Mixture of g Priors for Bayesian Variable Selection

Feng Liang, Rui Paulo *et al.*

Sheng Zhang

Department of Statistics,
University of Wisconsin Madison

April 30, 2010

Outline

Outline

Basic Setup

- Consider $Y \sim N(\boldsymbol{\mu}, I_n/\phi)$, where $Y = (y_1, y_2, \ldots, y_n)^T$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^T$, I_n is the $n \times n$ identity matrix, and ϕ is the precision parameter
- Potential centered predictors $X_1, \ldots, X_p$
- Only consider the case $n \geq p + 2$
- Index the model space by $\gamma_{p\times 1}$:

$$\gamma_j = \begin{cases} 0 & \text{if } X_j \text{ is excluded} \\ 1 & \text{if } X_j \text{ is included} \end{cases}$$

- Under model $\mathcal{M}_\gamma$: $\boldsymbol{\mu} = \mathbf{1}_n\alpha + X_\gamma\boldsymbol{\beta}_\gamma$

Key Idea of Bayesian Variable Selection

- Put priors on the unknowns $\boldsymbol{\theta}_\gamma = (\alpha, \beta_\gamma, \phi) \in \Theta_\gamma$
- Update prior probabilities of models $p(\mathcal{M}_\gamma)$ to

$$p(\mathcal{M}_\gamma|Y) = \frac{p(\mathcal{M}_\gamma)p(Y|\mathcal{M}_\gamma)}{\sum_\gamma p(\mathcal{M}_\gamma)p(Y|\mathcal{M}_\gamma)}$$

  where $p(Y|\mathcal{M}_\gamma) = \int_{\Theta_\gamma} p(Y|\boldsymbol{\theta}_\gamma, \mathcal{M}_\gamma)p(\boldsymbol{\theta}_\gamma|\mathcal{M}_\gamma)d\boldsymbol{\theta}_\gamma$, and $p(\mathcal{M}_\gamma)$ could be $1/2^p$
- Choose the model with greatest $p(\mathcal{M}_\gamma|Y)$

The Goal of the Paper

- $Y|\alpha, \beta_\gamma, \phi, \mathcal{M}_\gamma \sim N(1_n\alpha + X_\gamma\beta_\gamma, I_n/\phi)$

  $p(\alpha, \phi|\mathcal{M}_\gamma) = \frac{1}{\phi}$

  $\beta_\gamma|\phi, \mathcal{M}_\gamma \sim N(0, \frac{g}{\phi}(X_\gamma^T X_\gamma)^{-1})$ (Zellner's g prior)
- Several previous work involves choices of calibration of g
- g acts as a dimensionality penalty
- The goal of the paper is to propose a new family of priors for g, the hyper-g prior family, to guarantee:
  - robustness of mis-specification of g
  - a closed-form marginal likelihoods
  - computational efficiency
  - desirable consistency properties in model selection

Outline

Null-Based Bayes Factors (1)

- The bayes factor of comparing each of $\mathcal{M}_\gamma$ to a base model $\mathcal{M}_b$ is

$$\mathrm{BF}[\mathcal{M}_\gamma : \mathcal{M}_b] = \frac{p(Y|\mathcal{M}_\gamma)}{p(Y|\mathcal{M}_b)}$$

- To compare two models $\mathcal{M}_\gamma$ and $\mathcal{M}_{\gamma'}$,

$$\mathrm{BF}[\mathcal{M}_\gamma : \mathcal{M}_{\gamma'}] = \frac{\mathrm{BF}[\mathcal{M}_\gamma : \mathcal{M}_b]}{\mathrm{BF}[\mathcal{M}_{\gamma'} : \mathcal{M}_b]}$$

- The posterior probability could be written as

$$p(\mathcal{M}_\gamma|Y) = \frac{p(\mathcal{M}_\gamma)\mathrm{BF}[\mathcal{M}_\gamma : \mathcal{M}_b]}{\sum_{\gamma'} p(\mathcal{M}_{\gamma'})\mathrm{BF}[\mathcal{M}_{\gamma'} : \mathcal{M}_b]}$$

Null-Based Bayes Factors (2)

- $\mathcal{M}_b = \mathcal{M}_N$
- $H_0 : \beta_\gamma = 0$ *vs.* $H_0 : \beta_\gamma \neq 0$
- Recall $p(\alpha, \phi|\mathcal{M}_\gamma) = \frac{1}{\phi}$ and $\beta_\gamma|\phi, \mathcal{M}_\gamma \sim N(0, \frac{g}{\phi}(X_\gamma^T X_\gamma)^{-1})$
- Closed form of marginal likelihood:

  $$p(Y|\mathcal{M}_\gamma, g) = \frac{\Gamma((n-1)/2)}{\sqrt{(\pi)^{(n-1)}}\sqrt{n}} \| Y - \bar{Y} \|^{-(n-1)} \times \frac{(1+g)^{(n-1-p_\gamma)/2}}{[1+g(1-R^2_\gamma)]^{-(n-1)/2}}$$

  The null model $p(Y|\mathcal{M}_N)$ corresponds to $R^2_\gamma = 0$ and $p_\gamma = 0$
- $\mathrm{BF}[\mathcal{M}_\gamma : \mathcal{M}_N] = (1+g)^{(n-1-p_\gamma)/2}[1 + g(1 - R^2_\gamma)]^{-(n-1)/2}$

Paradoxes of fixed g Priors – Bartlett's Paradox

When $g \to \infty$ while n and p_γ are fixed:

$$\begin{aligned} \mathrm{BF}[\mathcal{M}_\gamma : \mathcal{M}_N] &= (1+g)^{(n-1-p_\gamma)/2}[1+g(1-R^2_\gamma)]^{-(n-1)/2} \\ &\to 0 \end{aligned}$$

This means, regardless of the information in the data, the Bayes factor always favors the null model, which is due to the large spread of the prior induced by the noninformative choice of g

Paradoxes of fixed g Priors – Information Paradox

- Suppose $\|\hat{\boldsymbol{\beta}}_\gamma\|^2 \to \infty$ so that $R^2_\gamma \to 1$ while n and p_γ are fixed
- Expect $\mathrm{BF}[\mathcal{M}_\gamma : \mathcal{M}_N] \to \infty$
- However, as $R^2_\gamma \to 1$,

$$\begin{aligned} \mathrm{BF}[\mathcal{M}_\gamma : \mathcal{M}_N] &= (1+g)^{(n-1-p_\gamma)/2}[1+g(1-R^2_\gamma)]^{-(n-1)/2} \\ &\to (1+g)^{(n-p_\gamma-1)/2} \end{aligned}$$

which is a constant!

Choices of g

- *Unit information prior*: $g = n$ (BF behaves like BIC)
- *Risk inflation criterion*: $g = p^2$ (minimax perspective)
- *Benchmark prior* : $g = max(n, p^2)$ (BRIC)
- *Local empirical Bayes* : the MLE of $p(Y|\mathcal{M}_\gamma, g)$ with the nonnegative constraint. $\hat{g}_\gamma^{EBL} = max(F_\gamma - 1, 0)$, where $F_\gamma = \frac{R_\gamma^2/p_\gamma}{(1-R_\gamma^2)/(n-1-p_\gamma)}$.
- *Global empirical Bayes*: $\hat{g}^{EBL} = \text{argmax}_{g>0} \sum_\gamma p(\mathcal{M}_\gamma) \frac{(1+g)^{(n-1-p_\gamma)/2}}{[1+g(1-R_\gamma^2)]^{(n-1)/2}}$

Choices of g and Information Paradox

- For fixed n and p,
  - The *Unit information prior*, *Risk inflation criterion* and the *Benchmark prior* do not solve the information paradox
  - The two EB approaches do have the desirable behavior
- *Theorem 1*: In the setting of the information paradox with fixed n, $p < n$ and $R^2_\gamma \to 1$, for both global and local EB estimate of g,

$$\begin{aligned} \mathrm{BF}[\mathcal{M}_\gamma : \mathcal{M}_N] &= (1+g)^{(n-1-p_\gamma)/2}[1+g(1-R^2_\gamma)]^{-(n-1)/2} \\ &\to \infty \end{aligned}$$

Proof: by direct checking

Outline

Desirable $\pi(g)$

- $g \sim \pi(g)$
- The Bayes factor $\mathrm{BF}[\mathcal{M}_\gamma : \mathcal{M}_N] = \int_0^\infty (1+g)^{(n-1-p_\gamma)/2}[1+g(1-R^2_\gamma)]^{-(n-1)/2}\pi(g)dg$
- The posterior mean $\boldsymbol{\mu}$ under $\mathcal{M}_\gamma \neq \mathcal{M}_N$: $\mathbb{E}[\boldsymbol{\mu}|\boldsymbol{\mu}_\gamma, Y] = \mathbf{1}_n\hat{\alpha} + \mathbb{E}\left[\frac{g}{1+g}|\mathcal{M}_\gamma, Y\right]X_\gamma\hat{\boldsymbol{\beta}}_\gamma$, where $\hat{\alpha}$ and $\hat{\beta}$ are least square estimates of α and β, and $\mathbb{E}\left[\frac{g}{1+g}\right.$ is regarded as a shrinkage factor
- The optimal Bayes estimate of $\boldsymbol{\mu}$ under the squared error loss: $\mathbb{E}[\boldsymbol{\mu}|Y] = \mathbf{1}_n\hat{\alpha} + \sum_{\gamma:\mathcal{M}_\gamma \neq \mathcal{M}_N} p(\mathcal{M}_\gamma|bY)\mathbb{E}\left[\frac{g}{1+g}|\mathcal{M}_\gamma, Y\right]X_\gamma\hat{\boldsymbol{\beta}}_\gamma$
- g appears everywhere: BF, posterior mean and prediction
- Want priors leading to tractable computation for these quantities, and consistent model selection and risk properties

Zellner-Siow Cauchy Priors

- Jeffreys (1961) rejected normal priors essentially for reasons related to BF paradoxes
- Cauchy prior is the simplest prior to satisfy basic consistency requirement for hypothesis testing
- The Zellner-Siow priors can be represented as a mixture of g priors with an Inv-Gamma$(1/2, n/2)$:

$$\pi(g) = \frac{(n/2)^{1/2}}{\Gamma(1/2)} g^{-3/2} e^{-n/(2g)}$$

- The corresponding integrals are are approximated by Laplace approximation
- As the model dimensionality increases, the accuracy of the approximation decreases

Hyper-g Priors (1)

- $\pi(g) = \frac{a-2}{2}(1+g)^{-a/2},\ g > 0$
- Only consider the case $a > 2$ when $\pi(g)$ is a proper prior
- This prior leads to the shrinkage factor $\frac{g}{1+g} \sim \text{Beta}(1, \frac{a}{2} - 1)$
- Value of $a \geq 4$ tends to put more mass on shrinkage values near 0, which is undesirable, hence only consider $2 < a \leq 4$
- When a $= 4$, $\frac{g}{1+g}$ has a uniform distribution
- When a $= 3$, most of the mass is near 1

Hyper-g Priors (2)

- Main advantage of hyper-g prior : leads to closed form of posterior distribution of g in terms of Gaussian hypergeometric function
- The posterior distribution of g:

$$\begin{aligned} p(g|Y,\mathcal{M}_\gamma) &= \frac{p_\gamma + a - 2}{2_2F_1((n-1)/2, 1; (p_\gamma + a)/2; R^2_\gamma)} \\ &\times (1+g)^{(n-1-p_\gamma-a)/2}[1+(1-R^2_\gamma)g]^{-(n-1)/2} \end{aligned}$$

- ${}_2F_1(a,b;c;z)$ is convergent for real $|z| < 1$ with $c > b > 0$ and for $z = \pm 1$ only if $c > a + b$ and $b > 0$
- To evaluate Gaussian hypergeometric function, numerical overflow is problematic for moderate to large n and large R^2_γ.

Hyper-g Priors (3)

Gaussian hypergeometric function appears in many quantities of interest:

- $\mathrm{BF}[\mathcal{M}_\gamma : \mathcal{M}_N] = \frac{a-2}{p_\gamma+a-2} {}_2F_1(\frac{n-1}{2}, 1; \frac{p_\gamma+a}{2}; R^2_\gamma)$
- $\mathbb{E}[g|\mathcal{M}_\gamma, \mathbf{Y}] = \frac{2}{p_\gamma+a-4} \frac{{}_2F_1((n-1)/2,2;(p_\gamma+a)/2;R^2_\gamma)}{{}_2F_1((n-1)/2,1;(p_\gamma+a)/2;R^2_\gamma)}$
- $\mathbb{E}[\frac{g}{1+g}|\mathcal{M}_\gamma, \mathbf{Y}] = \frac{2}{p_\gamma+a} \frac{{}_2F_1((n-1)/2,2;(p_\gamma+a)/2+1;R^2_\gamma)}{{}_2F_1((n-1)/2,1;(p_\gamma+a)/2;R^2_\gamma)}$

Outline

Overview

- The following three aspects of consistency are considered:
  1) the "information paradox" where $R^2_\gamma \longrightarrow 1$
  2) the asymptotic consistency of model posterior probabilities as $n \longrightarrow \infty$
  3) the asymptotic consistency for prediction
- The above are studied under the assumption of the true model

Consistency–Information Paradox (1)

Theorem 2: To resolve the information paradox for all n and $p < n$, it suffices to have

$$\int_0^\infty (1+g)^{(n-1-p_\gamma)/2}\pi(g)dg = \infty \quad \forall p_\gamma \leq p$$

In the case of minimal sample size ($n = p + 2$), it suffices to have $\int_0^\infty (1+g)^{1/2}\pi(g)dg = \infty$.

Proof: The Bayes factor $\mathrm{BF}[\mathcal{M}_\gamma : \mathcal{M}_N]$ is monotonic increasing function of R^2_γ. By monotone convergence theorem, it goes to $\int (1+g)^{(n-1-p_\gamma)/2}\pi(g)dg$ as $R^2_\gamma \to 1$. Hence the non-integrability of $(1+g)^{(n-1-p_\gamma)/2}\pi(g)$ is sufficient and necessary condition for resolving the "information paradox".

Consistency–Information Paradox (2)

- Zellner-Siow prior satisfies the condition
- When $a \leq n - p_\gamma + 1$, the hyper-g prior satisfies the condition
- Fixed g prior corresponds to the degenerate prior that is a point mass at a selected value of g, so no fixed choice of g solves the paradox

Consistency–Model Selection Consistency (1)

- Want: $\text{plim}_n p(\mathcal{M}_\gamma|Y) = 1$ when $\mathcal{M}_\gamma$ is the true model, where the probability measure is the sampling distribution under the assumption of true model
- Equivalently, $\text{plim}_n \text{BF}[\mathcal{M}_{\gamma'} : \mathcal{M}_\gamma] = 0$ for all $\mathcal{M}_{\gamma'} \neq \mathcal{M}_\gamma$
- Assumption: for $\mathcal{M}_{\gamma'}$ that doesn't contain $\mathcal{M}_\gamma$,

$$\lim_{n\to\infty} \frac{\beta_\gamma^T X_\gamma^T (I - P_{\gamma'}) X_\gamma \beta_\gamma}{n} = b_{\gamma'} \in (0, \infty) \quad (a)$$

  where $P_{\gamma'}$ is the projection matrix onto the span of $X_{\gamma'}$
- Fernandez et al. (2001) have shown the consistency for BRIC and BIC under the assumption

Consistency–Model Selection Consistency (2)

Theorem 3: Assume assumption (a) holds. When the true model is not the null model ($\mathcal{M}_\gamma \neq \mathcal{M}_N$), posterior probabilities under empirical Bayes, Zellner-Siow priors, and hyper-g priors are consistent for model selection; when $\mathcal{M}_\gamma = \mathcal{M}_N$, consistency still holds true for the Zellner-Siow prior, but does not hold for the hyper-g or local and global empirical Bayes.

- Z-S prior on g depends on n, while EB or hyper-g priors don't
- For EB and hyper-g priors, under $\mathcal{M}_N$, the null model is still the model with highest posterior probability, although it is bounded away from 1.
- Could consider EB and hyper-g priors as consistent in a weaker sense (under a 0-1 loss)
- The hyper-g/n prior is proposed to solve the inconsistency problem under $\mathcal{M}_N$: $\pi(g) = \frac{a-2}{2n}(1 + \frac{g}{n})^{-a/2}$

Consistency–Model Selection Consistency (3 (proof))

The following preliminary results from Fernandez et al. (2001) are cited without proof. Under the assumed true model $\mathcal{M}_\gamma$:

1) If $\mathcal{M}_\gamma$ is nested within or equal to a model $\mathcal{M}_{\gamma'}$, then

$$\text{plim}_{n\to\infty} \frac{\text{RSS}_{\gamma'}}{n} = \frac{1}{\phi} \qquad (R1)$$

2) For any model $\mathcal{M}_{\gamma'}$ that does not contain $\mathcal{M}_\gamma$, under the assumption (a),

$$\text{plim}_{n\to\infty} \frac{\text{RSS}_{\gamma'}}{n} = \frac{1}{\phi + b_{\gamma'}} \qquad (R2)$$

where $\text{RSS}_\gamma = (1 - R^2_\gamma)\| Y - \bar{Y}\|^2$ is the residual sum of squares

Consistency–Model Selection Consistency (4 (proof))

Firstly consider the consistency result for local EB estimate when $\mathcal{M}_\gamma \neq \mathcal{M}_N$. Note $R^2_\gamma \to c \in (0,1)$ when $\mathcal{M}_\gamma \cap \mathcal{M}_{\gamma'} \neq \emptyset$, we have:

$$\hat{g}^{EBL}_{\gamma'} = \Big[\frac{R^2_{\gamma'}/p_{\gamma'}}{(1-R^2_{\gamma'})/(n-1-p_{\gamma'})}\Big](1+o_p(1))$$

$$\mathrm{BF}_{EBL}[\mathcal{M}_{\gamma'}:\mathcal{M}_N] \sim_P \frac{1}{(1-R^2_{\gamma'})^{(n-1-p_{\gamma'})/2}} \frac{(n-1-p_{\gamma'})^{(n-1-p_{\gamma'})/2}}{(n-1)^{(n-1)/2}}$$

$$\mathrm{BF}_{EBL}[\mathcal{M}_{\gamma'}:\mathcal{M}_\gamma] \sim_p \frac{1}{n^{(P_{\gamma'}-p_\gamma)/2}} \Big(\frac{RSS_\gamma/n}{RSS_{\gamma'}/n}\Big)^{n/2}$$

Consistency–Model Selection Consistency (5 (proof))

a) $\mathcal{M}_\gamma \cap \mathcal{M}_{\gamma'} \neq \emptyset$ and $\mathcal{M}_\gamma \not\subseteq \mathcal{M}_{\gamma'}$. Apply (R1) and (a),

$$\text{plim}_{n\to\infty}\left(\frac{RSS_\gamma/n}{RSS_{\gamma'}/n}\right) = \lim_{n\to\infty}\left(\frac{1/\phi}{1/(\phi + b_{\gamma'})}\right)^{n/2} \to_p 0$$

hence $\text{BF}_{EBL}[\mathcal{M}_{\gamma'} : \mathcal{M}_\gamma] \to_p 0$

b) $\mathcal{M}_\gamma \subseteq \mathcal{M}_{\gamma'}$. Since

$$(RSS_\gamma/RSS_{\gamma'})^{n/2} \to_d exp(\chi^2_{p_{\gamma'}-p_\gamma}/2) \quad (Fernandez\ 2001)$$

together with the fact that $1/n^{(p_{\gamma'}-p_\gamma)/2} \to 0$, we have $\text{BF}_{EBL}[\mathcal{M}_{\gamma'} : \mathcal{M}_\gamma] \to_p 0$.

Consistency–Model Selection Consistency (6 (proof))

c) $\mathcal{M}_\gamma \cap \mathcal{M}_{\gamma'} = \emptyset$. In this case $nR^2_{\gamma'} \to_d \chi^2_{p_{\gamma'}}/(1+\phi b'_\gamma)$. Since

$$\begin{aligned} \mathrm{BF}_{EBL}[\mathcal{M}_{\gamma'} : \mathcal{M}_N] &= \frac{(1+g)^{(n-1-p_{\gamma'})/2}}{[1+(1-R^2_{\gamma'})g]^{(n-1)/2}} \\ &\leq (1-R^2_{\gamma'})^{-(n-1)/2} \end{aligned}$$

we have $\mathrm{BF}_{EBL}[\mathcal{M}_{\gamma'} : \mathcal{M}_\gamma] = O_p(1)$. On the other hand, since

$$\mathrm{BF}_{EBL}[\mathcal{M}_\gamma : \mathcal{M}_N] \sim^P (n-1)^{-p_\gamma/2}(1-R^2_\gamma)^{-n/2}$$

where the second term goes to ∞ exponentially fast, $\mathrm{BF}_{EBL}[\mathcal{M}_{\gamma'} : \mathcal{M}_\gamma] \to_p 0$

Consistency–Model Selection Consistency (7 (proof))

- Similarly we can get the consistency for global EB, Z-S prior, hyper-g prior and hyper-g/n priors, when $\mathcal{M}_\gamma \neq \mathcal{M}_N$
- When $\mathcal{M}_\gamma = \mathcal{M}_N$, only the Z-S prior is still consistent. The proof is similar with the case $\mathcal{M}_\gamma \neq \mathcal{M}_N$. The only difference is that $R^2_{\gamma'} \to 0$ if $\mathcal{M}_{\gamma'} \neq \mathcal{M}_N$

Consistency–Prediction Consistency (1)

- The optimal point estimator under the squared error loss is

$$\hat{Y}_n^\star = \hat{\alpha} + \sum_\gamma x_\gamma^{\star\top} \hat{\beta}_\gamma p(\mathcal{M}_\gamma | Y) \int_0^\infty \frac{g}{1+g} \pi(g|\mathcal{M}_\gamma, Y) dg$$

- $\hat{Y}_n^\star$ is consistent under prediction if

$$\text{plim}_n \hat{Y}_n^\star = \mathbb{E} Y^\star = \alpha + \mathbf{x}_\gamma^{\star\top} \beta_\gamma$$

Consistency–Prediction Consistency (2)

Theorem 4. $\hat{Y}_n^\star$ is consistent under empirical Bayes, the hyper-g, hyper-g/n and Zellner-Siow priors are consistent in prediction.

- When $\mathcal{M}_\gamma = \mathcal{M}_N$, $\|\hat{\beta}_\gamma\| \to 0$ by the consistency of LSE. Hence the prediction consistency of $\hat{Y}_n^\star$ follows
- When $\mathcal{M}_\gamma \neq \mathcal{M}_N$, $\pi(\mathcal{M}_\gamma|Y) \to 1$ by Theorem 3. Using the consistency of LSE, it suffices to show

$$\text{plim}_n \int_0^\infty \frac{g}{1+g} \pi(g|\mathcal{M}_\gamma, Y) dg = 1$$

The result follows by applying Laplace approximation

Outline

Discussion

- Advantage of mixture g priors
  - Solved some paradox issues
  - Perform as well as other default choices
- Limitation
  - Numerical problem for large n and large R^2_γ
  - Zellner-Siow priors require $p_\gamma < n - 2$, and hyper-g prior requires $p_\gamma < n - 3 - a$
- Future work
  - Consider using other priors on $P(\mathcal{M}_\gamma)$
  - Look into the case when X_γ is not of full rank
  - Large p small n problem

THANK YOU