On choosing the size of data perturbation in adaptive model selection

Hung Chen

Department of Mathematics, NTU
Joint work with and Su-Yun Huang and Chiuan-Fa Tang
2011 Taipei International Statistical Symposium

12/19/2011

Motivation: Try to understand LASSO

Consider orthogonal-components regression with iid normal errors. (i.e., $\mathbf{X}_k^T \mathbf{X}_k = \mathbf{I}$ for all k.)

- $\hat{\beta}_k \sim N(\beta_k, \sigma^2)$ and $\hat{\beta}_k$'s are independent.
- What can be achieved by Lasso can be understood easily by spacing and order statistics of $\hat{\beta}_k$.

Complexity of three modeling processes

Assume that $\beta_k = 0$ for all k and let $|\hat{\beta}|_{(j)}$ denote the jth order statistic of β_k's.
Three cases will be considered.

Case 1. Wavelet with hard thresholding: Individual pixels in an image are marked as object pixels if their value is greater than some threshold value and as background pixels otherwise.

Case 2. Nested linear models with C_p in which the predictors are pre-ordered such that the index of predictors are pre-assigned.

Case 3. All subset selection with C_p in which the assigned importance of predictors are data-driven.

Case 1: Select λ with hard thresholding.

Consider $Y_k = \beta_k + \epsilon_k$ where $\beta_k = 0$ for all k.

- $\hat{\beta}_k^2 \sim \sigma^2 \chi_1^2$ for all k. (For simplicity, consider $\sigma^2 = 1$ from now on.)
- Model selection:

$$|\hat{\beta}_k| = \begin{cases} 0 & \text{if } |\hat{\beta}_k(\lambda)| \leq \lambda \\ \hat{\beta}_k & \text{otherwise} \end{cases}$$

- When we decrease λ from $|\hat{\beta}|_{(K)}$ to $|\hat{\beta}|_{(1)}$, the number of kept predictors decreases from K to 1. Namely,

$$\hat{\beta}_k^{(m)} = I(|\hat{\beta}_k| \geq \lambda)\hat{\beta}_k, \quad k = 1, \ldots, K,$$

  where $I(\cdot)$ is the indicator function.
- Note that $RSS(M_k) - RSS(M_{k+1}) = |\hat{\beta}|_{(k+1)}^2$ which is not distributed according to χ_1^2.

Fixed penalty with model selection

Consider nested linear regression models
$\mathcal{M} = \{M_k, k = 1, \ldots, K\}$.

- For model M_k, $\beta_j \neq 0$ for $j \leq k_0$ and $\beta_j = 0$ for $j > k$.
  - β's are estimated by the **least square method** and
  - $\boldsymbol{\mu}$ is estimated by $\hat{\boldsymbol{\mu}}_k = P_k \mathbf{Y}$, where P_k is the projection matrix corresponding to model M_k.
  - Its residual sum of squares is defined as

$$RSS(M_k) = (\mathbf{Y} - \hat{\boldsymbol{\mu}}_k)^T (\mathbf{Y} - \hat{\boldsymbol{\mu}}_k).$$

- If AIC (Mallows' C_p) is used to score models, choose the model $\hat{M}$ by minimizing

$$RSS(M_k) + 2|M_k|\sigma^2$$

with respect to all competing models $\mathcal{M}$, where $|M_k|$ is the size of M_k.

Adaptive choice of penalty λ

Q: Is it possible to come up an adaptive choice of λ

$$\min_{M_k \in \mathcal{M}} RSS(M_k) + \lambda |M_k| \sigma^2$$

to achieve model selection consistency on flexible configuration of n and $\mathcal{M}$?

- C_p: Overfitting is not severe.
  - When true model is among $\mathcal{M}$ with $1 \leq k_0 \leq K$, Woodroofe (1982, *AS*) and Zhang (1992, *JASA*) gave a detailed analysis on the performance of C_p.
- BIC replaces 2 with $\log n$ which leads to a consistent selection.

Unbiased risk estimate, covariance inflation and GDF

What is the prediction error of the model $\hat{M}(\lambda)$ which is the minimizer of $RSS(M_k) + \lambda |M_k| \sigma^2$ with respect to all competing models $\mathcal{M} = \{M_k, k = 1, \ldots, K\}$?
Note that

$$\frac{1}{n}\left\{RSS(\hat{M}(\lambda)) + 2E\left[\varepsilon^T\left(\hat{\boldsymbol{\mu}}^{\hat{M}(\lambda)} - \boldsymbol{\mu}_0\right)\right]\right\}$$

is an **unbiased risk estimator** for each $\lambda > 0$. Define

$$g_0(\lambda) = \frac{2}{\sigma^2} E\left[\boldsymbol{\epsilon}^T\left(\hat{\boldsymbol{\mu}}^{\hat{M}(\lambda)} - \boldsymbol{\mu}_0\right)\right] = 2GDF.$$

- $g_0(\lambda)/2$ is defined as the generalized degrees of freedom (GDF) by Ye (1998, *JASA*).

Covariance Inflation (Efron, 1986)

Note that

$$\begin{aligned} RSS &= (\mathbf{Y}-\hat{\mu})^T(\mathbf{Y}-\hat{\mu}) \\ &= (\hat{\mu}-\mu_0)^T(\hat{\mu}-\mu_0) - 2\epsilon^T(\hat{\mu}-\mu_0) + \epsilon^T\epsilon. \end{aligned}$$

When $\hat{\mu} = P\mathbf{Y}$, we have $E\left[\epsilon^T(\hat{\mu}-\mu_0)\right] = \sigma^2 tr(P)$.
Larger model gives a better fit for this particular realization.
Note that

$$\begin{aligned} &(\mathbf{Y}^F-\hat{\mu})^T(\mathbf{Y}^F-\hat{\mu}) \\ &= (\mu_0-\hat{\mu})^T(\mu_0-\hat{\mu}) - 2(\epsilon^F)^T(\hat{\mu}-\mu_0) + (\epsilon^F)^T\epsilon^F \end{aligned}$$

and $E\left[(\epsilon^F)^T(\hat{\mu}-\mu_0)\right] = 0$.
The benefit of larger model is not realized in future prediction. (another realization).

Adaptive penalty selection

Shen and Ye (2002, *JASA*) proposed to choose $\lambda > 0$ to minimize the unbiased risk estimator

$$\hat{\lambda} = \mathit{argmin}_{\lambda>0} \left\{ RSS(\hat{M}(\lambda)) + g_0(\lambda)\sigma^2 \right\}.$$

The resulting selected model is denoted as $\hat{M}(\hat{\lambda})$.
As an attempt to understand their proposal, consider the situation

- BIC is consistent (no underfitting).
- nested competing models
- $\lambda \in [0, \log n]$

Adaptive model selection with nested linear regression models

Is

$$\hat{M}(\hat{\lambda}) = \hat{M}(\log n) = M_{k_0}$$

or $\hat{\lambda} = \log n$?

- Will the adaptive penalty selection in Shen and Ye (2002, *JASA*) increase the penalty ($\lambda = 2$) with C_p to BIC so that the probability of overfitting is reduced?

Note that

- $\hat{\beta}_k^2 \sim \sigma^2 \chi_1^2$
- $RSS(M_k) - RSS(M_{k+1}) = \hat{\beta}_{k+1}^2$ which is distributed according to $\sigma^2 \chi_1^2$

How do we compute $g_0(\lambda)$?

Calculate $g_0(\lambda)$.

It follows from the results of Spitzer (1956), Woodroofe (1982) and Zhang (1992) that, for all $\lambda \in [0, \log n]$,

$$g_0(\lambda) = 2 \sum_{j=1}^{K-K_0} \left[P(\chi^2_{j+2} > j\lambda) \right] + 2K_0.$$

- We just do a simulation study when $K - K_0 = 20$.
- Similar conclusion hold for $K - k_0 > 20$ by using theorem of approximation type instead of asymptotics.
- How big is the following item?

$$2 \sum_{j=21}^{K-k_0} \left[P(\chi^2_{j+2} > j\lambda) \right]$$

x[

Probability of correct selection: URE is not a cure.

When λ increases from 1.0 to 1.5, $g_0(\lambda)$ decreases from 25.35 to 11.22. (Drop rate is too fast).

$\lvert\hat{M}(\hat{\lambda})\rvert$	$\lambda \in [0, \log n]$	$\lambda = 2$	$\lvert\hat{M}(\hat{\lambda})\rvert$	$\lambda \in [0, \log n]$	$\lambda = 2$
k_0	0.5402	0.7130	$k_0 + 11$	0.0143	0.0022
$k_0 + 1$	0.0603	0.1120	$k_0 + 12$	0.0165	0.0020
$k_0 + 2$	0.0360	0.0565	$k_0 + 13$	0.0157	0.0015
$k_0 + 3$	0.0268	0.0348	$k_0 + 14$	0.0167	0.0012
$k_0 + 4$	0.0217	0.0236	$k_0 + 15$	0.0185	0.0012
$k_0 + 5$	0.0179	0.0154	$k_0 + 16$	0.0196	0.0010
$k_0 + 6$	0.0166	0.0114	$k_0 + 17$	0.0202	0.0006
$k_0 + 7$	0.0157	0.0080	$k_0 + 18$	0.0240	0.0004
$k_0 + 8$	0.0155	0.0062	$k_0 + 19$	0.0310	0.0005
$k_0 + 9$	0.0146	0.0046	$k_0 + 20$	0.0433	0.0004
$k_0 + 10$	0.0149	0.0036			

Probability of correct selection:

$\hat{M}(\hat{\lambda}) = M_{k_0+}$	$[0, \log n]$	$[0.5, \log n]$	$[1, \log n]$	$[1.5, \log n]$	$[2, \log n]$
0	0.5457	0.5457	0.5457	0.6483	0.7539
1	0.0565	0.0565	0.0565	0.0681	0.0807
2	0.0312	0.0312	0.0312	0.0386	0.0474
3	0.0262	0.0262	0.0262	0.0320	0.0348
4	0.0239	0.0239	0.0239	0.0283	0.0249
5	0.0188	0.0188	0.0188	0.0227	0.0166
6	0.0156	0.0156	0.0156	0.0190	0.0103
7	0.0134	0.0134	0.0134	0.0169	0.0071
8	0.0136	0.0136	0.0136	0.0157	0.0051
9	0.0140	0.0140	0.0140	0.0151	0.0041
10	0.0155	0.0155	0.0155	0.0132	0.0039
11	0.0155	0.0155	0.0155	0.0107	0.0022
12	0.0153	0.0153	0.0153	0.0106	0.0018
13	0.0163	0.0163	0.0163	0.0097	0.0018
14	0.0177	0.0177	0.0177	0.0080	0.0015
15	0.0185	0.0185	0.0185	0.0074	0.0012
16	0.0210	0.0210	0.0210	0.0070	0.0008
17	0.0242	0.0242	0.0242	0.0074	0.0005
18	0.0212	0.0212	0.0212	0.0069	0.0006
19	0.0307	0.0307	0.0307	0.0065	0.0005
20	0.0452	0.0452	0.0452	0.0079	0.0003

Some heuristics:

Consider the case that $K - k_0 = 20$ and $\lambda = 2$.

- For one realization, we have 3 observations which are greater than 2 and 2 observations fall between 1.5 and 2. (i.e. $V_1 = 4.7$, $V_9 = 2.6$, $V_{13} = 1.8$, $V_{14} = 7.2$, and $V_{15} = 1.7$.)
- Minimum of random process $\{S_j(2), 0 \leq j \leq 20\}$ occurs at $\hat{j}(2) = 1$ for this realization.
  - Include one extra predictor x_{k_0+1}. (Note that $S_0(2) = 0$.)
- Let $N(\lambda)$ denote the number of V_j which are greater than λ.
  - Note that $N(2) \sim Bin(20, 0.1573)$
- $S_j(2)$: positive drift
  - $\hat{j}(2)$ cannot be large.

AMS improves when $\lambda \geq 2$.

Adaptive choice of λ with GDF over $[0, \log n]$: $K - k_0 = 20$

- When $\lambda = 1$, it is expected that it leads an overfitting model with lots of superfluous covariates as comparing a χ^2_1 random variable) to 1 repeatedly.
- Finding by simulation: Adaptive penalty selection of λ over $[0, \log n]$ not only cannot improve over C_p but decreases the probability of selecting M_{k_0} to about 54% as shown in previous slides.

Adaptive choice of λ with GDF over $[2, \log n]$: $K - k_0 = 20$

$\hat{\lambda}$ is defined as follows:

$$\hat{\lambda} = \min_{\lambda \in [2, \log n]} \left\{ RSS(\hat{M}(\lambda)) + g_0(\lambda)\sigma^2 \right\}.$$

Refer to the table presented in next slide.

- Adaptive penalty selection improves C_p but not much.
- Improves over C_p by increasing the probability of correct selection in the range of 3% to 4%.

Probability of correct selection

The reduction from 0.1120 to 0.0838 can be supported by an approximation type of theorem. (It is around 0.0340.)

$\lvert\hat{M}(\hat{\lambda})\rvert$	$\lambda \in [2, \log n]$	$\lambda = 2$	$\lvert\hat{M}(\hat{\lambda})\rvert$	$\lambda \in [2, \log n]$	$\lambda = 2$
k_0	0.7484	0.7130	$k_0 + 11$	0.0022	0.0022
$k_0 + 1$	0.0838	0.1120	$k_0 + 12$	0.0020	0.0020
$k_0 + 2$	0.0505	0.0565	$k_0 + 13$	0.0015	0.0015
$k_0 + 3$	0.0341	0.0348	$k_0 + 14$	0.0012	0.0012
$k_0 + 4$	0.0233	0.0236	$k_0 + 15$	0.0011	0.0012
$k_0 + 5$	0.0153	0.0154	$k_0 + 16$	0.0010	0.0010
$k_0 + 6$	0.0113	0.0114	$k_0 + 17$	0.0006	0.0006
$k_0 + 7$	0.0080	0.0080	$k_0 + 18$	0.0004	0.0004
$k_0 + 8$	0.0062	0.0062	$k_0 + 19$	0.0005	0.0005
$k_0 + 9$	0.0046	0.0046	$k_0 + 20$	0.0004	0.0004
$k_0 + 10$	0.0036	0.0036			

Issue: Data perturbation with $\tau/\sigma = 0.1$

Probability of correct selection:

$\lvert\hat{M}(\hat{\lambda})\rvert$	$\lambda \in [0, \log n]$	$\lambda = 2$	$\lvert\hat{M}(\hat{\lambda})\rvert$	$\lambda \in [0, \log n]$	$\lambda = 2$
k_0	0.8015	0.7092	$k_0 + 11$	0.0022	0.0019
$k_0 + 1$	0.0758	0.1103	$k_0 + 12$	0.0016	0.0016
$k_0 + 2$	0.0355	0.0584	$k_0 + 13$	0.0014	0.0012
$k_0 + 3$	0.0223	0.0359	$k_0 + 14$	0.0010	0.0009
$k_0 + 4$	0.0138	0.0247	$k_0 + 15$	0.0008	0.0009
$k_0 + 5$	0.0097	0.0165	$k_0 + 16$	0.0013	0.0008
$k_0 + 6$	0.0065	0.0119	$k_0 + 17$	0.0016	0.0007
$k_0 + 7$	0.0053	0.0090	$k_0 + 18$	0.0027	0.0008
$k_0 + 8$	0.0027	0.0055	$k_0 + 19$	0.0042	0.0005
$k_0 + 9$	0.0036	0.0054	$k_0 + 20$	0.0042	0.0001
$k_0 + 10$	0.0023	0.0038			

GDF, data perturbation, and Stein's lemma

How do we compute GDF for general modeling process?

Definition

The GDF for a modeling procedure $\mathcal{M}$ are given by $GDF(\mathcal{M}) = \sum_{i=1}^{n} h_i^{\mathcal{M}}(\boldsymbol{\mu})$, where

$$\begin{aligned} h_i^{\mathcal{M}}(\boldsymbol{\mu}) &= \frac{\partial E_{\boldsymbol{\mu}}[\hat{\mu}_i^{\mathcal{M}}(\mathbf{Y})]}{\partial \mu_i} = \lim_{\delta \to 0} E_{\boldsymbol{\mu}} \left[\frac{\hat{\mu}_i^{\mathcal{M}}(\mathbf{Y} + \delta \mathbf{e}_i) - \hat{\mu}_i^{\mathcal{M}}(\mathbf{Y})}{\delta} \right] \\ &= \frac{1}{\sigma^2} E \left[\hat{\mu}_i^{\mathcal{M}}(\mathbf{Y})(Y_i - \mu_i) \right] = \frac{1}{\sigma^2} cov(\hat{\mu}_i^{\mathcal{M}}(\mathbf{Y}), Y_i - \mu_i), \end{aligned}$$

where $\mathbf{e}_i$ is the ith column of $\mathbf{I}_n$.

Data perturbation

Shen and Ye (2002) propose to compute $2GDF(\lambda)$ by data perturbation method. The procedure is as follows:
Repeat $t = 1, \ldots, T$.

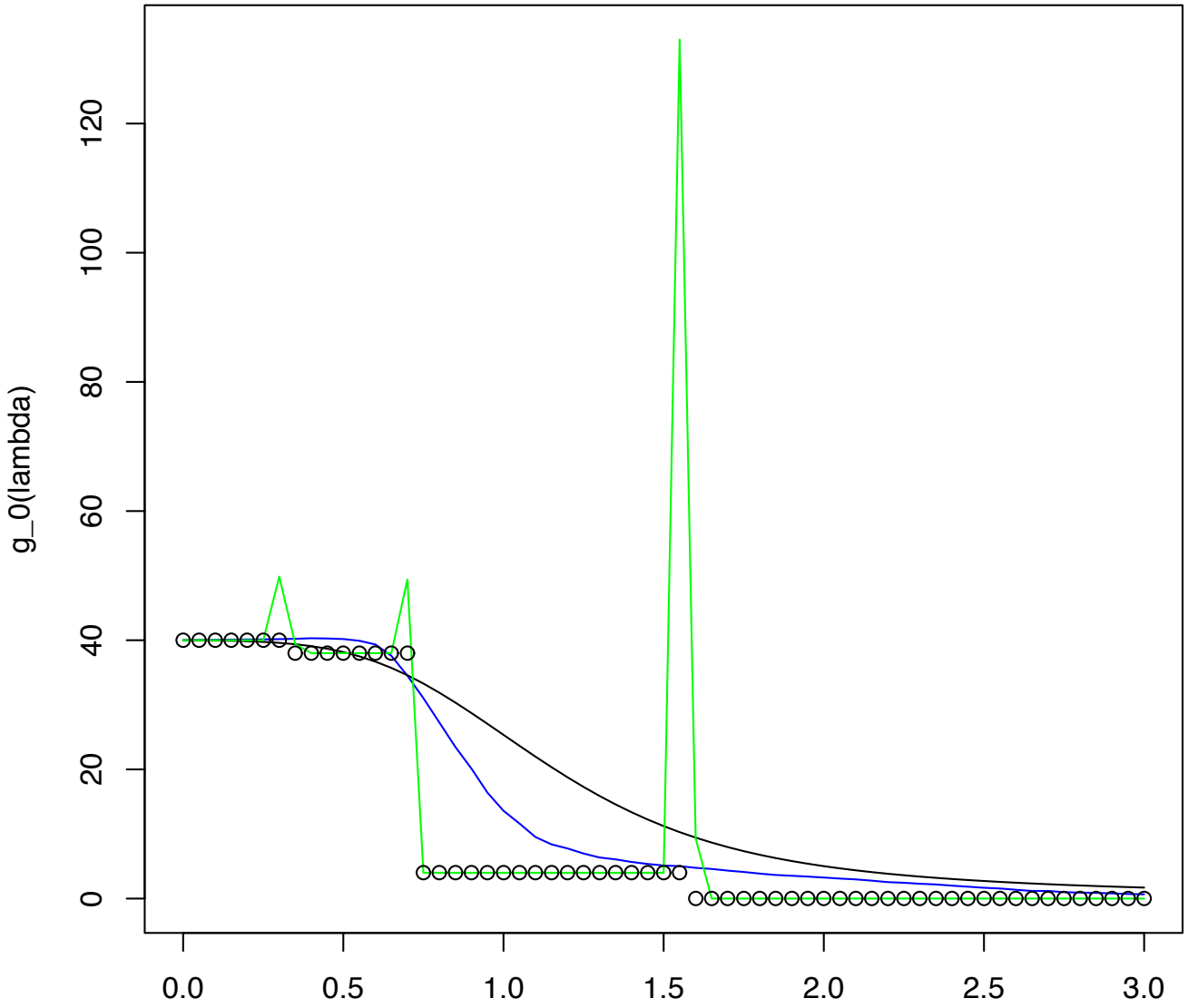
- Generate the perturbed dataset $\mathbf{y} + \boldsymbol{\delta}_t$ where $\boldsymbol{\delta}_t = (\delta_{t1}, \ldots, \delta_{tn}) \in R^n$, $1 \leq t \leq T$, from a normal distribution $N(0, \tau^2)$.
- Evaluate $\hat{\boldsymbol{\mu}}^{\mathcal{M}}(\mathbf{y} + \boldsymbol{\delta})$ based on the modeling procedure $\mathcal{M}$.

Calculate $\hat{h}_i^{\mathcal{M}}$ as the regression slope from

$$\hat{\mu}_i^{\mathcal{M}}(\mathbf{y} + \boldsymbol{\delta}) = \alpha + \hat{h}_i^{\mathcal{M}} \delta_{ti}, \quad t = 1, \ldots, T.$$

The estimate of $GDF(\mathcal{M})$ is $\sum_i \hat{h}_i^{\mathcal{M}}$.

nested regression: $\tau/\sigma = 0.01$(Green),0.5(Blue), and 100(Black)



When $\tau/\sigma \to 0$, what happens?

Answer:

- It is equivalent to adding constraint $\lambda \geq 2$ or do the adaptive choice of λ over $[2, \infty)$. This is the least amount of added penalty suggested in Mallows (1973, *Technometrics*).
  - Why? Refer to previous slide or next slide.
- sensitivity analysis: For another realization of $\mathbf{Y}$, are you comfortable with your modeling procedure leading to a different choice of model?
  - If not, this particular choice of λ cannot be a good one?

For subset regression with λ, it adds the penalty to achieve unbiased risk estimate for that choice of λ.

Replace $\lambda \in [0, \infty)$ by adaptively choice $\lambda \in [2, \infty)$

Recall that $K - k_0 = 20$ and $\lambda = 1.5$.

- For one realization, we have the reduction of RSS for adding one more predictor $4.7, 0.62, 0.24, 0.46, 1.2, 0.2, 0.8, 0.54, 2.6, 1.2, 0.022, 1.2, 1.8, 7.2, 1.7, 0.02, 1.3, 0.096, 0.31, 0.3$.
- Add up complexity with $\lambda = 1.5$. We have

$$-3.2, -2.32, -1.06, -0.02, 0.28, 1.58, 2.28, 3.24, \cdots .$$

  The minimum occurs at -3.2 so that one zero-coefficient predictor will be included.
- Note that $(\sqrt{4.7} - \sqrt{0.62})/0.1 \approx 13.8$.

It is equivalent to saying that this modeling procedure goes with this particular data realization will always choose one more predictor.

- Adding GDF is equivalent to replacing $\lambda = 1.5$ by $\lambda = 2.0$.

Research Agenda:

Choose τ/σ and the number of replications T to address the questions raised in Shen and Ye (2002) and Breiman (1992, 1995, 1996).

Recall that $\mathbf{Y} \sim N(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$ and $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$.

- Consider the commonly used linear estimator $\hat{\boldsymbol{\mu}}(\mathbf{Y}) = \mathbf{P}\mathbf{Y} = (\hat{\mu}_1(\mathbf{Y}), \ldots, \hat{\mu}_n(\mathbf{Y}))^T$ where $\mathbf{P} = (p_{ij})_{n \times n}$ is not necessarily to be a projection matrix.
  - Those $p'_{ij}s$ are constants which can depend on the choice of model but is independent of $\boldsymbol{\epsilon}$.

Data perturbation provides an unbiased estimate of $GDF(\mathcal{M})$ for all τ.

Evaluate the derivative of the smooth estimator $E[\hat{\boldsymbol{\mu}}(\mathbf{y}+\boldsymbol{\delta})]$.

- $\phi_\tau(x)$: density function of $N(0,\tau^2)$, $\phi'_\sigma(x) = -x\phi_\sigma(x)/\sigma^2$

$$\lim_{h\to 0} \frac{E[\hat{\mu}_i(\mathbf{y}+\boldsymbol{\delta}+h\mathbf{e}_j)] - E[\hat{\mu}_i(\mathbf{y}+\boldsymbol{\delta})]}{h}$$

$$= \lim_{h\to 0} \frac{1}{h} \int [\hat{\mu}_i(\mathbf{y}+\boldsymbol{\delta}+h\mathbf{e}_j) - \hat{\mu}_i(\mathbf{y}+\boldsymbol{\delta})] \prod_{\ell=1}^{n} \phi_\tau(\delta_\ell)\, d\boldsymbol{\delta}$$

$$= \lim_{h\to 0} \frac{1}{h} \int \left\{ \hat{\mu}_i(\mathbf{y}+\boldsymbol{\delta}) \left[\phi_\tau(\delta_j - h)) - \phi_\tau(\delta_j)\right] \prod_{\ell\neq j} \phi_\tau(\delta_\ell) \right\} d\boldsymbol{\delta}$$

$$= \int \left[\hat{\mu}_i(\mathbf{y}+\boldsymbol{\delta}) \frac{-d\phi(\delta_j)}{d\delta_j} \prod_{\ell\neq j} \phi_\tau(\delta_\ell) \right] d\boldsymbol{\delta}$$

$$= \frac{1}{\tau^2} \int [\hat{\mu}_i(\mathbf{y}+\boldsymbol{\delta}) \cdot \delta_j] \prod_{\ell=1}^{n} \phi_\tau(\delta_\ell)\, d\boldsymbol{\delta}.$$

cont.

Evaluate $E_{\mathbf{Y}}\left\{\lim_{h\to 0}\left[E_{\boldsymbol{\delta}}[\hat{\mu}_i(\mathbf{Y}+\boldsymbol{\delta}+h\mathbf{e}_j)]-E[\hat{\mu}_i(\mathbf{Y}+\boldsymbol{\delta})]\right]/h\right\}$.
Note that

$$
\begin{aligned}
& E\left[\lim_{h\to 0}\frac{E[\hat{\mu}_i(\mathbf{Y}+\boldsymbol{\delta}+h\mathbf{e}_j)]-E[\hat{\mu}_i(\mathbf{Y}+\boldsymbol{\delta})]}{h}\right] \\
= & \int\frac{1}{\tau^2}\left[\int[\hat{\mu}_i(\mathbf{y}+\boldsymbol{\delta})\cdot\delta_j]\prod_{\ell=1}^{n}\phi_\tau(\delta_\ell)\prod_{\ell=1}^{n}\phi_\sigma(y_\ell-\mu_\ell)\,d\boldsymbol{\delta}\right]d\mathbf{y} \\
= & \frac{1}{\tau^2}\int\left[\left(\int\hat{\mu}_i(\mathbf{y}+\boldsymbol{\delta})\prod_{\ell=1}^{n}\phi_\sigma(y_\ell-\mu_\ell)\,d\mathbf{y}\right)\cdot\delta_j\right]\prod_{\ell=1}^{n}\phi_\tau(\delta_\ell)\,d\boldsymbol{\delta} \\
= & \frac{1}{\tau^2}\int\left(\int\hat{\mu}_i((\boldsymbol{\mu}+\boldsymbol{\delta})+(\mathbf{y}-\boldsymbol{\mu}))\prod_{\ell=1}^{n}\phi_\sigma(y_\ell-\mu_\ell)\,d\mathbf{y}\right)\cdot\delta_j\prod_{\ell=1}^{n}\phi_\tau(\delta_\ell)\,d\boldsymbol{\delta} \\
= & \frac{1}{\tau^2}\int\{E[\hat{\mu}_i((\boldsymbol{\mu}+\boldsymbol{\delta})+(\mathbf{Y}-\boldsymbol{\mu}))]\}\cdot\delta_j\prod_{\ell=1}^{n}\phi_\tau(\delta_\ell)\,d\boldsymbol{\delta}
\end{aligned}
$$

cont.

$$
\begin{aligned}
& E\left[\lim_{h\to 0}\frac{E[\hat{\mu}_i(\mathbf{Y}+\boldsymbol{\delta}+h\mathbf{e}_j)]-E[\hat{\mu}_i(\mathbf{Y}+\boldsymbol{\delta})]}{h}\right] \\
= \quad & \frac{1}{\tau^2}E\left\{E[\hat{\mu}_i((\boldsymbol{\mu}+\boldsymbol{\delta})+(\mathbf{Y}-\mu))]\right\}\delta_j \\
= \quad & E\left[\lim_{h\to 0}\frac{E[\hat{\mu}_i((\boldsymbol{\mu}+\boldsymbol{\delta})+(\mathbf{Y}-\boldsymbol{\mu})+h\mathbf{e}_j)]-E[\hat{\mu}_i((\boldsymbol{\mu}+\boldsymbol{\delta})+(\mathbf{Y}-\boldsymbol{\mu}))]}{h}\right] \\
= \quad & E\left[\lim_{h\to 0}\frac{E[\hat{\mu}_i(\mathbf{Y}+\boldsymbol{\delta}+h\mathbf{e}_j)]-E[\hat{\mu}_i(\mathbf{Y}+\boldsymbol{\delta})]}{h}\right] \\
= \quad & \lim_{h\to 0}E\left[\frac{E[\hat{\mu}_i(\mathbf{Y}+\boldsymbol{\delta}+h\mathbf{e}_j)]-E[\hat{\mu}_i(\mathbf{Y}+\boldsymbol{\delta})]}{h}\right] \\
= \quad & \lim_{h\to 0}E\left[\frac{\hat{\mu}_i(\mathbf{Y}^*+h\mathbf{e}_j)-\hat{\mu}_i(\mathbf{Y}^*)}{h}\right].
\end{aligned}
$$

Here $\mathbf{Y}^* = \mathbf{Y}+\boldsymbol{\delta}$ and $\mathbf{Y}^* \sim N(\boldsymbol{\mu}, (\sigma^2+\tau^2)\mathbf{I_n})$.

cont.

By the same argument, we have

$$Cov(\hat{\mu}_i^{\mathcal{M}}(\mathbf{Y}), \epsilon_j) = \int \hat{\mu}_i^{\mathcal{M}}(\mathbf{y})(y_j - \mu_j) \prod_{\ell=1}^{n} \phi(y_\ell - \mu_\ell)\, d\mathbf{y}$$

$$= \lim_{h \to 0} \frac{E[\hat{\mu}_i^{\mathcal{M}}(\mathbf{Y} + h\mathbf{e}_j)] - E[\hat{\mu}_i^{\mathcal{M}}(\mathbf{Y})]}{h}.$$

It states that Stein's lemma can be applied to covariance inflation and the feasibility of data perturbation for giving an unbiased estimate of covariance inflation occurred in model selection.

- Address the effect that $Var(Y_i) = \sigma^2$ while $Var(Y_i^*) = \sigma^2 + \tau^2$. Consider

$$argmin_{M_k \in \mathcal{M}} \left\{ (\mathbf{y} + \boldsymbol{\delta})^T (\mathbf{I}_n - P_k)(\mathbf{y} + \boldsymbol{\delta}) + \lambda |M_k| (\sigma^2 + \tau^2) \right\}$$

where $|M_k|$ denotes the number of unknown parameters in model M_k.

GDF (unbiased risk estimate) or Data Perturbation?

- For any given $\tau > 0$, data perturbation method gives an unbiased estimator of GDF when $\boldsymbol{\mu}$ is a zero vector.
- Otherwise, it gives a biased estimator of GDF unless the fitting procedure is overfitting.

For nested regression with adaptive penalty selection, we can use a two-stage procedure.

- Stage 1: Apply tiny bootstrap (τ/σ is close to 1).
  - T must be large to find a proper range of λ.
- Stage 2: Apply little bootstrap ($\tau/\sigma - 1 \in [0.5, 1]$).

Summary: daptive penalty selection for nested regression

- When $\lambda \in [2, \log n]$, there are about 75% to choose the true model.
- The probability of selecting correct model decreases to 55% if $\lambda \in [1, 2) \cup [2, \log n]$.
  - Just adding GDF won't work for adaptive penalty selection.
- For data perturbation method with small τ/σ, it gives a measure on the stability of selected model to the small error added to the responses.
  - It is equivalent to force the range of λ should be at least 2 as C_p does.
- Choice of τ/σ needs more investigation.
- Data perturbation is more than a computational tool for calculate GDF or covariance inflation.

Issue 1: How should I pick up right τ and T?

Define

$$\mathbf{e}_k = \frac{(\mathbf{I}_n - \mathbf{P}_{k-1})\mathbf{x}_k}{\|(\mathbf{I}_n - \mathbf{P}_{k-1})\mathbf{x}_k\|} \quad \text{and} \quad \xi_k = \frac{\hat{\beta}_k}{\|(\mathbf{I}_n - \mathbf{P}_{k-1})\mathbf{x}_k\|} = \mathbf{Y}^T\mathbf{e}_k$$

for all $k = 1, \ldots, K$.
Denote $(\xi_{K_0+1}, \ldots, \xi_K)^T$ by $\boldsymbol{\xi}$ where $\boldsymbol{\xi} \sim N(0, \mathbf{I}_{K-K_0}\sigma^2)$.
Least-squares estimate based on the perturbed data:
$\hat{\boldsymbol{\beta}}^* = (\hat{\beta}_1^*, \ldots, \hat{\beta}_K^*)^T$. Then

$$\frac{\hat{\beta}_k^*}{\|(\mathbf{I}_n - \mathbf{P}_{k-1})\mathbf{x}_k\|} = (\mathbf{Y}+\boldsymbol{\delta})^T\mathbf{e}_k \quad \text{and} \quad \zeta_k = \frac{\hat{\beta}_k^* - \hat{\beta}_k}{\|(\mathbf{I}_n - \mathbf{P}_{k-1})\mathbf{x}_k\|} = \boldsymbol{\delta}^T\mathbf{e}_k$$

Two Markov chains:

Chain associated with perturbed data:

$$\hat{M}(\lambda)(\boldsymbol{\xi}+\boldsymbol{\zeta})$$
$$= argmax_{M_j \in \mathcal{M}} \left\{ \sum_{i=K_0}^{K} \left[\left(\frac{\xi_i}{\sqrt{\sigma^2+\tau^2}} + \frac{\zeta_i}{\sqrt{\sigma^2+\tau^2}} \right)^2 - \lambda \right] \right\}$$

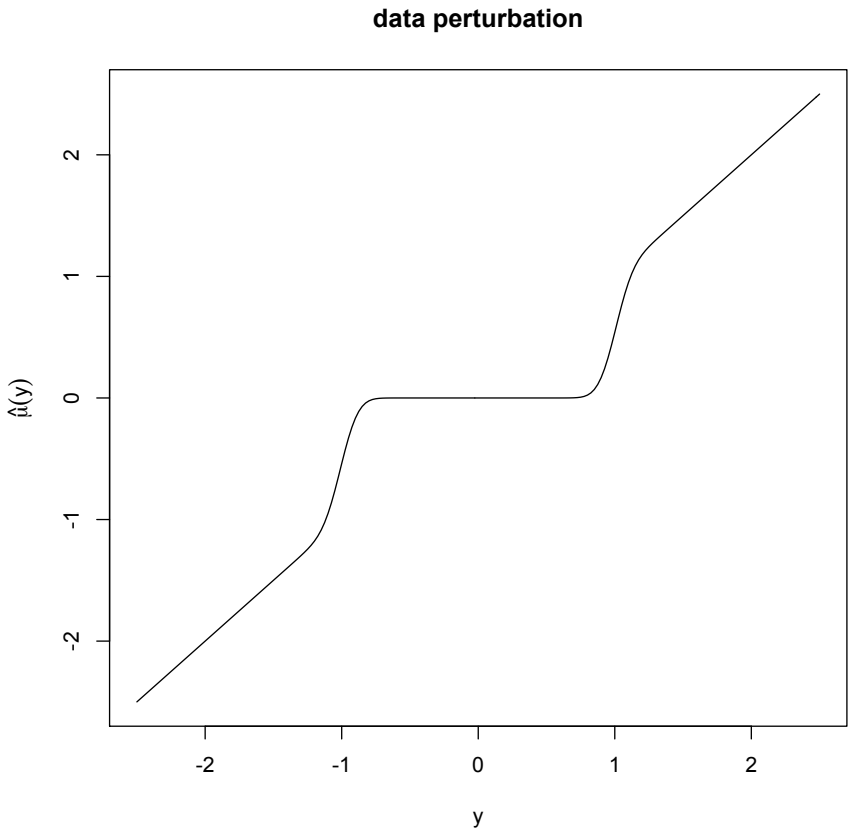
- small perturbation
- It loses symmetry. (noncentral χ^2)

Original chain:

$$\hat{M}(\lambda)(\boldsymbol{\xi}) = argmax_{M \in \mathcal{M}} \left\{ \sum_{i=K_0}^{K} \left[\left(\frac{\xi_i}{\sqrt{\sigma^2}} \right)^2 - \lambda \right] \right\}$$

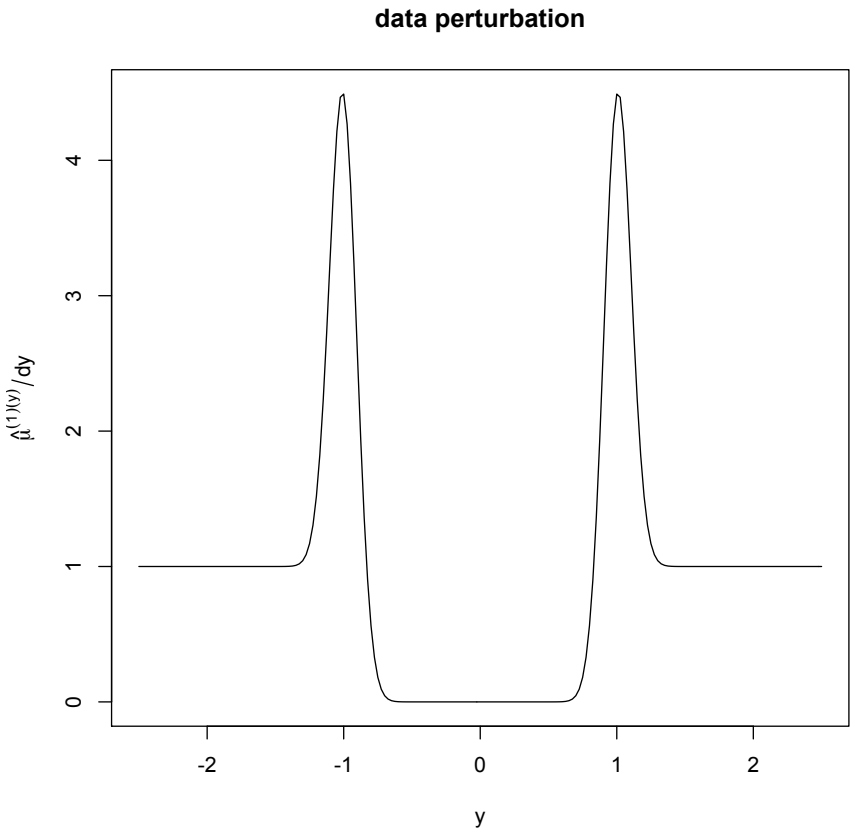
Unsettled Issues

Issue 2: hard thresholding: smoothing $\mu_\lambda^\tau(\cdot)$ with $\tau = 0.1$ and $\lambda = 1$



Unsettled Issues

hard thresholding: generalized derivative of $\mu_\lambda(\cdot)$ with $\tau = 0.1\sigma$ and $\lambda = 1$



Unsettled Issues

Red: $\tau = 0.01$, Yellow: $\tau = 0.1$, Green: $\tau = 0.5$, Blue: $\tau = 0.8$, Black: $\tau = 1$, $n = 1000$, $T = 1000$