VARIABLE SELECTION WITH ERROR CONTROL: ANOTHER LOOK AT STABILITY SELECTION

Richard Samworth
University of Cambridge

Joint work with Rajen Shah
What is Stability Selection?

Stability Selection (Meinshausen and 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 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

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

Stability Selection fixes $\tau \in [0, 1]$ and selects

$$\hat{S}_{n,\tau}^{SS} = \{k : \hat{\Pi}(k) \geq \tau\}.$$
Why subsets of size $\lfloor n/2 \rfloor$?

Both taking subsamples of size $m$ and bootstrap (with-replacement) sampling are examples of exchangeably weighted bootstrap schemes (Mason and Newton, 1992; Præstgaard and Wellner, 1993).

The sum of the weights is $n$ in both cases, and the variance of each component of the bootstrap weights is

$$\text{Var} \ \text{Bin}(n, 1/n) = 1 - 1/n \to 1.$$ 

For subsampling, the variance of each component is $n/m - 1$, which converges to 1 iff $m/n \to 1/2$. 
Error control

Meinshausen and Bühlmann (2010)

Assume that \( \{ \mathbb{1}_{k \in \hat{S}_{[n/2]}} : k \in N \} \) is exchangeable, and that \( \hat{S}_{[n/2]} \) is not worse than random guessing:

\[
\frac{\mathbb{E}(|\hat{S}_{[n/2]} \cap S|)}{\mathbb{E}(|\hat{S}_{[n/2]} \cap N|)} \geq \frac{|S|}{|N|}.
\]

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

\[
\mathbb{E}(|\hat{S}_{n,\tau}^{ss} \cap N|) \leq \frac{1}{2\tau - 1} \left( \frac{\mathbb{E}(|\hat{S}_{[n/2]}|)}{p} \right)^2.
\]
Error control discussion

In principle, this theorem helps the practitioner choose the tuning parameter $\tau$. However:

- The theorem requires two conditions, and the exchangeability assumption is very strong.
- There are too many subsets to evaluate $\hat{S}_{n,\tau}^{SS}$ 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 \(\hat{S}_{n,\tau}^{CPSS} = \{k : \hat{\Pi}_B(k) \geq \tau\}\).
Worst case error control bounds

Let \( p_{k,n} = P(k \in \hat{S}_n) \). 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

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

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

\[
E|\hat{N}_{n,\tau}^{CPSS} \cap H_\theta| \leq \frac{1 - \theta}{1 - 2\tau} E|\hat{N}_{\lfloor n/2 \rfloor} \cap H_\theta|.
\]
Illustration and discussion

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

- 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

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

Let

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

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

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

Thus

\[ \mathbb{P}\{\hat{\Pi}_B(k) \geq \tau\} \leq \mathbb{P}\left\{ \frac{1}{2} (1 + \tilde{\Pi}_B(k)) \geq \tau \right\} = \mathbb{P}\{\tilde{\Pi}_B(k) \geq 2\tau - 1\} \leq \frac{1}{2\tau - 1} p^2_{k, \lfloor n/2 \rfloor}. \]
Proof 2

Note that

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

It follows that

$$\mathbb{E} | \hat{S}_{n, \tau}^{\text{CPSS}} \cap L_\theta | = \mathbb{E}\left( \sum_{k: p_k, \lfloor n/2 \rfloor \leq \theta} 1_{\{k \in \hat{S}_{n, \tau}^{\text{CPSS}}\}} \right) = \sum_{k: p_k, \lfloor n/2 \rfloor \leq \theta} \mathbb{P}(k \in \hat{S}_{n, \tau}^{\text{CPSS}}) \leq \frac{1}{2\tau - 1} \sum_{k: p_k, \lfloor n/2 \rfloor \leq \theta} p_k^2, \lfloor n/2 \rfloor \leq \theta \leq \frac{\theta}{2\tau - 1} \mathbb{E} | \hat{S}_{\lfloor n/2 \rfloor} \cap L_\theta |.$$

September 12, 2013 - 12
Bounds with no assumptions whatsoever

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

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

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

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

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

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

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

where $D$ can be evaluated numerically.

Our simulations suggest $r = -1/2$ is a safe and sensible choice.
Extremal distribution under $\gamma$-concavity
$r = -\frac{1}{2}$ is sensible
Reducing the threshold $\tau$

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

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

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

Linear model $Y_i = X_i^T \beta + \epsilon_i$ with $X_i \sim N_p(0, \Sigma)$. Take $\Sigma$ Toeplitz with $\Sigma_{ij} = \rho^{|i-j|-p/2|-p/2}$. Let $\beta$ have sparsity $s$, with $s/2$ equally spaced within $[-1, -0.5]$ and $s/2$ equally spaced in $[0.5, 1]$. Fix $n = 200$, $p = 1000$.

Use Lasso and seek $\mathbb{E}|\hat{S}_{n, \tau}^{\text{CPSS}} \cap L_q/p| \leq l$. Fix $q = \sqrt{0.8lp}$ and for worst-case bound choose $\tau = 0.9$. Choose $\tilde{\tau}$ from $r$-concave bound, oracle $\tau^*$, and oracle $\lambda^*$ for Lasso $\hat{S}_{n, \lambda^*}^\lambda$. Compare

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

- CPSS can be used in conjunction with any variable selection procedure.

- We can bound the average number of low selection probability variables chosen by CPSS under no conditions on the model or original 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 \(\tau\) in an effective way.
References


discussion)*, 72, 417–473.
