Regularization of covariance matrix

Bo Chang

Graphical Models Reading Group

May 29, 2015

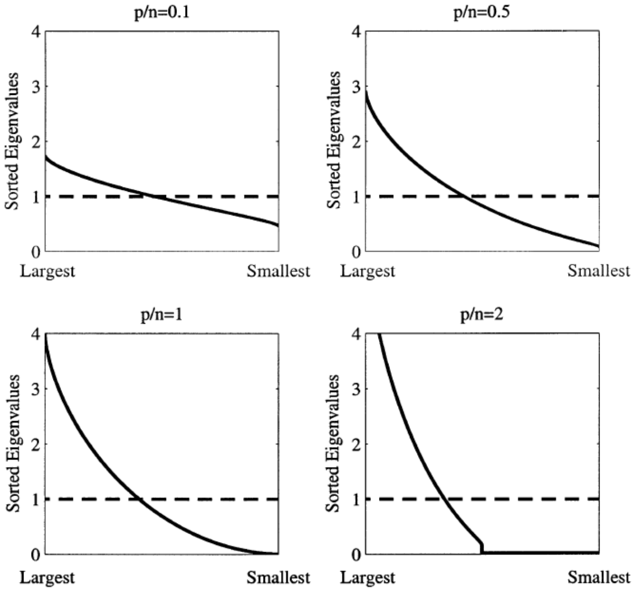
Distorted eigenstructure

- Sample covariance matrix

$$\mathbf{S} = \frac{1}{n}\mathbf{X}'\mathbf{X}.$$

- The eigenstructure of $\mathbf{S}$ tends to be systematically distorted unless p/n is small.
- Larger eigenvalues are overestimated; smaller eigenvalues are underestimated.

Distorted eigenstructure



Loss and risk functions

- Two commonly used loss functions when $n > p$

$$L_1(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}) = \mathrm{tr}(\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}) - \log|\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}| - p,$$

$$L_2(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}) = \mathrm{tr}[(\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1} - \mathbf{I})^2],$$

  where $\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Sigma}}(\mathbf{S})$ is an estimator.
- Risk functions

$$R_i(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}) = \mathbb{E}(L_i(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma})), \quad i = 1, 2.$$

- Among all the estimators $\hat{\mathbf{\Sigma}} = a\mathbf{S}$ where a is a scalar, $\mathbf{S}$ is optimal under L_1 and $\frac{n}{n+p+1}\mathbf{S}$ is optimal under L_2.

Shrinking sample eigenvalues

- The spectral decomposition of $\mathbf{S}$ is

$$\mathbf{S} = \mathbf{Q}\mathrm{diag}(\lambda_1, \ldots, \lambda_p)\mathbf{Q}',$$

where $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ are the eigenvalues of $\mathbf{S}$, and $\mathbf{Q}$ is an orthogonal matrix whose columns are corresponding eigenvectors.

- Stein (1956) proposed the class of Steinian shrinkage estimators:

$$\hat{\mathbf{\Sigma}} = \mathbf{Q}\mathrm{diag}(\varphi_1, \ldots, \varphi_p)\mathbf{Q}',$$

where $\varphi_j = \varphi_j(\lambda)$ estimates the jth largest eigenvalue of $\mathbf{\Sigma}$.

Shrinking sample eigenvalues

Stein's estimator

- $\hat{\mathbf{\Sigma}}_{\text{Stein}} = \mathbf{Q}\text{diag}(\varphi_1, \ldots, \varphi_p)\mathbf{Q}'$.
- $\varphi_j = \lambda_j / \alpha_j$, where

$$\alpha_j = \frac{n - p + 1 + 2\lambda_j \sum_{i \neq j} (\lambda_j - \lambda_i)^{-1}}{n}.$$

- $\hat{\mathbf{\Sigma}}_{\text{Stein}}$ approximately minimizes the L_1 risk.

Ledoit-Wolf shrinkage estimator

- Modified Frobenius norm and inner product:

$$\|\mathbf{A}\| = \sqrt{p^{-1}\mathrm{tr}(\mathbf{A}\mathbf{A}')}.$$

$$\langle \mathbf{A}_1, \mathbf{A}_2 \rangle = p^{-1}\mathrm{tr}(\mathbf{A}_1\mathbf{A}_2').$$

- Ledoit and Wolf (2004) used a modified Frobenius norm as the loss function.

$$L_3(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}) = \|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|^2 = p^{-1}\mathrm{tr}[(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})^2]$$

- To ensure non-singularity, they proposed a shrinkage estimator

$$\hat{\mathbf{\Sigma}}_{\mathrm{LW}} = \alpha_1\mathbf{I} + \alpha_2\mathbf{S}.$$

Ledoit-Wolf shrinkage estimator

- To minimize L_3 risk,

$$\hat{\mathbf{\Sigma}}_{\mathrm{LW}} = \frac{\beta^2}{\delta^2}\mu\mathbf{I} + \frac{\alpha^2}{\delta^2}\mathbf{S},$$

  where

$$\mu = \langle \mathbf{\Sigma}, \mathbf{I} \rangle, \quad \alpha^2 = \|\mathbf{\Sigma} - \mu\mathbf{I}\|^2,$$

$$\beta^2 = \mathbb{E}\|\mathbf{S} - \mathbf{\Sigma}\|^2, \quad \delta^2 = \mathbb{E}\|\mathbf{S} - \mu\mathbf{I}\|^2.$$

- $$\mathbb{E}\|\hat{\mathbf{\Sigma}}_{\mathrm{LW}} - \mathbf{\Sigma}\|^2 = \frac{\alpha^2\beta^2}{\delta^2}$$

- Since $\alpha^2 + \beta^2 = \delta^2$, $\hat{\mathbf{\Sigma}}_{\mathrm{LW}}$ is a convex combination of $\mu\mathbf{I}$ and $\mathbf{S}$.

Geometric interpretation

- A Hilbert space. Norm: $\sqrt{\mathbb{E}(\|\mathbf{A}\|^2)}$. Inner product: $\mathbb{E}(\langle \mathbf{A}_1, \mathbf{A}_2 \rangle)$.

$$\hat{\mathbf{\Sigma}}_{\mathrm{LW}} = \frac{\beta^2}{\delta^2}\mu\mathbf{I} + \frac{\alpha^2}{\delta^2}\mathbf{S},$$

$$\mu = \langle \mathbf{\Sigma}, \mathbf{I} \rangle, \alpha^2 = \|\mathbf{\Sigma} - \mu\mathbf{I}\|^2,\ \beta^2 = \mathbb{E}\|\mathbf{S} - \mathbf{\Sigma}\|^2, \delta^2 = \mathbb{E}\|\mathbf{S} - \mu\mathbf{I}\|^2.$$

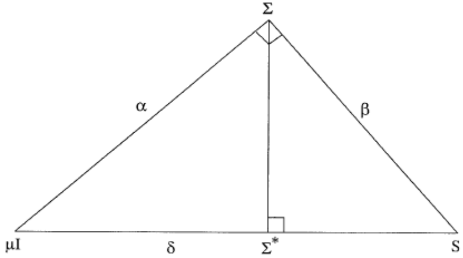


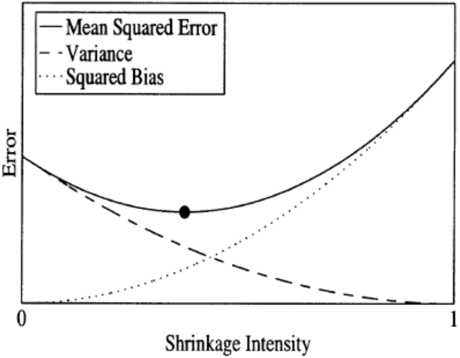
Fig. 1. Theorem 2.1 interpreted as a projection in Hilbert space.

Ledoit-Wolf shrinkage estimator

Bias-variance trade-off:

$$\mathbb{E}(\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|^2) = \mathbb{E}(\|\hat{\mathbf{\Sigma}} - \mathbb{E}(\hat{\mathbf{\Sigma}})\|^2) + \|\mathbb{E}(\hat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\|^2$$

- $\mu \mathbf{I}$: all bias no variance.
  $\mathbf{S}$: all variance no bias.

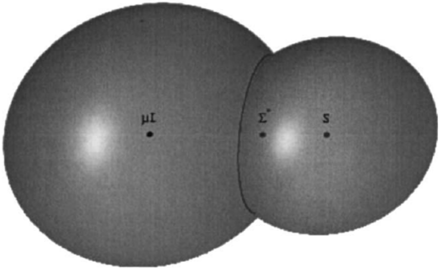


Ledoit-Wolf shrinkage estimator

Bayesian interpretation

- Prior information: $\mathbf{\Sigma}$ lies on the sphere centered around $\mu\mathbf{I}$ with radius α.
- Sample information: $\mathbf{\Sigma}$ lies on the sphere centered around $\mathbf{S}$ with radius β.

Bayesian Interpretation



Ledoit-Wolf shrinkage estimator

Shrinkage of sample eigenvalues

$$\hat{\mathbf{\Sigma}}_{\mathrm{LW}} = \frac{\beta^2}{\delta^2}\mu\mathbf{I} + \frac{\alpha^2}{\delta^2}\mathbf{S},$$

- Shrinking the sample eigenvalues towards their grand mean.
- Steinian shrinkage estimator:

$$\varphi_j = \varphi_j(\lambda_j) = \frac{\beta^2}{\delta^2}\mu + \frac{\alpha^2}{\delta^2}\lambda_j.$$

Ledoit-Wolf shrinkage estimator

$$\hat{\mathbf{\Sigma}}^*_{\mathrm{LW}} = \frac{b^2}{d^2} m\mathbf{I} + \frac{a^2}{d^2}\mathbf{S},$$

where

- $m = \langle \mathbf{S}, \mathbf{I} \rangle$ is a consistent estimator of μ,
- $d = \|\mathbf{S} - m\mathbf{I}\|^2$ is a consistent estimator of δ^2,
- $b^2 = \min(d^2, \bar{b}^2)$ is a consistent estimator of β^2, where

$$\bar{b}^2 = \frac{1}{n^2} \sum_{k=1}^{n} \|\mathbf{X}'_{k\cdot}\mathbf{X}_{k\cdot} - \mathbf{S}\|^2,$$

- $a^2 = d^2 - b^2$ is a consistent estimator of α^2.

Ridge estimation of correlation matrix

Warton (2008)

- The sample correlation matrix is regularized as

$$\hat{\mathbf{R}}_\alpha = \alpha\hat{\mathbf{R}} + (1-\alpha)\mathbf{I},$$

  where $\hat{\mathbf{R}}$ is the sample correlation matrix.
- Properties: shrinkage to $\mathbf{I}$, bias-variance trade-off.

Ridge estimation of correlation matrix

Estimation of α: K-fold cross validation

- Split the data into K parts $\mathbf{X}' = (\mathbf{X}'_1, \ldots, \mathbf{X}'_K)$.
- $\mathbf{X}_j$ is reserved as the validation data and all others $\mathbf{X}^{-j}$ are used as training data.
- Estimate α to maximize the cross-validated log-likelihood function.

$$\hat{\alpha} = \mathrm{argmax}_{\alpha} \sum_{j=1}^{K} \log L(\hat{\boldsymbol{\mu}}^{-j}, \hat{\boldsymbol{\Sigma}}_{\alpha}^{-j}; \mathbf{X}_j)$$

Condition number regularization

Won et al. (2013)

- The condition number of a positive definite matrix $\boldsymbol{\Sigma}$ is

$$\mathrm{cond}(\boldsymbol{\Sigma}) = \frac{\lambda_{\mathsf{max}}(\boldsymbol{\Sigma})}{\lambda_{\mathsf{min}}(\boldsymbol{\Sigma})}$$

- Constrained maximum likelihood estimation:

$$\begin{aligned} \text{maximize} \quad & l(\boldsymbol{\Sigma}) \\ \text{s.t.} \quad & \mathrm{cond}(\boldsymbol{\Sigma}) \leq \kappa_{\mathsf{max}}. \end{aligned}$$

Condition number regularization

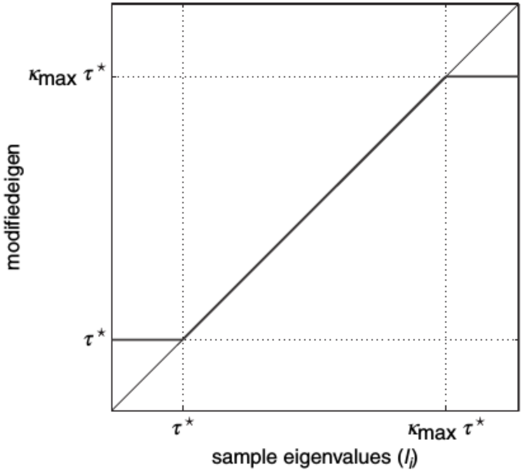
- Steinian shrinkage estimator:

$$\varphi_j = \min\{\max\{\tau^*, \lambda_j\}, \kappa_{\max}\tau^*\} = \begin{cases} \tau^*, & \lambda_j \leq \tau^*, \\ \lambda_j, & \tau^* < \lambda_j < \kappa_{\max}\tau^*, \\ \kappa_{\max}\tau^*, & \lambda_j \geq \kappa_{\max}\tau^*, \end{cases}$$

  for some τ^*, which is determined by the data and $\kappa_{\max}$.

- τ^* can be found exactly and easily in $O(p)$ time.

Condition number regularization



References

- Pourahmadi, M. (2011). Covariance estimation: The GLM and regularization perspectives. Statistical Science, 26(3), 369-387.
- Pourahmadi, M. (2013). High-Dimensional Covariance Estimation: With High-Dimensional Data. John Wiley & Sons.
- Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In Proceedings of the Third Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 399, pp. 197-206).
- Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. Journal of multivariate analysis, 88(2), 365-411.
- Warton, D. I. (2008). Penalized normal likelihood and ridge regularization of correlation and covariance matrices. Journal of the American Statistical Association, 103(481).
- Won, J. H., Lim, J., Kim, S. J., & Rajaratnam, B. (2013). Condition-number-regularized covariance estimation. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75(3), 427-450.

The End