta}^{(M)}, \bar{s}^2_M \bar{V}_M)$$

- Posterior hyperparameters for this case (where ι_n is vector of n ones)

$$\begin{aligned}
\bar{V}_M &= \left[(1+g_M)X_M^\top X_M\right]^{-1}, \quad \bar{\beta}^{(M)} = \bar{V}_M X_M^\top y \\
\bar{s}^2_M &= \frac{1}{n}\left[\frac{1}{1+g_M} y^\top P_{X_M} y + \frac{g_M}{1+g_M}(y - \bar{y}\iota_n)^\top (y - \bar{y}\iota_n)\right] \\
P_{X_M} &= I_n - X_M (X_M^\top X_M)^{-1} X_M^\top
\end{aligned}$$

Posteriors for BMA

- We have marginalized all other parameters to focus on β
- For any given model $M \in \mathcal{M}$, the marginal likelihood (evidence) $p(y \mid M)$ is available in closed form!

$$p(y \mid M) \propto \left(\frac{g_M}{1+g_M}\right)^{\frac{p_M}{2}} \left[\frac{y^\top P_{X_M} y}{1+g_M} + \frac{g_M}{1+g_M}(y - \bar{y}\iota_n)^\top (y - \bar{y}\iota_n)\right]^{-\frac{n-1}{2}}$$

- For model $M \in \mathcal{M}$ with prior model probabilities $p(M)$, we can compute

$$p(M \mid y) \propto p(y \mid M)p(M)$$

- We can now use $p(M \mid y)$ to compare and aggregate across models!

Posteriors for BMA: Example

- Only two potential covariates X_1 or X_2, so set $p = 2$
- $R = 2^p = 4$, such that we have 4 potential models to consider
- M_0 only includes a constant, M_1 includes X_1 only, M_2 includes X_2 only, and M_3 includes both

$$
\begin{aligned}
M_0 &: y_i = \beta_0 + u_i \\
M_1 &: y_i = \beta_0 + \beta_1 \cdot x_{1,i} + u_i \\
M_2 &: y_i = \beta_0 + \beta_2 \cdot x_{2,i} + u_i \\
M_3 &: y_i = \beta_0 + \beta_1 \cdot x_{1,i} + \beta_1 \cdot x_{2,i} + u_i
\end{aligned}
$$

Posteriors for BMA: Example

- Bayes factors for model comparisons:

$$BF_{01} = \frac{p(y \mid M_0)}{p(y \mid M_1)}, \quad BF_{30} = \frac{p(y \mid M_3)}{p(y \mid M_0)}, \quad BF_{13} = \frac{p(y \mid M_1)}{p(y \mid M_3)}$$

- Posterior for β_1 across model space:

$$\begin{aligned} p(\beta_1 \mid y) &= p(\beta_1 \mid y, M_0)p(M_0 \mid y) + p(\beta_1 \mid y, M_1)p(M_1 \mid y) \\ &\quad + p(\beta_1 \mid y, M_2)p(M_2 \mid y) + p(\beta_1 \mid y, M_3)p(M_3 \mid y) \\ &= p(\beta_1 \mid y, M_1)p(M_1 \mid y) + p(\beta_1 \mid y, M_3)p(M_3 \mid y) \end{aligned}$$

- Posterior variance for β_2 across model space:

$$\text{Var}(\beta_2 \mid y) = \text{Var}(\beta_2 \mid y, M_2)p(M_2 \mid y) + \text{Var}(\beta_2 \mid y, M_3)p(M_3 \mid y)$$

BMA in Large Dimensions

- We assumed p was small enough that we could compute all 2^p posteriors
- For $p > 40$, this is no longer feasible. What do we do then?
- Objective: calculate posterior expectations over *model space* $\mathcal{M}$
- Main issue: model space is too large to enumerate all candidates models and compute all 2^p posterior model probabilities $p(M \mid y)$
- Can compute $p(M \mid y) \propto p(y \mid M)p(M)$ for any given M, but cannot integrate across all $M \in \mathcal{M}$

Metropolis-Hastings in Large Dimensions

- Metropolis-Hastings algorithm objective: calculate posterior expectations over high-dimensional *parameter space* Θ
- Main issue: parameter space is too large to enumerate all candidate values and calculate posterior probabilities $p(\theta \mid y)$
- Can compute $p(\theta \mid y) \propto p(y \mid \theta)p(\theta)$ for any given θ, but cannot integrate across all $\theta \in \Theta$

Metropolis-Hastings in Large Dimensions

- Metropolis-Hastings algorithm objective: calculate posterior expectations over high-dimensional *parameter space* Θ
- Main issue: parameter space is too large to enumerate all candidate values and calculate posterior probabilities $p(\theta \mid y)$
- Can compute $p(\theta \mid y) \propto p(y \mid \theta)p(\theta)$ for any given θ, but cannot integrate across all $\theta \in \Theta$
- Both problems seem very related...
- Metropolis-Hastings solution: sample from the posterior $p(\theta \mid y)$ to compute posterior expectations across Θ
- Analogue idea: sample from the posterior of models $p(M \mid y)$ to compute posterior expectations across $\mathcal{M}$

BMA in Large Dimensions

- Madigan and York (1995) propose a random-walk Metropolis-Hastings algorithm to sample from the space of models
- Ingredients for Metropolis-Hastings: target kernel and proposal distribution
- For given $M \in \mathcal{M}$, can compute target kernel $p(y \mid M)p(M)$
- Need proposal distribution $g(M)$ over models $M \in \mathcal{M}$

BMA in Large Dimensions

- Madigan and York (1995) propose a random-walk Metropolis-Hastings algorithm to sample from the space of models
- Ingredients for Metropolis-Hastings: target kernel and proposal distribution
- For given $M \in \mathcal{M}$, can compute target kernel $p(y \mid M)p(M)$
- Need proposal distribution $g(M)$ over models $M \in \mathcal{M}$
- Recall equivalence between model $M \in \mathcal{M}$ and $\gamma = (\gamma_1, \ldots, \gamma_p) \in \{0,1\}^p$
- Can express all model uncertainty in terms of γ through its marginal posterior $p(\gamma \mid y)$
- Posterior expectations over $M \in \mathcal{M}$ are the same as those over $\gamma \in \{0,1\}^p$
- MCMC methods to aggregate over model space: Markov Chain Monte Carlo Model Composition (MC^3) methods

BMA with Model Composition

- Recall γ lists all covariates included in a given model
- Let $\gamma^{(s)}$ represent the current model in consideration
- Random-walk proposal $g(\gamma \mid \gamma^{(s)})$ over $\gamma \in \{0,1\}^p$:
  1. Set candidate $\gamma^* = \gamma^{(s)}$
  2. Choose a variable index $j \in \{1, \ldots, p\}$ uniformly at random; i.e., choose a covariate X_j at random
  3. If $\gamma_j^{(s)} = 1$, set $\gamma_j^* = 0$; otherwise, set $\gamma_j^* = 1$
- Based on a given model $\gamma^{(s)}$, either add or delete a covariate at random
- By iterating this process, we can get a sample of models $\gamma^{(1)}, \ldots, \gamma^{(S)}$

BMA with Model Composition: Implementation

1. Start from a baseline model $\gamma^{(0)}$
2. Across iterations $s = 1, \ldots, S$:
   - 2.1 Conditional on the model $\gamma^{(s)}$, draw candidate model γ^* from $g(\gamma \mid \gamma^{(s)})$
   - 2.2 Find M^* as the model implied by γ^* and $M^{(s)}$ as implied by $\gamma^{(s)}$
   - 2.3 Compute posterior model probabilities $p(M^* \mid y)$ and $p(M^{(s)} \mid y)$
   - 2.4 Calculate the acceptance rate

$$\alpha(M^{(s)}, M^*) = \min\left\{\frac{p(M^* \mid y)}{p(M^{(s)} \mid y)}, 1\right\} = \min\left\{\frac{p(y \mid M^*)p(M^*)}{p(y \mid M^{(s)})p(M^{(s)})}, 1\right\}$$

   - 2.5 Draw U uniformly between 0 and 1
   - 2.6 If $U \leq \alpha(M^{(s)}, M^*)$, set $M^{(s+1)} = M^*$. Otherwise, set $M^{(s+1)} = M^{(s)}$
   - 2.7 Using model $M^{(s+1)}$, fit β and remaining model parameters

BMA and Model Uncertainty

- Again, recall our objective: summarize and account for model uncertainty
- We now have a way of exploring the model space even when it is too large to go through all possible models
- We can look at similar quantities of interest as before:
  - Probability of inclusion for each covariate of interest
  - Most likely model and model size
  - Estimates aggregate information from all visited models
- Limitations:
  - If model space is too large, can only visit limited number of candidate models
  - Restrictive model choice as we required computation of marginal likelihood
  - Require large amounts of computation

UofG