Undirected graphical models

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"

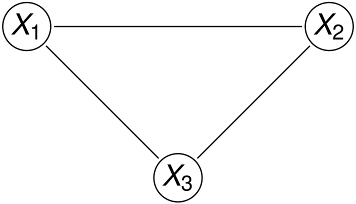
Gaussian “counterexample”

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

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

$\rightsquigarrow$ 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 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

(Meinshausen & Bühlmann, 2006)

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

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

for GGM:

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

nodewise regression

- 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 $\hat{E}$:

$$\text{or rule:} \qquad (j,k) \in \hat{E} \Longleftrightarrow j \in \hat{S}^{(k)} \text{ or } k \in \hat{S}^{(j)}$$

$$\text{and rule:} \qquad (j,k) \in \hat{E} \Longleftrightarrow j \in \hat{S}^{(k)} \text{ and } k \in \hat{S}^{(j)}$$

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):

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

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

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

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

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

- GLasso is computationally (much) slower than nodewise regression
  $O(np^3)$ computational complexity (for potentially dense problems)
- GLasso provides estimates of Σ^{-1} and also of Σ 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

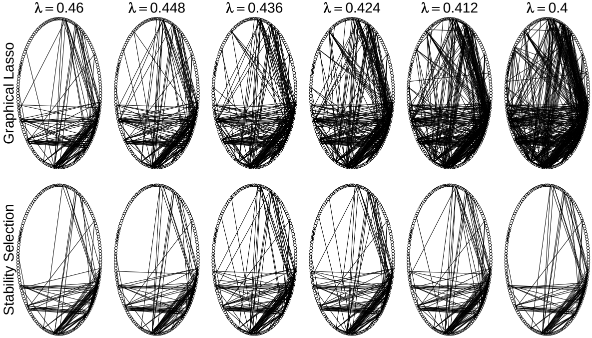
Tuning of the methods

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 $\mathbb{E}[V] \le v_0 = 30$



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 $F_j(u) = \mathbb{P}[X^{(j)} \leq u]$: monotone

$\leadsto$ a semiparametric Gaussian copula model

Lemma

Assume that (G, P) is a nonparanormal graphical model with f_j being differentiable for all $j = 1, \ldots, p$. 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 Σ^{-1} $\square$

we only have to estimate the non-zeroes of Σ^{-1}
but Σ is not the covariance matrix of $X = (X^{(1)}, \ldots, X^{(p)})$
Σ is the covariance matrix of the unknown $f_1(X^{(1)}), \ldots, f_p(X^{(p)})$

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:

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

this has provable guarantees in the case of a nonparanormal graphical model for estimating Σ^{-1}

as an important implication:
the rank-based version of GLasso exhibits some robustness for estimating the conditional independence pattern of $X \sim P$
that is: if the distribution is nonparanormal, it still works well and properly!

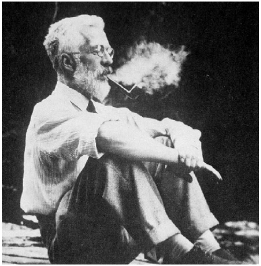
this is different and much better than:
GLasso works for estimating $\mathrm{Cov}(X)^{-1}$ even if $X \sim P$ is non-Gaussian
although this is true, if sufficient amount of moments exist for non-Gaussian P: zeroes of $\mathrm{Cov}(X)^{-1}$ do not encode conditional independencies!

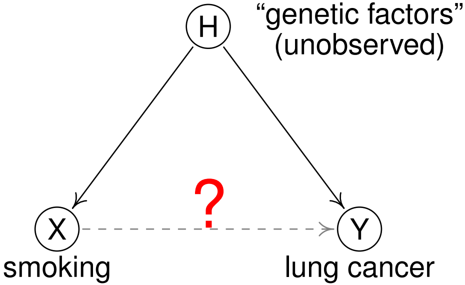
The danger of hidden confounding!

Lasso, Group Lasso, neural networks, neighborhood selection, GLasso,...
for (generalized) linear models, nonlinear models, undirected graphical models, ...

they all give “wrong” answers in presence of hidden confounding

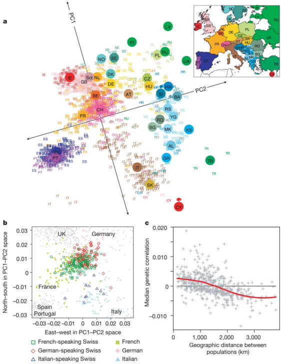
Does smoking cause lung cancer?



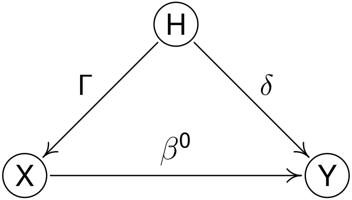


Genes mirror geography within Europe (Novembre et al., 2008)

SNP data plotted on first 2 principal components



confounding effects about geographical origin of data are found on the first principal components



$$Y \leftarrow X_{n\times p}\beta^0 + H\delta + \eta$$
$$X \leftarrow H_{n\times q}\Gamma + E$$

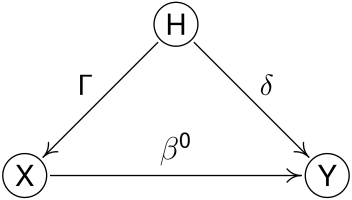
goal: infer β^0 from observations $(X_1, Y_1), \ldots, (X_n, Y_n)$

the population least squares principle leads to the parameter

$$\beta^* = \mathrm{argmin}_u \mathbb{E}[(Y - Xu)^2], \quad \beta^* = \beta^0 + \underbrace{b}_{\text{“bias”}}$$

$$\|b\|_2 \le \frac{\|\delta\|_2}{\sqrt{\text{“number of } X\text{-components affected by } H\text{”}}}$$

small “bias” if confounder has dense effects!
blessing of high dimensionality!



$$Y \leftarrow X_{n\times p}\beta^0 + H\delta + \eta$$
$$X \leftarrow H_{n\times q}\Gamma + E$$

goal: infer β^0 from observations $(X_1, Y_1), \ldots, (X_n, Y_n)$

the population least squares principle leads to the parameter

$$\beta^* = \mathrm{argmin}_u \mathbb{E}[(Y - Xu)^2], \quad \beta^* = \beta^0 + \underbrace{b}_{\text{“bias”}}$$

$$\|b\|_2 \le \frac{\|\delta\|_2}{\sqrt{\text{“number of } X\text{-components affected by } H\text{”}}}$$

small “bias” if confounder has dense effects!
blessing of high dimensionality!

perhaps more importantly: view this as

$$Y = X\beta^* + \varepsilon = X \underbrace{(\beta^0 + b)}_{\text{sparse + dense}} + \varepsilon,$$

$$\varepsilon = Y - \mathbb{E}[Y|x]$$

$\rightsquigarrow$ we should use high-dimensional methods for "sparse + dense" regression parameter vector

- Lava (Chernozhukov, Hansen & Liao, 2017)
- Spectral Deconfounding (Ćevid, Bühlmann & Meinshausen, 2020, Guo, Ćevid & Bühlmann, 2021)

similarly for undirected graphical modeling:

$$\mathrm{Cov}(X)^{-1} = \text{sparse matrix} + \text{low rank matrix}$$

$\rightsquigarrow$ use Gaussian likelihood for $\mathrm{Cov}(X)^{-1}$ but with penalty enforcing sparsity + low rank

(Chandrasekaran, Parrilo & Willsky, 2012) still lots of things to do!