Estimation with zeros in the precision matrix

Why is it convenient to have a sparse $\boldsymbol{\Sigma}^{-1}$?

- Because the quadratic form $\boldsymbol{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$ is faster to compute?
- Linear discrimination with a common covariance matrix $\boldsymbol{\Sigma}^{-1}$ in different groups.
- The parameters of $\boldsymbol{\Sigma}^{-1}$ are the canonical parameters of the multivariate normal density.

Pourahmadi and other sources: the quadratic form with $\boldsymbol{\Sigma}^{-1}$ is used in classification applications based on the covariance matrix.

Some references say that $\boldsymbol{\Sigma}^{-1}$ is harder to compute for higher dimensions.

However for p-factor models and other parsimonious dependence structures with a latent factor, $\boldsymbol{\Sigma}^{-1}$ has no zeros but can be obtained in closed form with at most inversion of a low-dimensional matrix. For p-factor, it is only required to invert a small $p \times p$ matrix $\boldsymbol{I}_p + \boldsymbol{A}^T\boldsymbol{D}^{-1}\boldsymbol{A}$, where $\boldsymbol{A}$ is a $d \times p$ matrix of loadings and $\boldsymbol{D}$ is a diagonal matrix of residual variances (pp 135–136, *Dependence modeling with copulas*).

Exponential family model

Data $(y_{i1}, \ldots, y_{id})$, $i = 1, \ldots, n$.

d-variate normal: exponential family with sufficient statistics $\overline{y}_j = n^{-1}\sum_{i=1}^n y_{ij}$, $n^{-1}\sum_{i=1}^n y_{ij}^2$ for $j = 1, \ldots, d$, and $n^{-1}\sum_{i=1}^n y_{ij}y_{ik}$ for $j < k$. σ^{jj}, σ^{jk} are canonical parameters (coefficients of the quadratic sufficient statistics).

Parsimonious if a smaller set of sufficient statistics explains the dependence, if $\sigma^{jk} = 0$, then $n^{-1}\sum_{i=1}^n y_{ij}y_{ik}$ is removed from the vector of sufficient statistics; that is, the correlation ρ_{jk} is a function of other correlations.

Some history:

- Dempster (1972) Covariance selection. *Biometrics* 28, 157–175 [Earliest(?) paper to suggest parsimonious models based on precision matrix or canonical parameters of multivariate normal].
- Wermuth: PhD student of Dempster, many papers and one book with DR Cox. [Examples in their papers and books report precision matrices, but choice of models also involve low-order partial correlations and recursive models (Bayesian networks)]
- Whittaker (1990). Graphical Models in Applied Multivariate Statistics. [The independence graph is a summary. Parameters are estimated by maximum likelihood. Maybe this book influenced the direction for graphical models.]
- Pourahmadi (2013). High-dimensional Covariance Estimation. [Mentions methods based on precision matrix, e.g., glasso, no mention of comparison with maximum likelihood.]

Numerical maximum likelihood given 0 positions in $\boldsymbol{\Sigma}^{-1}$

easier

Assume that data have been centred to have mean 0 so that mean parameters don't have to be estimated??

For maximum likelihood estimation with exponential family models, common numerical methods are iterative proportional fitting and Newton-Raphson. A complication here might be the positive definite constraint implicit in $\boldsymbol{\Sigma}^{-1}$ with some given zeros.

Software for doing this: not sure if code is in one of Wermuth's paper. Also see paper of Speed and Kiiveri (1986, Annals of Statistics, 14, 138–150) for an algorithm.

Iterative proportional fitting and other algorithms for fitting exponential family models are not difficult to implement in low dimensions.

For Homework 2, either code yourself (don't worry about efficient code for the first implementation) or find existing code, and write pseudo-code for your implementation. Also explain how you validate the correctness of your code.

How to decide on zero positions in $\mathbf{\Sigma}^{-1}$ for the graphical model.
A. Whittaker (1990, pp 156–157):

- Compute sample correlation or covariance matrix; gets the sample precision matrix from the matrix inverse, convert to a correlation matrix with 1 on diagonal; change sign of off-diagonal elements to get partial correlations. Note that the sample precision matrix doesn't exist if $d > n$ and it is unreliable if d/n is < 1 but close to 1.

proof in next week's class

- Set any partial correlation which in absolute value is below a tolerance. Draw the independence graph.

B. Regression

generalized inverse is not unique;

- Regress each variable on remainder, set those $\hat{\beta}$'s with small t ratios to zero. This approach might work only if $d < n$ because otherwise multiple regression estimates are not unique.
- With zero positions from the regressions, either apply ML given zero positions in $\mathbf{\Sigma}^{-1}$, or apply methods in Arnold, Castillo and Sarabia (1999), *Conditional Specification of Statistical Models*, Springer: Chapter 3 and Section 8.15 for multivariate distributions with conditional normal distributions given remaining variables.

C. glasso and ridge regression methods when $d > n$ or d/n is not small.

partial least squares can also handle d>n (in R package pos). Main idea is to reduce most betas to 0 anyways.

partial least squares can also handle d>n (in R package pos). Main idea is to reduce most betas to 0 anyways.
partial least squares can also handle d>n (in R package pos). Main idea is to reduce most betas to 0 anyways.

Implementation of Whittaker's approach
$\mathbf{\Sigma}^{-1} = (\sigma^{ij})$ is precision matrix: see previous handout.

$$\sigma^{\mathbf{ij}} \stackrel{\text{sgn}}{=} -\rho_{\mathbf{ij};\mathbf{T(i,j)}} = -\rho_{\mathbf{ij};\text{rest}}, \quad \mathbf{T(i,j)} = \{\mathbf{1},\ldots,\mathbf{d}\}\backslash\{\mathbf{i},\mathbf{j}\}$$

$$\rho_{\mathbf{ij};\text{rest}} = \frac{-\sigma^{\mathbf{ij}}}{\sqrt{\sigma^{\mathbf{ii}}\sigma^{\mathbf{jj}}}}$$

In R:

```
# rmat = sample correlation matrix
rinv=solve(rmat)  # precision matrix
pcmat=cov2cor(rinv)
pcmat=-pcmat
diag(pcmat)=1
# pcmat has partial correlation of each pair given the rest
```

Example, look at these partial correlations (of normal scores) for the abalone data.

Friedman, Hastie, Tibshirani. (2008). Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9, 432–441. [see also Pourahmadi's book, pp 126–130]

Lasso: Regression with L_1 penalty to encourage insignificant betas to be estimated as 0. There is a tuning parameter that affects the number of zeros in the regression coefficients.

Graphical lasso: iteratively apply lasso for each variable regressed on the remaining variables; preserve the positive definite condition to estimate $\mathbf{\Sigma}$ and a sparse $\mathbf{\Sigma}^{-1}$. Numerical observation for $d \ll n$: estimated $\mathbf{\Sigma}$ is not good based on discrepancy of fit for correlation matrix; there tends to be shrinkage of correlations to 0. R packages: `glasso` and `huge`.

Idea of graphical lasso: maybe OK for getting 0 positions but not to estimate covariance matrix (unless shrinking is desired for linear discrimination).