Bayesian Model Selection in High-Dimensional Settings

Valen E. JOHNSON and David ROSSELL

Kyunghee Kim and Sanghyun Lee

Department of Statistics

Sungkyunkwan University

November 30, 2021

Overview

1. Introduction
2. Main result
3. Computational Strategies
4. Simulation Studies
5. Disscussion

Introduction

Model Selection Procedures

- Local prior
- Frequentist methods
- Nonlocal prior

Local prior

- **Local prior**

  e.g. g-priors(Liang et al., 2008), intrinsic Bayes factors(Berger and Pericchi,1996), fractional Bayes factors(O'Hagan, 1995).

  Most current Bayesian model selection procedures employ this prior.

  Positive at null parameter values.

  Assign a posterior probability of 0 to the true model(Theorem 2).

Frequentist method

- **Frequentist method**

  e.g. SCAD(Fan and Li, 2001), adaptive Lasso(Zou, 2006),

  Elastic net algorithm(Zou and Hastie, 2005), and Dantzig selector(Candes and Tao, 2007)

  Fan and Peng(2004) showed a consistency property that identify the correct model to certain penalized-likelihood-based model selection procedure; where $p < O(n^{\frac{1}{3}})$

Nonlocal prior

- **Nonlocal prior**

  Identically zero whenever a model parameter is equal to its null value.

  Nonlocal prior's model selection have a consistency property which assign a posterior probability of 1 to the true model as the sample size n increases, $p = O(n)$.

  - Especially in **large sample settings**, model selection based on nonlocal prior densities are often better able to identify the correct model and have smaller prediction errors.

Nonlocal prior

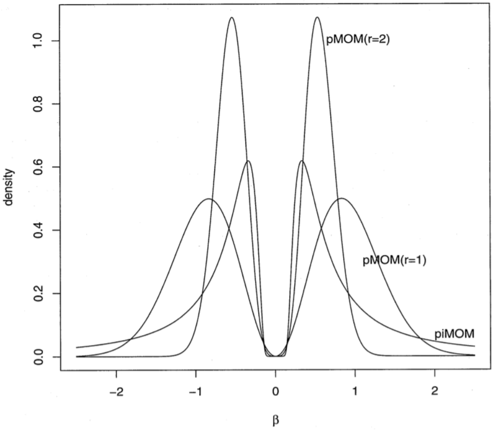


Figure 1. Nonlocal prior densities for a single regression coefficient. These densities correspond to the default nonlocal priors used in the simulation study in Section 4.

Consistency

- **Consistency**

  Consistency in this paper $\neq$ Pairwise consistency

Examples

Intrinsic Bayes factors become unbounded as p=O(n), and BIC p=O(n^α), $\alpha < 1$(Moreno, Giron and Casella, 2010).

Pairwise consistency

Bayes factor between the true model and any *single* model $k \in J$ as n increases.It is much weaker property than model consistency since it's possible to achive when the posterior probability of the true model approaches 0.

Nonlocal prior

- MOM and iMOM prior(Johnson and Rossel,2010).

  0 only when all components of the parameter vector are 0.

  1. MOM prior.

  $$\pi_M(\theta) = \frac{(\theta - \theta_0)^{2k}}{\tau_k}\pi_b(\theta),\ \tau_k = \int_\theta (\theta - \theta_0)^{2k}\pi_b(\theta)d\theta$$

  2. iMOM prior.

  $$\pi_I(\theta) = \frac{k\tau^{\nu/2}}{\Gamma(\nu/2k)}(\theta - \theta_0)^{2^{-(\nu+1)/2}} exp\Big[-\{\frac{(\theta-\theta_0)^2}{\tau}\}^{-k}\Big]$$

Nonlocal prior

We will put much stronger penalty on the regression vector.

1. pMOM prior (Product case of MOM prior).

$$\pi(\boldsymbol{\beta}|\tau,\sigma^2,r) = d_p(2\pi)^{-p/2}(\tau\sigma^2)^{-rp-p/2}|\mathbf{A}_p|^{\frac{1}{2}} \exp\left[-\frac{1}{2\tau\sigma^2}\boldsymbol{\beta}^{\boldsymbol{t}}\mathbf{A}_p\boldsymbol{\beta}\right]\prod_{i=1}^{p}\beta_i^{2r}$$

2. piMOM prior (Product case of iMOM prior).

$$\pi(\boldsymbol{\beta}|\tau,\sigma^2,r) = \frac{(\tau\sigma^2)^{rp/2}}{\Gamma(r/2)^p}\prod_{i=1}^{p}|\beta_i|^{-(r+1)}\exp\left(-\frac{\tau\sigma^2}{\beta_i^2}\right)$$

for $\tau > 0$, $\mathbf{A}_p$ a $p \times p$ nonsingular scale matrix, and r = 1,2,....

Main result

- Goal: To select the nonzero components of $\boldsymbol{\beta}$ from $\mathbf{Y}_n \sim N(\mathbf{X}_n\boldsymbol{\beta}, \sigma^2\mathbf{I}_n),\ p < n$.

  Sampling density is assumed to be $\mathbf{Y}_k|\boldsymbol{\beta}_k, \sigma^2 \sim N(\mathbf{X}_k\boldsymbol{\beta}_k, \sigma^2\mathbf{I}_k)$.

  Denote $\mathbf{t}$ the true model.

Theorem 1.	pMOM prior satisfies $p(\mathbf{t}|\mathbf{y}_n) \xrightarrow{\text{p}} 1$. with known σ^2
Corollary1.	pMOM prior satisfies $p(\mathbf{t}|\mathbf{y}_n) \xrightarrow{\text{p}} 1$. with unknown σ^2
Corollary2.	Assume the conditions of Corollary 1 hold with piMOM prior.
Theorem 2.	Local prior satisfies $p(\mathbf{t}|\mathbf{y}_n) \xrightarrow{\text{p}} 0$..

Table: Theorems for consistency

Theorem 1

- Theorem 1.

$$
\begin{aligned}
p(\mathbf{t}|\mathbf{y}_n) &= \frac{p(\mathbf{t})m_{\mathbf{t}}(\mathbf{y}_n)}{\sum_{\mathbf{k}\in J} p(\mathbf{k})m_{\mathbf{k}}(\mathbf{y}_n)} \\
&= \frac{p(\mathbf{t})m_{\mathbf{t}}(\mathbf{y}_n)}{\sum_{\mathbf{t}\subset\mathbf{k}} p(\mathbf{k})m_{\mathbf{k}}(\mathbf{y}_n) + \sum_{\mathbf{t}\not\subset\mathbf{k}} p(\mathbf{k})m_{\mathbf{k}}(\mathbf{y}_n) + p(\mathbf{t})m_{\mathbf{t}}(\mathbf{y}_n)} \\
&= \left[\sum_{\mathbf{t}\subset\mathbf{k}} \frac{p(\mathbf{k})m_{\mathbf{k}}(\mathbf{y}_n)}{p(\mathbf{t})m_{\mathbf{t}}(\mathbf{y}_n)} + \sum_{\mathbf{t}\not\subset\mathbf{k}} \frac{p(\mathbf{k})m_{\mathbf{k}}(\mathbf{y}_n)}{p(\mathbf{t})m_{\mathbf{t}}(\mathbf{y}_n)} + 1\right]^{-1} \xrightarrow{\text{p}} 1, r \geq 2.
\end{aligned}
$$

Where, $m_{\mathbf{k}}(\mathbf{y}_n) = d_k(2\pi)^{-n/2}\tau^{-k/2-rk}(\sigma^2)^{n/2-rk}[\frac{|A_k|}{|C_k|}]^{1/2}exp[-\frac{R_k}{2\sigma^2}]E_k(\prod_{i=1}^{k}\beta_{k_i}^{2r})$
$\frac{m_{\mathbf{k}}(\mathbf{y}_n)}{m_{\mathbf{t}}(\mathbf{y}_n)} \xrightarrow{\text{p}} exp(-\infty) \xrightarrow{\text{p}} 0$. where, $R_k = y_n^t(I_n - X_k C_k^{-1} X_k^t)y_n$ and $C_k = X_k^t X_k + \frac{1}{\tau}A_k$.

Theorem 1 (Corollary1, Corollary2)

- Corollary 1.

  Consider the case σ is unknown: Set $\sigma^2 \sim IG(\alpha, \psi)$

  Then,

  $$m_{\mathbf{k}}(\mathbf{y}_n) = d_k (2\pi)^{-n/2} 2^{\nu/2} \tau^{-rk-k/2} \left[\frac{|A_k|}{|C_k|}\right]^{1/2} \frac{\psi^{\alpha}}{\Gamma(\alpha)} (\nu_k s_k^2)^{-\nu_k/2} \Gamma(\tfrac{\nu_k}{2}) E_k^T(\textstyle\prod_{i \in k} \beta_i^{2r}),\ \ r \geq 2$$

- Corollary 2.

  Consider the case piMOM prior.

  Established from the consistency of pMOM priors under similar conditions, $\mathbf{A}_k = \mathbf{I}_k$.

Theorem 2

- Theorem 2.

$$p\left[\frac{m_{\mathbf{k}}(\mathbf{y}_n)}{m_{\mathbf{t}}(\mathbf{y}_n)} > n^{1/2}(\frac{c}{M})^{t/2}\sqrt{\frac{2\pi\sigma^2}{M}}c_L\right] \xrightarrow{\text{a.s.}} 1, \frac{\pi_k^L(\gamma_k)}{\pi_t^L(\beta_t)} \geq c_L > 0.$$

$$p(\mathbf{t}|\mathbf{y}_n) = \frac{p(\mathbf{t})m_{\mathbf{t}}(\mathbf{y}_n)}{\sum_{\mathbf{k}\in J} p(\mathbf{k})m_{\mathbf{k}}(\mathbf{y}_n)}$$

Posterior probability of the true model goes 0 whenever the following conditions apply

- Local prior densities are imposed
- The number of possible covariates is greater than $O(\sqrt{n})$.
- The relative prior probabilities assigned to all models are bounded away from 0.

Simulation Interests and Challenging

- We are interested in not only identifying the most probable model, but also assessing its probability and the probability of any other high-probability models.
    - i.e. We need to evaluate $p(\mathbf{t}|\mathbf{y}_n) = \frac{p(\mathbf{t})m_{\mathbf{t}}(\mathbf{y}_n)}{\sum_{\mathbf{k}\in J} p(\mathbf{k})m_{\mathbf{k}}(\mathbf{y}_n)}$.
- Problem : Evaluating a marginal density $m_{\mathbf{t}}(\mathbf{y}_n) = \int f(\mathbf{y}_n|\boldsymbol{\beta}_{\mathbf{t}})\pi(\boldsymbol{\beta}_{\mathbf{t}})d\boldsymbol{\beta}_{\mathbf{t}}$ may require numerical evaluation of a potentially high-dimensional integral.
    - Even piMOM has no analytic expressions.
- Solution : Laplace Approximation

Laplace Approximation

- Tierney and Kadane (1986) :

$$
\begin{aligned}
E(g(\theta)|y) &= \int g(\theta)f(y|\theta)\pi(\theta)d\theta \\
&= \int g(\theta)e^{l(\theta)}\pi(\theta)d\theta \qquad \text{where} \quad l(\theta) = \log f(y|\theta) \\
&= \int e^{n\cdot\frac{1}{n}\{\log g(\theta)+\log\pi(\theta)+l(\theta)\}}d\theta \\
&= \int e^{nL(\theta)}d\theta \qquad \text{where} \quad L(\theta) = \frac{1}{n}\{\log g(\theta) + \log\pi(\theta) + l(\theta)\}
\end{aligned}
$$

Here, let $\hat{\theta}$ is the mode of $L(\theta)$ and Σ is the inverse of Hessian matrix.

Laplace Approximation

- Tierney and Kadane (1986) : Cont'd

$$\int e^{nL(\theta)} d\theta = \int \exp\left[nL(\hat{\theta}) - \frac{n}{2}(\theta - \hat{\theta})^T \Sigma^{-1}(\theta - \hat{\theta})\right] d\theta \qquad \text{by Taylor's Expansion}$$

$$= exp\left[nL(\hat{\theta})\right] \cdot \left(\sqrt{\frac{2\pi}{n}}\right)^p \left(\det \Sigma\right)^{\frac{1}{2}}$$

$\hat{\theta}$: Newton-Raphson method by setting $\theta_{n+1} = \theta_n - \dfrac{L^{'}(\theta_n)}{L^{''}(\theta_n)}$

- More details in Luke Tierney and Joseph B.Kadane (1986)

Laplace Approximation marginal density for pMOM

- marginal density for pMOM :

$$m_{\mathbf{k}}(\mathbf{y}_n) = \frac{\Gamma(\frac{\nu_k}{2})\psi^{\alpha}2^{\frac{\nu_k}{2}}(2\psi + \mathbf{y}^T\mathbf{y} - \tilde{\boldsymbol{\beta}}_k^T\mathbf{C}_k\tilde{\boldsymbol{\beta}}_k)^{-\frac{\nu_k}{2}}}{\Gamma(\alpha)\left[(2r-1)!!\right]^k(2\pi)^{\frac{n}{2}}\tau^{\frac{k}{2}+rk}}$$

$$\times\ \frac{(\prod_{i\in k}(\beta_i^*)^{2r})exp\Big\{-\frac{\nu_k-2}{2\nu_k}(\boldsymbol{\beta}_k^* - \tilde{\boldsymbol{\beta}}_k)^T\frac{\mathbf{C}_k}{s_k^2}(\boldsymbol{\beta}_k^* - \tilde{\boldsymbol{\beta}}_k)\Big\}}{|\mathbf{C}_k + 2r\frac{\nu_k s_k^2}{(\nu_k-2)}D(\tilde{\boldsymbol{\beta}}_k^*)|^{\frac{1}{2}}}$$

where $D(\boldsymbol{\beta}_k^*)$ is the diagonal matrix with entry (i,j) given by $1/(\boldsymbol{\beta}_k^*)^2$

and $\boldsymbol{\beta}_k^* = \text{argmax}_{\boldsymbol{\beta}_k}\left\{N\left(\boldsymbol{\beta}_{\mathbf{k}};\tilde{\boldsymbol{\beta}}_k,\frac{\nu_k}{\nu_k-2}s_k^2\mathbf{C}_k^{-1}\right)\prod_{i\in k}\beta_i^{2r}\right\}$

Laplace Approximation marginal density for piMOM

- marginal density for piMOM :

$$m_{\mathbf{k}}(\mathbf{y}_n) = \frac{\psi^{\alpha}(2\tau)^{\frac{k}{2}}}{(2\pi)^{\frac{n}{2}}\Gamma(\alpha)} \frac{e^{f(\boldsymbol{\beta}_k^*,\eta^*)}}{|V(\boldsymbol{\beta}_k^*,\eta^*)|^{\frac{1}{2}}}, \quad \text{where} \quad (\boldsymbol{\beta}_k^*, \eta^*) = \text{argmax}_{(\boldsymbol{\beta}_k,\eta)} f(\boldsymbol{\beta}_k, \eta), \ \eta = \log(\sigma^2).$$

$$f(\boldsymbol{\beta}_k, \eta) = -\frac{2\psi + (\mathbf{y}_n - \mathbf{X}_k\boldsymbol{\beta}_k)^T\mathbf{X}_k^T\mathbf{X}_k(\mathbf{y}_n - \mathbf{X}_k\boldsymbol{\beta}_k)}{2e^{\eta}} - \frac{\eta(n-k+2\alpha)}{2} - \sum_{i\in k} \frac{\tau e^{\eta}}{\beta_i^2} + \log(\beta_i^2),$$

and $V(\boldsymbol{\beta}_k, \eta)$ is a $(k+1) \times (k+1)$ matrix with the following blocks :

$$V_{11} = -e^{-\eta}\mathbf{X}_k^T\mathbf{X}_k - \text{diag}(6\tau e^{\eta}\boldsymbol{\beta}_k^{-4} - 2\boldsymbol{\beta}_k^{-2}),$$

$$V_{12} = \frac{2\tau e^{\eta}}{\boldsymbol{\beta}_k^3} + e^{-\eta}(\mathbf{X}_k^T\mathbf{X}_k\boldsymbol{\beta}_k - \mathbf{X}_k^T\mathbf{y}_n),$$

$$V_{22} = -\frac{2\psi + (\mathbf{y}_n - \mathbf{X}_k\boldsymbol{\beta}_k)^T\mathbf{X}_k^T\mathbf{X}_k(\mathbf{y}_n - \mathbf{X}_k\boldsymbol{\beta}_k)}{2e^{-\eta}} - \sum_{i\in k} \frac{\tau e^{\eta}}{\beta_i^2}$$

MCMC Scheme

1. Choose an initial model $\mathbf{k}^{curr}$
2. For $i = 1, \ldots, p$,
    2.1 Define model $\mathbf{k}^{cand}$ by excluding or including β_i from model $\mathbf{k}^{curr}$, according to whether β_i is currently included or excluded from $\mathbf{k}^{curr}$.

    2.2 Compute

    $$r = \frac{m_{\mathbf{k}^{cand}}(y)p(\mathbf{k}^{cand})}{m_{\mathbf{k}^{cand}}(y)p(\mathbf{k}^{cand}) + m_{\mathbf{k}^{curr}}(y)p(\mathbf{k}^{curr})}$$

    2.3 Draw $u \sim U(0,1)$. If $r > u$, define $\mathbf{k}^{curr} = \mathbf{k}^{cand}$
3. Repeat step 2 until a sufficiently long chain is acquired.

Variable Selection Prior

- Scott and Berger(2010) : (a fully Bayesian method)

$$p(\mathbf{k}|\gamma) = \gamma^k (1-\gamma)^{n-k}, \quad \gamma \sim \text{Beta}(\zeta_0, \zeta_1).$$

$$\text{Then, } p(\mathbf{k}) = \int_0^1 p(\mathbf{k}|\gamma)\pi(\gamma)d\gamma$$

$$\text{In general, set } \zeta_0 = \zeta_1 = 1, \text{ then } p(\mathbf{k}) = \frac{1}{p+1}\binom{p}{k}^{-1}$$

Setting Priors and some parameters

- revisited :

  1. pMOM :

$$\pi(\boldsymbol{\beta}|\tau,\sigma^2,r) = d_p(2\pi)^{-p/2}(\tau\sigma^2)^{-rp-p/2}|\mathbf{A}_p|^{1/2}\exp\left[-\frac{1}{2\tau\sigma^2}\boldsymbol{\beta}^T\mathbf{A}_p\boldsymbol{\beta}\right]\prod_{i=1}^{p}\beta_i^{2r}$$

  2. piMOM :

$$\pi(\boldsymbol{\beta}|\tau,\sigma^2,r) = \frac{(\tau\sigma^2)^{rp/2}}{\Gamma(r/2)^p}\prod_{i=1}^{p}|\beta_i|^{-(r+1)}exp\left(-\frac{\tau\sigma^2}{\beta_i^2}\right)$$

- Need to specify : $\tau,\ \sigma^2,\ \mathbf{A}_p,\ r$ (Prior for $\boldsymbol{\beta}$ is already prepared.)

Meaning of τ

- revisited :

  1. MOM prior.

  $$\pi_M(\theta) = \frac{(\theta-\theta_0)^{2k}}{\tau_k}\pi_b(\theta),\ \tau_k = \int_\theta (\theta-\theta_0)^{2k}\pi_b(\theta)d\theta$$

     - base density $\pi_b(\beta_i)$: $\beta_i \sim N(0, \sigma^2\tau\lambda_i)$

  2. iMOM prior.

  $$\pi_I(\theta) = \frac{k\tau^{\nu/2}}{\Gamma(\nu/2k)}(\theta-\theta_0)^{2^{-(\nu+1)/2}} exp\Big[-\{\frac{(\theta-\theta_0)^2}{\tau}\}^{-k}\Big]$$

     - This have functional forms that are related to inverse gamma density functions, which means that their behaviour near θ_0 is similar to the behaviour of an inverse gamma density near 0.

Setting τ

- Valen E. Johnson (2010) :
  - pMOM(r=1) : $\tau = 0.348$
  - pMOM(r=2) : $\tau = 0.072$
  - piMOM : $\tau = 0.113$
  - At these values, the nonlocal priors assign 0.99 marginal prior probability to $|\beta_i| \geq 0.2\sigma$.
- In actual applications, the choice of τ should be determined after a subjective evaluation of the magnitude of substantively important effect sizes.

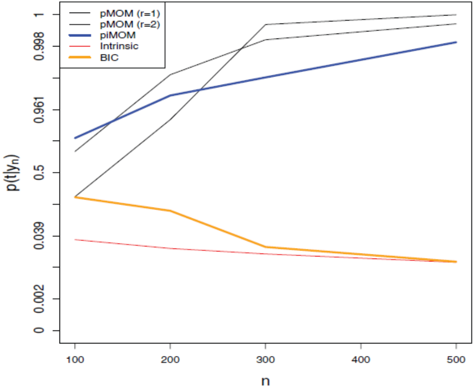
Setting σ^2, $\mathbf{A}_p$ and r

- $\sigma^2 \sim$ IG(0.001,0.001)
  - It is a non-informative prior.
  - It is known that posterior model probabilities are insensitive to the choice of the inverse-gamma density parameters which are both much smaller than 1.
- $\mathbf{A}_p = \mathbf{I}_p$
  - Actually, this is for computational advantages when using Laplace approximation.
- pMOM(r=1), pMOM(r=2), piMOM
  - pMOM(r=1) are not always guaranteed to provide consistent model selection under the assumptions of Theorem 1. However, this often leads to better finite sample properties.
  - piMOM does not have *r* since it canceled out after Laplace approximation.

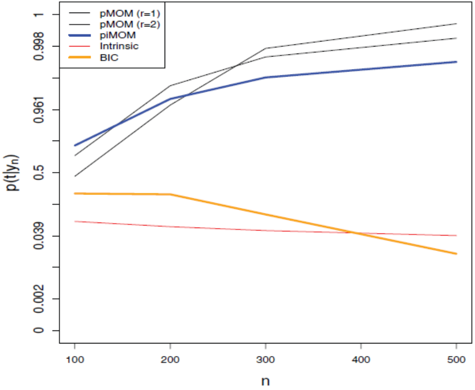
Comparison of Bayesian Model Selection $p(\mathbf{t}|\mathbf{y}_n)$

- $\sigma^2 = 1$

$\rho = 0$

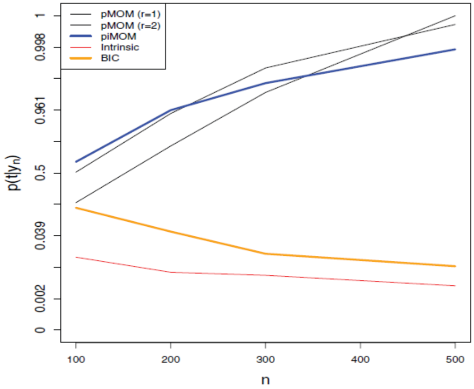


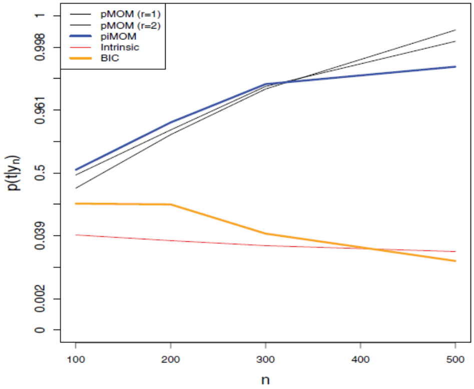
$\rho = 0.25$



Comparison of Bayesian Model Selection $p(\mathbf{t}|\mathbf{y}_n)$

- $\sigma^2 = 1.5$





Comparison of Bayesian Model Selection $p(\mathbf{t}|\mathbf{y}_n)$

- $\sigma^2 = 2$

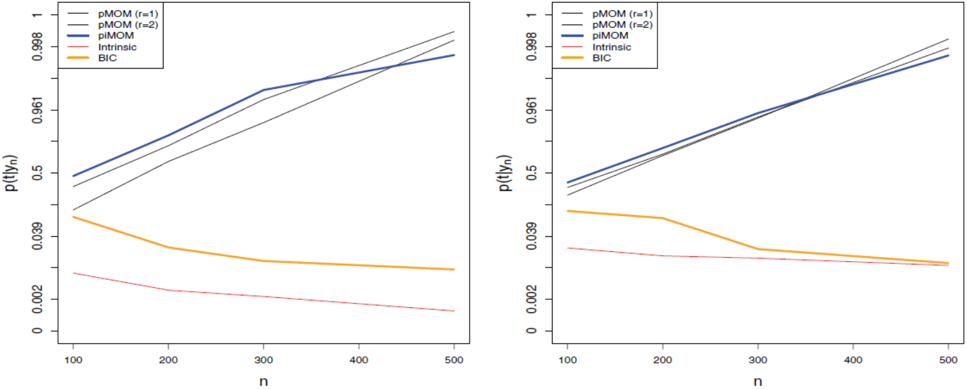
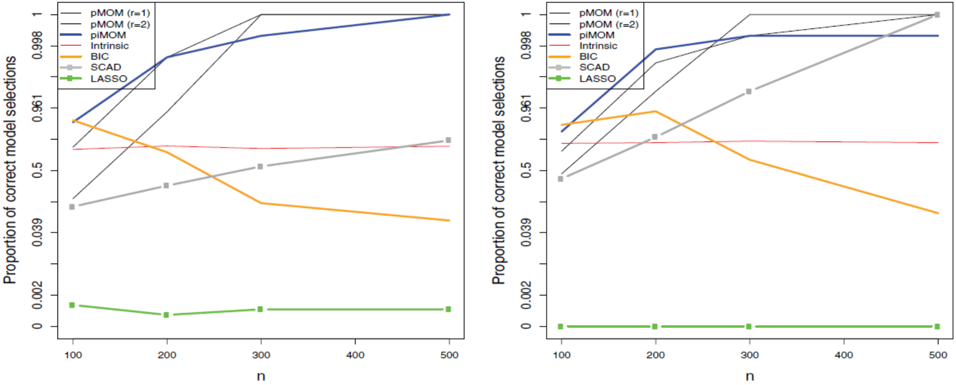


Figure 2. $p(\mathbf{t}|\mathbf{y}_n)$ versus n. Top: $\sigma^2 = 1$; middle: $\sigma^2 = 1.5$; bottom: $\sigma^2 = 2$.

Comparison to Penalized Likelihood Selection $p(\hat{\mathbf{t}} = \mathbf{t})$

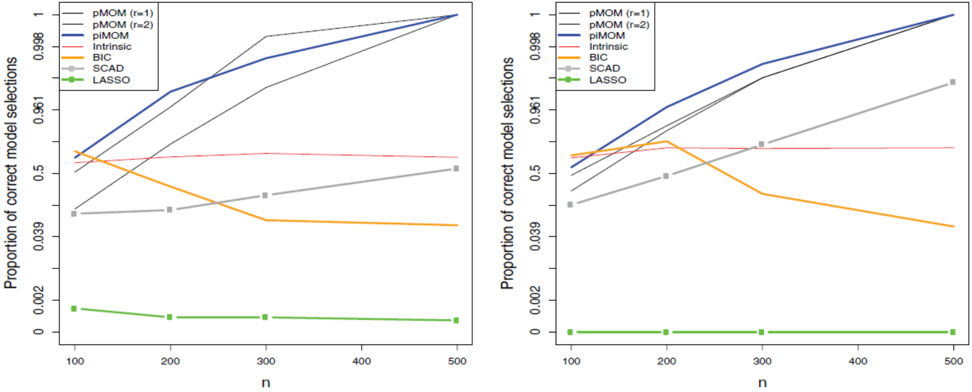
- $\sigma^2 = 1$



$\rho = 0$ $\rho = 0.25$

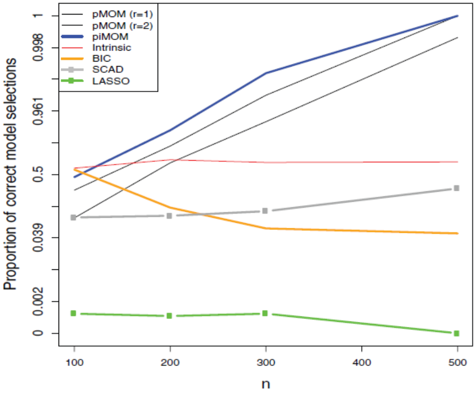
Comparison to Penalized Likelihood Selection $p(\hat{\mathbf{t}} = \mathbf{t})$

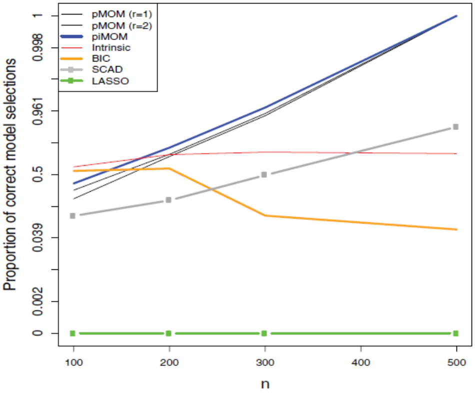
- $\sigma^2 = 1.5$



Comparison to Penalized Likelihood Selection $p(\hat{\mathbf{t}} = \mathbf{t})$

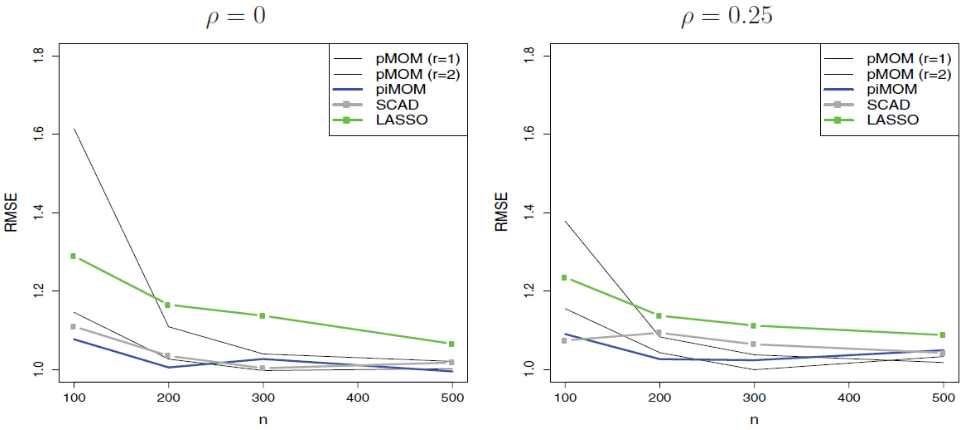
- $\sigma^2 = 2$





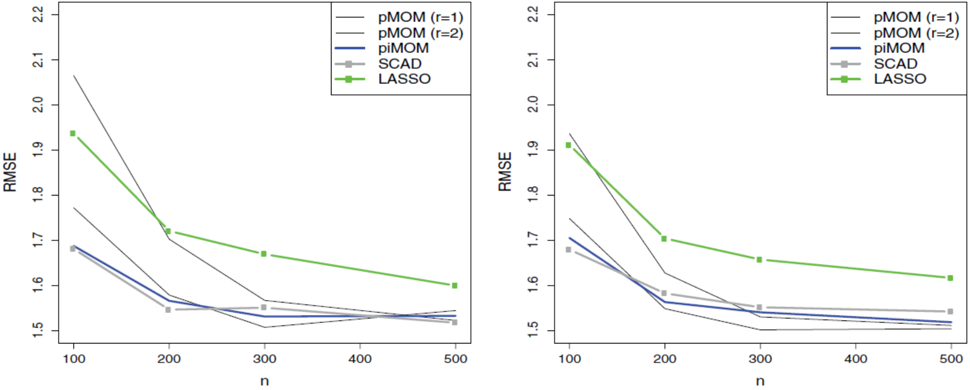
Comparison to Penalized Likelihood Selection : RMSE

- $\sigma^2 = 1$



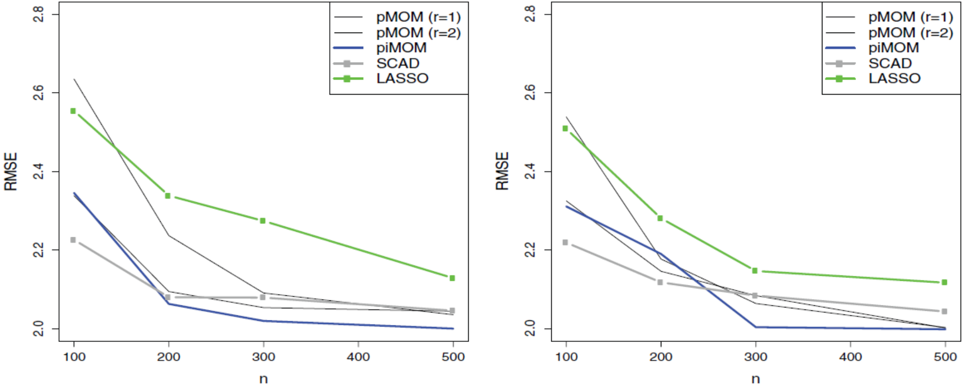
Comparison to Penalized Likelihood Selection : RMSE

- $\sigma^2 = 1.5$



Comparison to Penalized Likelihood Selection : RMSE

- $\sigma^2 = 2$



Summary of results

- remarks :
  - model consistency
    1. Nonlocal priors have shown the model consistency, even pMOM(r=1).
  - true model probability
    1. $p(\hat{\mathbf{t}} = \mathbf{t})$ went to 1 when using nonlocal priors, whereas the others stayed around 0.5, especially LASSO showed poor performance.
  - RMSE
    1. Typically, piMOM has slightly smaller RMSE than SCAD.
    2. When $n \geq 200$, pMOM(r=1) outperformed SCAD.
    3. The pMOM(r=2) was generally not competitive with any procedure except the LASSO.
- reasons :
  1. piMOM has the heaviest tails and so the smallest biases.
  2. pMOM(r=2) has the lighter tails at larger values of $\boldsymbol{\beta}$.

Limitation of this paper

1. In MCMC step, the truly nonzero regression coefficients were the last variables to be considered for inclusion in the initial model to avoid bias in the initial updates of the chain toward the true model. In practice, we don't know what the truly nonzero coefficients.
2. When using pMOM(r=1), this had only practical consistency in finite sample size, not theoritical consistency, but showed low RMSE. In contrast, pMOM(r=2) satisfied theoretical model consistency, but had large RMSE.

Thank you