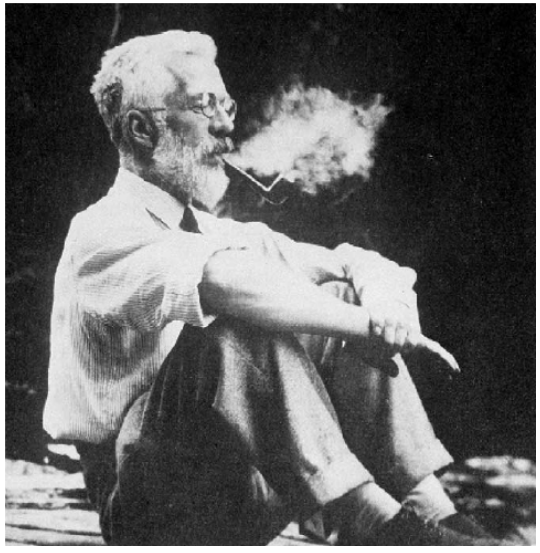
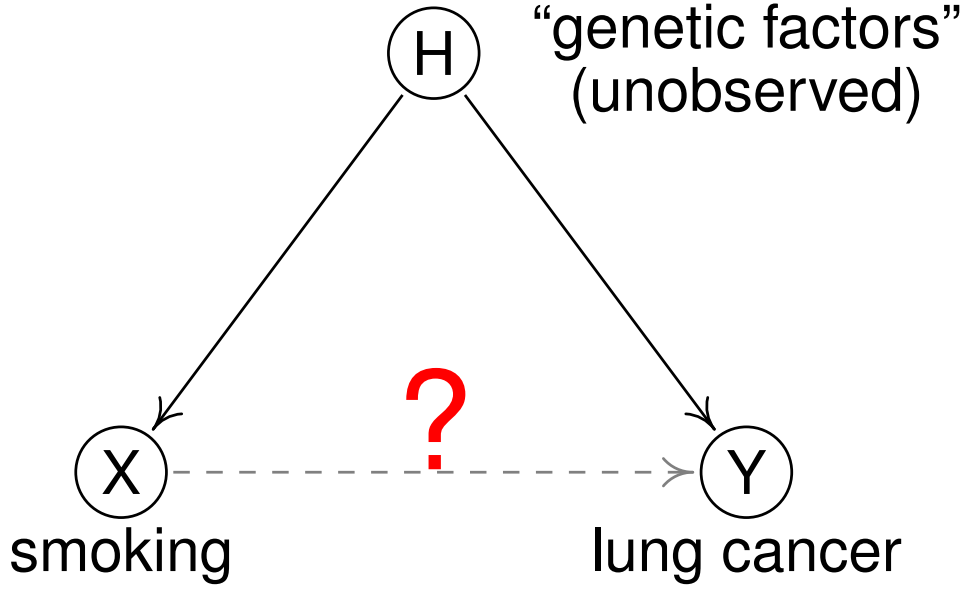
The danger of hidden confounding!

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

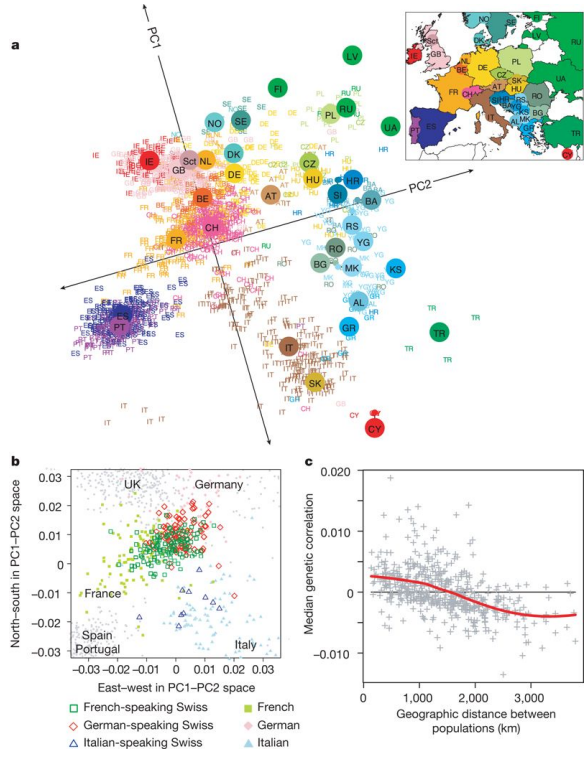
Does smoking cause lung cancer?



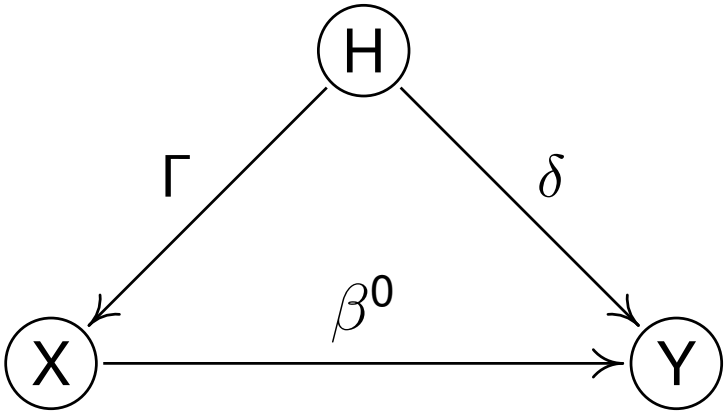


Genes mirror geography within Europe (Novembre et al., 2008)

SNP data plotted on first 2 principal components



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



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

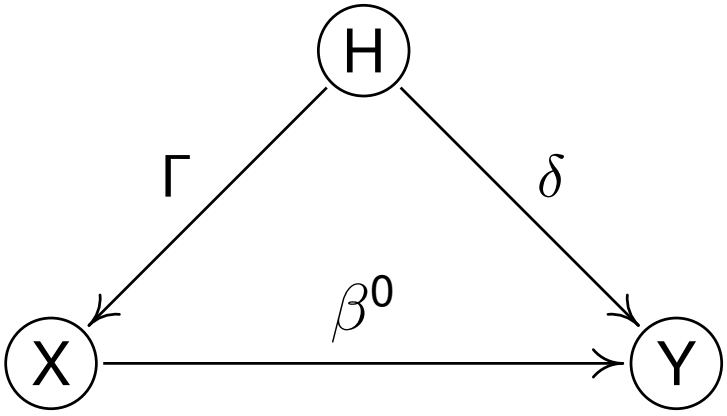
goal: infer β^0 from observations $(X_1, Y_1), \ldots, (X_n, Y_n)$

the population least squares principle leads to the parameter

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

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

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

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

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

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, 2022)

“Trick”: Spectral transformations

motivation: when using e.g. Lasso for the non-sparse problem with $\beta^* = \beta^0 + b$, even when aiming for β^*: due to non-sparsity,

a bias term $\|Xb\|_2^2/n \asymp O(1)$ enters despite that $\|b\|_2^2 \asymp p^{-1}$

strategy: linear transformation $F : \mathbb{R}^n \to \mathbb{R}^n$

$$\tilde{Y} = FY,\ \tilde{X} = FX,\ \tilde{\varepsilon} = F\varepsilon,$$
$$\tilde{Y} = \tilde{X}\beta^* + \tilde{\varepsilon} = \tilde{X}\beta^0 + \tilde{X}b + \tilde{\varepsilon}$$

and use e.g. Lasso for $\tilde{Y}$ versus $\tilde{X}$ such that

- $\|\tilde{X}b\|_2^2/n$ small
- $\|\tilde{X}\beta^0\|_2^2/n$ “large”
- $\tilde{\varepsilon}$ remains “of order $O(1)$”

Spectral transformations

which transform singular values of X will achieve

- $\|\tilde{X}b\|_2^2/n$ small
- $\tilde{X}\beta^0$ "large"
- $\tilde{\varepsilon}$ remains "of order $O(1)$

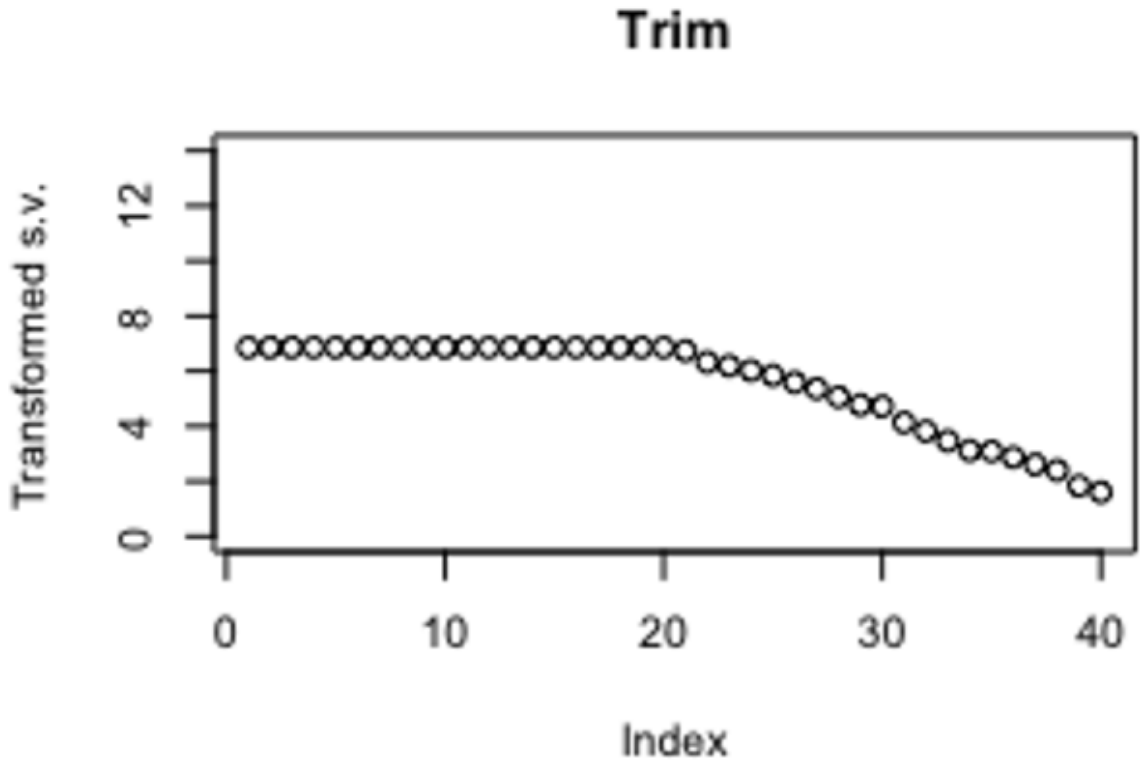
consider SVD of X:

$$
\begin{aligned}
&X = UDV^T, \\
&U_{n\times n}, V_{p\times n},\ U^TU = V^TV = I, \\
&D = \mathrm{diag}(d_1, \ldots, d_{\min(n,p)}),\ d_1 \geq d_2 \geq \ldots \geq d_{\min(n,p)} \geq 0
\end{aligned}
$$

map d_i to $\tilde{d}_i$: spectral transformation is defined as

$$
\begin{aligned}
& F = U\mathrm{diag}(\tilde{d}_1/d_1, \ldots, \tilde{d}_n/d_n)U^T \\
\rightsquigarrow \quad & \tilde{X} = U\tilde{D}V^T
\end{aligned}
$$

“Trick”: Trim (spectral) transform (Ćevid, PB & Meinshausen, 2018)

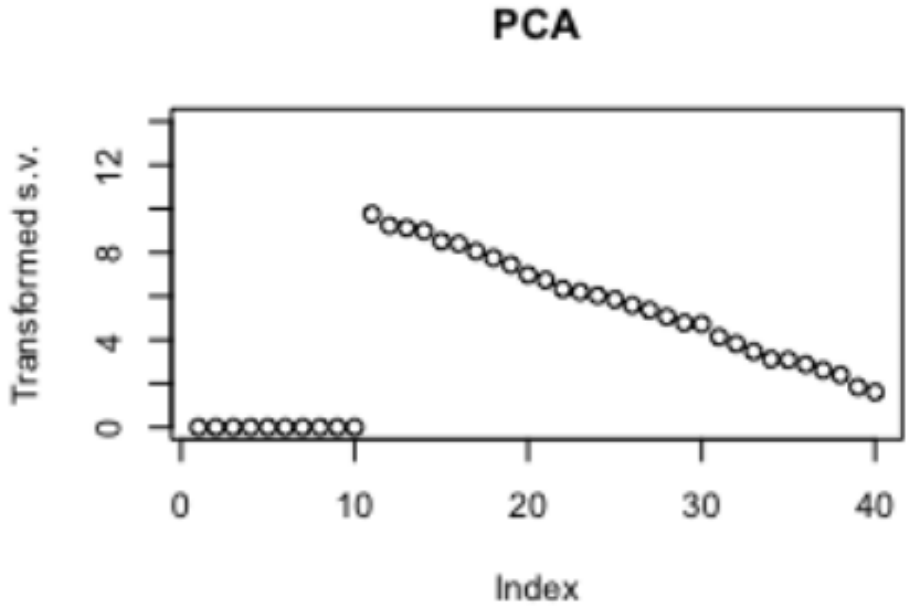


$$\tilde{d}_i = \min(d_i, \tau) \text{ with } \tau = d_{\lfloor n/2 \rfloor}$$

no tuning parameter!

PCA adjustment – used a lot in practice

is also a
spectral transformation



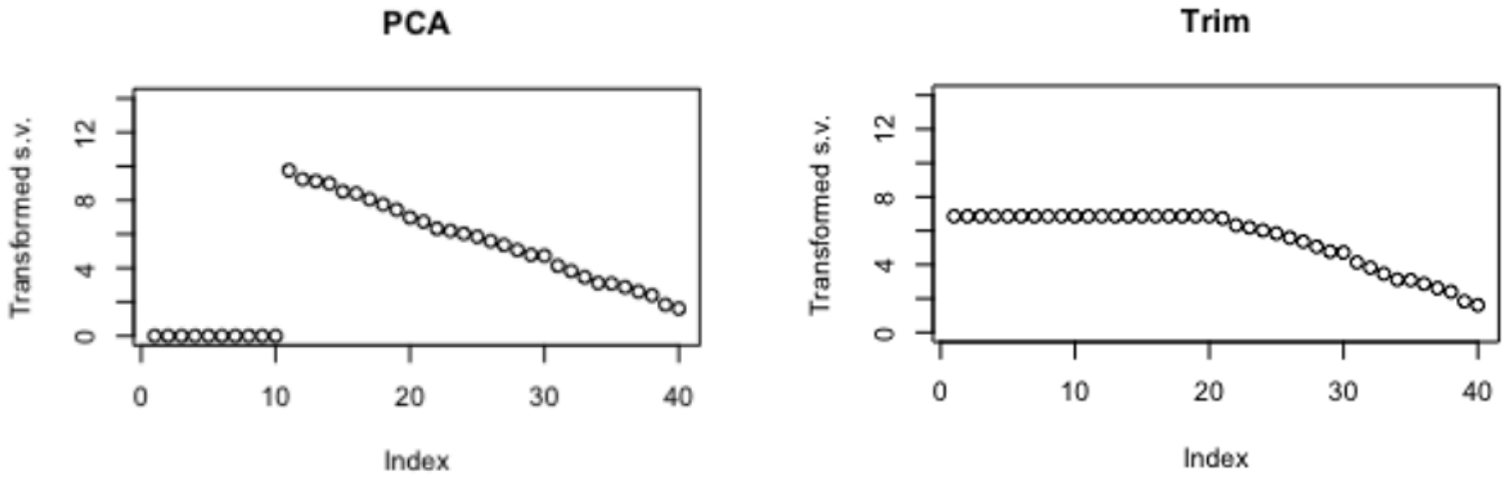
$\hat{H} =$ first k principal components, $\tilde{Y} = (I - \Pi_{\hat{H}})Y$, $\tilde{X} = (I - \Pi_{\hat{H}})X$

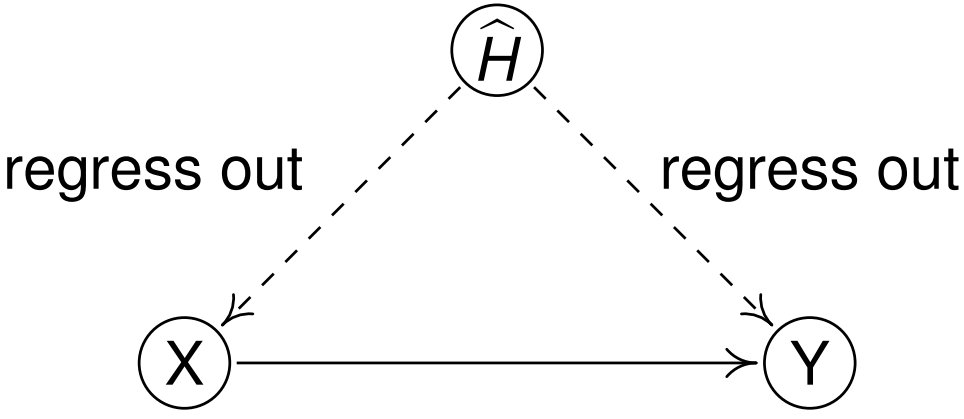
i.e., regressing out the first k principal components

under dense confounding and $k \geq q = \dim(H)$:
$\hat{H} \approx H$ up to rotation (Bai, 2003)

PCA adjustment – used a lot in practice

PCA adjustment does the “opposite” of trim transform





Some (subtle) theory

consider confounding model

$$Y = X\beta^0 + H\delta + \eta,$$
$$X = H\Gamma + E$$

Theorem (Ćevid, PB & Meinshausen, 2018)

- $\dim(H) = q$ not too fast growing
- Γ must spread to $O(qp\log(p)^{3/2}/n)$* components of X

*simplified (Guo, Ćevid & PB, 2021)

components of Γ and δ are i.i.d. sub-Gaussian r.v.s (but then thought as fixed)

Then, for Lasso on trim-transformed $\tilde{X}$, $\tilde{Y}$ and with suitable λ:

$$\|\hat{\beta} - \beta^0\|_1 = O_P\left(\frac{\sigma s_0}{\lambda_{\min}(\Sigma)}\sqrt{\frac{\log(p)}{n}}\right) \quad (p \geq n \to \infty)$$

same optimal rate of Lasso as without confounding variables

Other techniques 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)