Variable selection with error control: Another look at Stability Selection

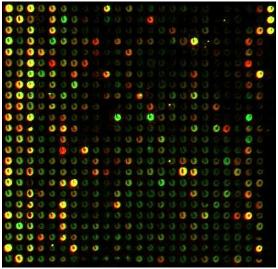
Richard J. Samworth and Rajen D. Shah
University of Cambridge

RSS Journal Webinar
25 October 2017

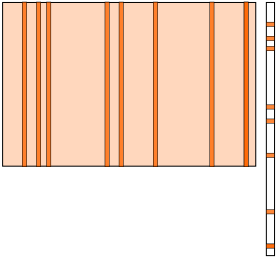
High-dimensional data

Many modern applications, e.g. in genomics, can have the number of predictors p greatly exceeding the number of observations n.

In these settings, variable selection is particularly important.



(a) Microarray data



(b) Sparsity

What is Stability Selection

- Stability Selection (Meinshausen & Bühlmann, 2010) 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 key feature of Stability Selection is the error control provided in the form 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 indices S of some components of Z_i as being 'signal variables', and the rest N as 'noise variables'.

E.g. $Z_i = (X_i, Y_i)$, with covariate vector $X_i \in \mathbb{R}^p$, 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

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

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 of Stability Selection

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 no worse than random guessing:

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

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 allows to user to choose τ based on the expected number of false positives they are willing to tolerate. However:

- The theorem requires two conditions, and the exchangeability assumption is very strong
- There are too many subsets to evaluate $\hat{S}^{\mathrm{SS}}_{n,\tau}$ exactly when $n \geq 30$
- The bound tends to be rather weak.

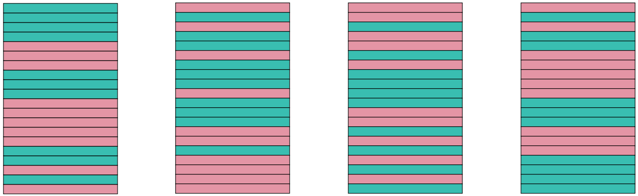
Complementary Pairs Stability Selection (CPSS)

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

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

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



Worst case error control bounds

Define the *selection probability* of variable k to be $p_{k,n} = \mathbb{P}(k \in \hat{S}_n)$.

We can divide our variables into those that have low and high selection probabilities: for $\theta \in [0,1]$, let

$$L_\theta := \{k : p_{k,\lfloor n/2 \rfloor} \leq \theta\} \qquad \text{and} \qquad H_\theta := \{k : p_{k,\lfloor n/2 \rfloor} > \theta\}.$$

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

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

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

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

Illustration and discussion

Suppose $p = 1000$ and $q := \mathbb{E}|\hat{S}_{\lfloor n/2 \rfloor}| = 50$. 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

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

Proof

Let

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

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

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

Thus

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

Proof 2

It follows that

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

$$\leq \frac{1}{2\tau - 1} \sum_{k:p_{k,\lfloor n/2 \rfloor} \leq \theta} p^2_{k,\lfloor n/2 \rfloor} \leq \frac{\theta}{2\tau - 1} \mathbb{E}|\hat{S}_{\lfloor n/2 \rfloor} \cap L_\theta|,$$

where the final inequality follows because

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

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?

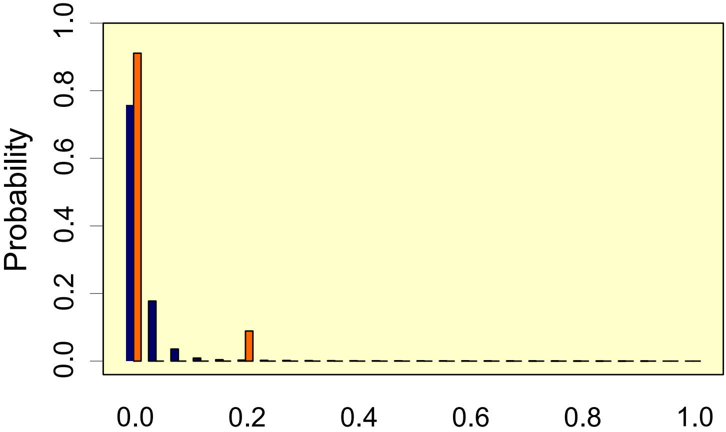


Figure : Typical and extremal pmfs of $\tilde{\Pi}_{25}(k)$ for a low selection probability variable k.

Can we improve on Markov's inequality?

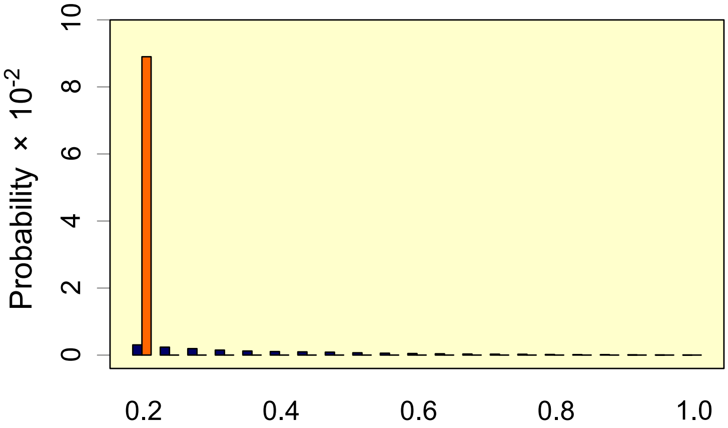


Figure : Typical and extremal pmfs of $\tilde{\Pi}_{25}(k)$ for a low selection probability variable k.

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

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

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

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

Extremal distribution under unimodality

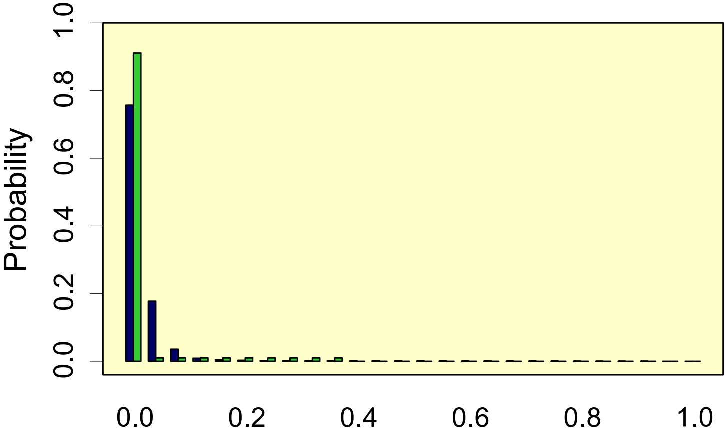


Figure : Typical and extremal pmfs of $\tilde{\Pi}_{25}(k)$ for a low selection probability variable k.

Extremal distribution under unimodality

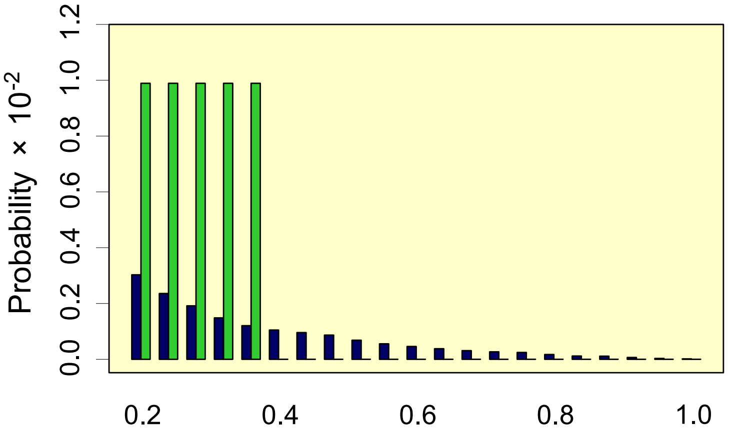


Figure : Typical and extremal pmfs of $\tilde{\Pi}_{25}(k)$ for a low selection probability variable k.

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 f^r is convex on I.

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 $\tau = (\frac{1}{2}, 1]$,

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

where D can be evaluated numerically.

Our simulations suggest $r = -1/2$ is a reasonable choice.

Extremal distribution under $-1/2$-concavity

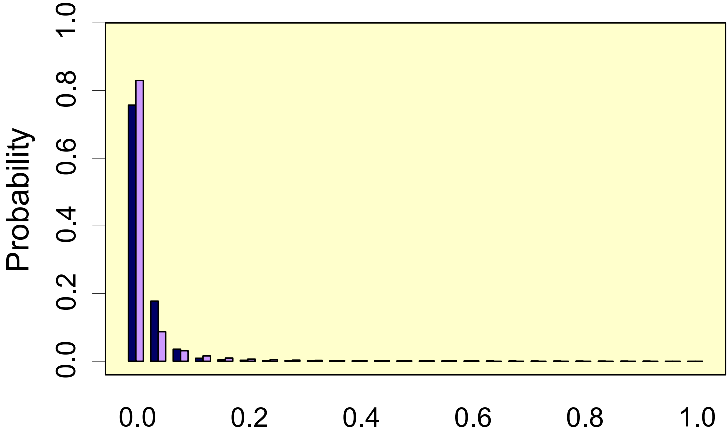


Figure : Typical and extremal pmfs of $\tilde{\Pi}_{25}(k)$ for a low selection probability variable k.

Extremal distribution under $-1/2$-concavity

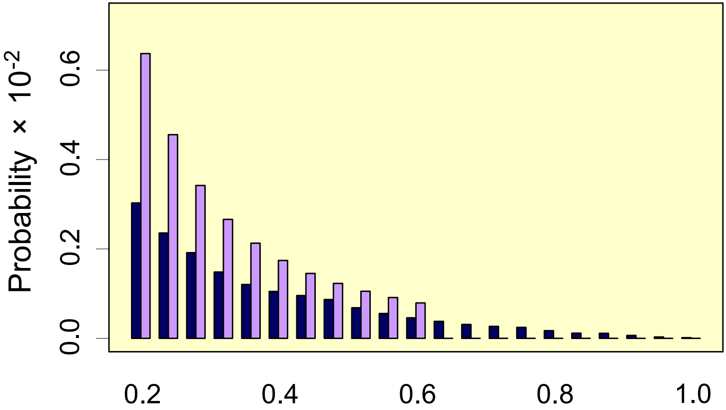
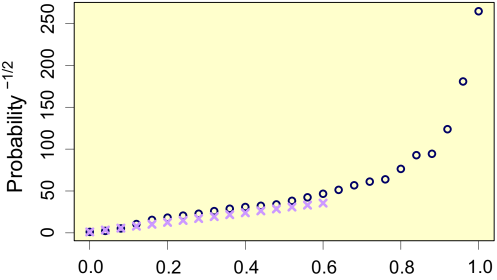
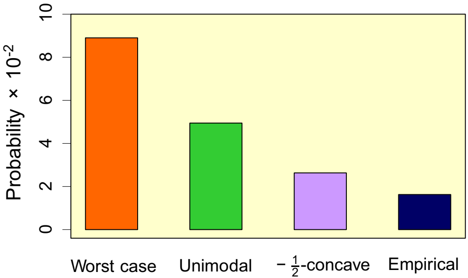


Figure : Typical and extremal pmfs of $\tilde{\Pi}_{25}(k)$ for a low selection probability variable k.

$r = -1/2$ is sensible



Typical and extremal pmfs raised to the power $-1/2$.



Tail probabilities from 0.2 onwards.

Reducing the threshold τ

Suppose $\hat{\Pi}_B(k)$ is $-1/4$-concave, and that $\tilde{\Pi}_B(k)$ is $-1/2$-concave for all $k \in L_\theta$. Then

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

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

Improved bounds

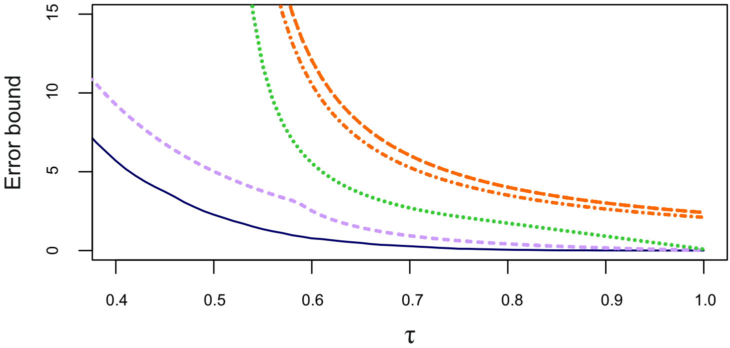


Figure : Comparison of the bounds on $\mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,\tau} \cap L_{q/p}|$ where $p = 1000$, $q = 50$ showing the M & B (dashes), worst case (dot dash), unimodal and r-concave bounds, and the true value for a simulated example.

Simulation study

- Linear model $Y_i = X_i^T \beta + \varepsilon_i$ with $X_i \in N_p(0, \Sigma)$.
- Toeplitz covariance $\Sigma_{ij} = \rho^{||i-j|-p/2|-p/2}$.
- β has sparsity s with $s/2$ equally spaced within $[-1, -0.5]$ and $s/2$ equally spaced within $[0.5, 1]$.
- $n = 200$, $p = 1000$.
- Use Lasso and seek $\mathbb{E}|\hat{S}_{n,\tau}^{\mathrm{CPSS}} \cap L_{q/p}| \leq \ell$. Fix $q = \sqrt{0.8\ell p}$ and for worst-case bound choose $\tau = 0.9$.
- Choose $\tilde{\tau}$ from r-concave bound, oracle τ^*, and oracle λ^* for Lasso $\hat{S}_n^{\lambda^*}$.

Compare

$$\frac{\mathbb{E}|\hat{S}_{n,0.9}^{\mathrm{CPSS}} \cap S|}{\mathbb{E}|\hat{S}_{n,\tau^*}^{\mathrm{CPSS}} \cap S|}, \quad \frac{\mathbb{E}|\hat{S}_{n,\tilde{\tau}}^{\mathrm{CPSS}} \cap S|}{\mathbb{E}|\hat{S}_{n,\tau^*}^{\mathrm{CPSS}} \cap S|} \quad \text{and} \quad \frac{\mathbb{E}|\hat{S}_n^{\lambda^*} \cap S|}{\mathbb{E}|\hat{S}_{n,\tau^*}^{\mathrm{CPSS}} \cap S|}.$$

Simulation results

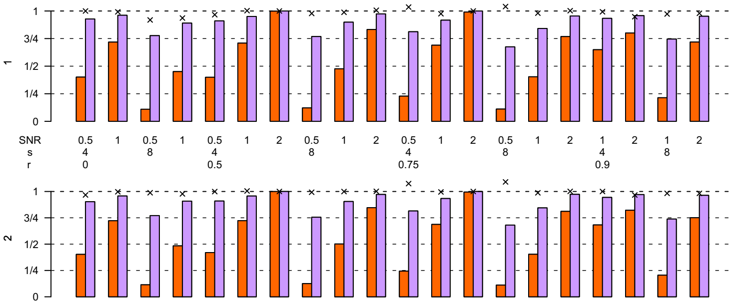


Figure : Expected number or true positives using worst case and *r*-concave bounds, and an oracle Lasso procedure (crosses), as a fraction of the expected number of true positives for an oracle CPSS procedure. The *y*-axis label gives the desired error control level ℓ.

Summary

- CPSS can be used with any variable selection procedure.
- We can bound the average number of low selection probability variables chosen by CPSS with no conditions on the model or original selection procedure needed.
- Under mild conditions e.g. unimodality or r-concavity, the bounds can be strengthened, yielding tight error control.
- This allows the user to choose the threshold τ in an effective way.
- `R` packages: `mboost` and `stabsel`.

Thank you for listening.