Ridge estimation of a Gaussian graphical model

Wessel N. van Wieringen
w.vanwieringen@vumc.nl
Dept. of Epidemiology and Biostatistics, VUmc
& Dept. of Mathematics, VU University
Amsterdam, The Netherlands

Contributors (in various degrees):
Carel Peeters, Anders Bilgrau, Viktorian Miok, Saskia Wilting,
Poul Svante Eriksen, Martin Bogsted.
**Aim**

*Molecular biology* aims to understand the molecular processes that occur in the cell. That is, which molecules present in the cell interact, and how is this coordinated?

For many cellular processes, it is unknown which genes play what role.

**This talk**
Small contribution of statistics to this endeavour.
**Pathways**

*Pathway* = chain of chemical reactions (that processes a signal)

≈ a set of genes believed to carry one function

Knowledge of pathways is incomplete and biased.

Various repositories contain pathway representations:

- KEGG
- BioCarta
- GennMapp
- Reactome

**Question**

- Reliability of these repositories?
- Which is most reliable? To what end?
Pathways

KEGG: p53 signalling pathway

Pathways

Pathway is represented by a network: $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

- **node**: a gene.
- **edge**: interaction between two genes. (interaction = \ldots)
Guassian graphical model
**Definition**

Let $Y$ be a $p$-dimensional random variable and $K = \{1, \ldots, p\}$ the corresponding set of nodes. The *conditional independence graph* of $Y$ is the undirected graph $G = (V, E)$ where $(j_1, j_2)$ is not in the edge set $E$ if and only if:

$$Y_{j_1} \perp Y_{j_2} \mid Y_{K \setminus \{j_1, j_2\}}$$

**Example**

$$Y_1 \perp Y_2 \mid Y_3$$
Gaussian graphical model

Aim

Assume: \[ Y \sim \mathcal{N}(0_{p \times 1}, \Sigma) \]

Define: \[ \Omega \equiv \Sigma^{-1} \]

Then: \[ (\Omega)_{1,2} = 0 \iff Y_1 \perp Y_2 \mid Y_3, \ldots, Y_p \]
Gaussian graphical model

**Partial correlation**

Element of precision matrix proportional to partial correlation:

$$ \rho(Y_a, Y_b \mid Y_c) = \frac{\text{Cov}(Y_a, Y_b \mid Y_c)}{\sqrt{\text{Var}(Y_a \mid Y_c)} \sqrt{\text{Var}(Y_b \mid Y_c)}} $$
High-dimensionality & networks

In network reconstruction throughout assumed: $p > n$.

Study effect of high-dimensionality in the estimation of partial correlations.

Network of 10 genes.

Visualization of true partial correlations structure:
Gaussian graphical model

Problem
Partial correlation estimates inflate with decreasing sample size.
**Solution**

**Penalization**: discourage “large” parameter estimates.

![Diagram showing Gaussian graphical model with constraint and parameter estimate points]
Ridge precision estimation
Ridge precision estimation

Instead of an ad-hoc fix for the inversion of the covariance matrix, one may also look for a ridge estimate which minimizes:

\[
\mathcal{L}(S, \Sigma^{-1}) - \frac{1}{2} \lambda_2 \| \Sigma^{-1} \|_2^2
\]

When writing \( \Omega = \Sigma^{-1} \) the ridge penalty is:

\[
\| \Omega \|_2^2 = \sum_{j_1, j_2 = 1}^{p} [(\Omega)_{j_1, j_2}]^2
\]
Ridge precision estimation

Again, this penalized estimation problem can be viewed as an constrained estimation problem.

For a 2x2 precision matrix the constraint is:

\[
\begin{align*}
\left[\Omega_{11}\right]^2 + 2\left[\Omega_{12}\right]^2 + \left[\Omega_{22}\right]^2 & \leq c(\lambda_2) \\
\end{align*}
\]

For plot: equal diagonal elements.
Ridge precision estimation

There exists an explicit expression of ridge estimator of the covariance matrix. Hereto consider the ridge loss function:

$$\log(|\Omega|) - \text{tr}(S\Omega) - \frac{1}{2}\lambda_2\text{tr}(\Omega\Omega^T)$$

Equation of the derivative w.r.t. the precision matrix to zero yields the estimating equation:

$$\Omega^{-1} - S - \lambda_2\Omega = 0_{p \times p}$$

Matrix algebra then yields:

$$\hat{\Omega}(\lambda_2) = \left[\frac{1}{2}S + (\lambda_2I_{p \times p} + \frac{1}{4}S^2)^{1/2}\right]^{-1}$$
Ridge precision estimation

Thus:
\[
\hat{\Sigma}(\lambda_2) = \frac{1}{2} \mathbf{S} + (\lambda_2 \mathbf{I}_{p \times p} + \frac{1}{4} \mathbf{S}^2)^{1/2}
\]

For \( \lambda_2 = 0 \), we obtain:
\[
\hat{\Sigma}(0) = \frac{1}{2} \mathbf{S} + (\frac{1}{4} \mathbf{S}^2)^{1/2}
\]
\[
= \frac{1}{2} \mathbf{S} + \frac{1}{2} \mathbf{S} = \mathbf{S}
\]

For large enough \( \lambda_2 \):
\[
\hat{\Sigma}(\lambda_2) \approx \lambda_2 \mathbf{I}_{p \times p}
\]
Ridge precision estimation

The derived ridge covariance estimator is positive definite.

Consider the eigen-decomposition:

\[ S = \mathbf{V}\mathbf{D}\mathbf{V}^T \]

with \( \mathbf{V} \) and \( \mathbf{D} \) the eigenvalue and -vector matrices.

The eigenvalues of \( \hat{\Sigma}(\lambda_2) \) are then:

\[ d_j[\hat{\Sigma}(\lambda_2)] = \frac{1}{2}d_j + (\lambda_2 + \frac{1}{4}d_j^2)^{1/2} \]

\( S \) semi-positive definite: all eigenvalues non-negative.
\( \rightarrow \) those of ridge covariance estimator all positive.
Ridge precision estimation

Effect of ridge penalty on the 95% confidence interval.
Ridge precision estimation

Effect of $\lambda_2$ on the partial correlation estimates

Non-monotone decrease with $\lambda_2$. 
Ridge precision estimation

Bias approximation
The bias of the ridge ML covariance estimate can be approximated using a binomial expansion of the square root:

\[
\hat{\Sigma}(\lambda_2) = \sqrt{\lambda_2} \left[ (I_p + U^2)^{1/2} + U \right]
\]

\[
= \sqrt{\lambda_a} U + \sqrt{\lambda_a} \sum_{k=0}^{\infty} \binom{1/2}{k} U^{2k}
\]

with \( U = (S - \lambda_2 T) / (2\sqrt{\lambda_2}) \).

For its evaluation:
→ limit the expansion to the desired degree,
→ replace powers of \( S \) by their expectation from the Wishart.
Ridge precision estimation

Bias approximation
This approach approximates $E[\hat{\Sigma}(\lambda_2)]$ by, e.g.:

$$\frac{1}{2} \Sigma + \sqrt{\lambda_a} I_{p \times p} + \frac{1}{8 \sqrt{\lambda_2}} \left[ \frac{n + 1}{n} \Sigma^2 + \frac{1}{n} \text{tr}(\Sigma) \Sigma \right]$$

Compare the bias approximation of ridge covariance estimator to the estimated bias over many samples:

$$\widehat{\text{Bias}}[\hat{\Sigma}(\lambda_2)] = \frac{1}{K} \left[ \sum_{k=1}^{K} \hat{\Sigma}^{(k)}(\lambda_2) - \Sigma \right]$$
Ridge precision estimation

Bias approximation

Results for two elements:
Ridge precision estimation

The ridge covariance estimator is biased but consistent.

**Proposition**
Let $S_n$ be the sample covariance matrix from a sample $Y_1, \ldots, Y_n$ drawn from a centered $p$-variate Gaussian. Denote by $\lambda_{2,n}$ a nonnegative random variable that converges almost surely to zero. Then:

$$\lim_{n \to \infty} E \left( \left\| \hat{\Sigma}_n(\lambda_{2,n}) - \Sigma \right\|_F^2 \right) = 0$$

Subscript $n$ explicates the dependence on the sample size.
Ridge precision estimation

**Network reconstruction**

1) *Choice of penalty parameter*
   
   LOOCV and approximate LOOCV (faster but unstable for small $n$).

2) *Edge selection*
   
   Partial correlation distribution modeled by mixture:
   
   $$(\eta_0) f_0 + (1 - \eta_0) f_E$$
   
   with $f_0(\cdot)$ and $f_E(\cdot)$ the densities of partial correlations of absent and present edges. Posterior probability of partial correlation stemming from $f_E(\cdot)$ used to decide on edge presence.
Ridge precision estimation

*R-package*

*rggs2ridges*  
(available from CRAN)
Comparisons

Many ways to regularize covariance estimation. Some of these estimators are often referred to as “ridge” estimates:

\[ S + \lambda_a I_{p \times p} \quad \text{for} \quad \lambda_a > 0 \]

and:

\[ (1 - \lambda_a)S + \lambda_a T \quad \text{for} \quad \lambda_a \in (0, 1) \]

where \( T \) is some nonrandom, pos. definite matrix.

Both are not derived from a penalized loss function, but are simply ad-hoc fixes to resolve the singularity of the estimate.
Comparisons

Ad-hoc vs. ridge

The two penalized covariance (precision) estimators do not differ dramatically (for low-dimensional data).

![correlations](image1)

![part. correlations](image2)
Comparisons

Ad-hoc vs. ridge

Compare the eigenvalues of the ad-hoc and ridge estimator of the covariance matrix.

Recall: the eigen-decomposition of the ad-hoc estimator of the covariance matrix is:

\[
\Sigma_{\text{ad hoc}}(\lambda_a) = V(\lambda_a I_{p \times p} + D)V^T
\]

For the comparison map the penalty parameters \( \lambda_2 \) and \( \lambda_a \) to the same scale by setting \( \lambda_2 = (\lambda_a)^2 \).
Comparisons

Ad-hoc vs. ridge

Writing \( d_j = (\mathbf{D})_{jj} \) it is easily seen that:

\[
\lambda_a + d_j \geq \frac{1}{2} d_j + \sqrt{\lambda_a^2 + \frac{1}{4} d_j^2}
\]

Thus, the ad-hoc estimator shrinks the eigenvalues of the sample covariance matrix more than the ridge estimator.

In turn, this eigenvalue inequality implies:

\[
\mathcal{L}[\mathbf{S}, \hat{\Omega}^{\text{ad hoc}}(\lambda_a)] \leq \mathcal{L}[\mathbf{S}, \hat{\Omega}^{\text{ridge}}(\lambda_a^2)]
\]

That is, the ridge estimator yields a larger likelihood.
Comparisons

Ad-hoc vs. ridge (loss comparison)

chain topology ($p=100$)
Comparisons

What about lasso? Everybody loves Ridge!
Comparisons

Setting
- GE data of 3 pathways from 5 breast cancer data sets
- Subsampling with sample sizes $n=5, 10, 25$
- 100 subsamples for each pathway–dataset–size combination
- Penalty determined by LOOCV

Loss
- Standardized precision estimates
- Population defined on all samples
- Frobenius and quadratic loss
- *Ridge estimator yields lower loss*
- ‘True’ conditional dependencies based on consensus truth: top 100 $\alpha$ in at least 4 out of 5 datasets by both ridge and glasso

- *Ridge estimator is more sensitive*

- Edge defined ‘stable’ when selected in the union of the top 100 $\alpha$ over the respective subsample sizes $n=5, 10, 25$

- *Ridge yields more stable networks*
Incorporation of prior information through a target

Why a target? (I)
Both the ad-hoc and ridge covariance estimator converge to:

\[ \hat{\Sigma}(\lambda_2) \approx \lambda_2 I_{p \times p} \quad \text{for large enough } \lambda_2. \]

Its inverse (the precision matrix) converges to the zero matrix. Also the diagonal elements! Consequently, the partial correlation of this matrix are undefined.

Hence, in the limit the penalized precision estimate equals a semi-positive definite matrix.
Why a target? (II)

Poor resemblance inferred network and repository:

If signal-to-noise ratio is poor, why not provide a hint.
Ridge & target
To ensure the ridge precision estimate converges to a positive definite target matrix $T$, the latter is incorporated in the penalty:

$$\frac{1}{2} \lambda_2 \text{tr}[(\Omega - T)(\Omega - T)^T]$$

Clearly, the penalty is minimized for $\Omega = T$.

Intuition
One expects that, for large $\lambda_2$, the maximization of the penalized log-likelihood requires the minimization of the penalty: the optimum moves close to $T$. 
Prior information I

Estimator

The log-likelihood augmented with this “target”-penalty is maximized by:

\[
\left\{ \frac{1}{2}(S - \lambda_2 T) + \left[ \lambda_2 I_{p \times p} + \frac{1}{4}(S - \lambda_2 T)^2 \right]^{1/2} \right\}^{-1}
\]

For generalized ridge precision estimator one can show e.g. that:

\[
\lim_{\lambda_2 \to \infty} \hat{\Omega}(\lambda_2) = T
\]

and \( \hat{\Omega}(\lambda_2) \succeq 0 \) for all \( \lambda_2 > 0 \).
Prior information I

Effect of the target

Simulation:
- Define a banded $p \times p$ precision matrix, $p = 100$.
- Draw $n = 10$ samples.
- Determine optimal lambda by LOOCV.
- Estimate precision matrix with and without target. Target is true precision.
Prior information I

Effect of the target

True network

zoom
Prior information I

Effect of the target

Null target
opt. $\lambda_2 = 276$

Target = true precision
opt. $\lambda_2 = 2525612760$
Prior information I

Effect of the target

True network

null target network

inf. target network
Prior information I

Use of the target
Estimator:

$$\left\{ \frac{1}{2}(S - \lambda_2 T) + \left[ \lambda_2 I_{p \times p} + \frac{1}{4}(S - \lambda_2 T)^2 \right]^{1/2} \right\}^{-1}$$

$T = C \hat{\Sigma}_{pilot}^{-1}$

Pilot data (GEO) + topology

Estimated from data ≈ “empirical Bayes”.
Incorporation of prior information through constraints
Knowledge on edge absence
If there a microRNA and an mRNA do not have a complementary (sub)sequence, there cannot be a direct edge connecting the two.
Prior information II

Known absence of edges
Absence of an edges $\leftrightarrow$ zero element of $\Omega$

*Problem*: ridge estimator does not yield zeros (as lasso).

But ...
$\rightarrow$ known zeros can be incorporated:

$$\max_{(\Omega)_{j_1,j_2}=0 \text{ for } (j_1,j_2) \not\in \mathcal{E}} \log(|\Omega|) - \text{tr}(S\Omega) - \lambda_2 f_{\text{pen}}(\Omega, T)$$

Only for *decomposable* graphs. Not strictly necessary, but:
$\rightarrow$ good initial estimate available,
$\rightarrow$ fast inverse.
Prior information II

Known absence of edges
Decomposable graph

graph

support of precision matrix
Prior information II

Known absence of edges
Decomposable graph

**graph**

**support of precision matrix**
Constrained estimation

Reformulate constrained to unconstrained problem through reparametrization of precision matrix:

\[
\max_x \log |\Omega(x)| - \text{tr}[S\Omega(x)] - \lambda_2 f^{\text{pen}}[\Omega(x), T]
\]

in which \(x\) is a vector with all non-zero, nonredundant elements of \(\Omega\).

This unconstrained estimation problem can be solved by means of Newton-Raphson or the like.
Prior information II

Known absence of edges

Compare initial guess and final estimator

\[ \text{loss gain: } -50.33 \text{ vs. } -50.32 \]

\[ \text{loss gain: } -120.32 \text{ vs. } -117.46 \]
Prior information II

Known absence of edges
Illustration on 3-banded chain graph \((p=50, n=10)\)

*without known absent edges*

*with known absent edges*
All nice ...

... but to what end?
Pathway reconstruction

Visualization is important!

Question
Which genes interact? E.g. do genes RB1 and E2F1 interact?
Pathway reconstruction
Visualization is important!
Two instances of same network:
“The most striking property of cancer proteins is the increased frequency of interactions they participate in. This observation indicates an underlying evolutionary pressure to which cancer genes, as genes of central importance, are subjected.”
All nice ...

**Node analysis**

*Hub* \(\approx\) many connections.

- gene with high degree
- gene with low degree

**Hypothesis**

Hubs are disease genes.

Infer network and compare to *census of human cancer genes*\(^*\) from:

Futreal et al. (2004), *Nature Reviews Cancer*.

**Question**: role of the hub?

Hypothesis not confirmed.
Node analysis

Measure influence between gene and rest by *mutual information*:

\[
\mathcal{I}(Y_{\setminus j}; Y_j) = \mathcal{H}(Y_{\setminus j}) - \mathcal{H}(Y_{\setminus j} | Y_j) = \log\{\|\text{Var}(Y_{\setminus j})\|\} - \log\{\|\text{Var}(Y_{\setminus j} | Y_j)\|\}
\]

*Hypothesis*

Hubs are influential

Measure of information shared between two random variables.
All nice ...

**Path analysis**

Understand the covariance between genes 3 and 4 by decomposition into the paths that propagate signals between these genes.

Which path contributes most to covariance?
Path analysis

The covariance between two nodes can be decomposed into the contributions of the paths connecting these nodes.

The covariance between nodes $j_1$ and $j_2$ equals:

$$\left( \Sigma \right)_{j_1,j_2} = \sum_{P \in \mathcal{P}_{j_1j_2}} (-1)^{r+1} \frac{\det(\Omega_{\setminus P,\setminus P})}{\det(\Omega)} \prod_{s=2}^{r} (\Omega)_{p_{s-1},p_s}$$

where $\mathcal{P}_{j_1j_2}$ the set of all paths from $j_1$ to $j_2$ and

$$P = \left\{ (p_1 = j_1, p_2), (p_2, p_3), \ldots, (p_{r-1}, p_r = j_2) \right\}$$

a path of length $r$ from $j_1$ to $j_2$. 
All nice ...

Path analysis

# E2F1 and RB1
node1 <- 5
node2 <- 11
pathStats <- GGMpathStats(sparseP, node1, node2,
    nodecol="red", VBcolor="orange")

Covariance between node pair : -0.0103

<table>
<thead>
<tr>
<th>path</th>
<th>length</th>
<th>contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5--11</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>5--7--11</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>5--10--7--11</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5--10--1--7--11</td>
<td>4</td>
</tr>
</tbody>
</table>

Sum path contributions : -0.0103
All nice ...

Path analysis

Top mediating and moderating paths are plotted.

*Mediating path:* contribution has same sign as observed covariance

*Moderating path:* contribution has opposite sign as observed covariance
Extensions
(in development)

Further topics

Network differences

Group A

\[
\begin{align*}
Y_1 & \sim \mathcal{N}(0_{p \times 1}, \Sigma_a) \\
Y_2 & \sim \mathcal{N}(0_{p \times 1}, \Sigma_a) \\
Y_3 & \sim \mathcal{N}(0_{p \times 1}, \Sigma_a) \\
Y_4 & \sim \mathcal{N}(0_{p \times 1}, \Sigma_a) \\
Y_5 & \sim \mathcal{N}(0_{p \times 1}, \Sigma_a) \\
Y_6 & \sim \mathcal{N}(0_{p \times 1}, \Sigma_a)
\end{align*}
\]

Group B

\[
\begin{align*}
Y_1 & \sim \mathcal{N}(0_{p \times 1}, \Sigma_b) \\
Y_2 & \sim \mathcal{N}(0_{p \times 1}, \Sigma_b) \\
Y_3 & \sim \mathcal{N}(0_{p \times 1}, \Sigma_b) \\
Y_4 & \sim \mathcal{N}(0_{p \times 1}, \Sigma_b) \\
Y_5 & \sim \mathcal{N}(0_{p \times 1}, \Sigma_b) \\
Y_6 & \sim \mathcal{N}(0_{p \times 1}, \Sigma_b)
\end{align*}
\]
Further topics

Network differences

**Fused ridge penalty:**

\[ \lambda_2 \sum_{g=1}^{G} \| \Omega_g \|_2^2 + \]

\[ \lambda_f \sum_{g_1, g_2 = 1}^{G} \| \Omega_{g_1} - \Omega_{g_2} \|_2^2 \]

Could include target.
Further topics

Dynamic networks

The VAR(1) model:

\[ Y_t = A Y_{t-1} + \varepsilon_t \]

with \( \varepsilon_t \sim \mathcal{N}(0_{p \times 1}, \Omega_{\varepsilon}^{-1}) \)

If \( p > n \), augment log-likelihood with ridge penalty:

\[ \lambda_a \| A \|_2^2 + \lambda_\omega \| \Omega_{\varepsilon} \|_2^2 \]

possibly including target, null structure, et cetera.
Further topics

Dynamic networks
This material is provided under the Creative Commons Attribution/Share-Alike/Non-Commercial License.

See http://www.creativecommons.org for details.