# 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 $\leadsto$ 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 $\beta_j^0$
  (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

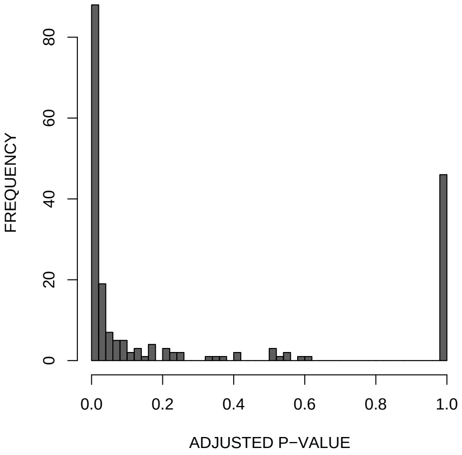
$$P_{\text{raw},j} \text{ from } t\text{-test based on } X_{I_2}^{\hat{S}(I_1)}, Y_{I_2}$$

$P_{\text{raw},j}$ 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 $\beta(M) = (X_M^T X_M)^{-1} X_M^T \mathbb{E}[Y]$

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 $\tilde{P}_{\mathrm{corr},j} \le \alpha$
$\leadsto$ FWER $= \mathbb{P}[V > 0] \le \alpha$
assuming that the raw p-values $P_{\mathrm{raw},j}$ are valid

(e.g. screening property holds)

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) $\gamma$-quantile function

*Proposition 11.1 (Bühlmann and van de Geer, 2011)*
Assume that the raw p-values $P_{\mathrm{raw},j}$ are valid.
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 $\gamma$:

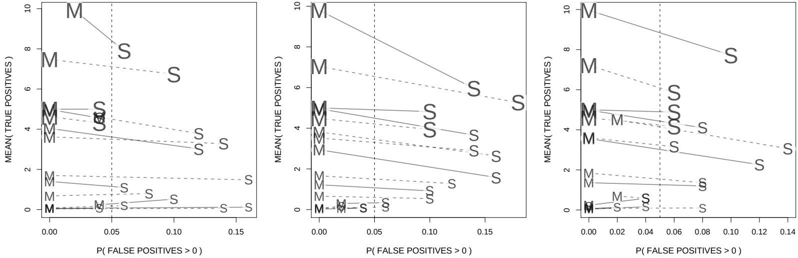
$$P_j = \min\{ \underbrace{(1 - \log \gamma_{\min})}_{\text{price to optimize over } \gamma} \inf_{\gamma \in (\gamma_{\min}, 1)} Q_j(\gamma), 1\} \ (j = 1, \ldots, p).$$

*Theorem 11.1 (Bühlmann and van de Geer (2011))*
Assume that the raw p-values $P_{\mathrm{raw},j}$ are valid.
For any $\gamma_{\min} \in (0, 1)$, $P_j$ are P-values which control the FWER

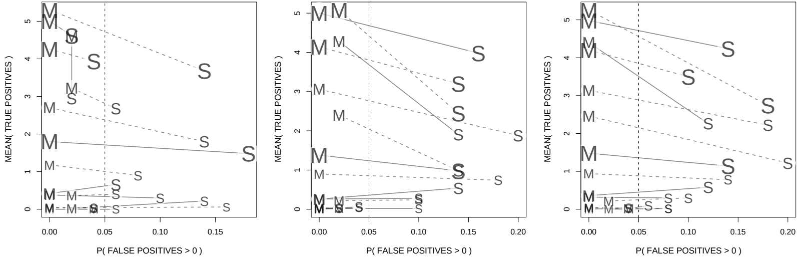
that is: reject $H_{0,j} : \beta_j^0 = 0$ if and only if $P_j \leq \alpha$ for all $j = 1, \ldots, p$
$\leadsto$ FWER = $\mathbb{P}[V > 0] \leq \alpha$.

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 selected 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

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

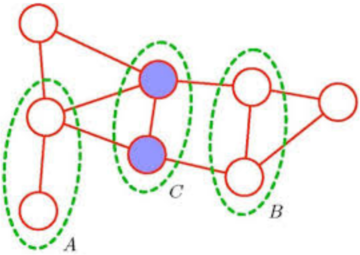
Global Markov property

(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: $X^{(j)} \perp X^{(k)} | X^{(V \setminus \{j,k\})}$

$\square$

$\rightsquigarrow$ 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”

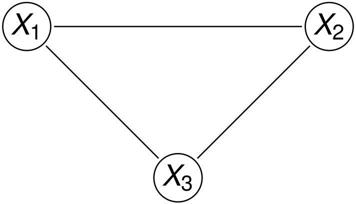
$$\begin{tikzcd} X_1 \arrow[rr, "\alpha"] \arrow[dr, "\beta"'] & & X_2 \arrow[dl, "\gamma"] \\ & X_3 & \end{tikzcd}$$

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

$\rightsquigarrow$ 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 $\alpha, \beta, \gamma$)

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:

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

nodewise regression (Meinshausen & Bühlmann, 2006)

- run Lasso for every node variable $X^{(j)}$ versus all others $\{X^{(k)};\ k \neq j\}$ $(j = 1, \ldots, p)$
- estimated active set $\hat{S}^{(j)} = \{r;\ \hat{\beta}_r^{(j)} \neq 0\}$ $(j = 1, \ldots, p)$
- estimate edges in $\hat{E}$ :

$$\begin{aligned} \text{or rule:} \quad & (j,k) \in \hat{E} \Longleftrightarrow j \in \hat{S}^{(k)} \text{ or } k \in \hat{S}^{(j)} \\ \text{and rule:} \quad & (j,k) \in \hat{E} \Longleftrightarrow j \in \hat{S}^{(k)} \text{ and } k \in \hat{S}^{(j)} \end{aligned}$$

just run Lasso $p$ times: it's fast!

(given the difficulty of the problem)

$O(np^2 min(n,p))$ computational complexity

and it has "near-optimal" statistical properties

(slightly better than penalized MLE)

`R`-packages `huge` and also in `glasso` (and set 'approx = T')

# GLasso: regularized maximum likelihood estimation

data $X_1, \ldots X_n$ i.i.d. $\sim \mathcal{N}_p(\mu, \Sigma)$

goal: estimate $K = \Sigma^{-1}$ (precision matrix)

approach, called GLasso (Friedman, Hastie and Tibshirani, 2008):

$$\hat{K}, \hat{\mu} = \mathrm{argmin}_{K \succ 0, \mu} \left(-\text{log-likelihood}(K, \mu;\ X_1, \ldots, X_n) + \lambda \|K\|_1\right)$$

$$\hat{\mu} = n^{-1} \sum_{i=1}^{n} X_i \text{ decouples}$$

$$\hat{K} = \mathrm{argmin}_{K \succ 0} (\underbrace{-\text{log-likelihood}(K, \hat{\mu};\ X_1, \ldots, X_n)}_{\propto -\log(\det K) + \mathrm{trace}(\hat{\Sigma}_{\mathrm{MLE}} K)} + \lambda \|K\|_1)$$

$$\|K\|_1 = \sum_{j,k} |K_{j,k}| \text{ or } \sum_{j \neq k} |K_{j,k}|$$

$$\hat{\Sigma}_{\mathrm{MLE}} = n^{-1} \sum_{i=1}^{n} (X_i - \hat{\mu})(X_i - \hat{\mu})^T$$

- GLasso is computationally (much) slower than nodewise regression
  $O(np^3)$ computational complexity (for potentially dense problems)
- GLasso provides estimates of $\Sigma^{-1}$ and also of $\Sigma$ by inversion
- one can run a hybrid approach:
  nodewise selection first with estimated edge set $\hat{E}$
  GLasso restricted to $\hat{E}$ with $\lambda = 0$:
  that is, unpenalized MLE restricted to $\hat{E}$

  fast and accurate!

  analogous to Lasso-OLS hybrid in regression

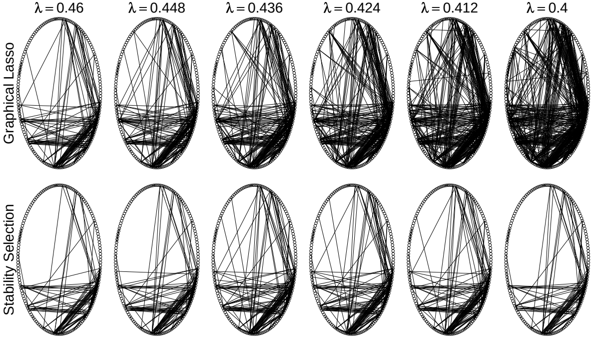
# Tuning of the methods

cross-validation of the (nodewise) likelihood

## and/or Stability Selection

$p = 160$ gene expressions, $n = 115$

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



# The nonparanormal graphical model

(Liu, Lafferty and Wasserman, 2009)

motivating question: are there other "interesting" distributions, besides the Gaussian, where conditional independence between two rv.'s is encoded as zero entries in a matrix?

nonparanormal graphical model:
$X$ has a nonparanormal distribution if there exist functions $f_j$ ($j = 1, \ldots, p$) such that

$$Z = f(X) = (f_1(X^{(1)}), \ldots, f_p(X^{(p)})) \sim \mathcal{N}_p(\mu, \Sigma)$$

w.l.o.g. $\mu = 0$ and $\Sigma_{jj} = 1$
$\leadsto$ $Z_j = f_j(X^{(j)}) \sim \mathcal{N}(0,1)$ and therefore:
$f_j(\cdot) = \Phi^{-1} F_j(\cdot)$ where $F_j(u) = \mathbb{P}[X^{(j)} \le u]$: monotone

$\leadsto$ a semiparametric Gaussian copula model

*Lemma*

Assume that $(G, P)$ is a nonparanormal graphical model with $f_j$s being differentiable. Then:

$$(j,k) \in E \iff X^{(j)} \not\perp X^{(k)} | X^{(V \setminus \{j,k\})} \iff \Sigma^{-1}_{j,k} \neq 0$$

Proof: the density of $X$ is

$$p(x) = \frac{1}{(2\pi)^{p/2} \det(\Sigma)^{1/2}} \exp(-\frac{1}{2}(f(x)-\mu)^T \Sigma^{-1}(f(x)-\mu)) \prod_{j=1}^{p} |f_j'(x_j)|$$

$\rightsquigarrow$ the density factorizes exactly as in the Gaussian case according to $\Sigma^{-1}$ $\square$

we only have to estimate the non-zeroes of $\Sigma^{-1}$
but $\Sigma$ is the covariance of the unknown $f(X)$...

the best proposal (Lue and Zhou, 2012):
rank-based!
compute empirical rank correlation of $X^{(1)}, \ldots, X^{(p)}$ with a bias correction from Kendall (1948)
denote this empirical rank correlation matrix as $\hat{R}$ (invariant under monotone $f_j$'s)

stick it into GLasso:

$$\hat{K} = \text{argmin}_{K \succ 0} - \log(\det K) + \operatorname{trace}(\hat{R}K) + \lambda \|K\|_1$$

this has provable guarantees in the case of a nonparanormal graphical model
robustness of GLasso by using rank-correlation as input matrix