Stability Selection: Theorem 10.1 in book

Assume:

- exchangeability condition:
  $\{1(j \in \hat{S}_\lambda), j \in S_0^c\}$ is exchangeable for all $\lambda \in \Lambda$
- $\hat{S}$ is not worse than random guessing

$$\frac{\mathbb{E}|S_0 \cap \hat{S}_\Lambda|)}{\mathbb{E}(|S_0^c \cap \hat{S}_\Lambda|)} \geq \frac{|S_0|}{|S_0^c|}.$$

Then, for $\pi_{\mathrm{thr}} \in (1/2, 1)$:

$$\mathbb{E}[V] \quad \leq \quad \frac{1}{2\pi_{\mathrm{thr}} - 1} \frac{q_\Lambda^2}{p}.$$

suppose we know q_Λ (see later)
strategy: specify $\mathbb{E}[V] = v_0$ (e.g. $= 5$)
$\leadsto$ for $\pi_{\mathrm{thr}} := \frac{1}{2} + \frac{q_\Lambda^2}{2pv_0}$: $\mathbb{E}[V] \leq v_0$

example: regression model with $p = 1000$ variables

$\hat{S}_\lambda =$ the top 10 variables from Lasso (e.g. the different λ from Lasso by CV and choose the top 10 variables with the largest absolute values of the corresponding estimated coefficients; if less than 10 variables are selected, take the selected variables) the value λ corresponds to the "top 10"; Λ is a singleton

we then know that $q_\Lambda = \mathbb{E}[|\hat{S}_\lambda(I)|] \leq 10$

For $\mathbb{E}[V] = v_0 := 5$ we then obtain

$$\pi_{\mathrm{thr}} = \frac{1}{2} + \frac{q_\Lambda^2}{2pv_0} = 0.5 + \frac{10^2}{2 * 1000 * 5} = 0.51$$

there is room to play around
recommendation: take $|\hat{S}_{\lambda}|$ rather large and stability selection will reduce again to reasonable size

when taking the "top 30", the threshold becomes

$$\pi_{\mathrm{thr}} = \frac{1}{2} + \frac{q_{\Lambda}^2}{2pv_0} = 0.5 + \frac{30^2}{2 * 1000 * 5} = 0.59$$

adding noise...
can always add (e.g. independent $\mathcal{N}(0,1)$) noise covariates
enlarged dimension p_{enlarged}

error control becomes better (for the same threshold)

$$\mathbb{E}[V] \quad \leq \quad \frac{1}{2\pi_{\mathrm{thr}} - 1} \frac{q_\Lambda^2}{p_{\mathrm{enlarged}}}$$

this sometimes helps indeed in practice – at the cost of loss in power

The assumptions for mathematical guarantees

not worse than random guessing

$$\frac{\mathbb{E}|S_0 \cap \hat{S}_\Lambda|)}{\mathbb{E}(|S_0^c \cap \hat{S}_\Lambda|)} \geq \frac{|S_0|}{|S_0^c|}$$

perhaps hard to check but very reasonable...

for Lasso in linear models it holds assuming the variable screening property
asymptotically: if beta-min and compatibility condition hold

exchangeability condition:
$\{1(j \in \hat{S}_\lambda), j \in S_0^c\}$ is exchangeable for all $\lambda \in \Lambda$

a restrictive assumption
but the theorem is very general, for any algorithm $\hat{S}$

a very special case where exchangeability condition holds: random equi-correlation design linear model

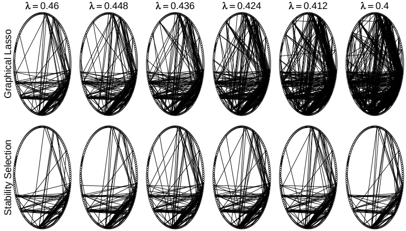
$$Y = X\beta^0 + \varepsilon,\ \mathrm{Cov}(X)_{i,j} \equiv \rho\ (i \neq j),\ \mathrm{Var}(X_j) \equiv 1 \forall j$$

distributions of $(Y, X^{(S_0)}, \{X^{(j)};\ j \in S_0^c\})$ and of $(Y, X^{(S_0)}, \{X^{(\pi(j))};\ j \in S_0^c\})$ are the same for any permutation $\pi : S_0^c \to S_0^c$

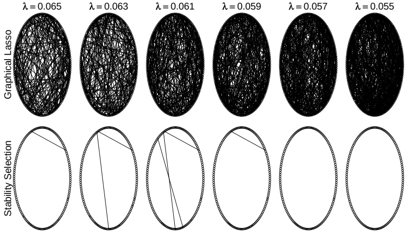
- distribution of $X^{(S_0)}, \{X^{(\pi(j))};\ j \in S_0^c\}$ is the same for all π (because of equi-correlation)
- distribution of $Y | X^{(S_0)}, \{X^{(\pi(j))};\ j \in S_0^c\}$ is the same for all π (because it depends only on $X^{(S_0)}$)
- therefore: distribution of $Y, X^{(S_0)}, \{X^{(\pi(j))};\ j \in S_0^c\}$ is the same for all π
  and hence exchangeability condition holds for any (measurable) function $\hat{S}_\lambda$

An illustration for graphical modeling

$p = 160$ gene expressions, $n = 115$
GLasso estimator, selecting among the $\binom{p}{2} = 12'720$ features
stability selection with $\mathbb{E}[V] \le v_0 = 30$



with permutation (empty graph is correct)



Stability Selection is extremely easy to use
and super-generic

the sufficient assumptions (far from necessary) for
mathematical guarantees are restrictive
but the method seems to work very well in practice

P-values based on multi sample splitting

(Ch. 11 in Bühlmann and van de Geer (2011))

Stability Selection

- uses subsampling many times – a good thing!
- provides control of the expected number of false positives rather than e.g. the familywise error rate $\rightsquigarrow$ we will "address" this with
  multi sample splitting and aggregation of P-values

familywise error rate (FWER):

$$\mathrm{FWER} = \mathbb{P}[V > 0], \; V \text{ number of false positives}$$

Fixed design linear model

$$Y = X\beta^0 + \varepsilon$$

instead of de-biased/de-sparsified method, consider the “older” technique (which is not statistically optimal but more generic and more in the spirit of stability selection) split the sample into two parts I_1 and I_2 of equal size $\lfloor n/2 \rfloor$

- use (e.g.) Lasso to select variables based on I_1: $\hat{S}(I_1)$
- perform low-dimensional statistical inference on I_2 based on data $(X_{I_2}^{(\hat{S}(I_1))}, Y_{I_2})$;
  for example using the t-test for single coefficients β_j^0
  (if $j \notin \hat{S}(I_1)$, assign the p-value 1 to the hypothesis $H_{0,j}:\ \beta_j^0 = 0$)

  due to independence of I_1 and I_2, this is a "valid" strategy (see later)

validity of the (single) data splitting procedure
consider testing $H_{0,j}: \beta_j^0 = 0$ versus $H_{A,j}: \beta_j^0 \neq 0$
assume Gaussian errors for the fixed design linear model :
thus, use the t-test on the second half of the sample I_2 to get a p-value

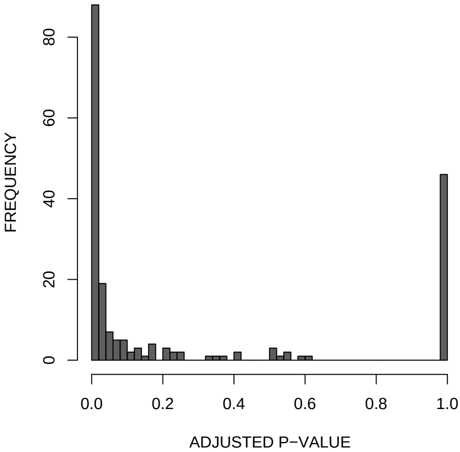
$$P_{\text{raw},j} \text{ from } t\text{-test based on } X_{I_2}^{(\hat{S}(I_1))}, Y_{I_2}$$

$P_{\text{raw},j}$ is a valid p-value (controlling type I error) for testing $H_{0,j}$ if $\hat{S}(I_1) \supseteq S_0$ (i.e., the screening property holds)

if the screening property does not hold: $P_{\text{raw},j}$ is still valid for $H_{0,j}(M): \beta_j(M) = 0$ where $M = \hat{S}(I_1)$ is a selected sub-model and $\beta(M) = ((X^{(M)})^T X^{(M)})^{-1}(X^{(M)})^T Y$

a p-value lottery depending on the random split of the data

motif regression $n = 287$, $p = 195$



$\rightsquigarrow$ should aggregate/average over multiple splits!

Multiple testing and aggregation of p-values

the issue of multiple testing:

$$\tilde{P}_j = \begin{cases} P_{\mathrm{raw},j} \text{ based on } Y_{I_2}, X_{I_2}^{(\hat{S}(I_1))} & , \text{if } j \in \hat{S}(I_1), \\ 1 & , \text{if } j \notin \hat{S}(I_1) \end{cases}$$

thus, we can have at most $|\hat{S}(I_1)|$ false positives
$\leadsto$ can correct with Bonferroni with factor $|\hat{S}(I_1)|$ (instead of factor p) to control the familywise error rate

$$\tilde{P}_{\mathrm{corr},j} = \min(\tilde{P}_j \cdot |\hat{S}(I_1)|, 1) \; (j = 1, \ldots, p)$$

decision rule: reject $H_{0,j}$ if and only if $\tilde{P}_{\mathrm{corr},j} \leq \alpha$
$\leadsto$ FWER $\leq \alpha$

the issue with P-value aggregation:

if we run sample splitting B times, we obtain P-values

$$\tilde{P}^{[1]}_{\mathrm{corr},j}, \ldots, \tilde{P}^{[B]}_{\mathrm{corr},j}$$

how to aggregate these dependent p-values to a single one?

for $\gamma \in (0,1)$ define

$$Q_j(\gamma) = \min\left\{ q_\gamma\big(\{\tilde{P}^{[b]}_{\mathrm{corr},j}/\gamma;\ b = 1,\ldots,B\}\big), 1\right\},$$

where $q_\gamma(\cdot)$ is the (empirical) γ-quantile function

Proposition 11.1 (Bühlmann and van de Geer, 2011)
For any $\gamma \in (0, 1)$, $Q_j(\gamma)$ are P-values which control the FWER

example: $\gamma = 1/2$
aggregate the p-values with the sample median and multiply by the factor 2

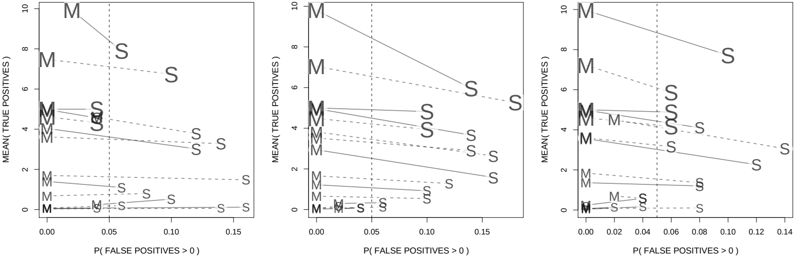
avoid choosing γ:

$$P_j = \min \left\{ \underbrace{\left(1 - \log \gamma_{\min}\right)}_{\text{price to optimize over } \gamma} \inf_{\gamma \in (\gamma_{\min}, 1)} Q_j(\gamma), 1 \right\} \ (j = 1, \ldots, p).$$

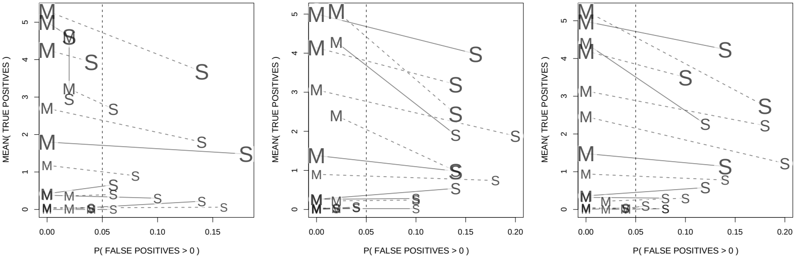
Theorem 11.1 (Bühlmann and van de Geer (2011))
For any $\gamma_{\min} \in (0, 1)$, P_j are P-values which control the FWER

the entire framework for p-value aggregation holds whenever the single p-values are valid ($\mathbb{P}[P_{\mathrm{raw},j} \leq \alpha] \leq \alpha$ under $H_{0,j}$)
has nothing to do with high-dimensional regression and sample splitting

$n = 100, p = 100$



$n = 100, p = 1000$



one can also adapt the method to control the False Discovery Rate (FDR) multi sample splitting and p-value construction:

- is very generic, also for “any other” model class
- is powerful in terms of multiple testing correction: we only correct for multiplicity from $|\hat{S}(I_1)|$ variables
- it relies in theory on the screening property of the selector in practice: it is a quite competitive method!
- Schultheiss et al. (2021): can improve multi sample splitting by multi carve methods, based on “technology” from selective inference

Undirected graphical models

(Ch. 13 in Bühlmann and van de Geer (2011))

- graph G:
  set of vertices/nodes $V = \{1, \ldots, p\}$
  set of edges $E \subseteq V \times V$
- random variables $X = X^{(1)}, \ldots, X^{(p)}$ with distribution P
  identify nodes in V with components of X

graphical model: (G, P)

pairwise Markov property:
P satisfies the pairwise Markov property (w.r.t. G) if

$$(j, k) \notin E \Longrightarrow X^{(j)} \perp X^{(k)} | X^{(V \setminus \{j,k\})}$$

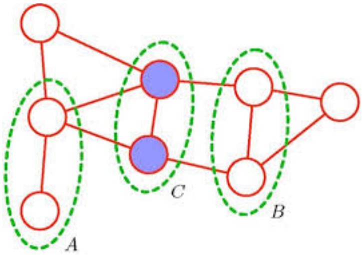
Global Markov property

(stronger property than pairwise Markov prop):

consider disjoint subsets $A, B, C \subseteq V$

P satisfies the global Markov property (w.r.t. G) if

$$A \text{ and } B \text{ are separated by } C \implies X^{(A)} \perp X^{(B)} | \underbrace{X^{(C)}}_{\text{only condition on subset } C}$$



global Markov property $\Longrightarrow$ pairwise Markov property

Proof:
consider $(j, k) \notin E$

denote by $A = \{j\}, B = \{k\}, C = V \setminus \{j, k\}$;
since $(j, k) \notin E$, $A = \{j\}$ and $B = \{k\}$ are separated by C

by the global Markov property: $X^{(j)} \perp X^{(k)} | X^{(V \setminus \{j,k\})}$

$\square$

$\leadsto$ global Markov property is more "interesting" consider graphical model (G, P)

if P has a positive and continuous density w.r.t. Lebesgue measure:
the global and pairwise Markov properties (w.r.t. G) coincide/are equivalent (Lauritzen, 1996)

prime example: P is Gaussian

the Markov properties imply some conditional independencies from graphical separation

for example with pairwise Markov property:

$$(j,k) \notin E \Longrightarrow X^{(j)} \perp X^{(k)} | X^{(V \setminus \{j,k\})}$$

how about reverse relation ?

$$(j,k) \in E \overbrace{\Longrightarrow}^{?} X^{(j)} \not\perp X^{(k)} | X^{(V \setminus \{j,k\})}$$

can we interpret existing edges?

in general: no! (unfortunately)

in some special cases:

$$(j,k) \in E \implies X^{(j)} \not\perp X^{(k)} | X^{(V \setminus \{j,k\})}$$

prime example: P is Gaussian

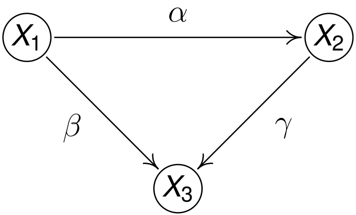
$$(j,k) \in E \iff X^{(j)} \not\perp X^{(k)} | X^{(V \setminus \{j,k\})}$$

for A and B not separated by C: in general not true that

$$X^{(A)} \not\perp X^{(B)} | X^{(C)}$$

... due to possible strange cancellations of “edge weights”

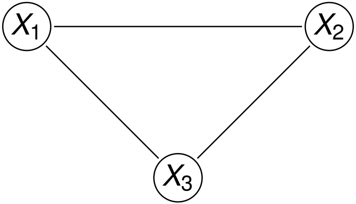
Gaussian "counterexample"



$$
\begin{aligned}
&X^{(1)} \leftarrow \varepsilon^{(1)},\\
&X^{(2)} \leftarrow \alpha X^{(1)} + \varepsilon^{(2)},\\
&X^{(3)} \leftarrow \beta X^{(1)} + \gamma X^{(2)} + \varepsilon^{(3)},\\
&\varepsilon^{(1)}, \varepsilon^{(2)}, \varepsilon^{(3)} \text{ i.i.d. } \mathcal{N}(0,1)
\end{aligned}
$$

$\rightsquigarrow$ a Gaussian distribution P
for $\beta + \alpha\gamma = 0$: $\mathrm{Corr}(X_1, X_3) = 0$ that is: $X^{(1)} \perp X^{(3)}$

it is a Gaussian Graphical Model where P is Markov w.r.t. the following graph



we know that $X^{(1)} \perp X^{(3)}$ (for special constellations of α, β, γ)

take $A = \{1\}, B = \{3\}, C = \emptyset$
although A and B are not separated (by the emptyset)
since there is a direct edge
it does not hold that $X^{(1)} \not\perp X^{(3)}$ (conditional on $\emptyset$, i.e., marginal)

Gaussian Graphical Model

conditional independence graph (CIG):
(G, P) satisfies the pairwise Markov property

Gaussian Graphical Model (GGM):
a conditional independence graph with P being Gaussian

for simplicity, assume mean zero: $P \sim \mathcal{N}_p(0, \Sigma)$

we know already that edges are equivalent to conditional dependence given all other variables

for a GGM:

$$(j, k) \in E \Longleftrightarrow (\Sigma^{-1})_{jk} \neq 0$$

Neighborhood selection: nodewise regression

$$X^{(j)} = \beta_k^{(j)} X^{(k)} + \sum_{r \neq j,k} \beta_r^{(j)} X^{(r)} + \varepsilon^{(j)},\ j = 1 \ldots, p$$

$$X^{(k)} = \beta_j^{(k)} X^{(j)} + \sum_{r \neq k,j} \beta_r^{(k)} X^{(r)} + \varepsilon^{(k)}$$

for GGM:

$$(j,k) \in E \Longleftrightarrow \beta_k^{(j)} \neq 0 \ \Longleftrightarrow \beta_j^{(k)} \neq 0$$