Log Covariance Matrix Estimation

Xinwei Deng

Department of Statistics

University of Wisconsin-Madison

Joint work with Kam-Wah Tsui (Univ. of Wisconsin-Madsion)

Outline

- Background and Motivation
- The Proposed Log-ME Method
- Simulation and Real Example
- Summary and Discussion

Background

- Covariance matrix estimation is important in multivariate analysis and many statistical applications.
- Suppose $x_1, \ldots, x_n$ are i.i.d. p-dimensional random vectors $\sim N(0, \Sigma)$. Let $S = \sum_{i=1}^{n} x_i x_i^{'}/n$ be the sample covariance matrix. The negative log-likelihood function is proportional to

$$L_n(\Sigma) = -\log|\Sigma^{-1}| + \mathrm{tr}[\Sigma^{-1}S]. \tag{1}$$

- Recent interests of p is large or $p \approx n$. S is not a stable estimate.
  - The largest eigenvalues of S overly estimate the true eigenvalues.
  - When $p > n$, S is singular and the smallest eigenvalue is zero. How to estimate Σ^{-1}?

Recent Estimation Methods on Σ or Σ^{-1}

- Reduce number of nonzeros estimates of Σ or Σ^{-1}.
  - Σ: Bickel and Levina (2008), using thresholding.
  - Σ^{-1}: Yuan and Lin (2007), l_1 penalty on Σ^{-1}.
    Friedman et al., (2008), Graphical Lasso.
    Meinshausen and Buhlmann (2006), Reformulated as regression.
- Shrinkage estimates of the covariance matrix.
  - Ledoit and Wolf (2006), $\rho\Sigma + (1-\rho)\mu I$.
  - Won et al. (2009), control the condition number (largest eigenvalue/smallest eigenvalue).

Motivation

- Estimate of Σ or Σ^{-1} needs to be positive definite.
  - The mathematical restriction makes the covariance matrix estimation problem challenging.
- Any positive definite Σ can be expressed as a matrix exponential of a real symmetric matrix A.

$$\Sigma = \exp(A) = I + A + \frac{A^2}{2!} + \cdots$$

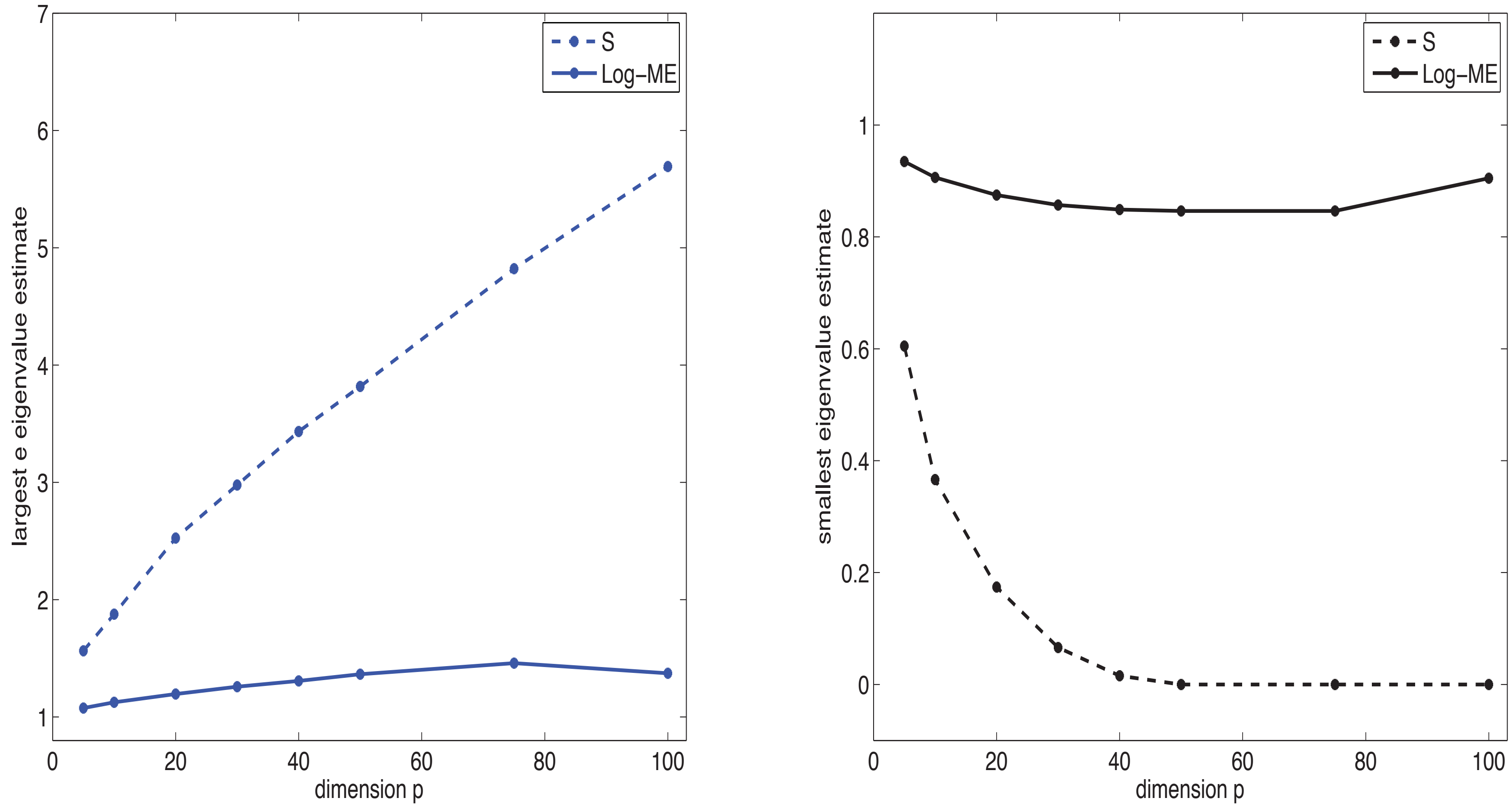
  - Expressing the likelihood function in terms of $A \equiv log(\Sigma)$ releases the mathematical restriction.
- Consider the spectral decomposition of $\Sigma = TDT'$ with $D = diag(d_1, \ldots, d_p)$. Then $A = TMT'$ with $M = diag(\log(d_1), \ldots, \log(d_p))$.

Idea of the Proposed Method

- Leonard and Hsu (1992) used this log-transformation method to estimate Σ by approximating the likelihood using Volterra integral equation.
  - Their approximation based on on S being nonsingular $\Rightarrow$ not applicable when $p \geq n$.
- We extend the likelihood approximation to the case of singular S.
- Regularize the largest and smallest eigenvalues of Σ *simultaneously*.
- An efficient iterative quadratic programming algorithm to estimate A ($\log \Sigma$).
- Call the resulting estimate "Log-ME", Logarithm-transformed Matrix Estimate.

A Simple Example

- Experiment: simulate x_i's from $N(0,I)$, $i = 1,\ldots,n$ where $n = 50$.
- For each p varying from 5 to 100, consider the the largest and smallest eigenvalues of the covariance matrix estimate.
- For each p, repeat the experiment 100 times and compute the average of the largest eigenvalues and the average of the smallest eigenvalues for
  - The sample covariance matrix.
  - The Log-ME covariance matrix estimate



The averages of the largest and smallest eigenvalues of covariance matrix estimates over the dimension p.The true eigenvalues are all equal to 1.

The Transformed Log-Likelihood

- In terms of the covariance matrix logarithm A, the negative log-likelihood function in (1) becomes

$$L_n(A) = \text{tr}(A) + \text{tr}[\exp(-A)S]. \tag{2}$$

- The problem of estimating a positive definite matrix Σ now becomes a problem of estimating a real symmetric matrix A.
- Because of the matrix exponential term $\exp(-A)S$, estimating A by directly minimizing $L_n(A)$ is nontrivial.
- **Our approach:** Approximate $\exp(-A)S$ using the Volterra integral equation (valid even for S singular case).

The Volterra Integral Equation

- The Volterra integral equation (Bellman, 1970, page 175) is

$$\exp(At) = \exp(A_0 t) + \int_0^t \exp(A_0(t-s))(A-A_0)\exp(As)ds. \tag{3}$$

- Repeatedly applying (3) leads to

$$\begin{aligned}\exp(At) = {} & \exp(A_0 t) + \int_0^t \exp(A_0(t-s))(A-A_0)\exp(A_0 s)ds \\ & + \int_0^t \int_0^s \exp(A_0(t-s))(A-A_0)\exp(A_0(s-u))(A-A_0)\exp(A_0 u)duds \\ & + \text{cubic and higher order terms,}\end{aligned} \tag{4}$$

where $A_0 = \log(\Sigma_0)$ and Σ_0 is an initial estimate of Σ.

- The expression of $\exp(-A)$ can be obtained by letting $t = 1$ in (4) and replacing A, A_0 in (4) with $-A, -A_0$.

Approximation to the Log-Likelihood

- The term $\mathrm{tr}[\exp(-A)S]$ can be written as

$$\begin{aligned}\mathrm{tr}[\exp(-A)S] =& \mathrm{tr}(S\Sigma_0^{-1}) - \int_0^1 \mathrm{tr}[(A-A_0)\Sigma_0^{-s}S\Sigma_0^{s-1}]ds \\ &+ \int_0^1 \int_0^s \mathrm{tr}[(A-A_0)\Sigma_0^{u-s}(A-A_0)\Sigma_0^{-u}S\Sigma_0^{s-1}]duds \\ &+ \text{cubic and higher order terms.}\end{aligned} \tag{5}$$

- By leaving out the higher order terms in (5), we approximate $L_n(A)$ by using $l_n(A)$:

$$\begin{aligned}l_n(A) =& \mathrm{tr}(S\Sigma_0^{-1}) - \left[\int_0^1 \mathrm{tr}[(A-A_0)\Sigma_0^{-s}S\Sigma_0^{s-1}]ds - \mathrm{tr}(A)\right] \\ &+ \int_0^1 \int_0^s \mathrm{tr}[(A-A_0)\Sigma_0^{u-s}(A-A_0)\Sigma_0^{-u}S\Sigma_0^{s-1}]duds.\end{aligned} \tag{6}$$

Explicit Form of $l_n(A)$

- The integrations in $l_n(A)$ can be analytically solved through the spectral decomposition of $\Sigma_0 = T_0 D_0 T_0'$.

- **Some Notation:**
  - Here $D_0 = diag(d_1^{(0)}, \ldots, d_p^{(0)})$ with $d_i^{(0)}$'s as the eigenvalues of Σ_0.
  - $T_0 = (t_1^{(0)}, \ldots, t_p^{(0)})$ with $t_i^{(0)}$ as the corresponding eigenvector for $d_i^{(0)}$.
  - Let $B = T_0'(A - A_0)T_0 = (b_{ij})_{p\times p}$, and $\tilde{S} = T_0' S T_0 = (\tilde{s}_{ij})_{p\times p}$.

- The $l_n(A)$ can be written as a function of b_{ij}:

$$l_n(A) = \sum_{i=1}^{p} \frac{1}{2}\xi_{ii} b_{ii}^2 + \sum_{i<j} \xi_{ij} b_{ij}^2 + 2\sum_{i=1}^{p}\sum_{j\neq i} \tau_{ij} b_{ii} b_{ij} + \sum_{k=1}^{p} \sum_{i<j, i\neq k, j\neq k} \eta_{kij} b_{ik} b_{kj} - \left[\sum_{i=1}^{p} \beta_{ii} b_{ii} + 2\sum_{i<j} \beta_{ij} b_{ij}\right], \quad (7)$$

up to some constant. Getting $B \leftrightarrow$ Getting A.

Some Details

- For the linear term,

$$\beta_{ii} = \frac{\tilde{s}_{ii}}{d_i^{(0)}} - 1, \ \beta_{ij} = \frac{\tilde{s}_{ij}(d_i^{(0)} - d_j^{(0)})/(d_i^{(0)} d_j^{(0)})}{(\log d_i^{(0)} - \log d_j^{(0)})}.$$

- For the quadratic term,

$$\xi_{ii} = \frac{\tilde{s}_{ii}}{d_i^{(0)}},$$

$$\xi_{ij} = \frac{\tilde{s}_{ii}/d_i^{(0)} - \tilde{s}_{jj}/d_j^{(0)}}{\log d_j^{(0)} - \log d_i^{(0)}} + \frac{(d_i^{(0)}/d_j^{(0)} - 1)\tilde{s}_{ii}/d_i^{(0)} + (d_j^{(0)}/d_i^{(0)} - 1)\tilde{s}_{jj}/d_j^{(0)}}{(\log d_j^{(0)} - \log d_i^{(0)})^2},$$

$$\tau_{ij} = \left[\frac{1/d_j^{(0)} - 1/d_i^{(0)}}{(\log d_j^{(0)} - \log d_i^{(0)})^2} + \frac{1/d_i^{(0)}}{\log d_j^{(0)} - \log d_i^{(0)}} \right] \tilde{s}_{ij},$$

$$\eta_{kij} = \left[\frac{1/d_i^{(0)} - 1/d_j^{(0)}}{\log(d_k^{(0)}/d_j^{(0)}) \log(d_j^{(0)}/d_i^{(0)})} + \frac{1/d_j^{(0)} - 1/d_i^{(0)}}{\log(d_k^{(0)}/d_i^{(0)}) \log(d_i^{(0)}/d_j^{(0)})} + \frac{2/d_k^{(0)} - 1/d_i^{(0)} - 1/d_j^{(0)}}{\log(d_k^{(0)}/d_i^{(0)}) \log(d_k^{(0)}/d_j^{(0)})} \right] \tilde{s}_{ij}$$

The Log-ME Method

- Propose a regularized method to estimate Σ by using the approximate log-likelihood function $l_n(A)$.
- Consider the penalty function $\|A\|_F^2 = \text{tr}(A^2) = \sum_{i=1}^{p}(\log(d_i))^2$, where d_i is the eigenvalue of the covariance matrix Σ.
  - If d_i goes to zero or diverges to infinity, the value of $\log(d_i)$ goes to infinity in both cases.
  - Such a penalty function can *simultaneously* regularize the largest and smallest eigenvalues of the covariance matrix estimate.
- Estimate Σ, or equivalently A, by minimizing

$$l_{n,\lambda}(B) \equiv l_{n,\lambda}(A) = l_n(A) + \lambda \text{tr}(A^2), \tag{8}$$

where λ is a tuning parameter.

An Iterative Algorithm

- The $l_{n,\lambda}(B)$ depends on an initial estimate Σ_0, or equivalently, A_0.
- Propose to iteratively use $l_{n,\lambda}(B)$ to obtain its minimizer $\hat{B}$:

  Algorithm:

  Step 1: Set an initial covariance matrix estimate Σ_0, a positive definite matrix.

  Step 2: Use the spectral decomposition $\Sigma_0 = T_0 D_0 T_0'$, and set $A_0 = \log(\Sigma_0)$.

  Step 3: Compute $\hat{B}$ by minimizing $l_{n,\lambda}$ in (10). Then obtain $\hat{A} = T_0 \hat{B} T_0' + A_0$, and update the estimate of Σ by

  $$\hat{\Sigma} = \exp(\hat{A}) = \exp(T_0 \hat{B} T_0' + A_0).$$

  Step 4: Check if $\|\hat{\Sigma} - \Sigma_0\|_F^2$ is less than a pre-specified positive tolerance value. Otherwise, set $\Sigma_0 = \hat{\Sigma}$ and go back to **Step 2**.

- Set an initial Σ_0 in **Step 1** to be $S + \varepsilon I$.

Simulation Study

- Six different covariance models of $\Sigma = (\sigma_{ij})_{p \times p}$ are used for comparison,
  - Model 1: Homogeneous model with $\Sigma = I$.
  - Model 2: MA(1) model with $\sigma_{ii} = 1, \sigma_{i,i-1} = \sigma_{i-1,i} = 0.45$.
  - Model 3: Circle model with $\sigma_{ii} = 1, \sigma_{i,i-1} = \sigma_{i-1,i} = 0.3$, $\sigma_{1,p} = \sigma_{p,1} = 0.3$.
- Compare four estimation methods: the banding estimate (Bickel and Levina, 2008), the LW estimate (Ledoit and Wolf, 2006), the Glasso estimate (Yuan and Lin, 2007), and the CN estimate (Won et al., 2009).
- Consider two loss functions to evaluate the performance of each method,

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

$$\Delta_1 = |\hat{d}_1/\hat{d}_p - d_1/d_p|,$$

where d_1 and d_p are the largest and smallest eigenvalue of Σ. Denote $\hat{d}_1$ and $\hat{d}_p$ to be their estimates.

Simulation Results

Averages and standard errors from 100 runs in the case of $n = 50, p = 50$.

	Log-ME		Banding		LW		Glasso		CN	
Model	KL	Δ_1	KL	Δ_1	KL	Δ_1	KL	Δ_1	KL	Δ_1
1	0.08	0.22	1.31	1.74	0.10	0.18	2.11	1.19	0.22	0.09
	(0.00)	(0.00)	(0.04)	(0.52)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)
2	12.75	15.19	912.02*	343.60	13.11	15.73	14.67	15.67	13.68	16.62
	(0.02)	(0.05)	(882.90)	(152.82)	(0.02)	(0.04)	(0.03)	(0.03)	(0.02)	(0.02)
3	4.85	1.56	3.72	5.62	4.70	2.10	7.27	1.82	4.88	2.71
	(0.01)	(0.01)	(0.13)	(0.39)	(0.01)	(0.03)	(0.02)	(0.02)	(0.01)	(0.02)

Note: The value marked with $*$ means it is affected by the matrix singularity.

Portfolio Optimization of Stock Data

- Apply the Log-ME method in an application of portfolio optimization.
- In mean-variance optimization, the risk of a portfolio $w = (w_1, \ldots, w_p)$ is measured by the standard deviation $\sqrt{w^T \Sigma^{-1} w}$, where $w_i \geq 0$ and $\sum_i^p w_i = 1$.
- The estimated minimum variance portfolio optimization problem is

$$\min_{w} w^T \hat{\Sigma}^{-1} w \tag{9}$$

$$\text{s.t. } \sum_{i}^{p} w_i = 1,$$

where $\hat{\Sigma}$ is an estimate of the true covariance matrix Σ.

- An accurate covariance matrix estimate $\hat{\Sigma}$ can lead to a better portfolio strategy.

The Setting-up

- Consider the weekly returns of $p = 30$ components of the Dow Jones Industrial Index from January 8th, 2007 to June 28th, 2010.
- Use the first $n = 50$ observations as the training set, the next 50 observations as the validation set, and *the remaining* 83 observations for the test set.
- Let X_{ts} be the test set and S_{ts} be the sample covariance matrix of X_{ts}. The performance of a portfolio w is measured by the *realized return*

$$R(w) = \sum_{x \in X_{ts}} w^T x,$$

  and the *realized risk*

$$\sigma(w) = \sqrt{w^T S_{ts} w}.$$

- The optimal portfolio $\tilde{w}$ is computed with $\hat{\Sigma}$ estimated by the Log-ME method, the CN method (Won et al., 2009) and the S, separately.

The Comparison Results

Table 1. The comparison of the realized return and the realized risk.

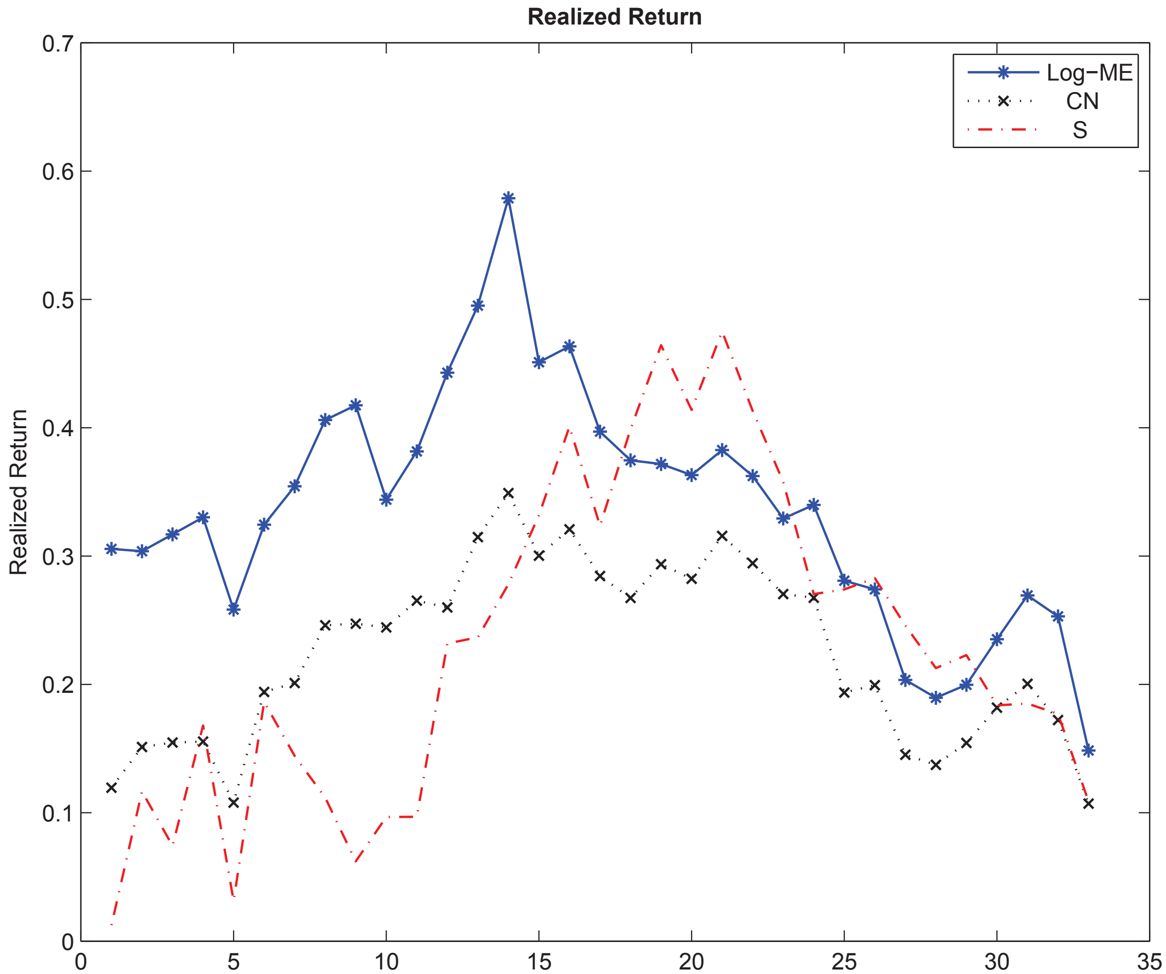
	Log-ME	CN	S
Realized return $R(\tilde{w})$	0.218	0.123	0.059
Realized risk $\sigma(\tilde{w})$	0.029	0.024	0.035

- The Log-ME method produced a portfolio with a larger realized return but smaller realized risk.

Comparison in Different Periods

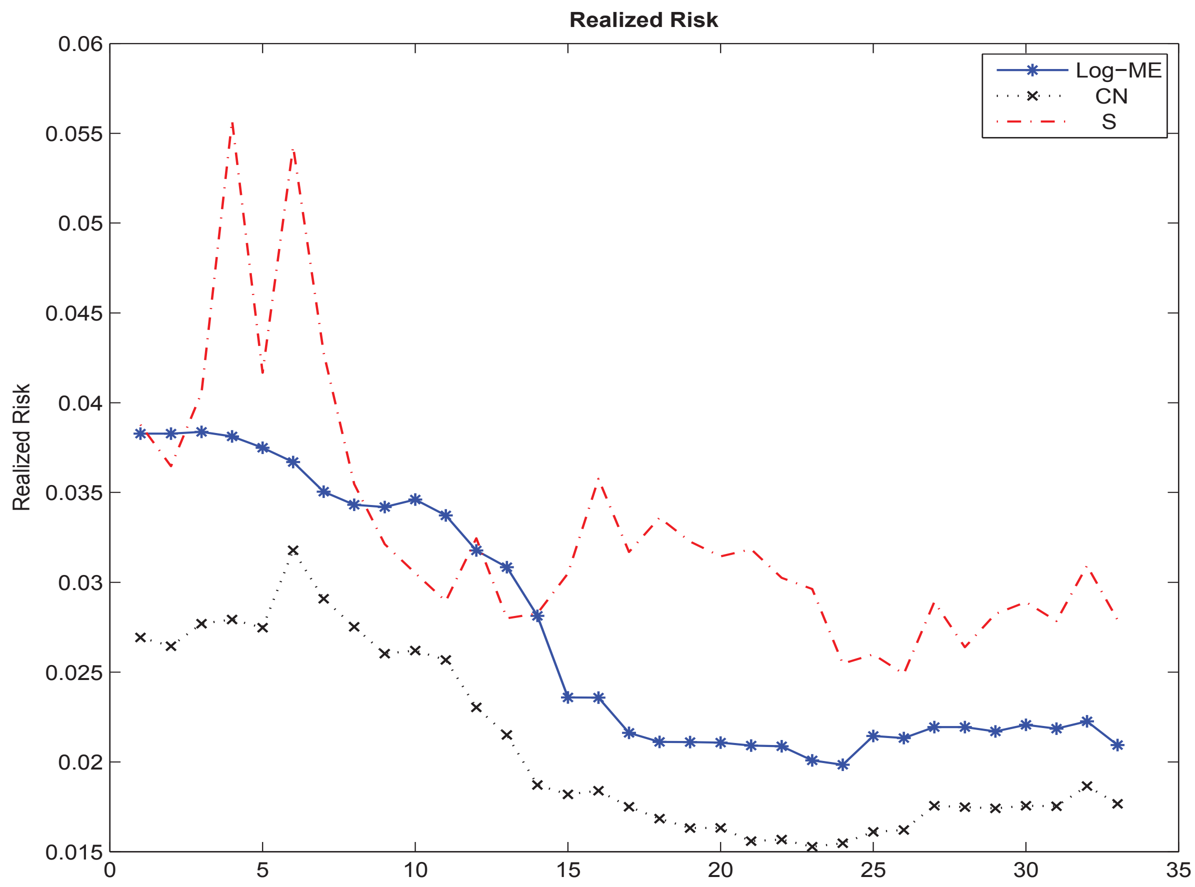
- Consider the portfolio strategy using the Log-ME method for various covariance matrix estimation methods.
- Given a stating week, use the first 50 observations as the training set, the next 50 observations as a validation set, and *the third* 50 *observations* as a test set.
- Shift the starting week one ahead every time, and evaluate the portfolio strategy of 33 different consecutive test periods.
- The optimal portfolio $\tilde{w}$ is computed with $\hat{\Sigma}$ estimated by the Log-ME method, the CN method and the sample covariance matrix method, separately.

The Realized Returns



The proposed Log-ME covariance matrix estimate can lead to higher returns.

The Realized Risks



The log-ME method has relatively higher risks than the CN method, but it provides much larger realized returns than the CN method.

Summary

- Estimate the covariance matrix through its matrix logarithm based on a penalized likelihood function.
- The Log-ME method regularizes the largest and smallest eigenvalues simultaneously by imposing a convex penalty.
- Other penalty functions can be considered to improve the estimation in different perspectives.
- Extend to Bayesian covariance matrix estimation for the large-p-small-n problem.

Thank you!

The Log-ME Method (Con't)

- Note that $\mathrm{tr}(A^2) = \mathrm{tr}((T_0 B T_0^{'} + A_0)^2)$ is equivalent to $\mathrm{tr}(B^2) + 2\mathrm{tr}(B\Gamma)$ up to some constant, where $\Gamma = (\gamma_{ij})_{p\times p} = T_0^{'} A_0 T_0$.

- In terms of B, the function $l_{n,\lambda}(A)$ becomes

$$
\begin{aligned}
l_{n,\lambda}(B) = & \sum_{i=1}^{p} \frac{1}{2}\xi_{ii} b_{ii}^2 + \sum_{i<j} \xi_{ij} b_{ij}^2 + 2\sum_{i=1}^{p}\sum_{j\neq i} \tau_{ij} b_{ii} b_{ij} + \sum_{k=1}^{p} \sum_{i<j, i\neq k, j\neq k} \eta_{kij} b_{ik} b_{kj} \\
& - \left(\sum_{i=1}^{p} \beta_{ii} b_{ii} + 2\sum_{i<j} \beta_{ij} b_{ij} \right) \\
& + \lambda \left[\frac{1}{2}\sum_{i=1}^{p} b_{ii}^2 + \sum_{i<j} b_{ij}^2 + \sum_{i=1}^{p} \gamma_{ii} b_{ii} + 2\sum_{i<j} \gamma_{ij} b_{ij} \right].
\end{aligned}
\tag{10}
$$

- The $l_{n,\lambda}(B)$ is still a quadratic function of $B = (b_{ij})$.

The CN Method

- The CN method is to estimate Σ with a constraint on its condition number (Won et al., 2009).

- They consider $\hat{\Sigma} = T\text{diag}(\hat{u}_1^{-1}, \ldots, \hat{u}_p^{-1})T'$, where T is from the spectral decomposition of $S = T\text{diag}(l_1, \ldots, l_p)T'$.

- The $\hat{u}_1, \ldots, \hat{u}_p$ are obtained by solving the constraint optimization

$$\min_{u, u_1, \ldots, u_p} \quad \sum_i^p (l_i u_i - \log u_i)$$
$$s.t. \quad u \leq u_i \leq \kappa_{max} u, \; i = 1, \ldots, p,$$

  where κ_{max} is a tuning parameter.

- The tuning parameter is computed through an independent validation set.