1 Problem 1

1. In order to prove that Σ is a valid covariance matrix, it is sufficient to show that Σ is positive semi-definite. According to Sylvester's criterion, Σ is positive definite if all of its leading principal minors must be positive, which is equivalent to all the diagonal elements obtained from Gaussian Elimination process must be positive. A proof of all the diagonal elements obtained by Gaussian Elimination process are positive is presented in Figure 1. Thus Σ is a positive definite matrix and a valid covariance matrix.

2. Since all diagonal entries of the upper triangle matrix resulted from the Gaussian Elimination are positive, $det(\Sigma) > 0$ thus Σ is invertible. Since Σ is positive semi-definite as we proved previously, apply Cholesky decomposition we get $\Sigma = LL^T$, where L is a lower triangular matrix with positive diagonal elements. In order to find Σ^{-1}, we first find L^{-1} then $\Sigma^{-1} = L^{-T}L^{-1}$. The detailed process can be found in Figure 2 and 3.

3. According to the definition of statistical correlation:

$$\begin{aligned} \rho(y_i, y_j) &= \frac{\sigma(y_i, y_j)}{\sqrt{\Sigma^y_{ii}\Sigma^y_{jj}}} \\ &= \frac{exp(-|y_i - y_j|/l)}{\sqrt{\sigma(y_i, y_i)\sigma(y_j, y_j)}} \\ &= \frac{exp(-|y_i, y_j|/l)}{\sqrt{1 * 1}} \\ &= exp(-|y_i, y_j|/l) \end{aligned} \tag{1}$$

4. Please refer to Figure 4.

5. Please refer to Figure 5.

6. Assume the original $x_1 < ... < x_N$ is generated through uniform Gaussian distribution with zero-mean. First find the empirical covariance matrix Γ from $y_1, ..., y_N$ which is an estimate of Σ. Then sample a set of x_i from the multivariate Gaussian with zero-mean and covariance matrix Γ. This solution is not in closed form and not unique because the result of different sampling from the multivariate Gaussian is different.

2 Problem 2

1. N/A

2. Maximize $logdet\Theta - tr(S\Theta)$ subject to $||\Theta||_1 < t$, where $t > 0$ is a tuning factor. That is, maximize the Gaussian log-likelihood of the data subject to the constraint of L_1(lasso) penalty.

3. The optimization objective of PCA can be viewed from two different perscpectives. The first perspective is to find a low-dimensional subspace such that when the data is reconstructed from the subspace, the reconstruction error is minimized. The second perspective is to find linear projection such that the variance in the subspace is maximized. Compared to PCA, Graph Lasso method is not concerned with reconstructing the data but only concerned about preserving the maximum correlation information between data while keeping the inverse covariance matrix sparse in order to maintain computational convenience. The two methods are similar in that they both try to preserve as much information (variance and correlation respectively) from original data while keeping computation complexity managable. However, the two methods simplifies data in two different ways: reducing dimension and keeping inverse covariance matris sparse.

4. Let $f = logdet\Theta - tr(S\Theta) - tr(\Theta^2)$. Take the derivative of f:

$$\frac{d}{d\Theta}f = \frac{1}{det\Theta}\frac{d}{dX}detX - \frac{d}{dX}tr(\Theta(S-\Theta)) \tag{2}$$

5. The adjacency matrix A and precision matrix Λ are exactly the same except the diagonal of A is consists of all 0s while the diagonal of Λ is consists of all 1s. They are both sparse because $Pr(a_{ij} = 1)$ is relativly small thus most of the entries are 0s. Thus the resulting adjacency graph is also sparse, in some cases some points are disconnected from all other points subject to initial samplingi, as can be observed from Figure 5. In contrast, the covariance matrix $\Sigma_0 = \Lambda_0^{-1}$ is relatively dense, as can be observed from the left most matrix in Figure 5.

6. According to the sample covariance and precision in the right two matrices of Figure 6, the sample covariance is almost identical to "true" covariance matrix converted from true precision with slight differences since the sample covariance is estimated directly from the dataset. However, the sample precision matrix is much denser than "true" precision matrix because it is computed as the inverse of sample covariance. Slight difference between sample covariance and "true" covariance makes it impossible to reproduce the "true" precision matrix from sample covariance but it is easy to identify the 1 entries in "true" precision matrix from the sample precision matrix although the areas around those entries are blurry (with small non-zero entries).

7.

8.

9. The Graphical Lasso method estimates a sparse inverse covariance matrix by maximizing the Gaussian log-likelihood of the data. It controls the number of zeros (sparsity) in the inverse covariance matrix by imposing a L_1 (lasso) penalty. The algorithm finds the inverse covariance matrix by solving a lasso problem with coordinate descent procedure in each iteration until the resulting matrix converges. Graphical Lasso is much faster compared to other algorithms that solves the same problem.

   The reconstructed adjacency and covariance matrix are depicted in Figure 7.

10. The first application is estimating the inverse covariance matrix of large-scale gene expression data from a multivariate Gaussian distribution. The second application is to learn the correlation of futures contract data with the estimated inverse covariance matrix and discover the graphical structure of how the price of commodities influence each other. The third application is to estimate the precision matrix of stocks in the US stock market in order to optimize portfolio management. PCA/PPCA cannot be used in these scenarios because true covariance matrix is hard to efficiently estimate in terms of large amount of data. The covariance matrix is large and dense because the data such as stocks and gene expression are closely related, thus the entries of covariance matrix cannot be forced to be sparse. Thus with Graphic Lasso it is much more efficient to estimate a sparse precision matrix and the performance is much faster than PCA/PPCA because of the usage of coordinate descent procedure. The primary usage of PCA/PPCA is to reduce the dimensionality of the data and preserve as much variance as possible, thus its not good for discovering correlation between data.

1. Let $\vec{\Sigma}_1$ be represented as $\begin{bmatrix} 1 & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & & & & \\ a_{31} & & & & \\ \vdots & & & \ddots & \\ a_{n1} & & & & 1 \end{bmatrix}$

where $a_{12} = a_{21}$, $a_{13} = a_{31}$, $\cdots$ $a_{1n} = a_{n1}$ since $\sigma(x, x') = \sigma(x', x)$ according to definition.

Use Gaussian Elimination on $\vec{\Sigma}_1$:

Step 1: $\vec{\Sigma}_{11} = \begin{bmatrix} 1 & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & 1-a_{12}^2 & a_{23}-a_{12}a_{13} & \cdots & \\ 0 & a_{32}-a_{12}a_{31} & 1-a_{13}^2 & & \\ \vdots & & & \ddots & \\ 0 & & & & \end{bmatrix}$

Since $\sigma(x, x') = \exp(-|x - x'|/\ell)$

if $x < x' < x''$, then

$$\sigma(x, x'') = \sigma(x, x') \cdot \sigma(x', x'')$$
$$= \exp(-|x - x'|/\ell) \cdot \exp(-|x' - x''|/\ell)$$
$$= \exp(-|x - x''|/\ell)$$

Thus $a_{13} = a_{12} \cdot a_{23}$, $\vec{\Sigma}_{11}$ can be simplified as:

$\vec{\Sigma}_{11} = \begin{bmatrix} 1 & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & 1-a_{12}^2 & a_{23}(1-a_{12}^2) & \cdots & \\ 0 & a_{23}(1-a_{12}^2) & 1-a_{13}^2 & & \\ \vdots & & & \ddots & \\ 0 & & & & \end{bmatrix}$

Step 2: use the $\sigma(x, x'') = \sigma(x, x') \cdot \sigma(x', x'')$ property

$\vec{\Sigma}_{12} = \begin{bmatrix} 1 & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & 1-a_{12}^2 & a_{23}(1-a_{12}^2) & \cdots & \\ 0 & 0 & 1-a_{23}^2 & \cdots & \\ \vdots & \vdots & a_{24}(1-a_{23}^2) & & \\ 0 & 0 & \vdots & & \end{bmatrix}$

we can use mathematical induction to prove that the diagonal entry on the nth row is $1 - a_{n,n+1}^2$

since e^{-x} is monotonically decreases when $x \geq 0$,

$\sigma(x, x')$ has max value 1 when $x = x'$

since $a_{n,n+1} = \sigma(x_n, x_{n+1})$, $x_n < x_{n+1}$, $a_{n,n+1} < 1$, $a_{n,n+1}^2 < 1$

Thus $\vec{\Sigma}_1$ is a positive semi-definite matrix because all of its pivots are possitive.

Figure 1: Problem 1.1

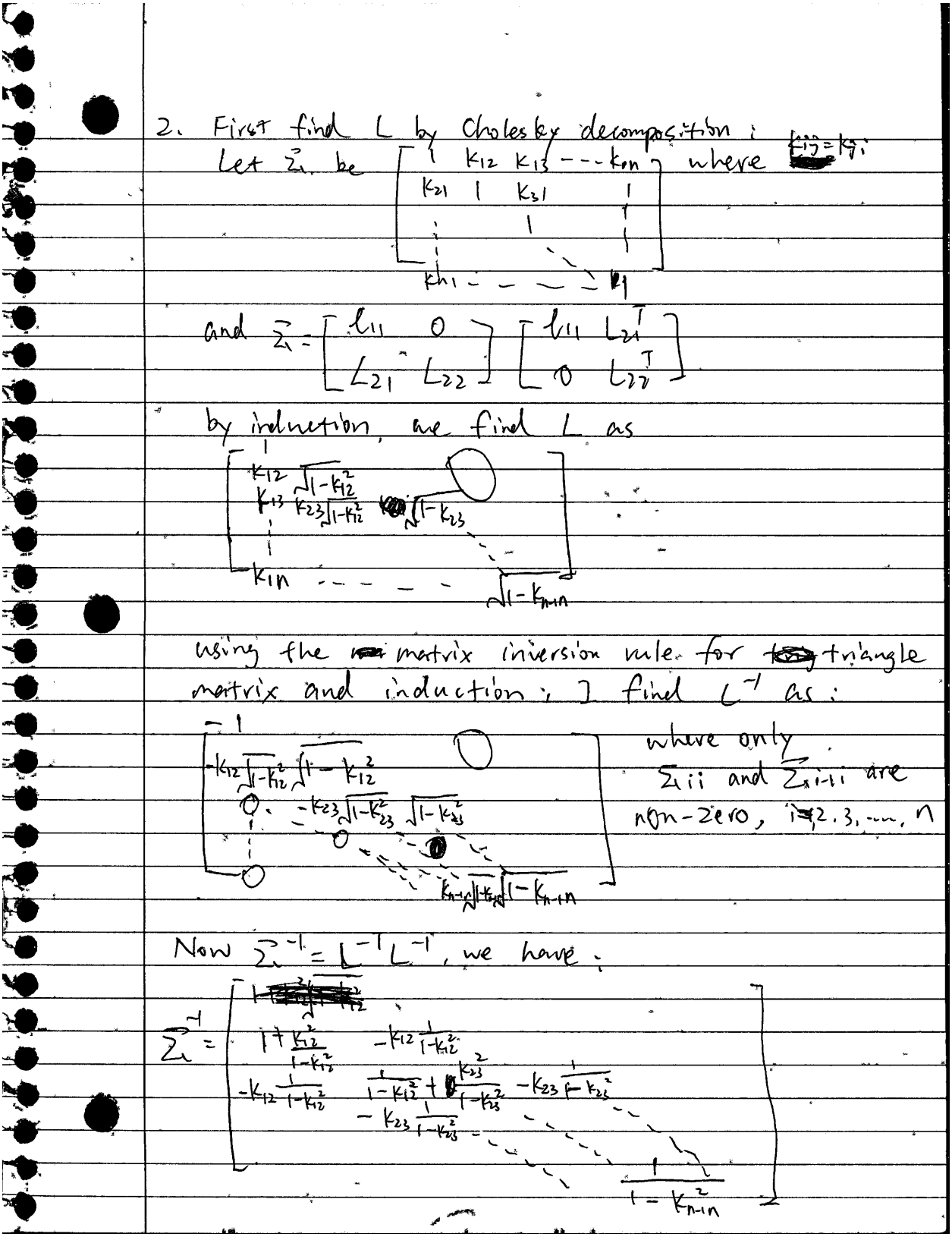

Figure 2: Problem 1.2

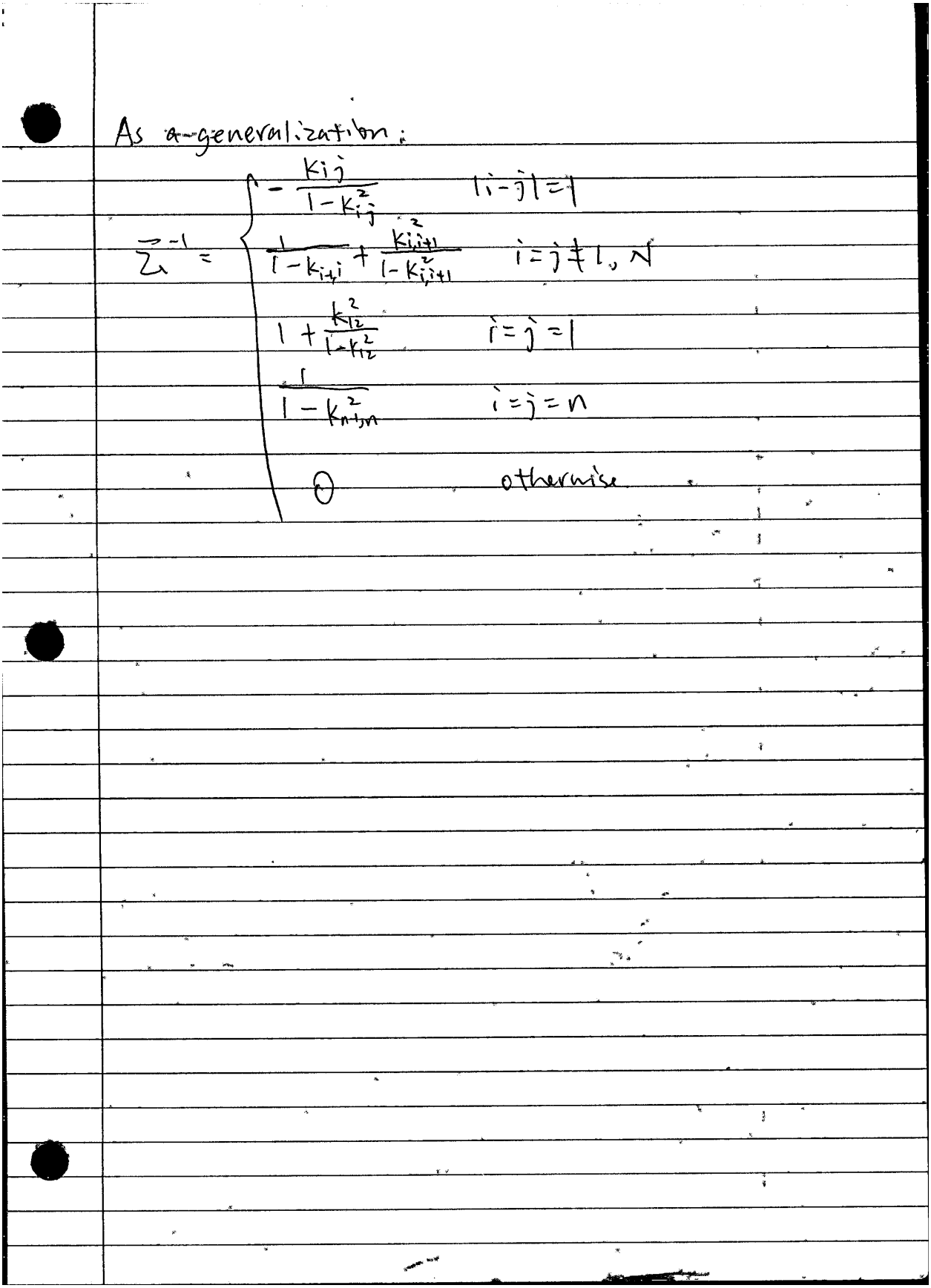

Figure 3: Problem 1.2

4. Since y_i, y_k, y_j are samples from multivariate Gaussian distribution $y \sim N(0, \Sigma)$, $y_i, y_j | y_k$ can be decomposed as

$$\begin{bmatrix} y_k \\ y_i \\ y_j \end{bmatrix} \sim N\left(0, \begin{bmatrix} K & K_1^T \\ K_1 & K_2 \end{bmatrix}\right)$$

where $K = \sigma(k,k) = 1$

$$K_1 = (\sigma(i,k), \sigma(j,k))^T$$

$$K_2 = \begin{pmatrix} 1 & \sigma(i,j) \\ \sigma(i,j) & 1 \end{pmatrix}$$

Since the conditional distribution of (y_i, y_j) given y_k itself is Gaussian-distributed, we can derive that

$$y_i, y_j | y_k \sim N(K_1 K^{-1}(y_k), K_2 - K_1 K^{-1} K_1^T)$$

The covariance matrix of $y_i, y_j | y_k$ is thus

$$\Lambda = K_2 - K_1 K^{-1} K_1^T$$

$$= \begin{pmatrix} 1 & \sigma(i,j) \\ \sigma(i,j) & 1 \end{pmatrix} - \begin{pmatrix} \sigma(i,k) \\ \sigma(j,k) \end{pmatrix} (\sigma(i,k), \sigma(j,k))$$

$$= \begin{pmatrix} 1 & \sigma(i,j) \\ \sigma(i,j) & 1 \end{pmatrix} - \begin{pmatrix} \sigma^2(i,k) & \sigma(i,j) \\ \sigma(i,j) & \sigma^2(j,k) \end{pmatrix}$$

$$= \begin{pmatrix} 1-\sigma^2(i,k) & 0 \\ 0 & 1-\sigma^2(j,k) \end{pmatrix}$$

According to the definition of correlation

$$\rho(y_i, y_j | y_k) = \frac{\mathrm{cov}(y_i, y_j | y_k)}{\sigma(i,i|k)\,\sigma(j,j|k)}$$

According to Λ, $\mathrm{cov}(y_i, y_j | y_k) = 0$ when $i \neq j$

thus $\rho(y_i, y_j | y_k) = 0$

Figure 4: Problem 1.4

5. Similar to the previous problem, decompose y^* and $Y=(Y_1, Y_2 \dots Y_n)$

$$\begin{pmatrix} Y \\ y^* \end{pmatrix} \sim N\left(0, \begin{bmatrix} C & C_1^T \\ C_1 & C_2 \end{bmatrix}\right)$$

where $C = \vec{Z}_1$, the original covariance matrix

$C_1 = (\sigma(1,*), \sigma(2,*) \dots \sigma(n,*))$

$C_2 = 1$

since the joint density of $[Y\ Y^*]^T$ is still Gaussian, we have:

$$Y^* | Y \sim N(C_1 C^{-1} Y,\ C_2 - C_1 C^{-1} C_1^T)$$

1) $\exists i : x^* = x_i$, $i = 1, \dots, n$, then

$C_1 = \vec{Z}_i$: the ith row of $\vec{Z}_i$

Thus $C_1 \cdot \vec{Z}_i^{-1} = (0, 0, \dots \sigma(i,i) \dots 0)$

$\hat{y}^* = C_1 C^{-1} Y = Y_i$ and density $(y^*) = 1$

2) when $x^* \neq x_i$, $\forall i = 1, \dots, n$

First compute $C_1 \vec{Z}_i^{-1}$ and C_1 is orthogonal to all columns of $\vec{Z}_i^{-1}$ except columns i, $i+1$ that $x_i < x_* < x_{i+1}$

$$C_1 \vec{Z}_i^{-1} = \left(0, \dots, \frac{\sigma(i,*)}{1+\sigma(i-1,i)} - \frac{\sigma(i,i+1)}{1+\sigma(i,i+1)} \cdot \frac{\sigma(i+1,*)}{1+\sigma(i,i+1)} \cdot \frac{\sigma(i+1,i+2)}{1+\sigma(i+1,i+2)}\right)$$

~~let the non-zero elements be~~ a_i, a_{i+1} $0, \dots$)

Thus

$$\hat{y}_* = a_i Y_i + a_{i+1} Y_{i+1}$$

density $(Y_*) = 1 - (a_i \sigma(i,*) + a_{i+1} \sigma(i+1,*)$

Figure 5: Problem 1.5

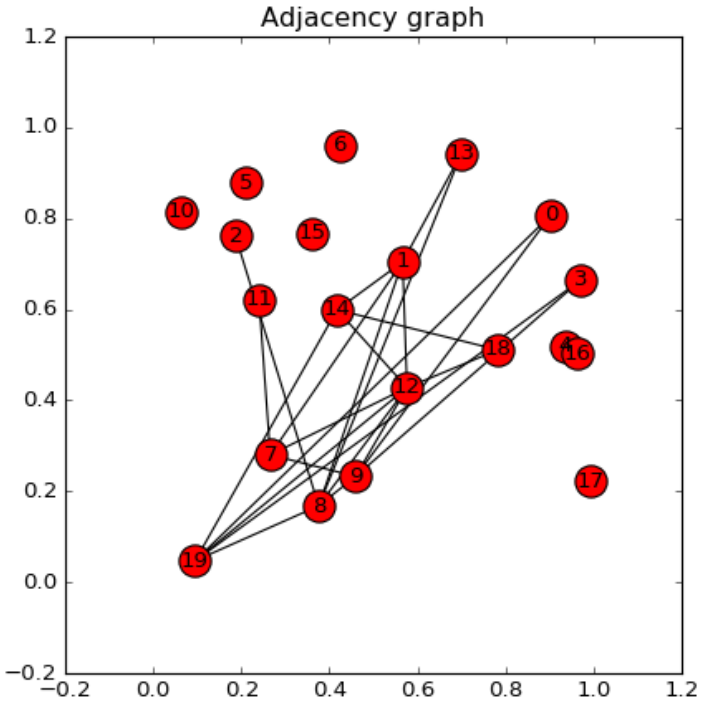

Figure 6: Adjacency graph of matrix A.

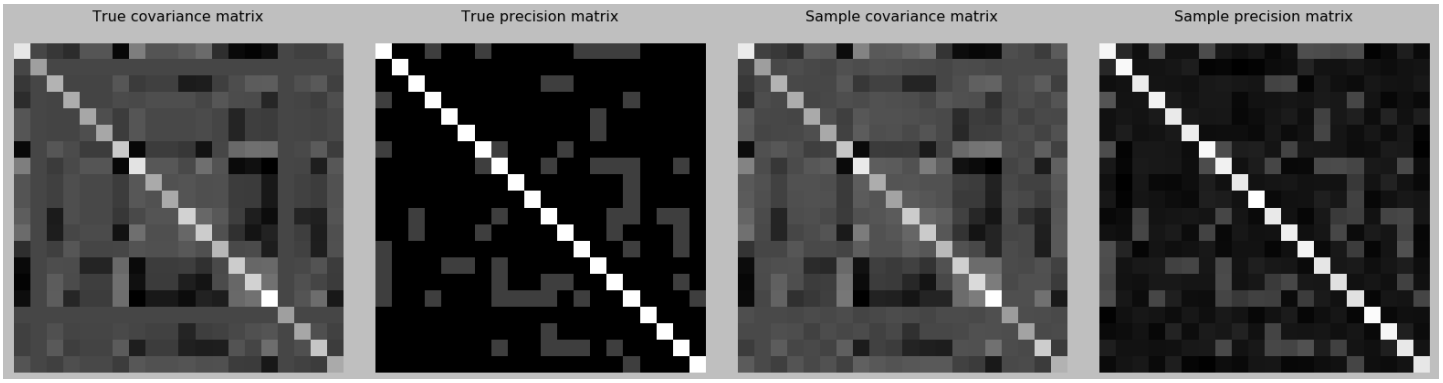

Figure 7: "True" covariance/precision matrix and sample covariance/precision matrix.

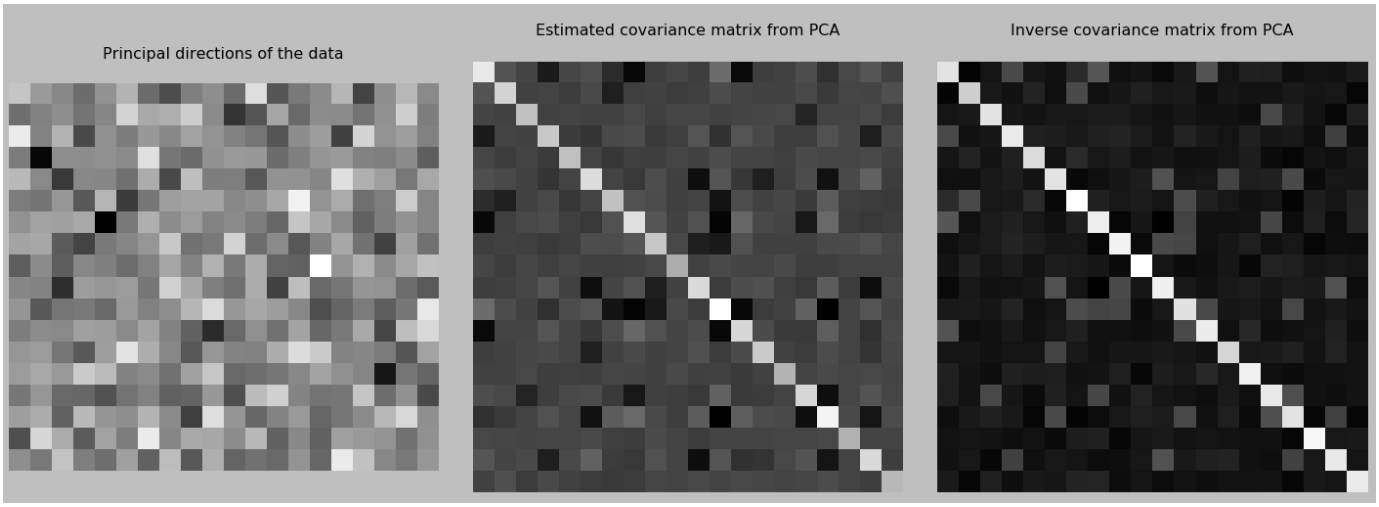

Figure 8: Principle directions and covariance/precision matrix of PCA.

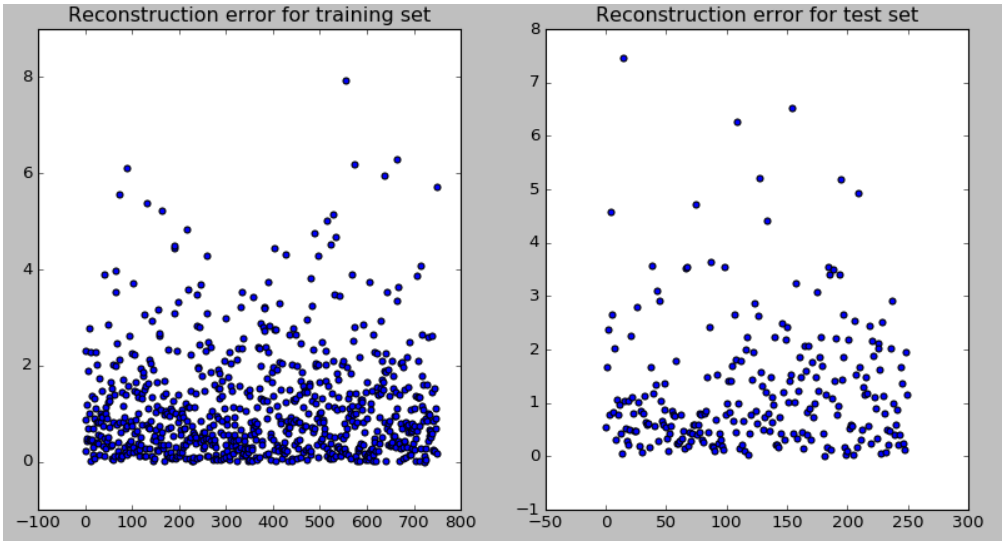

Figure 9: Reconstruction errors of training/test sets.

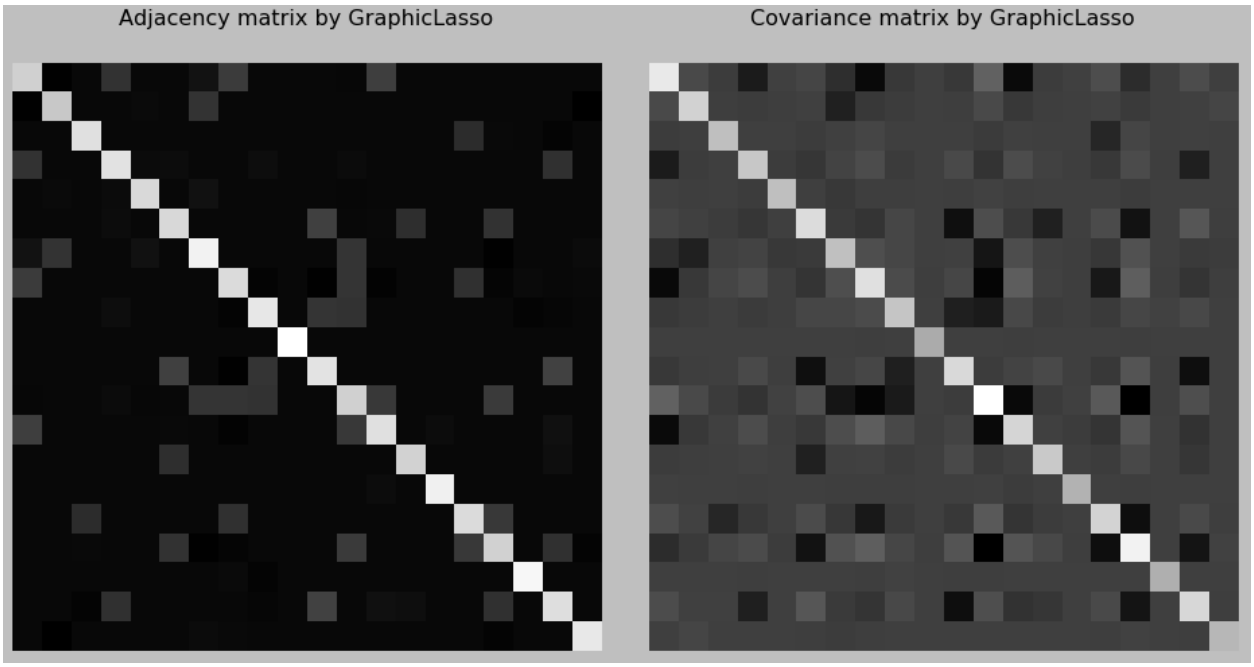

Figure 10: Adjacency and covariance matrices estimated by Graphical Lasso.