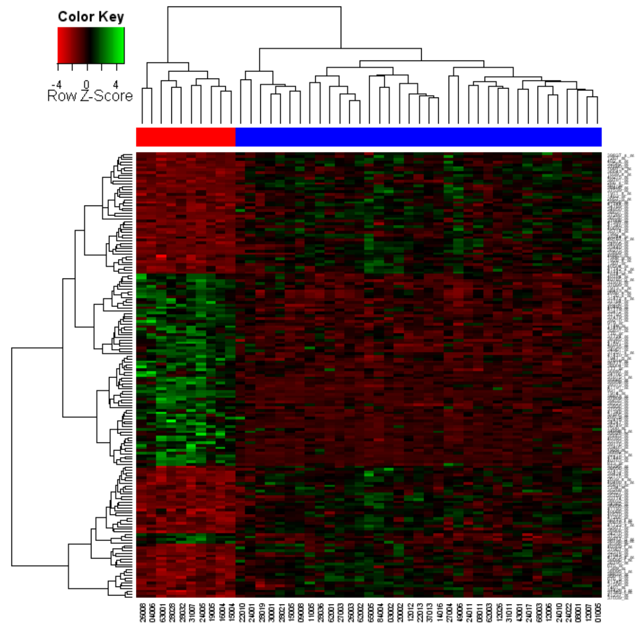
HIGH-DIMENSIONAL VARIABLE SELECTION



Richard Samworth

University of Cambridge

Joint work with Rajen Shah

Why is variable selection important?

Modern technology allows the collection and storage of data on previously unimaginable scales.

Appropriate statistical models and methods are required to extract useful information. When the model dimension p is larger than the sample size n, variable selection is essential for model interpretability.

20th Century data sets

- 1 *bit* is the amount of information stored by a digital device or other physical system that exists in one of two possible distinct states
- 1 *byte* = 8 bits, the number required to encode a single character of text
- 1 *Megabyte* $= 2^{20}$ bytes $\approx 10^6$ bytes encodes:
  - a 1024 $\times$ 1024 pixel bitmap image with 256 colours
  - 6 seconds of uncompressed CD audio
  - a typical English book volume in plain text format

21st Century data sets

- **1 *Terabyte* $= 2^{40}$ bytes $\approx 10^{12}$ bytes encodes:**
  - **the data collected in a single race from a Formula One car**
  - **0.5% of the U.S. Library of Congress**
  - **2% of the data collected by the Hubble space telescope in the last 20 years**

21st Century data sets

- **1 *Terabyte* $= 2^{40}$ bytes $\approx 10^{12}$ bytes encodes:**
  - **the data collected in a single race from a Formula One car**
  - **0.5% of the U.S. Library of Congress**
  - **2% of the data collected by the Hubble space telescope in the last 20 years**
- **1 *Exabyte* $= 2^{60}$ bytes $\approx 10^{18}$ bytes encodes:**
  - **one day's worth of data from the new SKA telescope initiative**
  - **the data from the hippocampi of 400 adult humans**
  - **the data from 40 days of Google searches**

Fields in which huge data sets arise

- Biological sciences: genetics, genomics, proteomics,...
- Text/document classification
- Neuroscience: fMRI, EEG, MEG, PET
- High energy physics: CERN Large Hadron Collider
- Astrophysics
- Communications networks: the internet

A (very) brief history of variable selection

Stein (1956) first realised the power of shrinkage in multi-dimensional problems.

Hoerl and Kennard (1970): Ridge regression (ℓ_2-penalisation)

Donoho and Johnstone (1994), Donoho et al. (1995): Sparsity (in wavelet estimation)

Tibshirani (1996): The Lasso, for simultaneous variable selection and parameter estimation

The Lasso

Consider the linear model

$$\underset{n\times 1}{Y} = \beta_0 \underset{n\times 1}{1_n} + \underset{n\times p}{X} \underset{p\times 1}{\beta} + \underset{n\times 1}{\epsilon}.$$

For $\lambda > 0$, the Lasso estimator is $\hat{\beta}_\lambda$, where $(\hat{\beta}_0, \hat{\beta}_\lambda)$ minimises

$$\frac{1}{2n}\|Y - \beta_0 1_n - X\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

over $(\beta_0, \beta) \in \mathbb{R} \times \mathbb{R}^p$.

Other penalties

More recent work has often focused on alternative penalty functions, e.g.

$$\frac{1}{2n}\|Y - \beta_0 1_n - X\beta\|^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|).$$

Examples include SCAD (Fan and Li, 2001), the elastic net (Zou and Hastie, 2005), MCP (Zhang, 2010). Other work assumes different structures, e.g. pseudo-likelihood models (Fan, S. and Wu, 2009), Group Lasso (Yuan and Lin, 2006).

Stability Selection

Meinshausen and Bühlmann (2010)

Stability Selection is a very general technique designed to improve the performance of a variable selection algorithm.

It is based on aggregating the results of applying a selection procedure to subsamples of the data.

A particularly attractive feature of Stability Selection is the error control provided by an upper bound on the expected number of falsely selected variables.

A general model for variable selection

Let $Z_1, \ldots, Z_n$ be i.i.d. random vectors. We think of the indices S of some components of Z_i as being 'signal variables', and others N as being 'noise variables'.

E.g. $Z_i = (X_i, Y_i)$, with covariate $X_i \in \mathbb{R}^p$, response $Y_i \in \mathbb{R}$ and log-likelihood of the form

$$\sum_{i=1}^{n} L(Y_i, X_i^T \beta),$$

with $\beta \in \mathbb{R}^p$. Then $S = \{k : \beta_k \neq 0\}$ and $N = \{k : \beta_k = 0\}$.

Thus $S \subseteq \{1, \ldots, p\}$ and $N = \{1, \ldots, p\} \setminus S$. A *variable selection procedure* is a statistic $\hat{S}_n := \hat{S}_n(Z_1, \ldots, Z_n)$ taking values in the set of all subsets of $\{1, \ldots, p\}$.

How does Stability Selection work?

For a subset $A = \{i_1, \ldots, i_{|A|}\} \subseteq \{1, \ldots, n\}$**, write**

$$\hat{S}(A) := \hat{S}_{|A|}(Z_{i_1}, \ldots, Z_{i_{|A|}}).$$

Meinshausen and Bühlmann defined

$$\hat{\Pi}(k) = \binom{n}{\lfloor n/2 \rfloor}^{-1} \sum_{\substack{A \subseteq \{1,\ldots,n\} \\ |A| = \lfloor n/2 \rfloor}} \mathbb{1}_{\{k \in \hat{S}(A)\}}.$$

Stability Selection fixes $\tau \in [0, 1]$ **and selects** $\hat{S}^{\mathrm{SS}}_{n,\tau} = \{k : \hat{\Pi}(k) \geq \tau\}$**.**

Error control

Meinshausen and Bühlmann (2010)

Assume that $\{\mathbb{1}_{\{k \in \hat{S}_{\lfloor n/2 \rfloor}\}} : k \in N\}$ **is exchangeable, and that** $\hat{S}_{\lfloor n/2 \rfloor}$ **is not worse than random guessing:**

$$\frac{\mathbb{E}(|\hat{S}_{\lfloor n/2 \rfloor} \cap S|)}{\mathbb{E}(|\hat{S}_{\lfloor n/2 \rfloor} \cap N|)} \geq \frac{|S|}{|N|}.$$

Then, for $\tau \in (\frac{1}{2}, 1]$**,**

$$\mathbb{E}(|\hat{S}^{\mathrm{SS}}_{n,\tau} \cap N|) \leq \frac{1}{2\tau - 1} \frac{(\mathbb{E}|\hat{S}_{\lfloor n/2 \rfloor}|)^2}{p}.$$

Error control discussion

In principle, this theorem helps the practitioner choose the tuning parameter τ. However:

- **The theorem requires two conditions, and the exchangeability assumption is very strong**
- **There are too many subsets to evaluate $\hat{S}^{\mathrm{SS}}_{n,\tau}$ when $n \geq 20$**
- **The bound tends to be rather weak.**

Complementary Pairs Stability Selection

Shah and S. (2013)

Let $\{(A_{2j-1}, A_{2j}) : j = 1, \ldots, B\}$ **be randomly chosen independent pairs of subsets of** $\{1, \ldots, n\}$ **of size** $\lfloor n/2 \rfloor$ **such that** $A_{2j-1} \cap A_{2j} = \emptyset$**.**

Define

$$\hat{\Pi}_B(k) := \frac{1}{2B} \sum_{j=1}^{2B} \mathbb{1}_{\{k \in \hat{S}(A_j)\}},$$

and select $\hat{S}^{\mathrm{CPSS}}_{n,\tau} = \{k : \hat{\Pi}_B(k) \geq \tau\}$**.**

Worst case error control bounds

Let $p_{k,n} = \mathbb{P}(k \in \hat{S}_n)$**. For** $\theta \in [0,1]$**, let** $L_\theta = \{k : p_{k,\lfloor n/2 \rfloor} \leq \theta\}$ **and** $H_\theta = \{k : p_{k,\lfloor n/2 \rfloor} > \theta\}$**.**

If $\tau \in (\frac{1}{2}, 1]$**, then**

$$\mathbb{E}|\hat{S}_{n,\tau}^{\mathrm{CPSS}} \cap L_\theta| \leq \frac{\theta}{2\tau - 1}\mathbb{E}|\hat{S}_{\lfloor n/2 \rfloor} \cap L_\theta|.$$

Moreover, if $\tau \in [0, \frac{1}{2})$**, then**

$$\mathbb{E}|\hat{N}_{n,\tau}^{\mathrm{CPSS}} \cap H_\theta| \leq \frac{1-\theta}{1-2\tau}\mathbb{E}|\hat{N}_{\lfloor n/2 \rfloor} \cap H_\theta|.$$

Illustration and discussion

Suppose $p = 1000$**, and** $q := \mathbb{E}|\hat{S}_{\lfloor n/2 \rfloor}| = 50$**. Then on average, CPSS with** $\tau = 0.6$ **selects no more than a quarter of the variables that have below average selection probability under** $\hat{S}_{\lfloor n/2 \rfloor}$**.**

- **The theorem requires no exchangeability or random guessing conditions**
- **It holds even when** $B = 1$
- **If exchangeability and random guessing conditions do hold, then we recover**

$$\mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,\tau} \cap N| \leq \frac{1}{2\tau - 1}\Big(\frac{q}{p}\Big)\mathbb{E}|\hat{S}_{\lfloor n/2 \rfloor} \cap L_{q/p}| \leq \frac{1}{2\tau - 1}\Big(\frac{q^2}{p}\Big).$$

Proof

Let

$$\tilde{\Pi}_B(k) := \frac{1}{B}\sum_{j=1}^{B} \mathbb{1}_{\{k\in\hat{S}(A_{2j-1})\}}\mathbb{1}_{\{k\in\hat{S}(A_{2j})\}},$$

and note that $\mathbb{E}\{\tilde{\Pi}_B(k)\} = p_{k,\lfloor n/2\rfloor}^2$**. Now**

$$0 \leq \frac{1}{B}\sum_{j=1}^{B}\bigl\{1-\mathbb{1}_{\{k\in\hat{S}(A_{2j-1})\}}\bigr\}\bigl\{1-\mathbb{1}_{\{k\in\hat{S}(A_{2j})\}}\bigr\} = 1-2\hat{\Pi}_B(k)+\tilde{\Pi}_B(k).$$

Thus

$$\begin{aligned}\mathbb{P}\{\hat{\Pi}_B(k) \geq \tau\} \leq \mathbb{P}\bigl\{\tfrac{1}{2}(1+\tilde{\Pi}_B(k)) \geq \tau\bigr\} &= \mathbb{P}\{\tilde{\Pi}_B(k) \geq 2\tau-1\} \\ &\leq \frac{1}{2\tau-1}p_{k,\lfloor n/2\rfloor}^2.\end{aligned}$$

Proof 2

Note that

$$\mathbb{E}|\hat{S}_{\lfloor n/2 \rfloor} \cap L_\theta| = \mathbb{E}\biggl(\sum_{k: p_{k,\lfloor n/2 \rfloor} \leq \theta} \mathbb{1}_{\{k \in \hat{S}_{\lfloor n/2 \rfloor}\}} \biggr) = \sum_{k: p_{k,\lfloor n/2 \rfloor} \leq \theta} p_{k,\lfloor n/2 \rfloor}.$$

It follows that

$$\begin{aligned} \mathbb{E}|\hat{S}_{n,\tau}^{\mathrm{CPSS}} \cap L_\theta| &= \mathbb{E}\biggl(\sum_{k: p_{k,\lfloor n/2 \rfloor} \leq \theta} \mathbb{1}_{\{k \in \hat{S}_{n,\tau}^{\mathrm{CPSS}}\}} \biggr) = \sum_{k: p_{k,\lfloor n/2 \rfloor} \leq \theta} \mathbb{P}(k \in \hat{S}_{n,\tau}^{\mathrm{CPSS}}) \\ &\leq \frac{1}{2\tau - 1} \sum_{k: p_{k,\lfloor n/2 \rfloor} \leq \theta} p_{k,\lfloor n/2 \rfloor}^2 \leq \frac{\theta}{2\tau - 1} \mathbb{E}|\hat{S}_{\lfloor n/2 \rfloor} \cap L_\theta|. \end{aligned}$$

Bounds with no assumptions whatsoever

If $Z_1, \ldots, Z_n$ are not identically distributed, the same bound holds, provided in L_θ we redefine

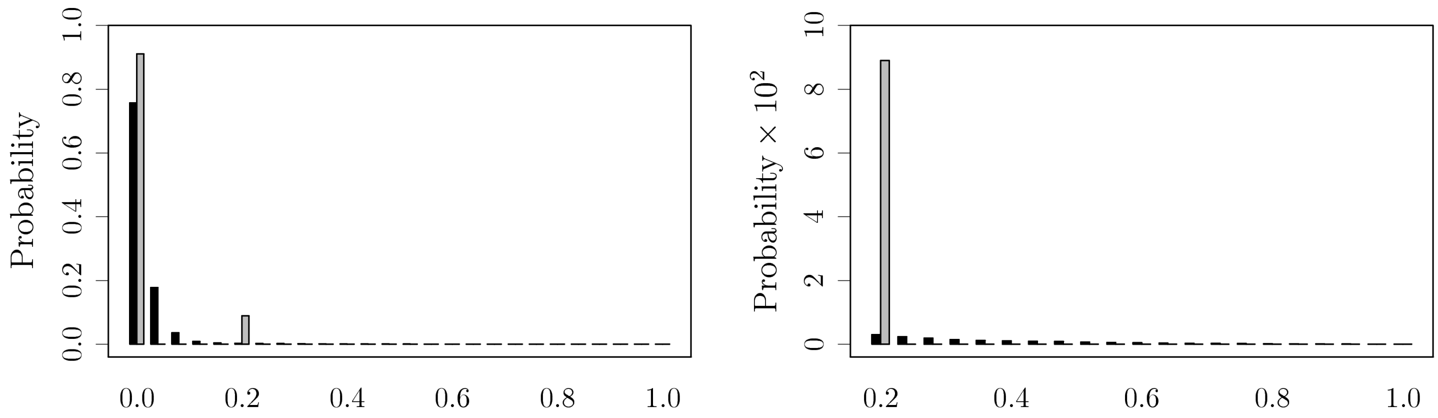
$$p_{k,\lfloor n/2 \rfloor} = \binom{n}{\lfloor n/2 \rfloor}^{-1} \sum_{|A|=n/2} \mathbb{P}\{k \in \hat{S}_{\lfloor n/2 \rfloor}(A)\}.$$

Similarly, if $Z_1, \ldots, Z_n$ are not independent, the same bound holds, with $p^2_{k,\lfloor n/2 \rfloor}$ as the average of

$$\mathbb{P}\{k \in \hat{S}_{\lfloor n/2 \rfloor}(A_1) \cap \hat{S}_{\lfloor n/2 \rfloor}(A_2)\}$$

over all complementary pairs A_1, A_2.

Can we improve on Markov's inequality?



Improved bound under unimodality

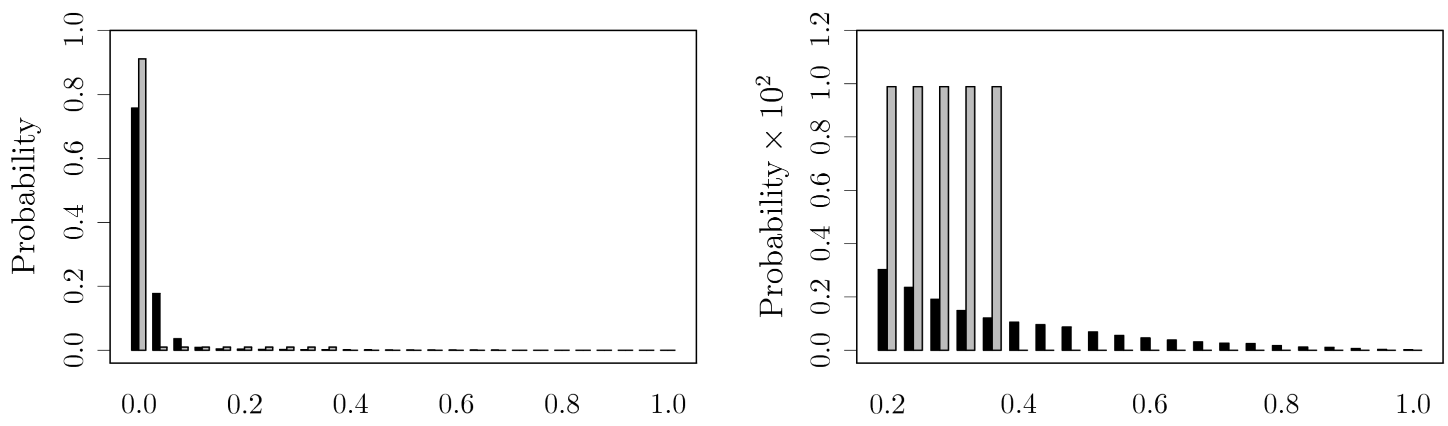
Suppose that the distribution of $\tilde{\Pi}_B(k)$ is unimodal for each $k \in L_\theta$. If $\tau \in \{\frac{1}{2} + \frac{1}{B}, \frac{1}{2} + \frac{3}{2B}, \frac{1}{2} + \frac{2}{B}, \ldots, 1\}$, then

$$\mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,\tau} \cap L_\theta| \leq C(\tau, B)\,\theta\,\mathbb{E}|\hat{S}_{\lfloor n/2 \rfloor} \cap L_\theta|,$$

where, when $\theta \leq 1/\sqrt{3}$,

$$C(\tau, B) = \begin{cases} \dfrac{1}{2(2\tau - 1 - 1/2B)} & \textbf{if } \tau \in (\min(\frac{1}{2} + \theta^2, \frac{1}{2} + \frac{1}{2B} + \frac{3}{4}\theta^2), \frac{3}{4}] \\ \dfrac{4(1 - \tau + 1/2B)}{1 + 1/B} & \textbf{if } \tau \in (\frac{3}{4}, 1]. \end{cases}$$

Extremal distribution under unimodality



The r-concavity constraint

r-concavity provides a continuum of constraints that interpolate between unimodality and log-concavity.

A non-negative function f on an interval $I \subset \mathbb{R}$ is r*-concave* with $r < 0$ if for every $x, y \in I$ and $\lambda \in (0, 1)$,

$$f(\lambda x + (1 - \lambda)y) \geq \{\lambda f(x)^r + (1 - \lambda)f(y)^r\}^{1/r};$$

equivalently iff f^r is convex. A pmf f on $\{0, 1/B, \ldots, 1\}$ is r-concave if the linear interpolant to $\{(i, f(i/B)) : i = 0, 1, \ldots, B\}$ is r-concave. The constraint becomes weaker as r increases to 0.

Further improvements under r-concavity

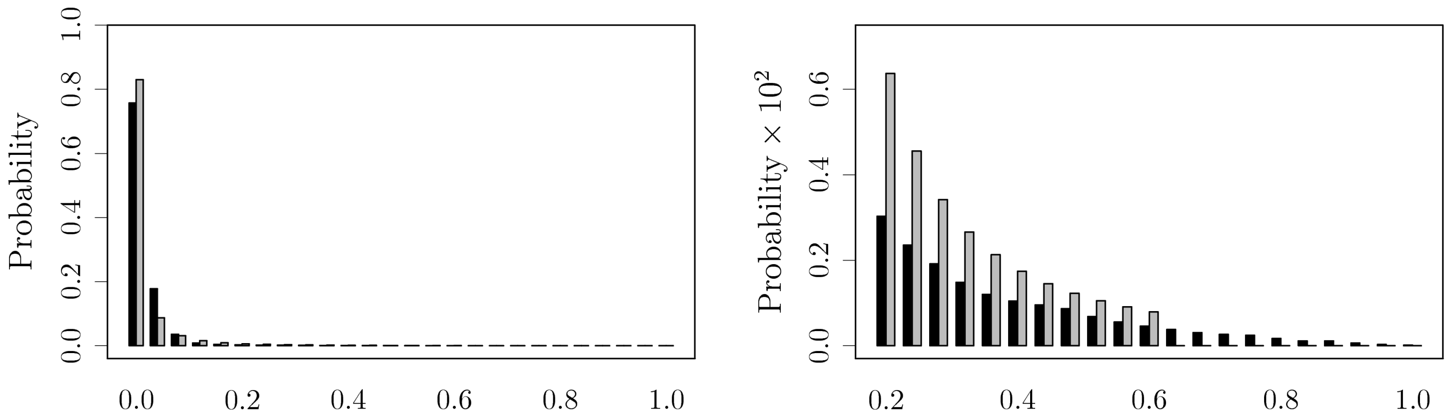
Suppose $\tilde{\Pi}_B(k)$ is r-concave for all $k \in L_\theta$. Then for $\tau \in (\frac{1}{2}, 1]$,

$$\mathbb{E}|\hat{S}_{n,\tau}^{\mathrm{CPSS}} \cap L_\theta| \leq D(\theta^2, 2\tau - 1, B, r)|L_\theta|,$$

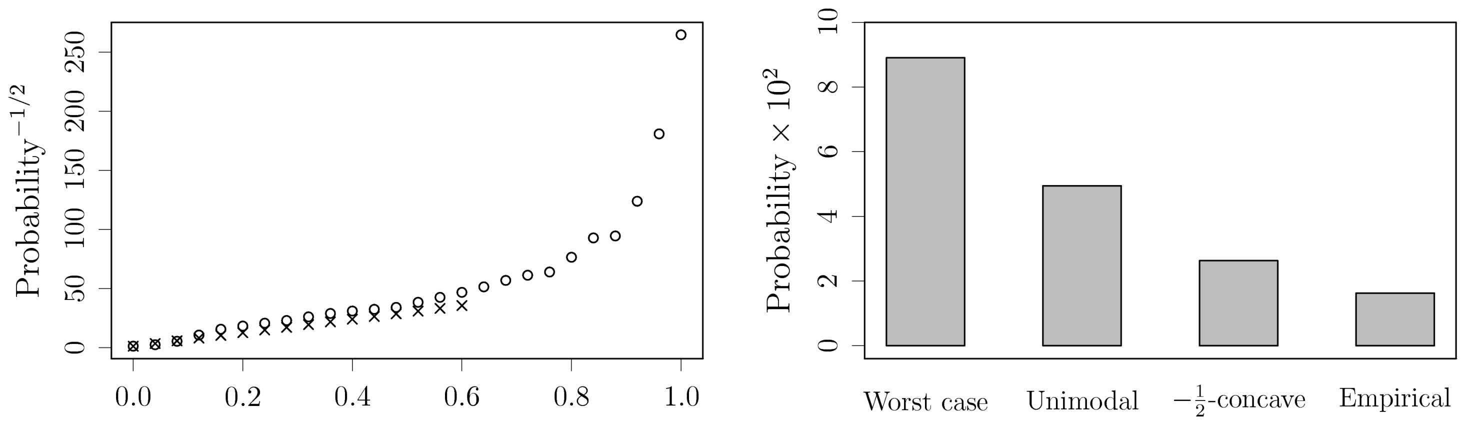
where D can be evaluated numerically.

Our simulations suggest $r = -1/2$ is a safe and sensible choice.

Extremal distribution under r-concavity



$r = -1/2$ is sensible



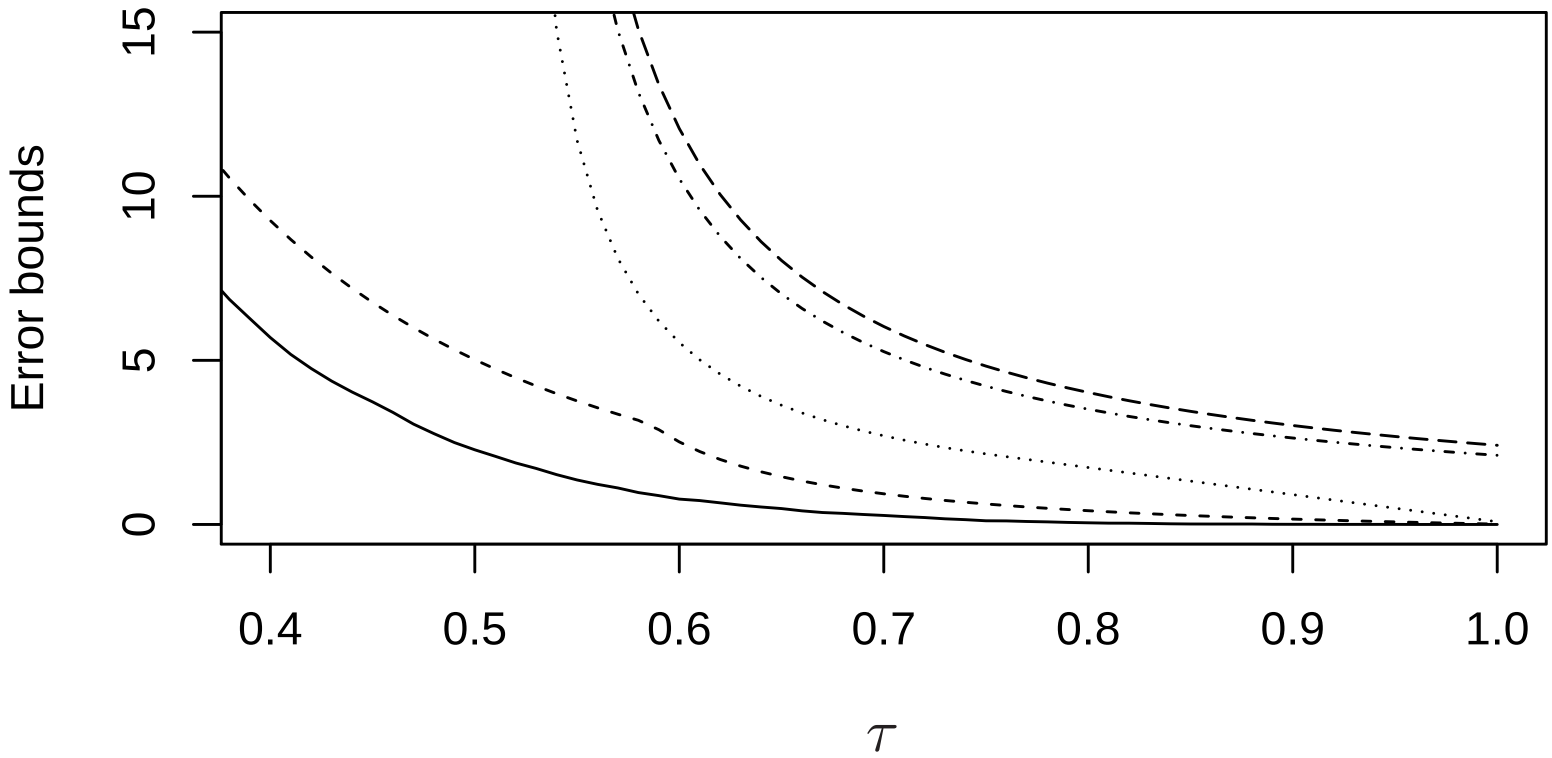
Reducing the threshold τ

Suppose $\tilde{\Pi}_B(k)$ **is** r**-concave for all** $k \in L_\theta$**, and that** $\hat{\Pi}_B(k)$ **is** $-1/4$**-concave for all** $k \in L_\theta$**. Then**

$$\mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,\tau} \cap L_\theta| \leq \min\{D(\theta^2, 2\tau - 1, B, -1/2), D(\theta, \tau, 2B, -1/4)\}|L_\theta|,$$

for all $\tau \in (\theta, 1]$**. (We take** $D(\cdot, t, \cdot, \cdot) = 1$ **for** $t \leq 0$**.)**

Improved bounds



Summary

- Variable selection is one of the most important problems in modern, high-dimensional statistics
- CPSS can be used in conjunction with any variable selection procedure to improve its performance.
- We can bound the average number of low selection probability variables chosen by CPSS under no conditions on the model or base selection procedure
- Under mild conditions, e.g. r-concavity, the bounds can be strengthened, yielding tight error control.
- This allows the practitioner to choose the threshold τ in an effective way.

References

- Donoho, D. L. and Johnstone, I. M. (1994) Ideal spatial adaptation by wavelet shrinkage, *Biometrika*, 81, 425–455.
- Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995) Wavelet Shrinkage: Asymptopia? *J. Roy. Statist. Soc., Ser B (with discussion)*, 57, 301–69.
- Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96, 1348-1360.
- Fan, J., Samworth, R. and Wu, Y. (2009) Ultrahigh dimensional feature selection: beyond the linear model, *J. Machine Learning Research*, 10, 2013–2038.
- Hoerl, A. E. and Kennard, R. W. (1970), Ridge regression, biased estimation for nonorthogonal problems, *Technometrics*, 12, 55–67.
- Meinshausen, N. and Bühlmann, P. (2010) Stability selection, *J. Roy. Statist. Soc., Ser. B (with discussion)*, 72, 417–473.
- Shah, R. D. and Samworth, R. J. (2013) Variable selection with error control: Another look at Stability Selection, *J. Roy. Statist. Soc., Ser. B*, 75, 55–80.
- Stein, C. (1956) Inadmissibility of the usual estimator for the mean of a multivariate distribution, *Proc. Third Berkeley Symp. Math. Statist. Prob.*, 1, 197-206.

- **Tibshirani, R. (1996) Regression shrinkage and selection via lasso. *J. Roy. Statist. Soc., Ser. B.*, 58, 267-288.**

- **Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables, J. Roy. Statist. Soc., Ser. B, 68, 49–67.**

- **Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net, *J. Roy. Statist. Soc., Ser. B*, 301–320.**