A new Bayesian variable selection criterion based on a g-Prior extension for $p > n$

Yuzo Maruyama
and Edward George

CSIS, The University of Tokyo, Japan
Department of Stat, University of Pennsylvania

Overview: Our recommendable Bayes factor

$$
\begin{cases}
\left\{ \overline{\mathrm{sv}}[X_\gamma] \times \|\hat{\beta}^{MP}_{LSE}[\gamma]\| \right\}^{-n+1} & \text{if } q_\gamma \geq n-1 \\
\dfrac{d_{q_\gamma}^{q_\gamma}(1-R_\gamma^2)^{-\frac{n-q_\gamma}{2}+\frac{3}{4}} B(\frac{q_\gamma}{2}+\frac{1}{4}, \frac{n-q_\gamma}{2}-\frac{3}{4})}{\overline{\mathrm{sv}}[X_\gamma]^{q_\gamma}(1-R_\gamma^2+d_{q_\gamma}^2\|\hat{\beta}_{LSE}[\gamma]\|^2)^{\frac{1}{4}+\frac{q_\gamma}{2}} B(\frac{1}{4}, \frac{n-q_\gamma}{2}-\frac{3}{4})} \\
\quad \text{if } q_\gamma \leq n-2
\end{cases}
$$

- A criterion based on full Bayes
- but we need no MCMC
- An exact closed form by using a special prior
- applicable for $p > n$ as well as $n > p$
- model selection consistency and good numerical performance

Full model

- $Y|\{\alpha, \beta, \sigma^2\} \sim N_n(\alpha 1_n + X\beta, \sigma^2 I)$
- α: an intercept parameter
- $1_n = (1, 1, \ldots, 1)'$
- $X = (X_1, \ldots, X_p)$: an $n \times p$ standardized design matrix rank $X = \min(n-1, p)$
- β: a $p \times 1$ vector of unknown coefficients
- σ^2: an unknown variance

Since there is usually a subset of useless regressors in the full model, we would like to choose a good sub-model with only important regressors.

Submodel

- submodel $\mathcal{M}_\gamma$

  $$Y|\{\alpha, \beta_\gamma, \sigma^2\} \sim N_n(\alpha 1_n + X_\gamma \beta_\gamma, \sigma^2 I)$$
- Assume the intercept is always included
- X_γ: the $n \times q_\gamma$ matrix, rank $X_\gamma = \min(n-1, q_\gamma)$
  columns = the γth subset of $X_1, \ldots, X_p$
- β_γ: a $q_\gamma \times 1$ vector of unknown regression coefficients
- q_γ: the number of regressors of $\mathcal{M}_\gamma$
- The null model: The special case of sub-model

$$\mathcal{M}_N : \; Y|\{\alpha, \sigma^2\} \sim N_n(\alpha 1_n, \sigma^2 I)$$

Variable selection in the Bayesian framework

- It entails the specification of prior
  - on the models $\Pr(\mathcal{M}_\gamma)$
  - on parameters $p(\alpha, \beta_\gamma, \sigma^2)$ of each model
- Assumption: equal model space probability

$$\Pr(\mathcal{M}_\gamma) = \Pr(\mathcal{M}_{\gamma'}) \text{ for any } \gamma \neq \gamma'$$

- Choose $\mathcal{M}_\gamma$ as the best model which maximizes

$$\text{posterior prob. } \Pr(\mathcal{M}_\gamma | y) = \frac{m_\gamma(y)}{\sum_\gamma m_\gamma(y)}$$

- $m_\gamma(y)$: the marginal density under $\mathcal{M}_\gamma$
  larger $m_\gamma(y)$ is better!

Variable selection in the Bayesian framework

- the marginal density

$$m_\gamma(y) = \iiint p_y(y|\alpha, \beta_\gamma, \sigma^2) p(\alpha, \beta_\gamma, \sigma^2) d\alpha d\beta_\gamma d\sigma^2$$

- Recall that we consider Full Bayes method, which means the joint prior density $p(\alpha, \beta_\gamma, \sigma^2)$ does not depend on data unlike Empirical Bayes method.
- Bayes factor is often used for expression of $\Pr(\mathcal{M}_\gamma|y)$

$$\Pr(\mathcal{M}_\gamma|y) = \frac{\mathrm{BF}(\mathcal{M}_\gamma; \mathcal{M}_N)}{\sum_\gamma \mathrm{BF}(\mathcal{M}_\gamma; \mathcal{M}_N)}$$

$$\text{where} \quad \mathrm{BF}(\mathcal{M}_\gamma; \mathcal{M}_N) = \frac{m_\gamma(y)}{m_N(y)}$$

Priors

- The form of our joint density

$$p(\alpha, \beta_\gamma, \sigma^2) = p(\alpha)\ p(\sigma^2)\ p(\beta|\sigma^2)$$
$$= 1\ \times\ \sigma^{-2}\ \times\ \int p(\beta|g, \sigma^2) p(g) dg$$

- $1 \times \sigma^{-2}$: a popular non-informative prior
- improper but justificated because α and σ^2 are included in all submodels
- $p(\beta|g, \sigma^2)$ and $p(g)$

The original Zellner's g-prior

- prior of regression coefficients
- Zellner's (1986) g-prior is popular

$$p_{\beta_\gamma}(\beta_\gamma|\sigma^2, g) = N_{q_\gamma}(0, g\sigma^2(X'_\gamma X_\gamma)^{-1})$$

- It is applicable for the traditional situation $p + 1 < n$
  $\Rightarrow q_\gamma + 1 < n$ for any $\mathcal{M}_\gamma$
- There are many papers which use g-priors including George and Foster (2000, Biometrika) and Liang et al. (2008, JASA)

The beauty of the g-prior

- The marginal density of y given g and σ^2

$$\exp\left(\frac{g}{g+1}\left\{\max_{\alpha,\beta_\gamma}\log p(Y|\alpha,\beta_\gamma,\sigma^2) - \frac{q_\gamma}{2}\frac{g+1}{g}\log(g+1)\right\}\right)$$

- Under known σ^2,

$$g^{-1}(g+1)\log(g+1) = 2, \text{ or } \log n$$

  leads to AIC by Akaike (1974) and BIC by Schwarz (1978) respectively
- several studies: how to choose g based on non-full Bayesian method

Many regressors case ($p > n$)

- In modern statistics, treating (very) many regressors case ($p > n$) becomes more and more important
- the original Zellner's g-prior is not available
- R^2 is always 1 in the case where $q_\gamma \geq n - 1$
  $\Rightarrow$ naive AIC and BIC methods do not work
- When we do not use the original g-prior, Bayesian method is available in many regressors case
  for example $\beta \sim N(0, \sigma^2 \lambda I)$
- inverse-gamma conjugate prior for σ^2 are also available

Many regressors case ($p > n$)

- The integral with respect to λ still remains in $m_\gamma(y)$ as long as the full Bayes method is considered.
- Needless to say, it should be calculated by numerical methods like MCMC or by approximation like Laplace method.
- We do not have comparative advantage in numerical methods,,,,,
- We like exact analytical results very much.

A variant of Zellner's g-prior

- a special variant of g-prior which enables us to
  - not only calculate the marginal density analytically (closed form!!)
  - but also treat many regressors case
- [KEY] singular value decomposition of X_γ

$$X_\gamma = U_\gamma D_\gamma W'_\gamma = \sum_{i=1}^{r} d_i[\gamma] u_i[\gamma] w'_i[\gamma]$$

- r: rank of $X = \min(q_\gamma, n-1)$
- the $n-1$ is from "X is the centered matrix"
- singular values $d_1[\gamma] \geq \cdots \geq d_r[\gamma] > 0$

A special variant of g-prior

$$p_\beta(\beta|g,\sigma^2) = \begin{cases} \prod_{i=1}^{n-1} p_i(w_i'\beta|g,\sigma^2) \times \overbrace{p_\#(W_\#'\beta)}^{arbitrary} \\ \qquad \text{if } q \ge n \\ \prod_{i=1}^{q} p_i(w_i'\beta|g,\sigma^2) \ \ \text{if } q \le n-1 \end{cases}$$

$$p_i(\cdot|g,\sigma^2) = N(0, \frac{\sigma^2}{d_i^2}\{\nu_i(1+g)-1\})$$

$W_\#$: a $q \times (q-r)$ matrix from the orthogonal complement of W

c.f. original g-prior $\ p_\beta(\beta|g,\sigma^2) = \prod_{i=1}^{q} p_i(w_i'\beta|g,\sigma^2)$ if $q \le n-1$

$$p_i(\cdot|g,\sigma^2) = N(0, g\frac{\sigma^2}{d_i^2})$$

A special variant of g-prior

- $\nu_1, \ldots, \nu_r$ ($\nu_i \geq 1$) where $r = \min\{n-1, q\}$ hyperparameters we have to fix
- $q \leq n-1 \Rightarrow (Z'Z)^{-1}$ exists
  $\nu_1 = \cdots = \nu_q = 1 \Rightarrow$ the original Zellner's prior
- the descending order $\nu_1 \geq \cdots \geq \nu_r$ like

  $$\nu_i = d_i^2/d_r^2 \quad \text{(our recommendation)}$$

  for $1 \leq i \leq r$ is reasonable for our purpose
- numerical experiment and the estimation after selection support the choice

Sketch of the calculation of the marginal density

- we have prepared all of priors except for g (we will give a prior of g later)
- the marginal density of y given g

  = the marignal density after the integration w.r.t. α, β, σ^2

$$m_\gamma(y|g) = C(n, y) \left\{ (g+1)(1 - R_\gamma^2) + \mathrm{GR}_\gamma^2 \right\}^{-(n-1)/2} \times \frac{(1+g)^{-r/2+(n-1)/2}}{\prod_{i=1}^{r} \nu_i^{1/2}}$$

where GR_γ^2 means the "generalized" R_γ^2

$$GR_\gamma^2 = \sum_{i=1}^{r} \frac{(u_i'\{y - \bar{y}1_n\})^2}{\nu_i \|y - \bar{y}1_n\|^2}$$

Many regressors case

- rank of $X = r = n-1$, $R^2_\gamma = 1$
- $m_\gamma(y|g)$ does not depend on g

$$m_\gamma(y) = m_\gamma(y|g) = C(n,y)\prod_{i=1}^{n-1}\nu_i^{-1/2}\left(GR^2_\gamma\right)^{-(n-1)/2}$$

- If $\nu_1 = \cdots = \nu_{n-1} = 1$, GR^2_γ just becomes 1 and hence $m_\gamma(y) = C(n,y)$
- it does not work for model selection because it always takes the same value in many regressors case
- That is why the choice of ν is important.

few regressors case ($q \leq n-2$)

- $p_g(g) = \{B(a+1, b+1)\}^{-1} g^b (1+g)^{-a-b-2}$
- it is proper if $a > -1$ and $b > -1$
- Liang et al (2008, JASA) "hyper-g priors" $b = 0$

$$p_g(g) = (a+1)^{-1}(g+1)^{-a-2}$$

- $b = (n-5-r)/2 - a$ is for getting a closed simple form of the marginal density
- $-1 < a < -1/2$ is for well-defining the marginal density of every sub-model
- The median $a = -3/4$ is our recommendation

Sketch of the calculation of the marginal density

- When $b = (n-5)/2 - r/2 - a$, the beta function takes the integration w.r.t. g

$$\int m_\gamma(y|g)p(g)dg = \frac{C(n,y)B(q/2+a+1,b+1)(1-R_\gamma^2+\mathrm{GR}_\gamma^2)^{-(n-1)/2+b+1}}{\prod_{i=1}^r \nu_i^{1/2} B(a+1,b+1)(1-R_\gamma^2)^{b+1}}$$

- When $b \neq (n-5)/2 - r/2 - a$, there remains an integral with R_γ^2 and GR_γ^2 in $m_\gamma(y)$
  $\Rightarrow$ the need of MCMC or approximation
- Liang et al (2008, JASA) $b = 0$, $\nu_1 = \cdots = \nu_r = 1$ the Laplace approximation

Our recommendable BF

- After insertion of our recommendable hyperparameters $a = -3/4$, $b = (n-5)/2 - r/2 - a$ and $\nu_i = d_i^2/d_r^2$
  Our criterion BF$[\mathcal{M}_\gamma; \mathcal{M}_N]$= $m_\gamma(y)/m_N(y)$ becomes

$$\begin{cases} \left\{ \overline{\mathrm{sv}}[X_\gamma] \times \|\hat{\beta}^{MP}_{LSE}[\gamma]\| \right\}^{-n+1} & \text{if } q_\gamma \geq n-1 \\ \dfrac{d_{q_\gamma}^{q_\gamma}(1-R_\gamma^2)^{-\frac{n-q_\gamma}{2}+\frac{3}{4}} B(\frac{q_\gamma}{2}+\frac{1}{4}, \frac{n-q_\gamma}{2}-\frac{3}{4})}{\overline{\mathrm{sv}}[X_\gamma]^{q_\gamma}(1-R_\gamma^2+d_{q_\gamma}^2\|\hat{\beta}_{LSE}[\gamma]\|^2)^{\frac{1}{4}+\frac{q_\gamma}{2}} B(\frac{1}{4}, \frac{n-q_\gamma}{2}-\frac{3}{4})} \\ \quad \text{if } q_\gamma \leq n-2 \end{cases}$$

- It is exactly proportional to the posterior probability
- based on fundamental aggregated information of y and X_γ

Our recommendable BF

- $\hat{\beta}_{LSE}[\gamma]$: the normal LSE
- $\hat{\beta}^{MP}_{LSE}[\gamma]$: the LSE using the Moore-Pennrose inverse matrix of X_γ

$$\hat{\beta}^{MP}_{LSE}[\gamma] = \sum_{i=1}^{n-1} \frac{w_i[\gamma]u_i'[\gamma](y-\bar{y}1_n)}{d_i[\gamma]\|y-\bar{y}1_n\|} = \frac{X_\gamma^-(y-\bar{y}1_n)}{\|y-\bar{y}1_n\|}$$

- $\overline{\mathrm{sv}}[X_\gamma]$: the geometric mean of the singular values of X_γ

$$\overline{\mathrm{sv}}[X_\gamma] = \left\{\prod_{i=1}^{r} d_i[\gamma]\right\}^{1/r}$$

one of the most important scalar of design matrix X

Interpretation of many regressors case

- $\hat{\beta}^{MP}_{LSE}[\gamma]$: the minimizer of $\|\beta\|$ among the solutions

$$\text{of the equation } \frac{y - \bar{y}1_n}{\|y - \bar{y}1_n\|} = X_\gamma \beta$$

  under each submodel $\mathcal{M}_\gamma$
- $\|\hat{\beta}^{MP}_{LSE}[\gamma]\|$ itself is not comparable beyond the submodel
- $\overline{\mathrm{sv}}[X_\gamma] \times \|\hat{\beta}^{MP}_{LSE}[\gamma]\|$ is comparable
- the smallest $\overline{\mathrm{sv}}[X_\gamma] \times \|\hat{\beta}^{MP}_{LSE}[\gamma]\|$ means the best among the submodels $\mathcal{M}_\gamma$ which satisfies $q_\gamma \geq n - 1$

The estimation after selection

- In order to avoid the identifiability when $n < q$, we consider the estimator of $X\beta$

$$X\hat{\beta}_{BAYES} = \sum_{i=1}^{\min(q,n-1)} (u_i'v)u_i \left\{1 - \frac{E[(1+g)^{-1}|y]}{\nu_i}\right\}$$

$$X\hat{\beta}_{LSE} = \sum_{i=1}^{\min(q,n-1)} (u_i'v)u_i$$

- u_1: the normalized first principal component
- $\vdots \quad \vdots \quad \vdots \quad \vdots$
- $u_{\min(q,n-1)}$: the normalized last principal component

The estimation after selection

- The descending order $\nu_1 \geq \cdots \geq \nu_{\min(q,n-1)}$ is reasonable
- less important components get shrunk more!
- See Hastie, Friedman, Tibshirani's book.
- On the other hand, the original Zellner's g-prior cannot make such a reasonable effect

$$\left\{1 - E[(1+g)^{-1}|y]\right\} X\hat{\beta}_{LSE}$$

- This effect supports the descending order of ν

Model selection consistency

- the case where p is fixed and n is large
- Definition

$$\text{plim}_n p(\mathcal{M}_\gamma|y) = 1 \text{ if } \mathcal{M}_\gamma \text{ is the true model}$$

- A standard assumption: $\exists$ p.d. matrix H_γ s.t.

$$\lim \frac{1}{n} X'_\gamma X_\gamma = H_\gamma$$

- Our criterion has model selection consistency!

Numerical experiments

possible regressors $p = 16$

correlated case

$$\overbrace{x_1, x_2}^{\text{cor}=0.9}, \underbrace{x_3, x_4}_{\text{cor}=-0.7}, \overbrace{x_5, x_6}^{\text{cor}=0.5}, \underbrace{x_7, x_8}_{\text{cor}=-0.3} \sim N(0,1)$$

$$\overbrace{x_9, x_{10}}^{\text{cor}=0.1}, x_{11}, x_{12}, x_{13} \sim N(0,1), \; x_{14}, x_{15}, x_{16} \sim U(-1,1)$$

simple case $x_1, \ldots, x_{16} \sim N(0,1)$

Numerical experiments

$n = 30$ (hence so called $n > p$ case)

4 true models

$$Y = 1 + 2 \sum_{i \in \{\text{true}\}} x_i + \{\text{normal error term } N(0,1)\}$$

- full model ($q_T = 16$)
- $x_1, \ldots, x_{10}, x_{11}, x_{14}$ ($q_T = 12$)
- $x_1, x_2, x_5, x_6, x_9, x_{10}, x_{11}, x_{14}$ ($q_T = 8$)
- x_1, x_2, x_5, x_6 ($q_T = 4$)

Numerical experiments

competitors of our BF

$$\mathrm{AIC} = -2 \times \text{max. log likelihood} + 2(q+2)$$

$$\mathrm{AICc} = -2 \times \text{max. log likelihood} + 2(q+2)\frac{n}{n-q-3}$$

$$\mathrm{BIC} = -2 \times \text{max. log likelihood} + q \log n$$

ZE: $\mathrm{BF}[\mathcal{M}_\gamma; \mathcal{M}_N]$ with $a = -3/4$, $\nu_1 = \cdots = \nu_q = 1$
(the effect of descending order ν)

EB: empirical Bayes criterion: George and Foster (2000)

$$\max_g m_\gamma(y|g, \hat{\sigma}^2) \quad \hat{\sigma}^2 = \mathrm{RSS}/(n-q-1)$$

(the effect of full Bayes)

$N = 500$ bigger is better

		cor	simple		cor	simple
BF		0.71	0.98		0.73	0.86
ZE		0.40	0.94		0.63	0.87
EB	**16**	0.41	0.95	**12**	0.63	0.87
AIC		0.95	1.00		0.23	0.22
AICc		0.25	0.82		0.67	0.85
BIC		0.88	0.99		0.41	0.41
BF		0.69	0.77		0.66	0.68
ZE		0.68	0.78		0.67	0.69
EB	**8**	0.67	0.76	**4**	0.66	0.65
AIC		0.09	0.08		0.05	0.05
AICc		0.52	0.55		0.25	0.24
BIC		0.31	0.27		0.23	0.22

Table: Frequency of the top of the true model

Numerical experiments (findings)

- [correlated and simple] AIC and BIC are too bad for all except $q_T = 16$.
- [correlated and simple] AICc is bad for $q_T = 16$ and 4 while it is good for $q_T = 8, 12$.
- [simple] BF, ZE and EB are very similar. There is no effect of the extention of Zellner's g-prior with descending ν.
- [correlated] EB, ZE and BF are very similar for $q_T = 4, 8$, but BF is much better for $q = 12, 16$.

In summary, our BF is the best for most case and extremely stable. The extention of Zellner's g-prior with descending ν is quite effective.

Numerical experiments

(in-sample) predictive error of selected model

$$\frac{(\hat{y}_* - \alpha_T 1_n - X_T\beta_T)'(\hat{y}_* - \alpha_T 1_n - X_T\beta_T)}{n\sigma^2}$$

- X_T, α_T, β_T are true
- $\hat{y}_*$: $\bar{y}1_n + X_{\gamma*}\hat{\beta}_{\gamma*}$, $X_{\gamma*}$: selected
- $\hat{\beta}_{\gamma*}$: selected Bayes estimator in BC, ZE, EB
- $\hat{\beta}_{\gamma*}$: selected LSE in AIC, BIC, AICc

$N = 500$ smaller is better

		cor	simple		cor	simple
oracle		17/30(≃0.57)	17/30		13/30(≃0.43)	13/30
BF		0.70	0.57		0.52	0.45
ZE		1.02	0.66		0.59	0.45
EB	**16**	1.00	0.65	**12**	0.58	0.45
AIC		0.56	0.56		0.54	0.54
AICc		1.29	0.98		0.56	0.46
BIC		0.58	0.56		0.53	0.52
oracle		9/30(=0.3)	0.30		5/30(≃0.17)	0.17
BF		0.37	0.35		0.26	0.25
ZE		0.41	0.34		0.27	0.24
EB	**8**	0.41	0.35	**4**	0.27	0.25
AIC		0.51	0.51		0.48	0.48
AICc		0.42	0.39		0.36	0.35
BIC		0.46	0.45		0.39	0.38

Table: The in-sample predictive error (mean)

Numerical experiments

- 14 true regressors $x_1, x_2, \ldots, x_{10}, x_{11}, x_{12}, x_{14}, x_{15}$
- $n = 12 \Rightarrow n < q_T < p$ case
- non-identifiable model is true
- there is no competitors in ZE, EB, AIC, BIC, AICc
- The true model could not get the top at all

frequency of number of regressors of the selected model: identifiable model is always selected

	0-7	8-9	10-11	12-16
correlated	0.21	0.56	0.23	0
simple	0.26	0.54	0.20	0

Numerical experiments

the frequency of each regressors of the selected model among $N = 500$.

	x_1 (T)	x_2 (T)	x_3 (T)	x_4 (T)	x_5 (T)	x_6 (T)
correlated	0.67	0.61	0.43	0.47	0.63	0.59
simple	0.54	0.54	0.54	0.54	0.54	0.57
	x_7 (T)	x_8 (T)	x_9 (T)	x_{10} (T)	x_{11} (T)	x_{12} (T)
correlated	0.56	0.56	0.59	0.58	0.58	0.60
simple	0.55	0.55	0.54	0.56	0.52	0.50
	x_{13} (F)	x_{14} (T)	x_{15} (T)	x_{16} (F)		
correlated	0.40	0.41	0.47	0.40		
simple	0.34	0.54	0.58	0.39		

- averagely the true variables are selected more often

Where is the true model?

- the average of rank of each sub-models
- the true model is the top with respect to the average of ranks both in correlated case and in simple structure case
- (the average of rank of the true model)$/2^{16}$ is about 0.03
- Although our criterion has an ability to find a true model averagely, a smaller identifiable model is selected as the best

Where is the true model?

- The frequency of the true model among $(16 \times 15)/2 = 120$ candidates whose number of regressors is 14

	1st	1st-2nd	1st-3rd
correlated	0.14	0.22	0.26
simple	0.13	0.20	0.26

- Not bad!! If the true number of regressors is given, the analytical criterion $\overline{\mathsf{sv}}[X_\gamma] \times \|\hat{\beta}_{LSE}^{MP}[\gamma]\|$ works
- To our knowledge, there was no analytical criterion which is available when the number of regressors are the same and $R^2 = 1$.

Numerical experiment (findings)

- We assumed equal model space prior probability $\Pr(\mathcal{M}_\gamma) = 2^{-p}$
- Under the equal model space prior probability, the submodel which has identifiability is selected.
- When the larger (non-identifiable, non-sparse) model is expected, unequal model space prior probability may lead a choice of such a non-sparce reasonable sub-model
- $\Pr(\mathcal{M}_\gamma) = w^{q_\gamma}(1-w)^{p-q_\gamma}$
- $\Pr(\mathcal{M}_\gamma) \propto B(\alpha + q_\gamma, \beta + p - q_\gamma)$
- We just started considering this issue,,,

Summary and Future work

Summary

- BF with a beautiful closed form
- consistency for large n and fixed p
- very good numerical performance when $n > p$
- reasonable estimator of $X\beta$ after selection

Future Work

- find a reasonable unequal model space prior probability
- Comparison with some famous methods including elastic-net

FYI

The older version of our paper is in Arxiv.