Model Selection in Regression:

some new (?) thoughts on the old (?) problem

Felix Abramovich, Tel Aviv University

(joint work with Vadim Grinshtein, The Open University of Israel)

Anestis & His Friends

Villard de Lans, 24-25 March, 2011

Outline

1. Brief survey on model selection in regression
2. MAP selection rule:
   - derivation
   - relations to other existing counterparts
   - basic properties: oracle inequality, adaptive minimaxity
3. Computational aspects
4. Special case: Normal Means problem
5. Main take-away messages

Gaussian Linear Regression

Gaussian linear regression model with p possible predictors and n observations:

$$\mathbf{y} = \beta_1 \mathbf{x}_1 + \ldots + \beta_p \mathbf{x}_p + \epsilon = X\boldsymbol{\beta} + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I_n)$$

- $p < n$ – classical setting
- $p \gg n$ – modern setting

Key sparsity assumption: only some subset of predictors is really “relevant”.

Goal: to identify this “relevant subset” (the “best” model)

What is the “best” model?

The meaning of the “best” model depends on the particular goal at hand :

What is the “best” model?

The meaning of the “best” model depends on the particular goal at hand :

- identification of a true model

What is the “best” model?

The meaning of the “best” model depends on the particular goal at hand :

- identification of a true model
- estimation of coefficients β

What is the “best” model?

The meaning of the “best” model depends on the particular goal at hand :

- identification of a true model
- estimation of coefficients β
- estimation (prediction) of the mean vector $X\beta$

What is the “best” model?

The meaning of the “best” model depends on the particular goal at hand :

- identification of a true model
- estimation of coefficients $\boldsymbol{\beta}$
- estimation (prediction) of the mean vector $X\boldsymbol{\beta}$
- prediction of future observations

$$\mathbf{y} = X\boldsymbol{\beta} + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I_n)$$

For a given model M:

- $d_{j,M} = I\{x_j \in M\}, \quad D_M = diag(\mathbf{d}_M), \quad |M| = \sum_{j=1}^{p} d_{j,M} = tr(D_M)$
- OLS, MLE : $\hat{\boldsymbol{\beta}}_M = (D_M X' X D_M)^{+} D_M X' \mathbf{y}$ ($\hat{\beta}_{j,M} = 0$ if $d_{j,M} = 0$)
- Quadratic risk (MSE): $E||X\hat{\boldsymbol{\beta}}_M - X\boldsymbol{\beta}||^2 = \underbrace{||X\boldsymbol{\beta}_M - X\boldsymbol{\beta}||^2}_{bias^2} + \underbrace{\sigma^2|M|}_{variance}$

$$\mathbf{y} = X\boldsymbol{\beta} + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I_n)$$

For a given model M:

- $d_{j,M} = I\{x_j \in M\}, \quad D_M = diag(\mathbf{d}_M), \quad |M| = \sum_{j=1}^{p} d_{j,M} = tr(D_M)$
- OLS, MLE : $\hat{\boldsymbol{\beta}}_M = (D_M X' X D_M)^+ D_M X' \mathbf{y} \quad (\hat{\beta}_{j,M} = 0$ if $d_{j,M} = 0)$
- Quadratic risk (MSE): $E||X\hat{\boldsymbol{\beta}}_M - X\boldsymbol{\beta}||^2 = \underbrace{||X\boldsymbol{\beta}_M - X\boldsymbol{\beta}||^2}_{bias^2} + \underbrace{\sigma^2|M|}_{variance}$

The (ideally) best model (oracle) :

$$E||X\hat{\boldsymbol{\beta}}_M - X\boldsymbol{\beta}||^2 \rightarrow \min_M$$

(note that the true underlying model is not necessarily the best)

- Empirical risk (least squares)

$$RSS = ||\mathbf{y} - X\hat{\boldsymbol{\beta}}_M||^2 \to \min_M \ ?$$

Trivial solution: a saturated model...

- Empirical risk (least squares)

$$RSS = ||\mathbf{y} - X\hat{\boldsymbol{\beta}}_M||^2 \to \min_M \ ?$$

Trivial solution: a saturated model...

- Idea : *penalized* least squares with a complexity penalty

$$||\mathbf{y} - X\hat{\boldsymbol{\beta}}_M||^2 + Pen(|M|) \to \min_M$$

- Empirical risk (least squares)

$$RSS = ||\mathbf{y} - X\hat{\boldsymbol{\beta}}_M||^2 \to \min_M \ ?$$

Trivial solution: a saturated model...

- Idea : *penalized* least squares with a complexity penalty

$$||\mathbf{y} - X\hat{\boldsymbol{\beta}}_M||^2 + Pen(|M|) \to \min_M$$

- Key question: how to choose a “proper” penalty?

Complexity Penalties

- linear-type penalties $Pen(k) = 2\sigma^2\lambda k$

$\lambda = 1$ — C_p (Mallows, '73), AIC (Akaike, '73)

$\lambda = \ln n/2$ — BIC (Schwarz, '79)

$\lambda = \ln p$ — RIC (Foster & George, '94)

Complexity Penalties

- linear-type penalties $Pen(k) = 2\sigma^2\lambda k$

  $\lambda = 1$ C_p (Mallows, '73), AIC (Akaike, '73)

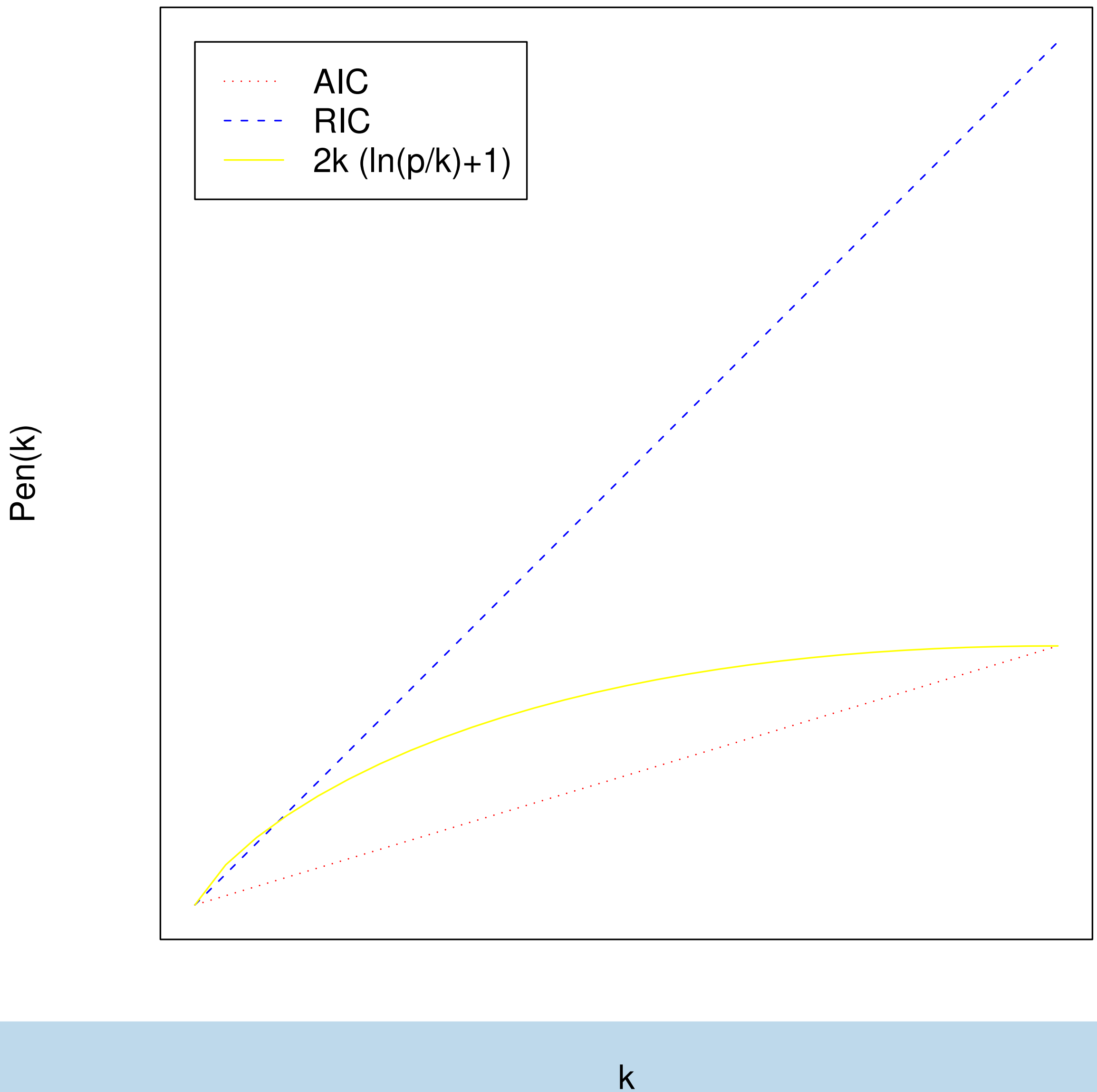
  $\lambda = \ln n/2$ BIC (Schwarz, '79)

  $\lambda = \ln p$ RIC (Foster & George, '94)

- $2k\ln(p/k)$-type nonlinear penalties $Pen(k) = 2\sigma^2\lambda k(\ln(p/k) + \zeta_{p,k})$, where $\zeta_{p,k}$ is "negligible"

  (Birgé & Massart, '01, '07; Johnstone, '02; Abramovich *et al.*, '06; Bunea, Tsybakov & Wegkamp, '07; Abramovich & Grinshtein, '10)

Complexity penalties



Bayesian approach

Why “to Bayes”?

Bayesian approach

Why “to Bayes”?

- (orthodox) Bayesians : since this is the (only) right way to do statistics!

Bayesian approach

Why "to Bayes"?

- (orthodox) Bayesians : since this is the (only) right way to do statistics!
- (intellectual) Bayesians : since this is the (only) right way to think statistics!

Why “to Bayes”?

- (orthodox) Bayesians : since this is the (only) right way to do statistics!
- (intellectual) Bayesians : since this is the (only) right way to think statistics!
- (orthodox) Frequentists : really, why?!!!

Bayesian approach

Why "to Bayes"?

- (orthodox) Bayesians : since this is the (only) right way to do statistics!
- (intellectual) Bayesians : since this is the (only) right way to think statistics!
- (orthodox) Frequentists : really, why?!!!
- (intellectual) Frequentists :

  provides intuition and interpretation for various frequentist procedures (e.g., ridge regression, spline smoothing)

  an efficient tool to obtain different types of estimators (e.g., shrinkage)

Bayesian approach to Model Selection

Model: $\mathbf{y} = X\boldsymbol{\beta} + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I_n)$

$rank(X) = r \leq \min(p, n)$ and any r columns of X are linearly independent

Bayesian approach to Model Selection

Model: $\mathbf{y} = X\boldsymbol{\beta} + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I_n)$

$rank(X) = r \leq \min(p, n)$ and any r columns of X are linearly independent

Prior:

Bayesian approach to Model Selection

Model: $\mathbf{y} = X\boldsymbol{\beta} + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I_n)$

$rank(X) = r \leq \min(p, n)$ and any r columns of X are linearly independent

Prior:

- $P(|M| = k) = \pi(k) > 0, \; k = 0, ..., r$

Bayesian approach to Model Selection

Model: $\mathbf{y} = X\boldsymbol{\beta} + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I_n)$

$rank(X) = r \leq \min(p, n)$ and any r columns of X are linearly independent

Prior:

- $P(|M| = k) = \pi(k) > 0, \ k = 0, ..., r$
- $P(M \big| |M| = k) = \binom{p}{k}^{-1}, \ k = 0, ..., r-1; \quad P(M \big| |M| = r) = 1$

Bayesian approach to Model Selection

Model: $\mathbf{y} = X\boldsymbol{\beta} + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I_n)$

$rank(X) = r \leq \min(p, n)$ and any r columns of X are linearly independent

Prior:

- $P(|M| = k) = \pi(k) > 0,\ k = 0, ..., r$
- $P(M \big| |M| = k) = \binom{p}{k}^{-1},\ k = 0, ..., r-1; \quad P(M \big| |M| = r) = 1$
- $\beta | M \sim N_p(0, \gamma\sigma^2 (D_M X' X D_M)^+)$ (g-prior – Zellner, '86)

Bayesian approach to Model Selection

Model: $\mathbf{y} = X\boldsymbol{\beta} + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I_n)$

$rank(X) = r \leq \min(p, n)$ and any r columns of X are linearly independent

Prior:

- $P(|M| = k) = \pi(k) > 0, \; k = 0, ..., r$
- $P(M \big| |M| = k) = \binom{p}{k}^{-1}, \; k = 0, ..., r-1; \quad P(M \big| |M| = r) = 1$
- $\beta | M \sim N_p(0, \gamma\sigma^2 (D_M X' X D_M)^+) \quad$ (g-prior – Zellner, '86)

Posterior:

$$P(M|\mathbf{y}) \propto \pi(|M|) \binom{p}{|M|}^{-1} (1+\gamma)^{-\frac{|M|}{2}} \exp\left\{ \frac{\gamma}{\gamma+1} \frac{\mathbf{y}' X D_M (D_M X' X D_M)^+ D_M X' \mathbf{y}}{2\sigma^2} \right\}$$

(without the binomial coefficient for $|M| = r$)

MAP rule

MAP rule :

$$P(M|\mathbf{y}) \propto \pi(|M|)\binom{p}{|M|}^{-1}(1+\gamma)^{-\frac{|M|}{2}} \exp\left\{\frac{\gamma}{\gamma+1}\frac{\mathbf{y}'XD_M(D_MX'XD_M)^{+}D_MX'\mathbf{y}}{2\sigma^2}\right\}$$

or, equivalently,

$$\underbrace{||\mathbf{y}-X\hat{\boldsymbol{\beta}}_M||^2}_{RSS} + \underbrace{2\sigma^2(1+1/\gamma)\ln\left\{\binom{p}{|M|}\pi^{-1}(|M|)(1+\gamma)^{\frac{|M|}{2}}\right\}}_{complexity\ penalty\ Pen(|M|)} \to \min_M$$

MAP model selector : penalized least squares with complexity penalty

$$Pen(|M|) = \begin{cases} 2\sigma^2(1+1/\gamma)\ln\left\{\binom{p}{|M|}\pi^{-1}(|M|)(1+\gamma)^{\frac{|M|}{2}}\right\} & |M| = 0,...,r-1 \\ 2\sigma^2(1+1/\gamma)\ln\left\{\pi^{-1}(r)(1+\gamma)^{\frac{r}{2}}\right\} & |M| = r \end{cases}$$

Examples of priors

A specific type of penalty depends on the choice of prior $\pi(|M|)$:

Examples of priors

A specific type of penalty depends on the choice of prior $\pi(|M|)$:

1. (truncated) binomial prior $B(p, \xi)$

$$Pen(k) = 2k\sigma^2(1+1/\gamma)\ln\left(\frac{1-\xi}{\xi}\sqrt{1+\gamma}\right) \sim 2k\sigma^2\ln\left(\frac{1-\xi}{\xi}\sqrt{\gamma}\right)$$ – linear penalty

- C_p, AIC: $\xi \sim \sqrt{\gamma}/(e+\sqrt{\gamma})$
- RIC: $\xi \sim \sqrt{\gamma}/(p+\sqrt{\gamma})$
- BIC: $\xi \sim \sqrt{\gamma}/(\sqrt{n}+\sqrt{\gamma})$

Examples of priors

A specific type of penalty depends on the choice of prior $\pi(|M|)$:

1. (truncated) binomial prior $B(p, \xi)$

   $Pen(k) = 2k\sigma^2(1+1/\gamma)\ln\left(\frac{1-\xi}{\xi}\sqrt{1+\gamma}\right) \sim 2k\sigma^2\ln\left(\frac{1-\xi}{\xi}\sqrt{\gamma}\right)$ – linear penalty

   - C_p, AIC: $\xi \sim \sqrt{\gamma}/(e+\sqrt{\gamma})$
   - RIC: $\xi \sim \sqrt{\gamma}/(p+\sqrt{\gamma})$
   - BIC: $\xi \sim \sqrt{\gamma}/(\sqrt{n}+\sqrt{\gamma})$

2. (truncated) geometric prior $\pi(k) \propto q^k$

   $Pen(k) = 2\sigma^2(1+1/\gamma)k(\ln(p/k)+\zeta(\gamma, q))$ – $2k\ln(p/k)$-type penalty

Oracle inequality

How good is MAP selector w.r.t. an oracle?

Oracle risk: $\inf_M E||X\hat{\boldsymbol{\beta}}_M - X\boldsymbol{\beta}||^2$

No estimator can attain a risk smaller than within $\ln(p)$-factor of that of an oracle (Foster & George, '94; Donoho & Johnstone, '95)

Assumption (P). *Assume that* $\pi(k) \leq \binom{p}{k} e^{-c(\gamma)k},\ k = 0, ..., r-1,$ *and* $\pi(r) \leq e^{-c(\gamma)r}$, *where* $c(\gamma) = 8(\gamma + 3/4)^2\ \ (\geq 9/2)$.

- holds for *any* $\pi(k)$ for all $k \leq pe^{-c(\gamma)}$
- for "sparse" priors $\pi(k) \approx 0$ for large k.

Oracle inequality (cont.)

Theorem (oracle inequality). *Let $\pi(k)$ satisfies Assumption (P) and, in addition, $\pi(0) \geq p^{-c},\ \pi(k) \geq p^{-ck},\ k = 1, ..., r$ for some $c > 0$. Then,*

$$E||X\hat{\boldsymbol{\beta}}_{\hat{M}} - X\boldsymbol{\beta}||^2 \leq c_2(\gamma) \ln p \, (\underbrace{\inf_M E||X\hat{\boldsymbol{\beta}}_M - X\boldsymbol{\beta}||^2}_{oracle\ risk} + \sigma^2)$$

for some $c_2(\gamma) \geq 2$.

Examples:

- binomial prior $B(p, 1/p)$ (RIC)
- geometric prior ($2k \ln(p/k)$-type penalty)

Risk bounds for sparse settings

Sparsity assumption : true model M_0 is sparse, i.e. $|M_0| = ||\boldsymbol{\beta}||_0 = p_0 \leq r$

Risk bounds for sparse settings

Sparsity assumption : true model M_0 is sparse, i.e. $|M_0| = ||\boldsymbol{\beta}||_0 = p_0 \leq r$

Theorem (upper bound). *Let the prior* $\pi(\cdot)$ *satisfy Assumption (P) and, in addition,* $\pi(p_0) \geq (p_0/(pe))^{cp_0}$ *if* $p_0 < r$ *and* $\pi(r) \geq e^{-cr}$ *if* $p_0 = r$ *for some* $c > c(\gamma)$. *Then,*

$$\sup_{\boldsymbol{\beta}:||\boldsymbol{\beta}||_0 \leq p_0} E||X\hat{\boldsymbol{\beta}}_{\hat{M}} - X\boldsymbol{\beta}||^2 \leq C_1(\gamma)\sigma^2 \min(p_0(\ln(p/p_0)+1), r)$$

Risk bounds for sparse settings

Sparsity assumption : true model M_0 is sparse, i.e. $|M_0| = ||\boldsymbol{\beta}||_0 = p_0 \leq r$

Theorem (upper bound). *Let the prior $\pi(\cdot)$ satisfy Assumption (P) and, in addition, $\pi(p_0) \geq (p_0/(pe))^{cp_0}$ if $p_0 < r$ and $\pi(r) \geq e^{-cr}$ if $p_0 = r$ for some $c > c(\gamma)$. Then,*

$$\sup_{\boldsymbol{\beta}:||\boldsymbol{\beta}||_0 \leq p_0} E||X\hat{\boldsymbol{\beta}}_{\hat{M}} - X\boldsymbol{\beta}||^2 \leq C_1(\gamma)\sigma^2 \min(p_0(\ln(p/p_0)+1), r)$$

Let $\tau[k]$ be the ratio between the minimal and maximal eigenvalues of all $k \times k$ submatrices of $X'X$ generated by any k columns of X.

Risk bounds for sparse settings

Sparsity assumption : true model M_0 is sparse, i.e. $|M_0| = ||\boldsymbol{\beta}||_0 = p_0 \le r$

Theorem (upper bound). *Let the prior* $\pi(\cdot)$ *satisfy Assumption (P) and, in addition,* $\pi(p_0) \ge (p_0/(pe))^{cp_0}$ *if* $p_0 < r$ *and* $\pi(r) \ge e^{-cr}$ *if* $p_0 = r$ *for some* $c > c(\gamma)$. *Then,*

$$\sup_{\boldsymbol{\beta}:||\boldsymbol{\beta}||_0 \le p_0} E||X\hat{\boldsymbol{\beta}}_{\hat{M}} - X\boldsymbol{\beta}||^2 \le C_1(\gamma)\sigma^2 \min(p_0(\ln(p/p_0)+1), r)$$

Let $\tau[k]$ be the ratio between the minimal and maximal eigenvalues of all $k \times k$ submatrices of $X'X$ generated by any k columns of X.

Theorem (minimax lower bound). *There exists* $C_2 > 0$ *such that*

$$\inf_{\hat{\mathbf{y}}} \sup_{\boldsymbol{\beta}:||\boldsymbol{\beta}||_0 \le p_0} E||\hat{\mathbf{y}} - X\boldsymbol{\beta}||^2 \ge \begin{cases} C_2\sigma^2\tau[2p_0]\, p_0(\ln(p/p_0)+1), & 1 \le p_0 \le r/2 \\ C_2\sigma^2\tau[p_0]\, r, & r/2 \le p_0 \le r \end{cases}$$

Raskutti *et al.* ('09), Rigollet & Tsybakov ('10) for $p_0 \le r/2$; Abramovich & Grinshtein ('10)

Asymptotic setup

"Classical" asymptotics : $n \to \infty$, p is fixed or, at most, $p_n \ll n$

"Modern" asymptotics : $n \to \infty$, $p_n \to \infty$ and it might be $p_n > n$ or even $p_n \gg n$

Sequences of designs $X_{n,p_n} = X_p$, coefficients vectors $\boldsymbol{\beta}_p$, priors $\pi_p(\cdot)$, etc.

$$\mathbf{y} = X_p \boldsymbol{\beta}_p + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_n)$$

$rank(X_p) = r \to \infty$ and any r columns of X_p are linearly independent ($\tau_p[r] > 0$)

Two types of design

upper bound : $C_1\sigma^2 \min(p_0(\ln(p/p_0)+1), r)$

lower bound : $$\begin{cases} C_2\sigma^2\tau_p[2p_0]\, p_0(\ln(p/p_0)+1), & 1 \le p_0 \le r/2 \\ C_2\sigma^2\tau_p[p_0]\, r, & r/2 \le p_0 \le r \end{cases}$$

Two types of design

upper bound : $C_1\sigma^2 \min(p_0(\ln(p/p_0)+1), r)$

lower bound : $\begin{cases} C_2\sigma^2\tau_p[2p_0]\, p_0(\ln(p/p_0)+1), & 1 \le p_0 \le r/2 \\ C_2\sigma^2\tau_p[p_0]\, r, & r/2 \le p_0 \le r \end{cases}$

Remark : lower bound depends on X_p only through $\tau_p[p_0]$ and $\tau_p[2p_0]$

Two types of design

upper bound : $C_1\sigma^2 \min(p_0(\ln(p/p_0)+1), r)$

lower bound : $\begin{cases} C_2\sigma^2\tau_p[2p_0]\, p_0(\ln(p/p_0)+1), & 1 \le p_0 \le r/2 \\ C_2\sigma^2\tau_p[p_0]\, r, & r/2 \le p_0 \le r \end{cases}$

Remark : lower bound depends on X_p only through $\tau_p[p_0]$ and $\tau_p[2p_0]$

- $\tau_p[r] \not\to 0$ – nearly-orthogonal design
- $\tau_p[r] \to 0$ – multicollinear design

Nearly-orthogonal design

- $p = O(r)$ and, therefore, $p = O(n)$

Nearly-orthogonal design

- $p = O(r)$ and, therefore, $p = O(n)$
- The minimax risk over $\mathcal{M}_{p_0} = \{M : |M| \leq p_0\}$ is of order $p_0(\ln(p/p_0) + 1)$

Nearly-orthogonal design

- $p = O(r)$ and, therefore, $p = O(n)$
- The minimax risk over $\mathcal{M}_{p_0} = \{M : |M| \leq p_0\}$ is of order $p_0(\ln(p/p_0) + 1)$
- Let

1. $\pi_p(k) \leq \binom{p}{k} e^{-c(\gamma)k},\ k = 0, ..., r-1$ and $\pi_p(r) \leq e^{-c(\gamma)r}$ (Assumption (P))
2. $\pi_p(k) \geq (k/(pe))^{c_1 k},\ k = 1, ..., r-1$ and $\pi_p(r) \geq e^{-c_2 r}, \quad c_1,\ c_2 > c(\gamma)$

Then, the MAP model selector is asymptotically minimax *simultnaneously* over all $\mathcal{M}_{p_0},\ 1 \leq p_0 \leq r$

Nearly-orthogonal design

- $p = O(r)$ and, therefore, $p = O(n)$
- The minimax risk over $\mathcal{M}_{p_0} = \{M : |M| \leq p_0\}$ is of order $p_0(\ln(p/p_0) + 1)$
- Let

  1. $\pi_p(k) \leq \binom{p}{k} e^{-c(\gamma)k},\ k = 0, ..., r-1$ and $\pi_p(r) \leq e^{-c(\gamma)r}$ (Assumption (P))
  2. $\pi_p(k) \geq (k/(pe))^{c_1 k},\ k = 1, ..., r-1$ and $\pi_p(r) \geq e^{-c_2 r}, \quad c_1,\ c_2 > c(\gamma)$

  Then, the MAP model selector is asymptotically minimax *simultnaneously* over all $\mathcal{M}_{p_0},\ 1 \leq p_0 \leq r$

- $||X_p\hat{\boldsymbol{\beta}}_p - X_p\boldsymbol{\beta}_p|| \asymp ||\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_p||$ – all the results remain true for estimating coefficients $\boldsymbol{\beta}_p$ (not true for multicollinear design!)

Examples of priors

- geometric prior ($2k\ln(p/k)$-type penalty)

Examples of priors

- geometric prior ($2k\ln(p/k)$-type penalty)
- no binomial prior $B(p,\xi)$ (hence, no linear penalty) can satisfy the conditions for both sparse ($p_0 \ll p$) and dense ($p_0 \sim p$) cases :

RIC ($\xi \sim 1/p$): $O(\sigma^2 p_0 \ln p) \sim O(\sigma^2 p_0(\ln(p/p_0)+1))$ for sparse cases

AIC ($\xi \sim const$): $O(\sigma^2 p) \sim O(\sigma^2 p_0(\ln(p/p_0)+1))$ for dense cases

Examples of priors

- geometric prior ($2k\ln(p/k)$-type penalty)

- no binomial prior $B(p,\xi)$ (hence, no linear penalty) can satisfy the conditions for both sparse ($p_0 \ll p$) and dense ($p_0 \sim p$) cases :

  RIC ($\xi \sim 1/p$): $O(\sigma^2 p_0 \ln p) \quad \sim \quad O(\sigma^2 p_0(\ln(p/p_0)+1))$ for sparse cases

  AIC ($\xi \sim const$): $O(\sigma^2 p) \quad \sim \quad O(\sigma^2 p_0(\ln(p/p_0)+1))$ for dense cases

- **Remark**: Lasso and Dantzig selectors – similar to RIC under stronger nearly-orthogonality restrictions (Bickel, Ritov & Tsybakov '09)

Multicollinear design

- Necessarily appears for $p \gg n$ setup

Multicollinear design

- Necessarily appears for $p \gg n$ setup
- There is a gap between upper and lower bounds

Multicollinear design

- Necessarily appears for $p \gg n$ setup
- There is a gap between upper and lower bounds
- Idea : exploit strong correlations between predictors to reduce the model's size (decrease the variance) without paying much extra price in bias – "blesssing of multicollinearity" (?)

Multicollinear design

- Necessarily appears for $p \gg n$ setup
- There is a gap between upper and lower bounds
- Idea : exploit strong correlations between predictors to reduce the model's size (decrease the variance) without paying much extra price in bias – "blesssing of multicollinearity" (?)
- MAP model selector indeed remains asymptotically minimax under certain additional constraints on X_p and $||\boldsymbol{\beta}_p||_\infty$ (see Abramovich & Grinshtein, '10 for technical detail)

Welcome to the real world...

1. Estimation of prior parameters and σ^2
   - fully Bayesian approach – priors on parameters
   - empirical Bayes – EM algorithm or its modifications (George & Foster, ’00)

Welcome to the real world...

1. Estimation of prior parameters and σ^2
   - fully Bayesian approach – priors on parameters
   - empirical Bayes – EM algorithm or its modifications (George & Foster, ’00)

2. MAP solution

$$RSS(M) + Pen(|M|) \rightarrow \min_{M}$$

combinatorical search (NP problem)!

Computational aspects

$$RSS(M) + Pen(|M|) = ||\mathbf{y} - X\hat{\boldsymbol{\beta}}_M||^2 + Pen(||\hat{\boldsymbol{\beta}}_M||_0) \rightarrow \min_M$$

Computational aspects

$$RSS(M) + Pen(|M|) = ||\mathbf{y} - X\hat{\boldsymbol{\beta}}_M||^2 + Pen(||\hat{\boldsymbol{\beta}}_M||_0) \rightarrow \min_M$$

- Greedy algorithms (forward selection, matching pursuit) – approximate the global solution by a stepwise sequence of local ones

Computational aspects

$$RSS(M) + Pen(|M|) = ||\mathbf{y} - X\hat{\boldsymbol{\beta}}_M||^2 + Pen(||\hat{\boldsymbol{\beta}}_M||_0) \to \min_M$$

- Greedy algorithms (forward selection, matching pursuit) – approximate the global solution by a stepwise sequence of local ones
- Convex relaxation methods (for linear penalties – Lasso, Dantzig selector) – replace the original combinatorial problem by a related convex program: e.g., Lasso replaces $||\hat{\beta}_M||_0$ in the linear penalty by $||\hat{\beta}_M||_1$

Computational aspects

$$RSS(M) + Pen(|M|) = ||\mathbf{y} - X\hat{\boldsymbol{\beta}}_M||^2 + Pen(||\hat{\boldsymbol{\beta}}_M||_0) \to \min_M$$

- Greedy algorithms (forward selection, matching pursuit) – approximate the global solution by a stepwise sequence of local ones
- Convex relaxation methods (for linear penalties – Lasso, Dantzig selector) – replace the original combinatorial problem by a related convex program: e.g., Lasso replaces $||\hat{\boldsymbol{\beta}}_M||_0$ in the linear penalty by $||\hat{\boldsymbol{\beta}}_M||_1$
- Stochastic search variable selection (SSVS) – exploits Bayesian nature of the selector by generating a sequence of models from the posterior distribution $P(M|\mathbf{y})$ (George & McCullogh, '93, '97)

Stochastic search variable selection

General idea : generate a sequence of models from the posterior distribution $P(M|\mathbf{y})$ or, equivalently, $P(\mathbf{d}_M|\mathbf{y})$

Key point : we need just the posterior mode, no need to generate the entire distribution of size 2^p. Models with highest posterior probabilities will appear more frequently and can be identified even for a relatively small ($\ll 2^p$) sample size

Gibbs sampler : generate a sequence of models (indicator vectors) $\mathbf{d}_1, ..., \mathbf{d}_M$ componentwise by sampling consecutively from the conditional distributions of $d_j|(\mathbf{d}_{(-j)}, \mathbf{y}) \sim B(1, P(d_j = 1|(\mathbf{d}_{(-j)}, \mathbf{y})),\ j = 1, ..., p$

Special case: Normal Means problem

$$y_i = \mu_i + \epsilon_i, \quad i = 1, ..., n, \quad \epsilon \overset{i.i.d.}{\sim} N(0, \sigma^2) \qquad (X = I_n)$$

Stein phenomenon: $\hat{\mu}_i = y_i$ ("naive" MLE estimate) is inadmissible!

James-Stein estimate: $\hat{\mu}_i^{JS} = \left(1 - \frac{n-2}{\sum_{j=1}^n y_j^2}\right)_+ y_i$

Key extra assumption: μ is "sparse" (to be quantified later).

Optimal strategy – thresholding (Donoho and Johnstone) : *keep* large y_i – they are "signal"; *kill* "small" y_i – they are "noise".

$$\hat{\mu}_i = \begin{cases} y_i, & |y_i| \geq \lambda \\ 0, & |y_i| < \lambda \end{cases}$$

(e.g., universal threshold $\lambda_U = \sigma\sqrt{2 \ln n}$ of Donoho and Johnstone)

MAP estimation

$$\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2+2\sigma^2(1+1/\gamma)\ln\left\{\binom{n}{k}\pi_n^{-1}(k)(1+\gamma)^{\frac{k}{2}}\right\}\to\min_{\hat{\mu},k}\qquad (k=||\hat{\mu}||_0)$$

which is equivalent to

1. $\sum_{i=k+1}^{n} y_{(i)}^2+2\sigma^2(1+1/\gamma)\ln\left\{\binom{n}{k}\pi_n^{-1}(k)(1+\gamma)^{\frac{k}{2}}\right\}\to\min_k$

2. $\hat{\mu}_i^*=\begin{cases} y_i, & |y_i|\geq |y|_{(\hat{k})} \\ 0, & \text{otherwise}\end{cases}$ – data-driven thresholding

Computationally simple: no need in combinatorical search

Sparsity

Assume that the unknown μ is "sparse". How to measure sparsity?

Sparsity

Assume that the unknown μ is "sparse". How to measure sparsity?

- l_0-balls. Number/proportion of non-zero components: $||\mu||_0 = \#\{i : \mu_i \neq 0,\ i = 1, ..., n\}$.

$$l_0[\eta] = \{\mu \in \mathbb{R}^n \ : ||\mu||_0 \leq \eta n\}$$

Sparsity

Assume that the unknown μ is "sparse". How to measure sparsity?

- l_0-balls. Number/proportion of non-zero components: $||\mu||_0 = \#\{i : \mu_i \neq 0,\ i = 1, ..., n\}$.

$$l_0[\eta] = \{\mu \in \mathbb{R}^n \ : ||\mu||_0 \leq \eta n\}$$

- weak l_p-balls. Proportion of "large" components:

$$m_p[\eta] = \{\mu \in \mathbb{R}^n \ : |\mu|_{(i)} \leq \sigma\eta(n/i)^{1/p},\ i = 1, ..., n\}$$

$$\frac{\#\{i : (\mu_i/\sigma) \geq \Delta\}}{n} \leq \left(\frac{\eta}{\Delta}\right)^p$$

Sparsity

Assume that the unknown μ is "sparse". How to measure sparsity?

- l_0-balls. Number/proportion of non-zero components: $||\mu||_0 = \#\{i : \mu_i \neq 0,\ i = 1, ..., n\}$.

$$l_0[\eta] = \{\mu \in \mathbb{R}^n \ : ||\mu||_0 \leq \eta n\}$$

- weak l_p-balls. Proportion of "large" components:

$$m_p[\eta] = \{\mu \in \mathbb{R}^n \ : |\mu|_{(i)} \leq \sigma\eta(n/i)^{1/p},\ i = 1, ..., n\}$$

$$\frac{\#\{i : (\mu_i/\sigma) \geq \Delta\}}{n} \leq \left(\frac{\eta}{\Delta}\right)^p$$

- (strong) l_p-balls. l_p-norm: $l_p[\eta] = \{\mu \in \mathbb{R}^n \ : \frac{1}{n}\sum_{i=1}^n |\mu_i|^p \leq (\sigma\eta)^p\}$

Adaptive optimality of MAP estimator

Sparsity Zones:

1. $\eta \not\to 0$ – dense case
2. $\eta \to 0$ – sparse case
3. $\eta < n^{-1/\min(2,p)}\sqrt{\log n} \quad (p > 0)$ – super-sparse case

Adaptive optimality of MAP estimator

Sparsity Zones:

1. $\eta \not\to 0$ – dense case
2. $\eta \to 0$ – sparse case
3. $\eta < n^{-1/\min(2,p)}\sqrt{\log n}$ $(p > 0)$ – super-sparse case

Theorem. *Assume Assumption (P) and that*

1. $\pi_n(0) \geq n^{-c_1}$ *for some* $c_1 > 0$

2. $\pi_n(k) \geq (k/n)^{c_2 k}$ *for all* $k = 1, \ldots, e^{-c(\gamma)}n$ *for some* $c_2 > 0$

3. $\pi_n(n) \sim e^{-c(\gamma)n}$

Then, the MAP estimator $\hat{\mu}^*$ *is asymptotically minimax simultaneously for all dense and sparse (though not super-sparse) balls, that is, for all* p *and* $\eta > n^{-1/\min(p,2)}\sqrt{\log n}$.

Main Take-Away Messages

AM1 MAP model selector implies a wide class of penalized least squares estimators with various complexity penalties

Main Take-Away Messages

AM1 MAP model selector implies a wide class of penalized least squares estimators with various complexity penalties

AM2
- Neither linear complexity penalties (e.g., AIC, RIC, BIC), nor Lasso and Dantzig estimators can “kill two birds with one stone” (sparse and dense cases) – bad news
- There exists the class of priors and associated nonlinear penalties (e.g., $2k \ln(p/k)$-type) that do yield such a wide adaptivity range – good news

Main Take-Away Messages

AM1 MAP model selector implies a wide class of penalized least squares estimators with various complexity penalties

AM2
- Neither linear complexity penalties (e.g., AIC, RIC, BIC), nor Lasso and Dantzig estimators can “kill two birds with one stone” (sparse and dense cases) – bad news
- There exists the class of priors and associated nonlinear penalties (e.g., $2k\ln(p/k)$-type) that do yield such a wide adaptivity range – good news

AM3 Multicollinearity – “curse” for model identification or coefficients estimation but may be “blessing” for mean vector estimation

Main Take-Away Messages

AM1 MAP model selector implies a wide class of penalized least squares estimators with various complexity penalties

AM2
- Neither linear complexity penalties (e.g., AIC, RIC, BIC), nor Lasso and Dantzig estimators can “kill two birds with one stone” (sparse and dense cases) – bad news
- There exists the class of priors and associated nonlinear penalties (e.g., $2k \ln(p/k)$-type) that do yield such a wide adaptivity range – good news

AM3 Multicollinearity – “curse” for model identification or coefficients estimation but may be “blessing” for mean vector estimation

AM4 SSVS can be an alternative computational tool for model selection procedures (further study is needed)

Thank You!