Parameterization of a covariance-matrix with unbalanced data

Helgi Tómasson
helgito@hi.is

September 3, 2024

Plan of talk

- Brief introduction
- Background of problem
- Why is a parameterization of a covariance matrix interesting?
- Numerical implementation of Choleski approach
- Numerical implementation of Givens-rotations approach
- Some statistical comments and intuition
- Final comments

Introduction

- A degree in mathematics from University of Iceland and Fil. Dr in statististic from University of Gothenburg (Sweden).
- A professor of econometrics at the faculty of economics at the University of Iceland
- Have given courses on general econometrics, time-series and computational methods
- Wrote a thesis on the computation of shrinkage (James-Stein, empirical Bayes) estimators in time-series models
- Shrinkage-estimators can be compared to pre-test estimators that are frequently used in practice.
- Do my own programming, Fortran, R, octave, Julia, etc.

Pre-test and Bayes estimator

- In applied statistical work a common practice is to do some test first and if the estimated parameter is considered "significant", the maximum-likelihood estimator is used.

$$H_0 : \mu = \mu_0, \quad \mu \neq \mu_0$$
$$\hat{\mu}_{PRE-TEST} = \mu_0 \mathrm{I}(H_0 \text{ not rejected}) + \hat{\mu}_{MLE} \mathrm{I}(H_0 \text{ rejected})$$

- If the model is: $\boldsymbol{X} \sim N(\mu, \sigma^2)$, where σ is known, and the prior is: $\mu \sim N(\mu_0, \tau^2)$, then the posterior has mean: $\mu_0 \frac{\sigma^2}{\sigma^2+\tau^2} + \hat{\mu}_{MLE} \frac{\tau^2}{\sigma^2+\tau^2},, \quad \hat{\mu}_{MLE} = X$.
- The key issue is that μ_0 is a reference-value for the unknown parameter.

- Both the pre-test approach and the Bayesian approach "shrink" the MLE-estimator towards a reference value μ_0.
- If *a priori*-information is weak, i.e.τ^2 is big, then the reference value has little impact.
- If the parameter μ is high-dimensional, the Bayesian estimator has better qualities than the MLE and $PRE - TEST$ in the mean-square-error sense if τ^2 is estimated from the data. The James-Stein estimators.
- Good *a priori* guess improves the MLE-estimator, if number of dimensions is higher than 3.
- I would like to compute something similar for estimators of the covariance-matrix.

The covariance-matrix

- What is covariance, or correlation(scaled covariance)? Essentially a geometric concept, an angle, it relates angles and length of vectors, i.e.:

  $$\boldsymbol{u} \cdot \boldsymbol{v} = \cos(\theta)\ ||\boldsymbol{u}||\ ||\boldsymbol{v}||,$$

  $\rho = \cos(\theta)$ measures linear the relationship of the vector.

- It is also a probabilistic concept:

  $$E(X_1 - \mu_1)(X_2 - \mu_2) = Cov(X_1, X_2) = \rho\sqrt{V(X_1)}\sqrt{V(X_2)},$$

  i.e. the correlation coefficient is scaled variance.

- On matrix-form:

$$\Sigma = \begin{bmatrix} V(X_1) & Cov(X_1, X_2) \\ Cov(X_1, X_2) & V(X_2) \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} = \underbrace{\begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}}_{\boldsymbol{\sigma}} \underbrace{\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}}_{Cor} \underbrace{\begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}}_{\boldsymbol{\sigma}}$$

- $-1 \leq \rho = \cos(\theta) \leq 1$ i.e. $-\pi \leq \theta \leq \pi$.
- In high dimensions admissible elements of the covariance/correlation matrix follow complicated restrictions.
- The matrix, Σ has to be positive-definite. (semi-positive for singular distributions).
- It might be sensible to write the covariance-matrix as a funcion of angles.

The Choleski approach

- Factorization of the correlation matrix. The non-singular correlation matrix can be written as $\boldsymbol{LL'}$
- A look at the Choleski algorithm shows that the cofficients are polar-coordinates.
- I.e. $\boldsymbol{L} = \boldsymbol{L}(\boldsymbol{\phi})$, where ϕ is a vector of angles.

$$\boldsymbol{\phi} = \begin{matrix} \phi_{21} & & & & \\ \phi_{31} & \phi_{32} & & & \\ \vdots & & \ddots & & \\ \phi_{n,1} & \cdots & \cdots & \cdots & \phi_{n,n-1} \end{matrix}$$

- I the correlation matrix is $n \times n$, the number of angles is $(n-1)n/2$. The angles are all in the interval $(0, \pi)$.

$$Cor = \begin{bmatrix} l_{11} & 0 & \cdots & \cdots & 0 \\ l_{21} & l_{22} & 0 & \cdots & \vdots \\ \vdots & \cdots & \ddots & \cdots & \vdots \\ l_{n1} & \cdots & \cdots & \cdots & l_{nn} \end{bmatrix} \begin{bmatrix} l_{11} & 0 & \cdots & \cdots & 0 \\ l_{21} & l_{22} & 0 & \cdots & \vdots \\ \vdots & \cdots & \ddots & \cdots & \vdots \\ l_{n1} & \cdots & \cdots & \cdots & l_{nn} \end{bmatrix}'$$

$\Sigma = \boldsymbol{\sigma}\, Cor \boldsymbol{\sigma}$, where $\boldsymbol{\sigma}$ is diagonal matrix consisting of square roots of the diagonal of Σ.

In two dimensions

$$Cor = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

$$\boldsymbol{L} = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix}.$$

If $\rho = cos(\phi)$, then $\boldsymbol{L} = \boldsymbol{L}(\phi)$.

$$\boldsymbol{L} = \begin{bmatrix} 1 & 0 \\ \cos(\phi) & \sin(\phi) \end{bmatrix}.$$

Extensions to n dimensions

$\mathbf{L} = \mathbf{L}(\phi_1, \ldots, \phi_{n(n-1)/2})$. One version is:

$$l_{ij} = \begin{cases} \cos(\phi_{ij}) \prod_{k=1}^{j-1} \sin(\phi_{ik}), & j = 1, \ldots i-1, \\ \prod_{k=1}^{j-1} \sin(\phi_{ik}), & j = i, \end{cases}$$

- Easily inverted , i.e. if the correlation matrix is known we can find the angles:

$$\phi_{ij} = \arccos\left[\frac{l_{ij}}{\sqrt{\sum_{k=j}^{i} l_{ik}^2}}\right].$$

- Calculus is easy:

$$\frac{\partial l_{ij}}{\partial \phi_{im}} = l_{ij}/\tan(\phi_{im}), \quad \text{for } m > j,$$
$$-l_{ij}\tan(\phi_{im}), \quad \text{for } m = j.$$

Some numerical concerns

- The matrix must not be singular or almost singular.
- Some of the angles will be poorly estimated.
- If an angle $\phi_{ij} = 0$, then the rest of that line is unindentified.
- The outcome is sensitive to the order of the variables in the vector.

My idea

- I want to be able to enforce the restrition of reduced-rank. E.g. a "single-factor" model.
- Restrictions of that type may be a sensible prior in a Bayesian approach.
- The ordering of variables in the observation vector should not matter.
- An approach might be to use singular-value-decomposition (SVD) and Givens-rotations. The SVD exist for all matrices.
- SVD and Givens rotations are smart computational devices.

- Pinheiro-Bates, give the following:

$$\Sigma = U \Lambda U'$$

$$U = G_1 G_2 \cdots G_{n(n-1)/2)}, \quad \text{where}$$

$$G_i[j,k] = \begin{cases} \cos(\phi_i), & \text{if} \quad j = k = m_1(i) \\ & \quad \text{or } j = k = m_2(i) \\ \sin(\phi_i), & \text{if} \quad j = m_1(i), k = m_2(i) \\ -\sin(\phi_i), & \text{if} \quad j = m_2(i), k = m_1(i) \\ 1, & \text{if} \quad j = k \neq m_1(i) \\ & \quad \text{and } j = k \neq m_2(i) \\ 0, & \text{otherwise} \end{cases}$$

$m_1(i) < m_2(i)$ integers in the range $(1, \ldots, n)$ and
$i = m_2(i) - m_1(i) + (m_1(i) - 1)(n - m_1(i)/2)$.

- The U matrix has the property $UU' = I$. The matrix Λ is diagonal with (semi-positive) values on the diagonal. The singular-values.
- For a given U, it is possible to invert this function, some ϕ_i' are in the interval $(-\pi, \pi)$ (the $\phi_{i+1,i}$'s) and the other in the interval $(-\pi/2, \pi/2)$.
- Calculus is easy.

$$
\mathbf{\Phi} = \begin{bmatrix} x & x & x & x & x \\ \phi_{21} & x & x & x & x \\ \phi_{31} & \phi_{32} & x & x & x \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_{n1} & \cdots & \cdots & \cdots & \phi_{n,n-1} \end{bmatrix}
$$

Similar to Choleski factorization but the angles have different meaning.

- If no singular values are equal and the matrix $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$ is order in decreasing order, the matrix U is unique up to signs of the columns. I decided to use $\det(U) = 1$, and the top row from second element are all positive.
- It easy to decide, e.g. that only some of the singular values are positive, rest 0.The a certain triangle of $\mathbf{\Phi}$ is undetermined and can be set to any value, e.g., 0.
- Enforcing restrictions, such as rank, as in factor-models is therefore trivial.

A numerical illustration

- Data from NBBO, trading in American markets January 2016. 10 frequently trading assets, 10 infrequently trading assets. Aim: guess of covariance matrix of innovations.
- Sample of most trading assets used, every transaction of the less traded assets.
- Assumed model is noisy random-walk.

$$y(t_i, k) = C_k \boldsymbol{X}(t_i) + \varepsilon_k \quad \text{measurement equation of asset } k \text{ at time } t_i$$

$$H = \begin{bmatrix} h_1^2 & 0 & \cdots & \cdots & 0 \\ 0 & h_2^2 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \cdots & \cdots & h_K^2 \end{bmatrix}, \quad h_k^2 = V(\varepsilon_k), \quad \text{measurement noise,}$$

$$\boldsymbol{X}(t), \text{ is a vector of true values at time } t,$$

$$C_k \text{ is a matrix that pick asset } k,$$

$$y(t_i, k) \text{ is log transaction price,} \quad \varepsilon_k \text{ measurment noise at time } t_i.$$

Here,$y(t,k)$ is the logarithm of the observed transaction price of asset k at time t_i. $\boldsymbol{X}(t_i)$ is the vector of (log of) the true values of the assets, ε_k is the deviation of traded price from true price. C_k is a matrix that picks coordinate k from the vector $\boldsymbol{X}(t_i)$. The true value is supposed to evolve in time by:

$$d\boldsymbol{X}(t) = d\boldsymbol{W}\boldsymbol{(t)}, \quad V(d\boldsymbol{W}(t)) = \boldsymbol{Q}dt, \quad \boldsymbol{W}(t),\text{Wiener process.}$$

The variance of the market micro-structure noise, H, is estimated by transaction which take place (almost) simultaneously. In the case of simultaneous trades $y(t_i, k)$ is the average of prices. The statistical problem is (mainly) to estimate the covariance of the innovations, $\boldsymbol{Q}$. Log-likelihood is calculated by means of Kalman filter.

Trading intensity

Asset	Count
AAPL	4605707
BAC	3763218
CHK	1073992
CSCO	2311239
EMCF	37
EXT	213
F	2126560
FB	2900320
FCX	2284878
GE	2999775
ICBK	75
KMDA	188
MSFT	3800629
PLBC	129
PME	188
SBB	291
SUNE	1641875

- For the high frequency trading a random sample was used, for the others every transaction was used.
- Many of the singular values of the estimated covariance matrix are very close to zero.
- That suggests that a factor model (reduced-rank covariance) is a good approximation.
- Even in the case of moderate dimension where all the singular values equal one results in an estimated matrix which is close to being singular.

A textbook factor model

$$\boldsymbol{r}_t = \boldsymbol{\alpha}_t + \boldsymbol{\beta}\boldsymbol{f}_t + \boldsymbol{\varepsilon}_t$$
$$V(\boldsymbol{r}_t) = \boldsymbol{\beta}\Sigma_f\boldsymbol{\beta}' + \boldsymbol{D}$$

- Factors, $\boldsymbol{f}$ could be observable or non-observable.
- An example of a single-factor model is Sharpe-CAPM:

$$r_{it} = \alpha_i + \beta_i r_{mt} + \varepsilon_{it}$$

- Bayesians mith want to set a prior on the number of factor or on partial coefficient using formulas of this type:

$$E(\boldsymbol{Y}|\boldsymbol{X}) = \boldsymbol{\mu}_{\boldsymbol{Y}} + \underbrace{\Sigma_{YX}\Sigma_{XX}^{-1}}_{\boldsymbol{\beta}}(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})$$

Applications and conclusions

- It is difficult to guess reasonable for the elements of a large covariance matrix. Perhaps it is easier guessing the values of the partial-correlation matrix (a function of the inverse of the covariance matrix). Reference value zero partial correlation can be sensible.
- By using angles and postive singular value enforcing a legal covariance matrix is trivial.
- A prior can easily be set and allowing small deviations, e.g. by means of penalty functions.
- The parameters are rotation angles and eigenvalues. It is intuitive to set a prior belief on these parameters.
- Other methods are plausible. E.g. start with a symmetric matrix and take the matrix-exponent. The outcome will allways be positive definite (Pinheiro-Bates, 1996). For a recent implementation see Hansen(2021).
- Choleski factorization may be easier for well behaved matrices. The Givens approach seems better for matrices that are close to being reduced rank. The Givens approach is also less