Bayesian model selection: prior distributions with philosophical, theoretical and practical appeal

David Rossell*, IRB Barcelona
Valen E. Johnson, M.D. Anderson Cancer Center

February 16, 2012

Statistics in the 21^{st} century

- From small p, moderate n to large p, small n
- High-dimensional applications usual: bioinformatics, image processing, finance, internet etc.
- Key to success: model parsimony

Goal: model selection, *i.e.* use $\mathbf{y}$ to guess M
($\mathbf{y}$: data, M: model that generated $\mathbf{y}$, θ: model parameters)

Bayesian approach

- Likelihood: $\mathbf{y}|\boldsymbol{\theta}, M \sim f(\mathbf{y}|\boldsymbol{\theta}, M)$
- Prior on model parameters: $\boldsymbol{\theta}|M \sim P(\boldsymbol{\theta}|M)$
- Prior on model space: $M \sim P(M)$

Posterior distribution: $P(M|\mathbf{y}) \propto m(\mathbf{y}|M)P(M)$

Bayesian approach

- Likelihood: $\mathbf{y}|\boldsymbol{\theta}, M \sim f(\mathbf{y}|\boldsymbol{\theta}, M)$
- Prior on model parameters: $\boldsymbol{\theta}|M \sim P(\boldsymbol{\theta}|M)$
- Prior on model space: $M \sim P(M)$

Posterior distribution: $P(M|\mathbf{y}) \propto m(\mathbf{y}|M)P(M)$

- $m(\mathbf{y}|M)$ is the integrated likelihood

$$m(\mathbf{y}|M) = \int f(\mathbf{y}|\boldsymbol{\theta}, M) dP(\boldsymbol{\theta}|M)$$

- Models M_1 & M_2 compared via $\text{BF}_{12} = \frac{m(\mathbf{y}|M_1)}{m(\mathbf{y}|M_2)}$

Intuition

1. $P(M)$ can favor parsimony (Scott & Berger, Ann Stat 2010)
2. $m(\mathbf{y}|M)$ "automatically" favors parsimony. Is it enough?

Intuition

1. $P(M)$ can favor parsimony (Scott & Berger, Ann Stat 2010)
2. $m(\mathbf{y}|M)$ "automatically" favors parsimony. Is it enough?

Example: $y_1, \ldots, y_n \sim N(\theta, 1)$. Test $\mathrm{H}_0 : \theta = 0$ vs $\mathrm{H}_1 : \theta \neq 0$.

- Under H_0: $\theta = 0$
- Under H_1: $\theta \sim N(0, \tau)$. By default set $\tau = 1$.
- $P(\mathrm{H}_0) = P(\mathrm{H}_1) = \frac{1}{2}$.

Simulate data for $\theta = 0.5$, $n = 100$.

- $\bar{y} = 0.49$, $P(\mathrm{H}_1|\mathbf{y}) = 0.99997$.
- $\bar{y} = 0.52$, $P(\mathrm{H}_1|\mathbf{y}) = 0.99992$.
- ...

Simulate data for $\theta = 0.5$, $n = 100$.

- $\bar{y} = 0.49$, $P(\mathrm{H}_1|\mathbf{y}) = 0.99997$.
- $\bar{y} = 0.52$, $P(\mathrm{H}_1|\mathbf{y}) = 0.99992$.
- ...

Simulate data for $\theta = 0$, $n = 100$.

- $\bar{y} = 0.01$, $P(\mathrm{H}_1|\mathbf{y}) = 0.091$.
- $\bar{y} = 0.07$, $P(\mathrm{H}_1|\mathbf{y}) = 0.113$.
- ...

Simpler model not enforced enough

Example: linear regression. BIC to find non-zero coef.

- Asymp. equiv to $P(\boldsymbol{\theta}|M) = N(\mathbf{0}, n\sigma^2(X_M'X_M)^{-1})$, $P(M) \propto 1$
- Complexity penalty is $p \times \log(n)$
- Finds true model as $n \longrightarrow \infty$, p fixed (O'Hagan & Forster, Bay Infer 2004)
- Can fail when p grows with n (Casella et al, Ann Stat 2009)

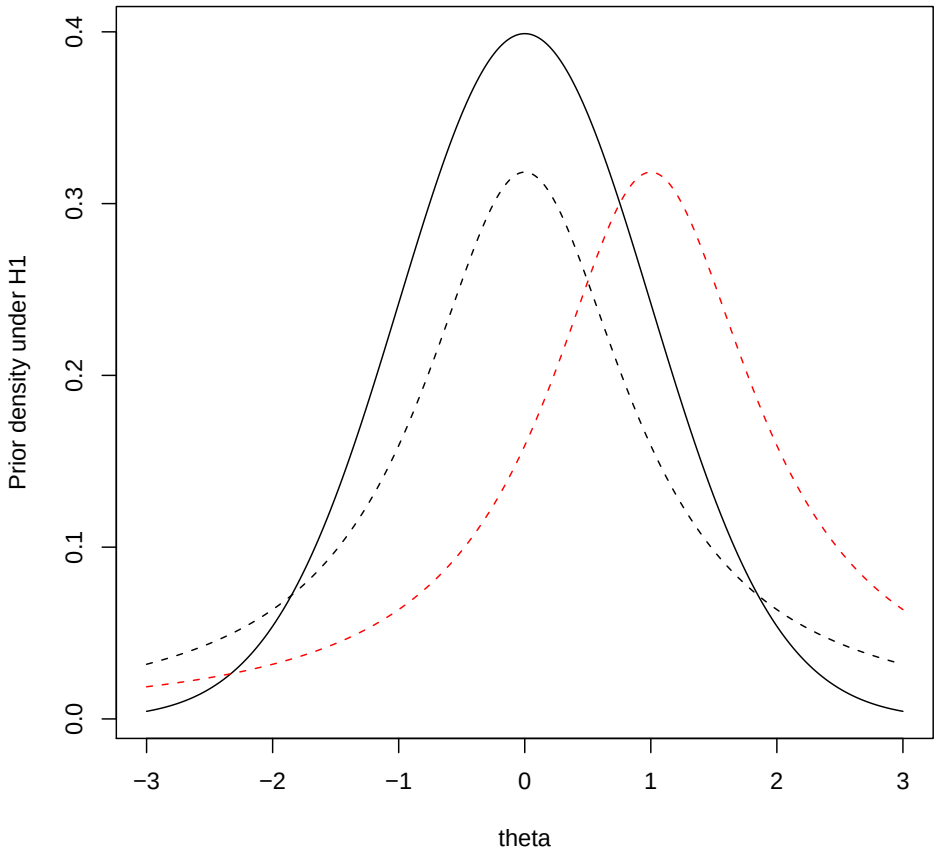
Example: linear regression. BIC to find non-zero coef.

- Asymp. equiv to $P(\boldsymbol{\theta}|M) = N(\mathbf{0}, n\sigma^2(X_M'X_M)^{-1})$, $P(M) \propto 1$
- Complexity penalty is $p \times \log(n)$
- Finds true model as $n \longrightarrow \infty$, p fixed (O'Hagan & Forster, Bay Infer 2004)
- Can fail when p grows with n (Casella et al, Ann Stat 2009)

The message: the prior matters, even for large n

Outline

Philosophically, $P(\theta|\mathrm{H}_1)$ should separate θ from $\mathbf{0}$. However,



Local priors

- Normal
- Cauchy
- Fractional BF
- Intrinsic BF
- ...

Non-local priors (Johnson & Rossell, JRSS-B 2010)

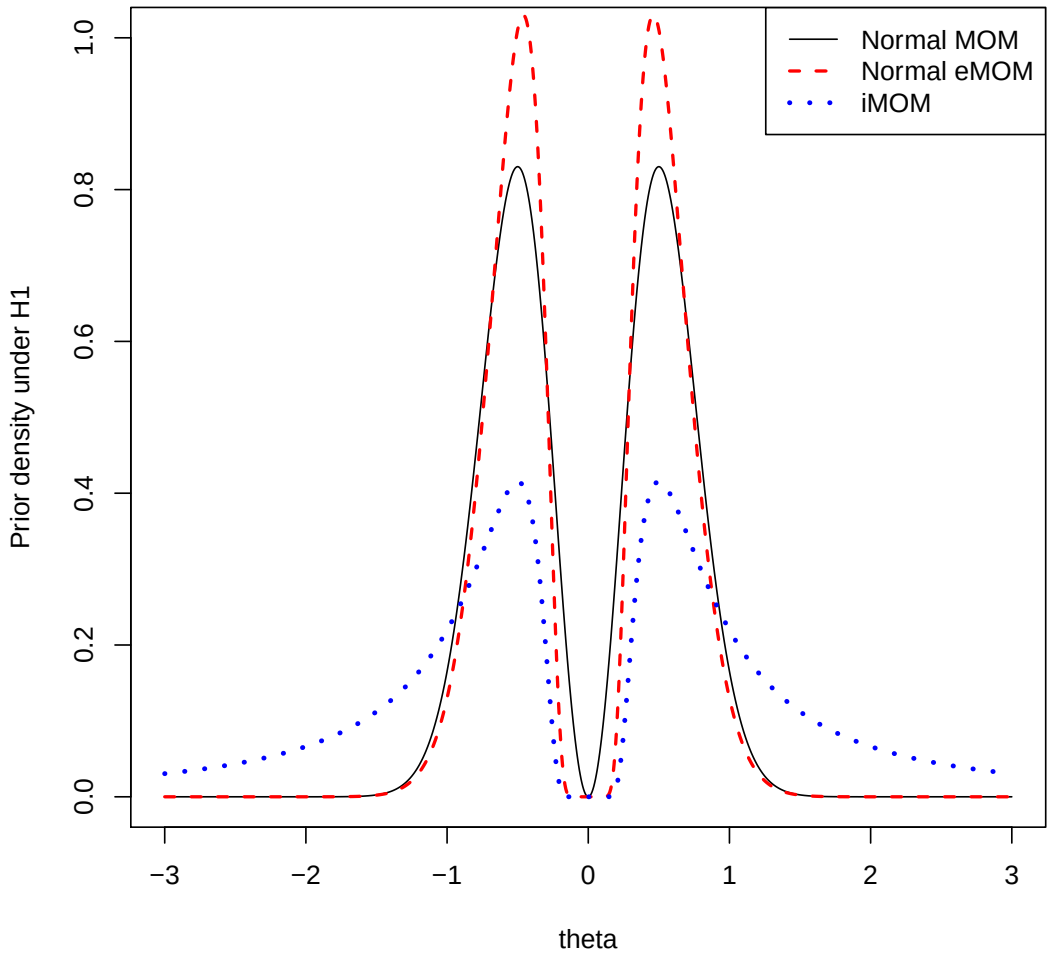
Prior density approaches zero as $\theta_i \longrightarrow 0$ for any i.

Non-local priors (Johnson & Rossell, JRSS-B 2010)

Prior density approaches zero as $\theta_i \longrightarrow 0$ for any i.

1. MOM: $\pi_m(\boldsymbol{\theta}) \propto \left(\prod_{i=1}^{p} \theta_i^{2r}\right) N(\boldsymbol{\theta}; \mathbf{0}, \tau\phi I)$
2. T MOM: $\pi_T(\boldsymbol{\theta}) \propto \left(\prod_{i=1}^{p} \theta_i^{2r}\right) T_{rp+.5a}(\boldsymbol{\theta}; \mathbf{0}, b\phi I)$
3. eMOM: $\pi_e(\boldsymbol{\theta}) \propto \left(\prod_{i=1}^{p} \exp\left\{-\frac{\tau}{\theta_i^2}\right\}\right) N(\boldsymbol{\theta}; \mathbf{0}, \tau\phi I)$, $\tau > 0$
4. iMOM: $\pi_i(\boldsymbol{\theta}) \propto \prod_{i=1}^{p} \exp\left\{-\frac{1}{\theta_i^2}\right\} (\theta_i^2)^{-\frac{\nu+d}{2}}$, $\nu \geq 1$

Non-local priors under $H_1 : \theta \neq 0$



Example: $y_1, \ldots, y_n \sim N(\theta, 1)$. Test $H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$.

- Simulations for $\theta = 0.5$, $n = 100$.

$\bar{y}$	$N(0,1)$	$\theta^2 N(0,1)$	iMOM
0.49	0.99997	0.99990	0.9986
0.52	0.99992	0.99970	0.9996

- Simulations for $\theta = 0$, $n = 100$.

$\bar{y}$	$N(0,1)$	$\theta^2 N(0,1)$	iMOM
0.01	0.091	0.0010	$7.83E^{-7}$
0.07	0.113	0.0019	$5.54E^{-6}$

Intuitive appeal

1. Model separation favors parsimony
2. Statistical significance vs Practical relevance

Practical issues

- Set prior dispersion
  1. Subjective elicitation
  2. Default values
  3. Objective Bayes (Consonni & LaRocca, Valencia Proc 2010)
- Computations
  1. Moderate p: closed form & Laplace approx.
  2. Large p: exact & approximate MCMC

Outline

Variable selection for fixed p

Theorem: under regularity cond., (Johnson & Rossell, JRSS-B 2010)

1. If $\theta_i \neq 0$: local & non-local BF grow expon.

$$\frac{1}{n} \log BF_n \xrightarrow{P} c, \qquad c > 0$$

2. If $\theta_i = 0$

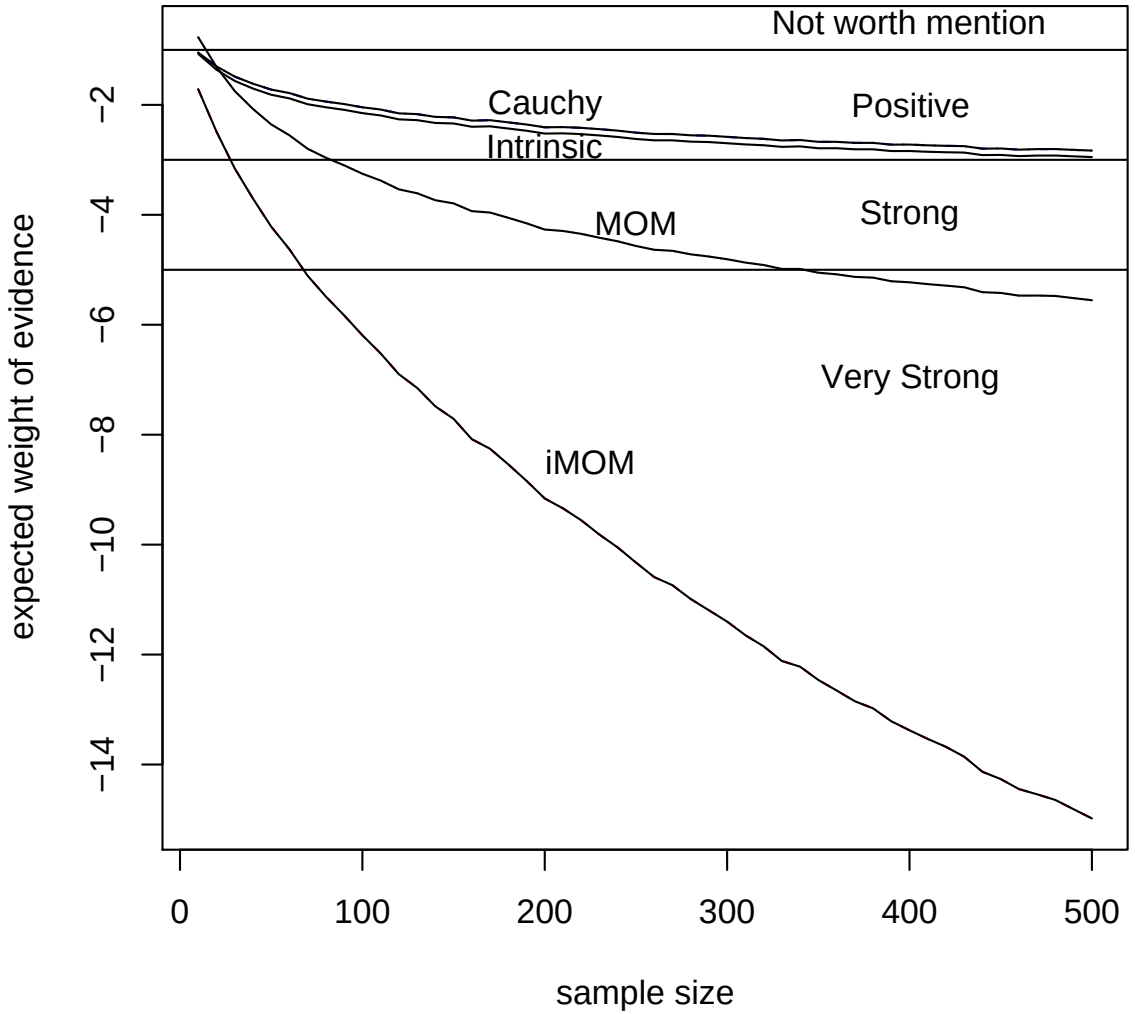
local BF$= O_p(n^{-1/2})$

MOM, T MOM: faster polynomial. $O_p(n^{-1/2-r})$

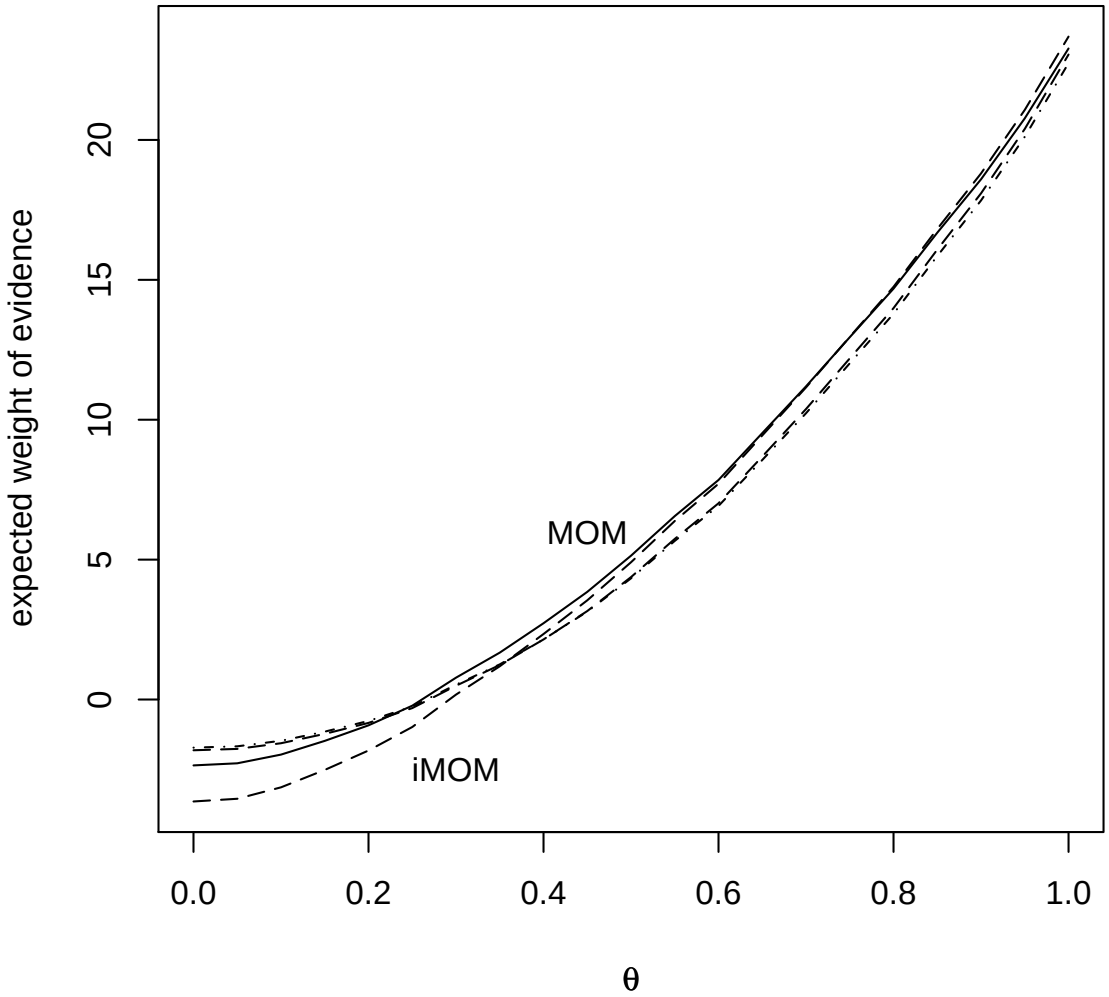
eMOM, iMOM: quasi-exp.

$$\frac{1}{\sqrt{n}} \log BF_n \xrightarrow{P} c, \quad c < 0$$

Example: $N(0, 1)$


Not worth mention
Cauchy
Intrinsic
Positive
MOM
Strong
Very Strong
iMOM
expected weight of evidence
−2
−4
−6
−8
−10
−12
−14
0
100
200
300
400
500
sample size

Example: $N(\theta, 1)$ ($\theta \neq 0$, $n = 50$)



Variable selection for $p = O(n^{\alpha})$

Theorem: (Johnson & Rossell, JASA 2012)

Let $\mathbf{y} \sim N(X\boldsymbol{\theta}, \sigma^2 I)$, M_t true model, $0.5 \leq \alpha < 1$. Assume

- $p \leq n$
- $X'X$ eigenvalues $nc_1 \leq \lambda_i \leq nc_2$
- $\frac{P(M_i)}{P(M_j)}$ bounded $\forall i, j$

Variable selection for $p = O(n^{\alpha})$

Theorem: (Johnson & Rossell, JASA 2012)

Let $\mathbf{y} \sim N(X\boldsymbol{\theta}, \sigma^2 I)$, M_t true model, $0.5 \leq \alpha < 1$. Assume

- $p \leq n$
- $X'X$ eigenvalues $nc_1 \leq \lambda_i \leq nc_2$
- $\frac{P(M_i)}{P(M_j)}$ bounded $\forall i, j$

Then

- $P(M_t|\mathbf{y}) \xrightarrow{P} 1$ for MOM ($r > 1$), eMOM & iMOM priors
- $P(M_t|\mathbf{y}) \xrightarrow{P} 0$ for local priors

Outline

Applications

- Hypothesis testing
- Variable selection
- Massive multiple testing
- Sequential clinical trials
- Infer graphical model structure
- ...

- θ: proportion responders to treatment
- $H_0 : \theta = 0.2$ vs $H_1 : \theta > 0.2$
- $P(H_0) = P(H_1) = 0.5$
- Stop if post prob > 0.9 for H_0 or H_1, up to 50 patients

Proportion of trials stopped to select correct hypothesis

Prior	$\theta = 0.2$	$\theta = 0.4$
Be(0.2,0.8) I($0.2 < \theta < 1$)	.43	.825
Be(1.2,1.8) I($0.2 < \theta < 1$)	.51	.823
iMOM I($0.2 < \theta < 1$)	.91	.812

Linear regression simulation

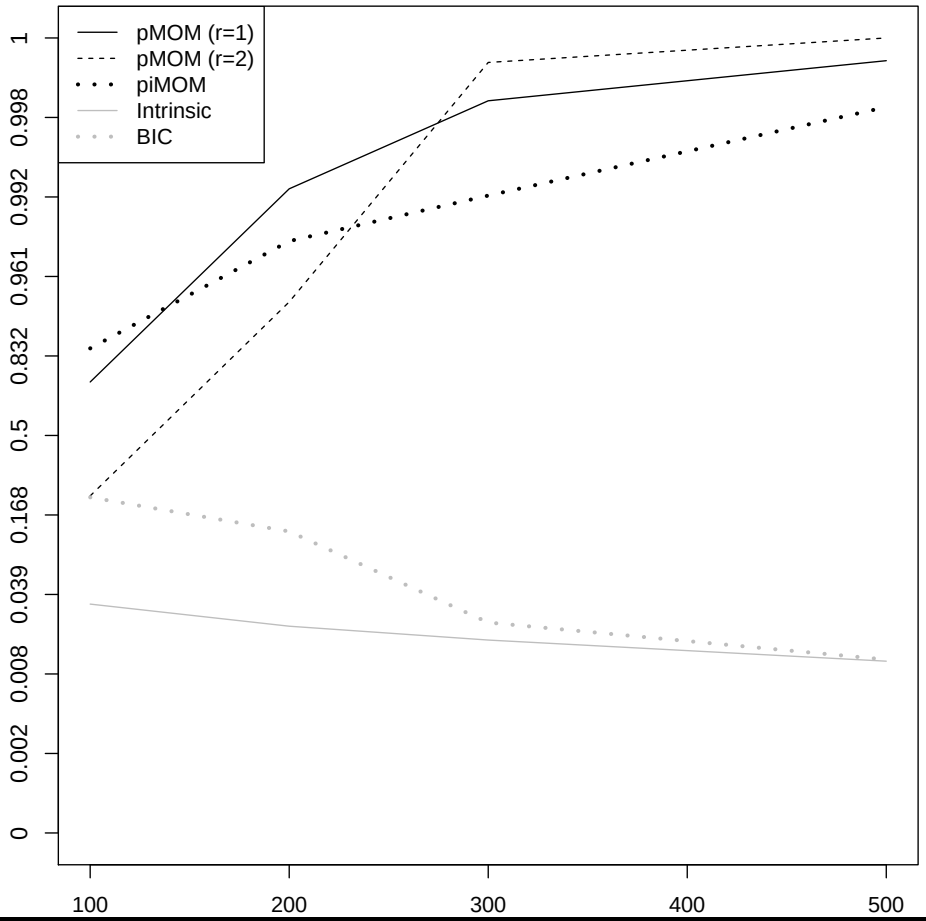
- 500 datasets with $p = n = 100, 200, 300, 500$
- $\boldsymbol{\theta} = (0, \ldots, 0, 0.6, 1.2, 1.8, 2.4, 3)$
- $\sigma^2 = 1, 1.5, 2$
- $\mathbf{x_i} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma})$, $\rho_{ij} = 0, 0.25$
- $P(M)$ =Beta-binomial (violates Theorem conditions)

Compare to

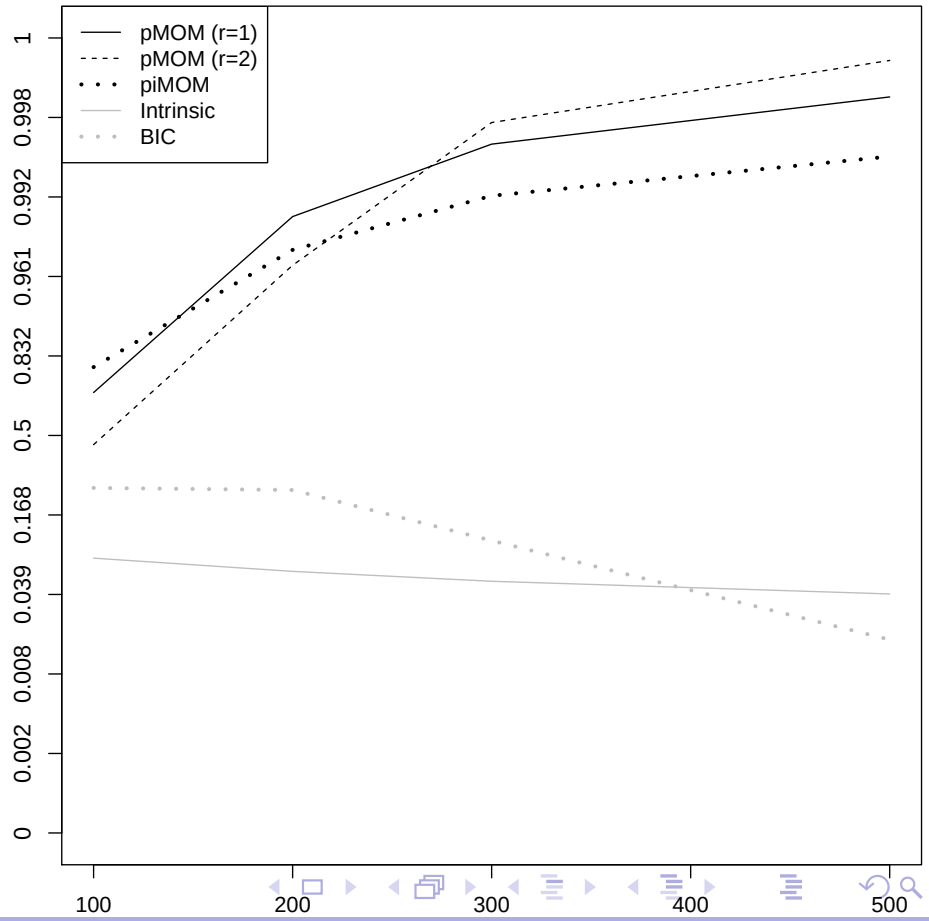
- Intrinsic BF (from below)
- BIC
- SCAD
- LASSO

Average $P(M_t|\mathbf{y})$

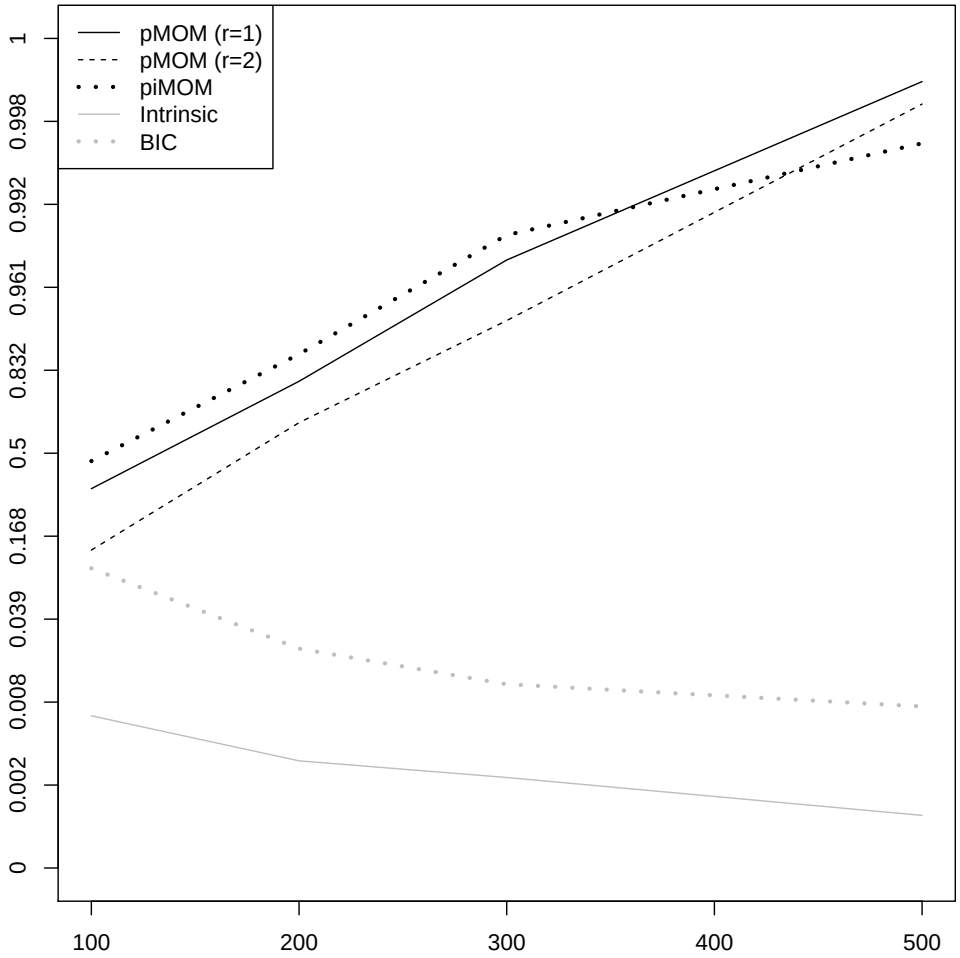
$\sigma^2 = 1$, $\rho_{ij} = 0$



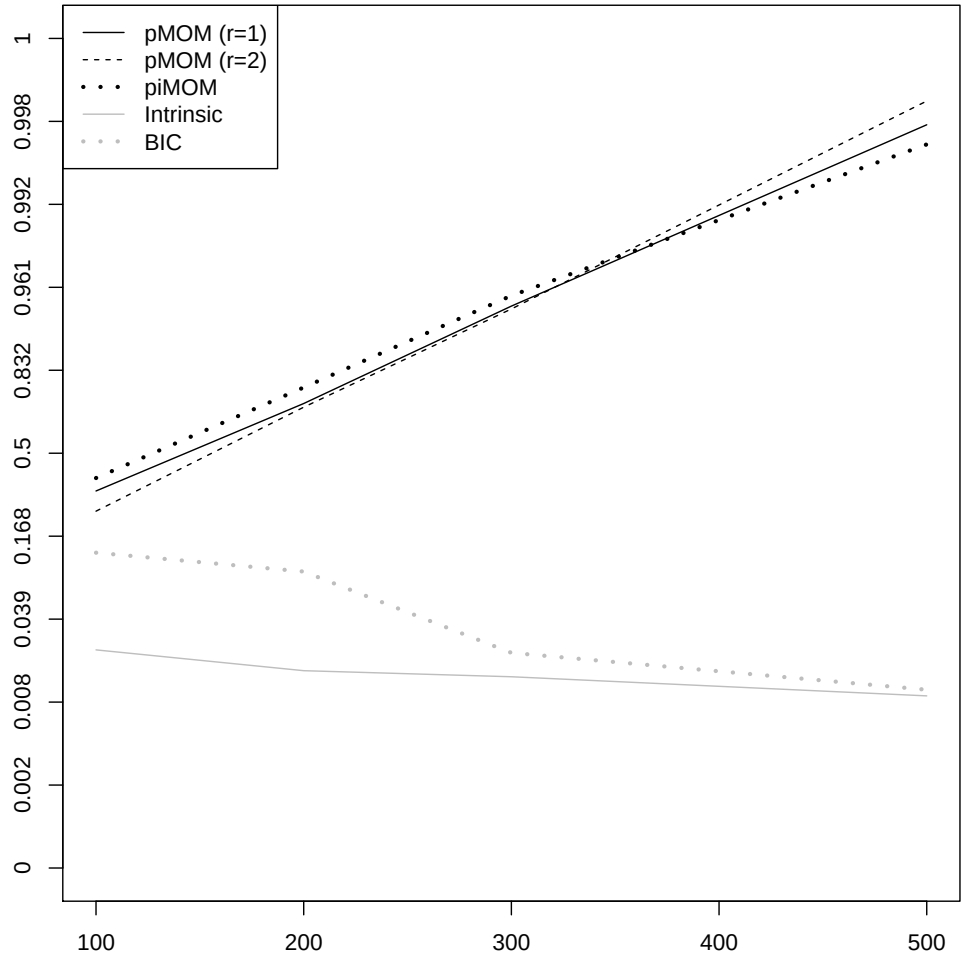
$\rho_{ij} = 0.25$



$\sigma^2 = 2,\ \rho_{ij} = 0$

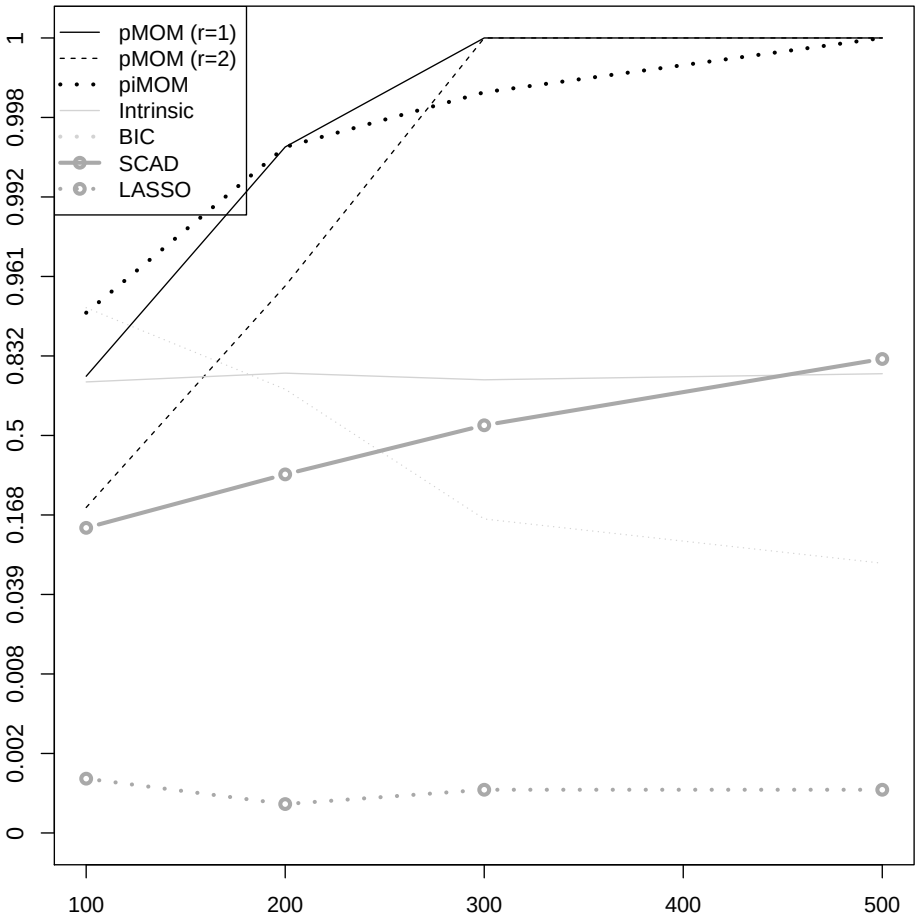


$\rho_{ij} = 0.25$

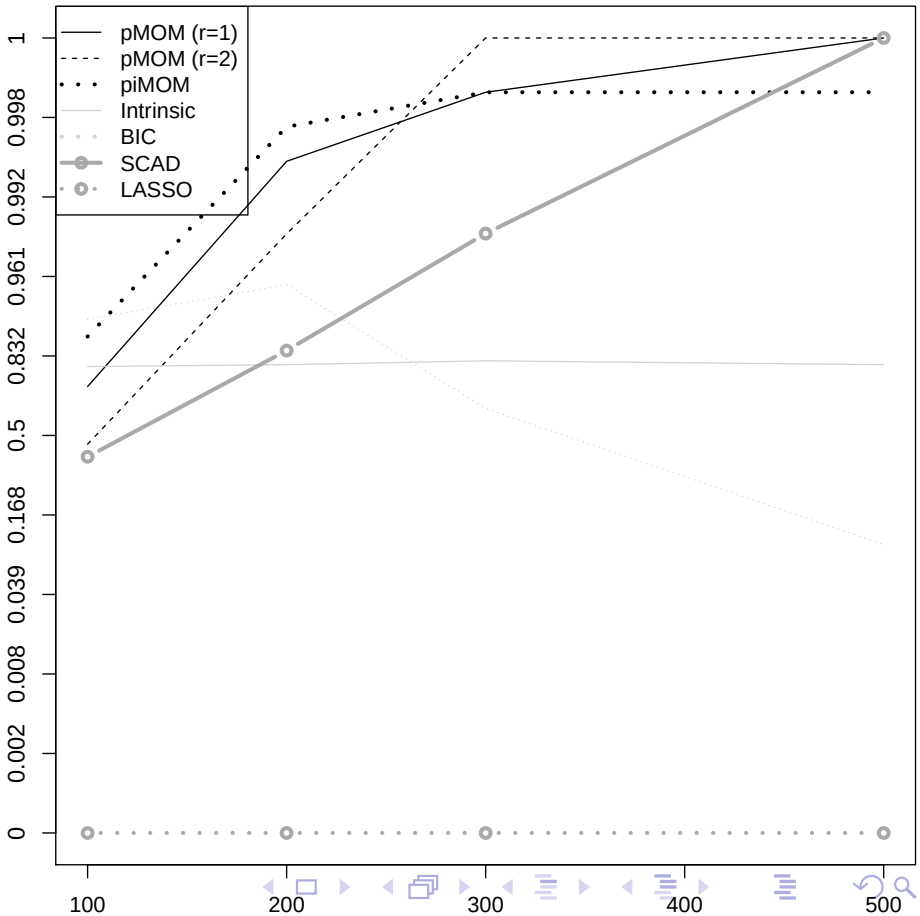


Correct model selections (posterior mode)

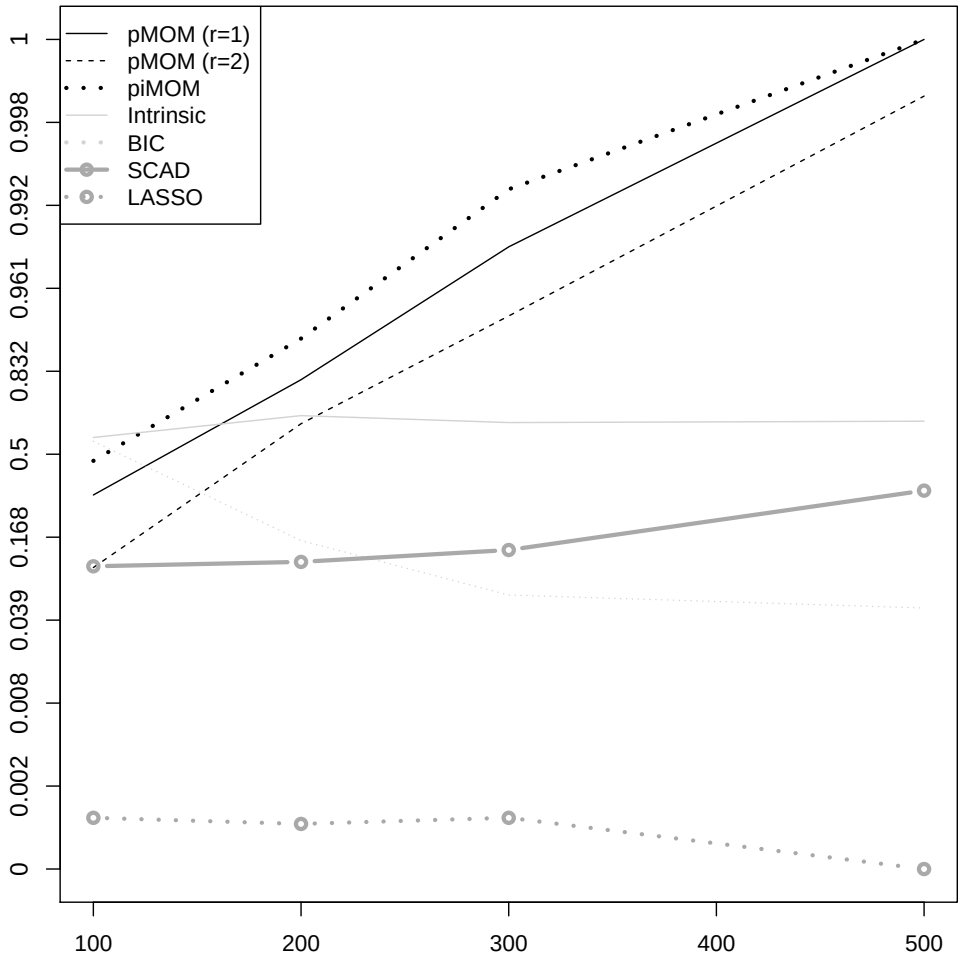
$\sigma^2 = 1,\ \rho_{ij} = 0$



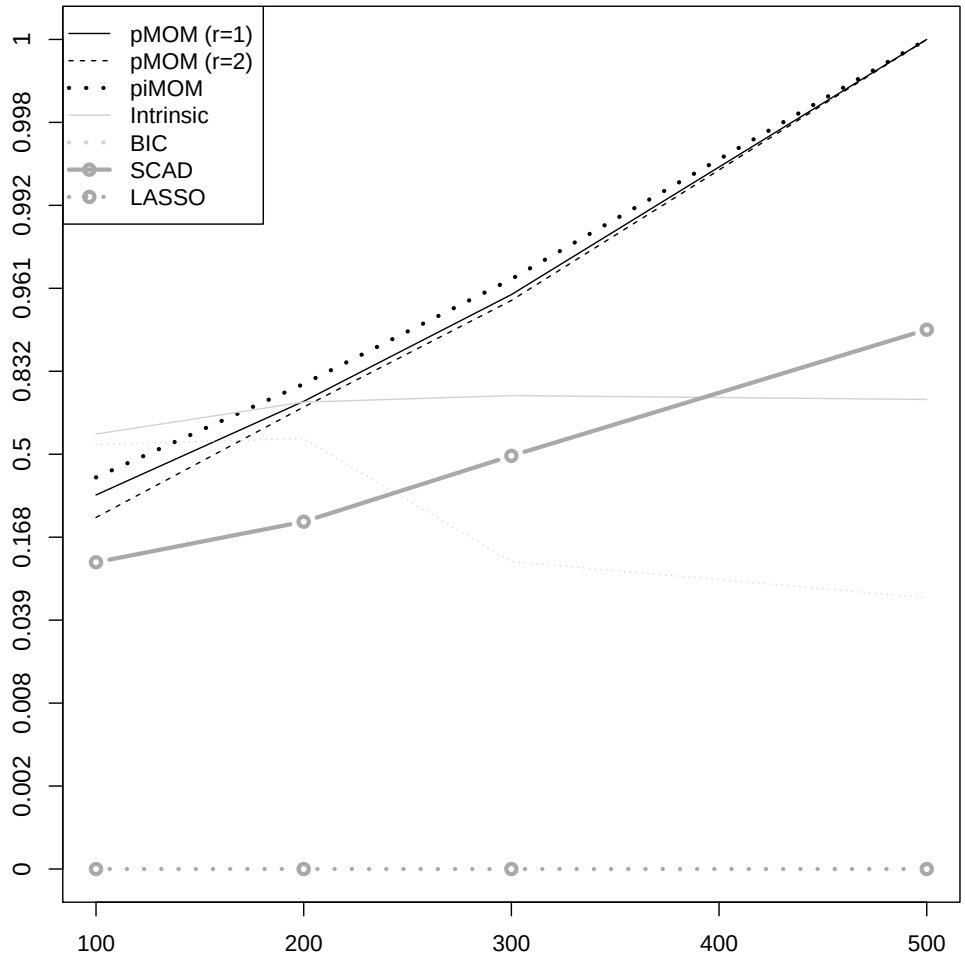
$\rho_{ij} = 0.25$



$\sigma^2 = 2,\ \rho_{ij} = 0$

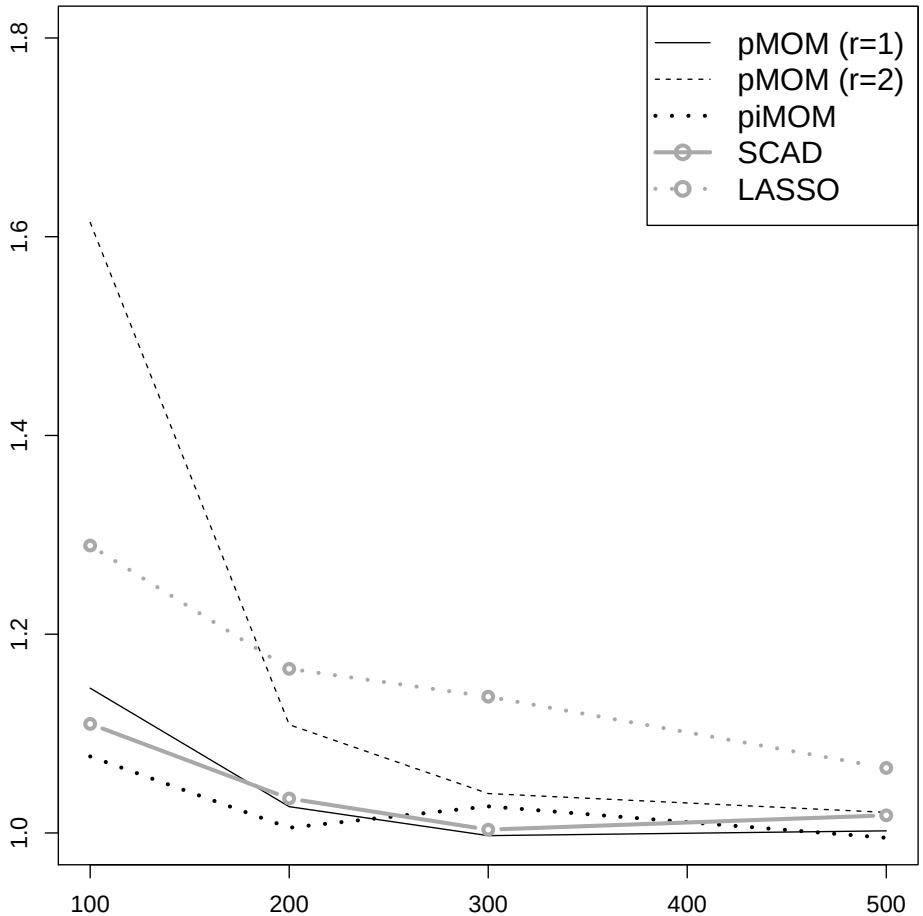


$\rho_{ij} = 0.25$

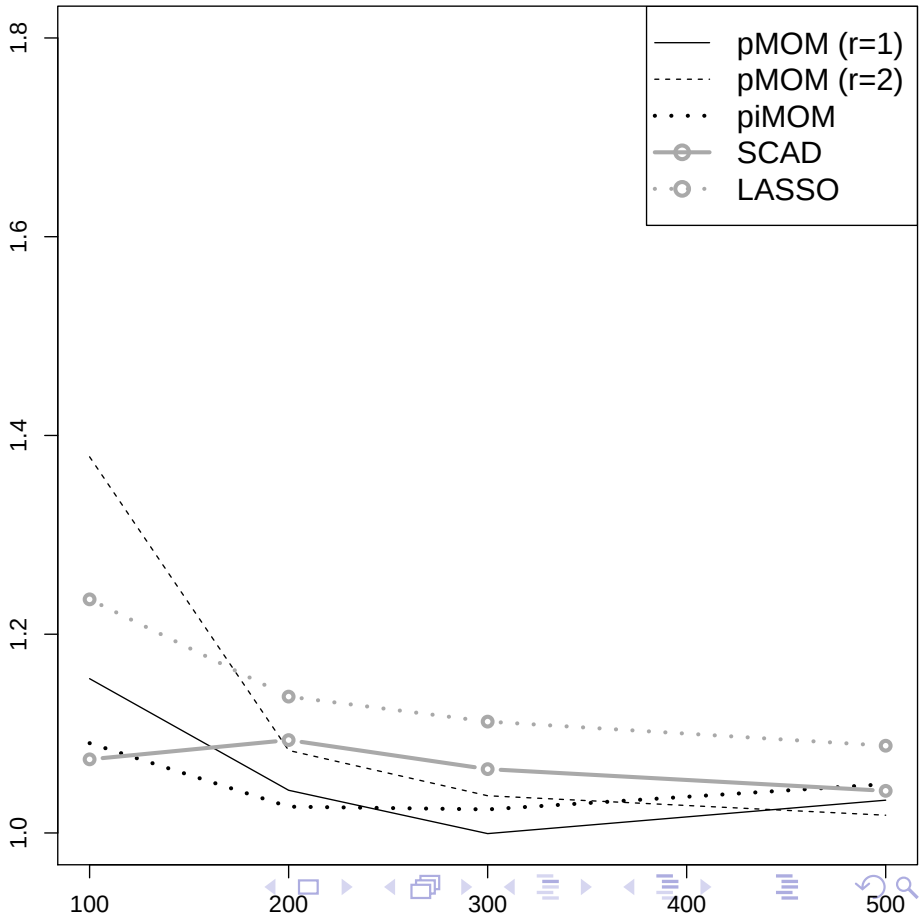


Prediction error (MSE)

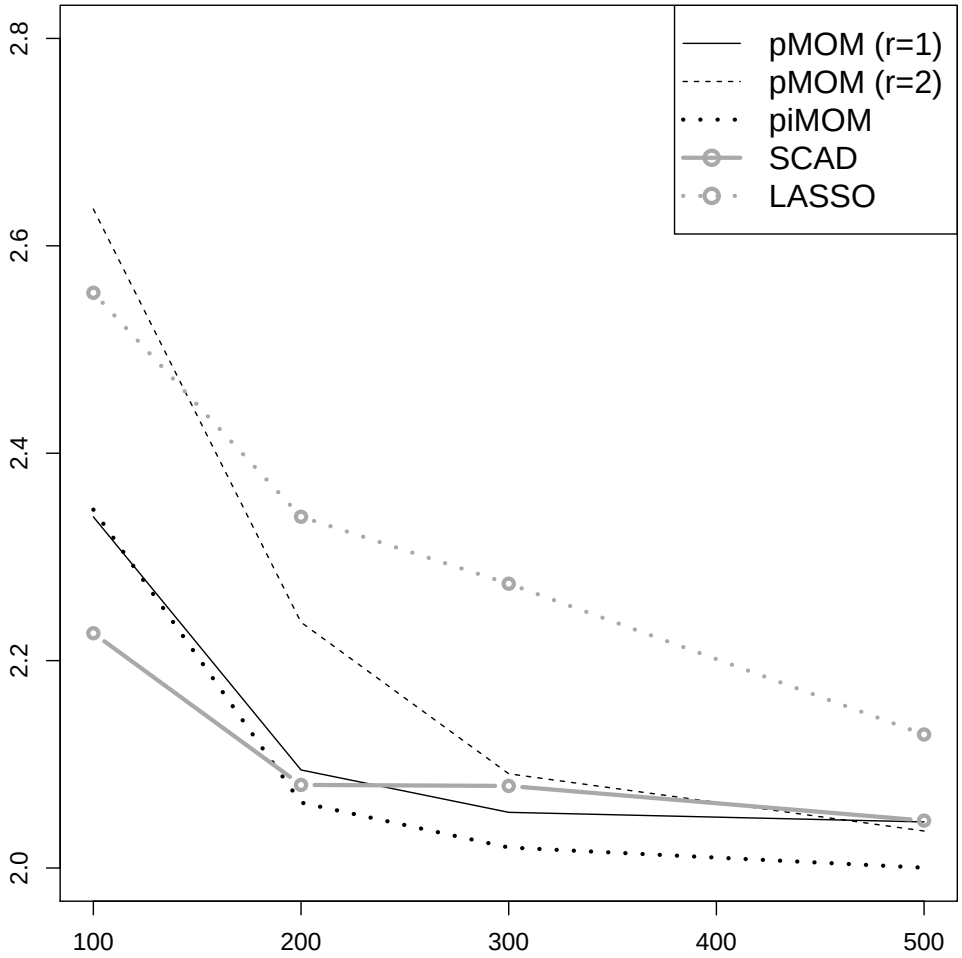
$\sigma^2 = 1,\ \rho_{ij} = 0$



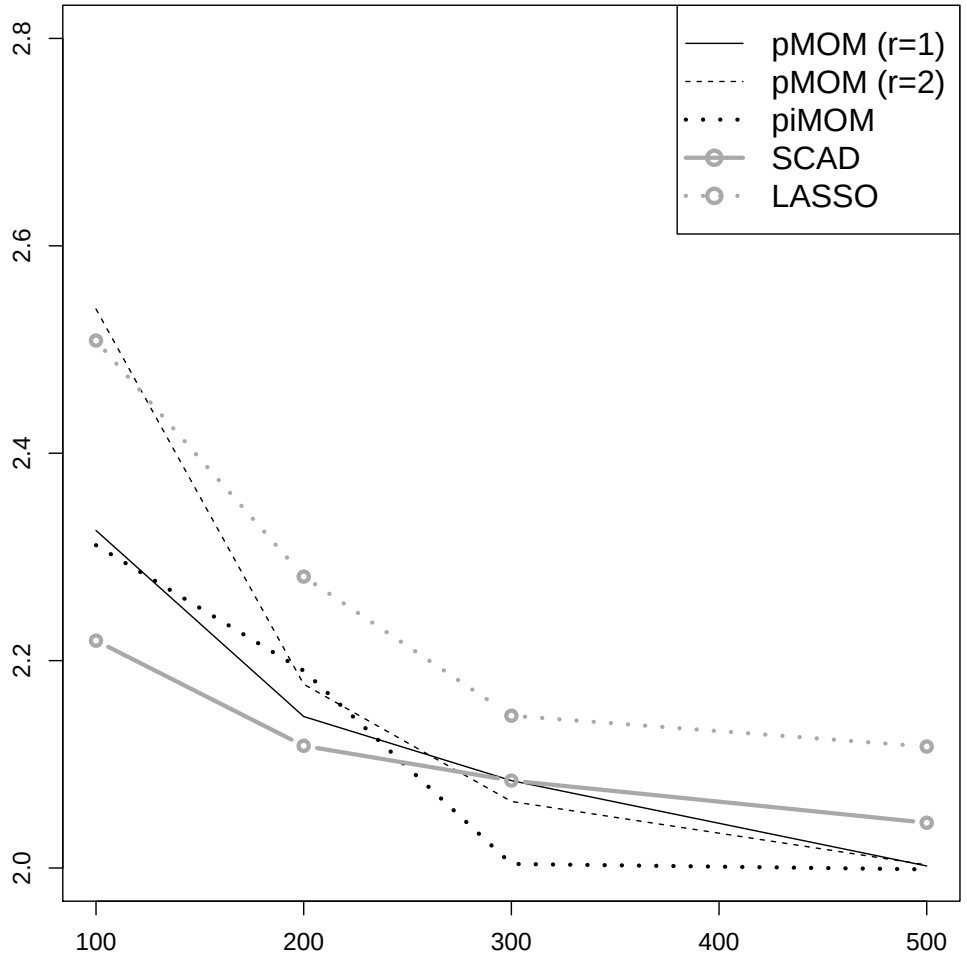
$\rho_{ij} = 0.25$



$\sigma^2 = 2,\ \rho_{ij} = 0$



$\rho_{ij} = 0.25$



Concluding remarks

- Philosophically appealing, good mathematical properties
- Model selection
  1. Faster convergence for fixed p
  2. Consistency when $p = O(n)$
  3. Improved sparsity & FDR control
- Model averaging
  1. Spread weight over smaller set of models
  2. Improved prediction in highly sparse setups
- Implemented in R package `mombf`