Covariance estimation with Cholesky decomposition and generalized linear model

Bo Chang

Graphical Models Reading Group

May 22, 2015

Modified Cholesky decomposition

- **Goal**: Find a re-parameterization of a covariance matrix that is *unconstrained* and *statistically interpretable*.
- Assume $Y = (Y_1, \ldots, Y_p)'$ is an ordered (time-ordered) random vector with mean 0 and covariance matrix $\mathbf{\Sigma}$.

$$Y_t = \sum_{j=1}^{t-1} \phi_{t,j} Y_j + \epsilon_t.$$

- Let $\sigma_t^2 = \mathrm{Var}(\epsilon_t)$ and

$$\mathrm{Cov}(\epsilon) = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2) = \mathbf{D}.$$

Modified Cholesky decomposition

- Rearranging

$$Y_t = \sum_{j=1}^{t-1} \phi_{t,j} Y_j + \epsilon_t,$$

we have $\mathbf{T}Y = \epsilon$, where

$$\mathbf{T} = \begin{pmatrix} 1 & & & & \\ -\phi_{2,1} & 1 & & & \\ -\phi_{3,1} & -\phi_{3,2} & 1 & & \\ \vdots & \vdots & & \ddots & \\ -\phi_{p,1} & -\phi_{p,2} & \cdots & -\phi_{p,p-1} & 1 \end{pmatrix}.$$

-

$$\mathrm{Cov}(\mathbf{T}Y) = \mathrm{Cov}(\epsilon) = \mathbf{T}\boldsymbol{\Sigma}\mathbf{T}' = \mathbf{D}.$$

Modified Cholesky decomposition

- Definition: For a positive-definite covariance matrix $\boldsymbol{\Sigma}$, its **modified Cholesky decomposition** is

$$\mathbf{T}\boldsymbol{\Sigma}\mathbf{T}' = \mathbf{D},$$

  where $\mathbf{T}$ is a unique unit lower-triangular matrix having ones on its diagonal and $\mathbf{D}$ is a unique diagonal matrix.
- Precision matrix can be written as

$$\boldsymbol{\Sigma}^{-1} = \mathbf{T}'\mathbf{D}^{-1}\mathbf{T}.$$

- $\mathbf{T}$ is unconstrained and statistically meaningful.
- $\mathbf{T}$ and $\mathbf{D}$ can be fitted by regressing a variable Y_t on its predecessors.

Sparse estimation

k-banding:

- $AR(k)$ model.

$$Y_t = \sum_{i=1}^{k} \phi_{t,t-i} Y_{t-i} + \epsilon_t$$

- The resulting estimate of the precision matrix is also k-banded.

Sparse estimation

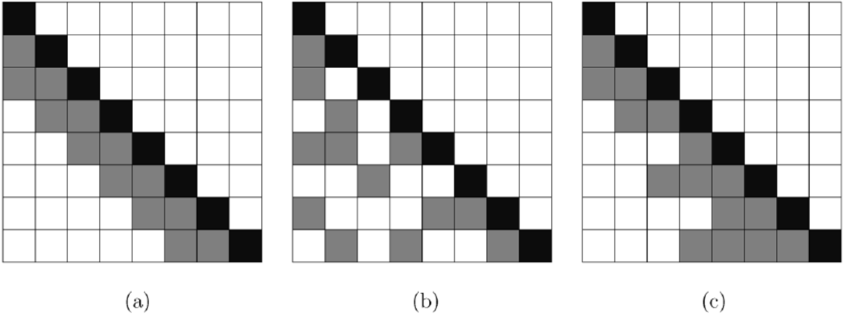


FIG. 1. *The placement of zeros in the Cholesky factor T*: (a) *Banding*; (b) *Lasso penalty of Huang et al.*; (c) *Adaptive banding*.

k-banding:

- Nonparametric estimation: Wu and Pourahmadi (2003) used local polynomial estimators to smooth the subdiagonals of $\mathbf{T}$.

$$\sum_{j=0}^{k} f_{j,p}(t/p)Y_{t-j} = \sigma_p(t/p)\varepsilon_t,$$

  where $f_{0,p}(\cdot) = 1$, $f_{j,p}(\cdot)$ and $\sigma_p(\cdot)$ are continuous functions on $[0, 1]$. ε_t are independent with mean 0 and variance 1.

- $$\phi_{t,t-j} = f_{j,p}(t/p), \quad \sigma_t = \sigma_p(t/p).$$

Sparse estimation

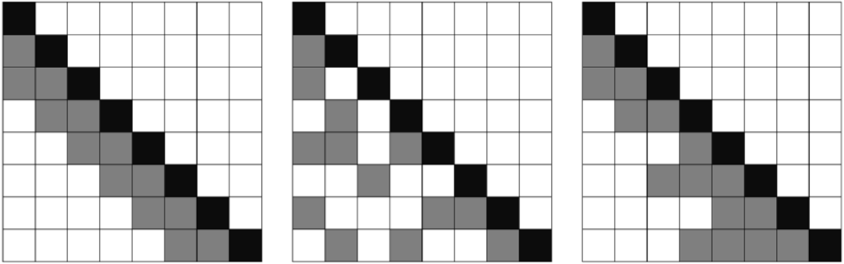
Lasso penalty: Huang et al. (2006)

- Minimize

$$n \log |\mathbf{\Sigma}| + n\mathrm{tr}(\mathbf{D}^{-1}\mathbf{TST}') + \lambda \sum_{t=2}^{p} \sum_{j=1}^{t-1} |\phi_{t,j}|.$$

- Zeros are placed in $\mathbf{T}$ with no regular patterns.
- Sparsity of the precision matrix is not guaranteed.

Sparse estimation



(a) (b) (c)

FIG. 1. *The placement of zeros in the Cholesky factor T*: (a) *Banding*; (b) *Lasso penalty of Huang et al.*; (c) *Adaptive banding.*

Sparse estimation

Nested lasso penalty / Adaptive banding: Levina et al. (2008)

- Minimize

$$n \log |\mathbf{\Sigma}| + n\mathrm{tr}(\mathbf{D}^{-1}\mathbf{TST}') + \lambda \sum_{t=2}^{p} P(\phi_t),$$

$$P(\phi_t) = |\phi_{t,t-1}| + \frac{|\phi_{t,t-2}|}{|\phi_{t,t-1}|} + \cdots + \frac{|\phi_{t,1}|}{|\phi_{t,2}|},$$

  where 0/0 is defined to be zero.
- Select the best model that regresses the jth variable on its k closest predecessors, where $k = k_j$ is dependent on j.

Sparse estimation

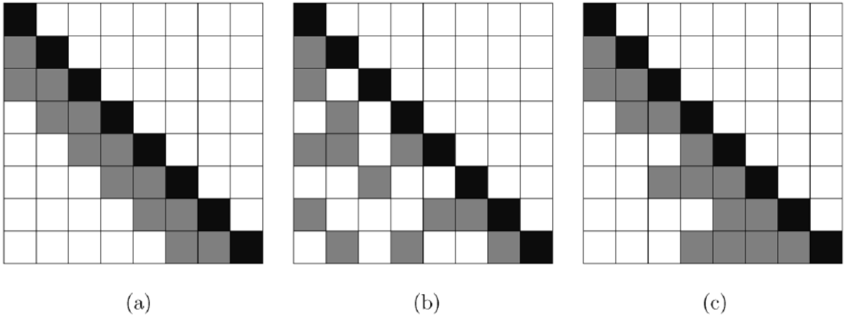


FIG. 1. *The placement of zeros in the Cholesky factor T*: (a) *Banding*; (b) *Lasso penalty of Huang et al.*; (c) *Adaptive banding*.

Sparse estimation

Forward adaptive banding: Leng and Li. (2011)

- Minimize modified BIC:

$$n \log |\mathbf{\Sigma}| + n\mathrm{tr}(\mathbf{D}^{-1}\mathbf{T}\mathbf{S}\mathbf{T}') + C_n \log(n) \sum_{j=1}^{p} k_j,$$

$$\text{s.t.} \quad k_j \leq \min\{n/(\log n)^2, j-1\},$$

  where k_j is the band length.
- Fit $AR(k_j)$ to obtain $\mathbf{T}$ and $\mathbf{D}$.

Cholesky decomposition: summary

- Cholesky decomposition is dependent on the order in which the variables appear in the random vector Y.
- It works when the variables have a natural ordering.

GLM for covariance matrices

- Another way to reduce number of covariance parameters is to use **covariates**, as in modeling the mean vector.
- Path of development: linear $\rightarrow$ log-linear $\rightarrow$ GLM.

Linear covariance models

- Linear covariance models (LCM):

$$\mathbf{\Sigma}^{\pm} = \alpha_1 \mathbf{U}_1 + \cdots + \alpha_q \mathbf{U}_q,$$

  where $\mathbf{U}_i$'s are some known symmetric basis matrices (covariates) and α_i's are unknown parameters.
- For $q = p^2$, any covariance matrix can be written as:

$$\mathbf{\Sigma} = (\sigma_{ij}) = \sum_{i=1}^{p} \sum_{j=1}^{p} \sigma_{ij} \mathbf{U}_{ij},$$

  where $\mathbf{U}_{ij}$ is matrix with 1 on (i, j)th position and 0 elsewhere.

Linear covariance models

- MLE: the score equation of α_i is

$$\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{U}_i) - \mathrm{tr}(\mathbf{S}\mathbf{\Sigma}^{-1}\mathbf{U}_i\mathbf{\Sigma}^{-1}) = 0,$$

  which can be solved by an iterative method.
- Constraint: α_i's are restricted so that the matrix is positive definite.
- Lack of interpretation.

Log-linear covariance models

- Log-linear covariance models:

$$\log \boldsymbol{\Sigma} = \alpha_1 \mathbf{U}_1 + \cdots + \alpha_q \mathbf{U}_q,$$

- α_i's are now unconstrained.

GLM via Cholesky decomposition

Pourahmadi (1999):

- Cholesky decomposition: $\mathbf{\Sigma}^{-1} = \mathbf{T}'\mathbf{D}^{-1}\mathbf{T}$.
- $\mathbf{T}$ and $\log \mathbf{D}$ are unconstrained.
- Parametric models for $\phi_{t,j}$ and $\log \sigma_t^2$:

$$\log \sigma_t^2 = z_t'\lambda, \quad \phi_{t,j} = w_{t,j}'\gamma,$$

  where z_t and $w_{t,j}$ are $q \times 1$ and $d \times 1$ vectors of covariates, λ and γ are parameters.
- Common covariates are powers of times and lags

$$z_t = (1, t, t^2, \ldots, t^{q-1})',$$

$$w_{t,j} = (1, t-j, (t-j)^2, \ldots, (t-j)^{d-1})'.$$

GLM via Cholesky decomposition

- Number of parameters: $q + d$.
- Computing MLE is relatively simple:

$$-2l(\lambda, \gamma) = n \log |\mathbf{D}| + n\mathrm{tr}(\mathbf{D}^{-1}\mathbf{TST}').$$

Given **D**, the MLE of **T** has a closed form. Similarly, given **T**, the MLE of **D** has a closed form.

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.
- Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. Biometrika, 86(3), 677-690.
- Huang, J. Z., Liu, N., Pourahmadi, M., & Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. Biometrika, 93(1), 85-98.
- Leng, C., & Li, B. (2011). Forward adaptive banding for estimating large covariance matrices. Biometrika, 98(4), 821-830.
- Levina, E., Rothman, A., & Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested Lasso penalty. The Annals of Applied Statistics, 2(1), 245-263.
- Wu, W. B., & Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. Biometrika, 90(4), 831-844.

The End