Banded precision matrix tests

Jiaming Shen

University of Manchester

jiaming.shen@postgrad.man.ac.uk

November 15, 2018

Overview

Introduction

Three approaches for test the bandwidth of the precision matrix is presented in this slides, which are An, Guo, & Liu, (2014) work,Cheng, Zhang, & Zhang (2017) and Lee & Lin (2018). The main differences between this two is There are two main differences in constructing the test statistic.

1. An et al. (2014) use the modified cholesky decomposition while Cheng et al. (2017) didn't.
2. Cheng et al. (2017) sum up non-banded entry of the residual covariance matrix, while An et al. (2014) only sum up k-sub-diagonal entry.

One main reason for such distinct is the Hypothesis are slightly different. The hypothesis for Cheng et al. (2017) is "Whether bandwidth is γ or not". The hypothesis for An et al. (2014) is "$K \leqslant k-1$" or "$K \leqslant k$".

Another approach by Lee & Lin (2018) is a similar method by An et al. (2014), however he used Bayesian methods rather than the frequency methods. Under the Bayesian set, the hypothesis is construct by bayes factor rather than a statistics follow a certain distribution.
Lee & Lee (2017) also showed a method using modified Cholesky decomposition to estimate precision matrix. k-banded Cholesky prior
Lee & Lin (2018) mainly concerned at the consistency of bandwidth selection and Bayes factor. (However, I am not too care about this, I mainly care about the methodology using to estimate the bandiwidth K and the precision matrices.)

Numbers of method will be benefit from the covariance structure modelling. A example is the mean-covariance model in longitudinal data by Pan & Pan (2017) and (Pourahmadi, 2000) analysis which involve the polynomial of time and lag-time to modelling the covariance matrices. Another example is Lee & Lin (2018) which use the Gaussian DAG Models which has band-structured covariance/precision matrices. An et al. (2014) directly modelling the precision matrix in the multivariable Normal distribution without any specification application of the model. Cheng et al. (2017) consider the Undirected Graph which may refer as Markov random field.

Not only for the LDA data, estimation of the covariance matrices is also important for principle component analysis (PCA), linear/quadratic distriminant analysis and multivariate analysis of variance (MANOVA)
n>>p case, sample covariance fails to converge to the true covariance matrix (Johnstone, manuscript, & 2004, n.d.).

Another from Lee & Lin (2018) should be considered is Banerjee & Ghosal (2015) . This paper is mainly concern about the graphical model structure, a sparse precision matrix of a Gaussian graphical model. The edge presents or absences in a graphical model describing conditional independence. A popular Non-Bayes method is graphical lasso. This paper use bayesian method instead, use posterior distribution to learn the covariance structure.G-Wishart prior is mentioned here.

A very common question is rised here, why choosing banded structure, how to permutate the variable order to make the covariance matrix is banded? How to link the sparsity and the order to make a good explanation of a covariance matrix.
Or in other words, how to explore the structure may be the main concern before the hypothesis test.
Another question, if the stucture is sparse, are there any possibility to do some permutation to the matrix, change the order of variables can transfer the matrix to "banded like" or blocked diagonal matrices.

The model specification:

Gaussian DAG model by (Lee & Lin, 2018) .
DAG: Directed acyclic graph.
Direct graph: $\mathcal{D} = (V, E)$,V is vertices $V = \{1, \ldots, p\}$, E is directed edges. For any $i, j \in V$, $(i, j) \in E$ as a directed edge $i \to j$. There is no cycle in a DAG model. Assume parent ordering is known, where $i < j$ holds for any parent i of j in a DAG $\mathcal{D}$. Gaussian DAG model over $\mathcal{D}$: $Y = (Y_1, \ldots, Y_p)^T \sim N_p\left(0, \Omega_n^{-1}\right)$ satisfies:

$$Y_i \perp \{Y_j\}_{j<i, j \in pa_i(\mathcal{D})} \mid \{Y_j\}_{j \in pa_i(\mathcal{D})}$$

Consider,

$$X_1, \ldots, X_n | \Omega_n \overset{i.i.d}{\sim} N_p\left(0, \Omega_n^{-1}\right) \tag{1}$$

Cholesky Decomposition (Slight different from Pourhmadi)
Do the Cholesky decomposition as following form :

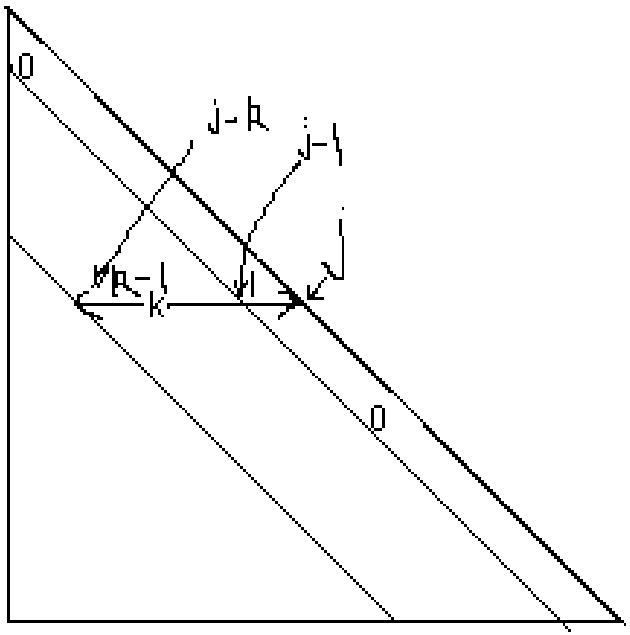
$$\Omega_n = (I_p - A_n)^T D_n^{-1} (I_p - A_n)$$

with $D_n = diag(d_j)$, $a_{jj} = 0$. Called A_n *Cholesky factor* is unique and lower triangular matrix. Result: $a_{jl} \neq 0$ iff $l \in pa_j(\mathcal{D})$. Description: a_{jl} not 0 iff l is parent of j, the edge $l \to j$ exists.

Model:

$$X_{i1}|d_1 \overset{i.i.d}{\sim} N(0, d_1)$$
$$X_{ij}|a_j^{(k)}, d_j, k \overset{ind}{\sim} N\left(\sum_{l=(j-k)_1}^{j-1} X_{il} a_{jl}, d_j\right), \quad j = 2, \ldots, p$$

That is, showed in above figure



A very useful result: For Cholesky decomposition upon, the bandwidth of A_n is k iff bandwidth of Ω_n is k.
Prior Distribution I'm not that concerned yet, however, should notice that it is the conjugate prior. So the posterior distribution can be calculated in a closed form up to some normalizing constant.
Assumption is important for the proof of estimation consistency, however, I regards this is "not neccessary" for the idea of method, so omit in this review. But I think the explaination in Section 2.4 is good and worth reading if involve the proof of consistency and other property.

1. Bandwidth selection consistency
2. Consistency of One-Sample Bandwidth Test Here construct a Bayesian bandwidth test for the testing problem:

$$H_0 : k \leqslant k^* \ vs \ \ H_1 : k > k^*$$

This is a composite hypothesis. The hypothesis test is based on the Bayes factor $B_{10}(X_n)$ defined by the **ratio of marginal likelihoods** ,

$$B_{10}(X_n) = \frac{p(X_n|H_1)}{p(X_n|H_0)}$$

Denote the prior under the hypothesis H_i as $\pi_i(A_n, D_n, k)$ for $i = 0, 1$. Using prior $\pi_0(k)$ and $\pi_1(k)$ such that

$$\begin{aligned} \pi_0(k) &= C_0^{-1}\pi(k), \quad k = 0, 1, \ldots, k^* \\ \pi_1(k) &= C_1^{-1}\pi(k), \quad k = k^\star + 1, \ldots, R_n \end{aligned}$$

with $C_0 = \sum_{k=0}^{k^*} \pi(k)$, $C_1 = \sum_{k=k^*+1}^{R_n} \pi(k)$.

Then the Bayes factor (the statistic we construct for Hypothesis), is in analytic form,

$$B_{10}(\mathbf{X}_n) = \frac{\sum_{k>k^*} \int p(\mathbf{X}_n|\Omega_n, k)\,\pi(\Omega_n|k)\,\pi_1(k)d\Omega_n}{\sum_{k\leq k^*} \int p(\mathbf{X}_n|\Omega_n, k)\,\pi(\Omega_n|k)\,\pi_0(k)d\Omega_n}$$
$$= \frac{\pi(k > k^*|\mathbf{X}_n)}{\pi(k \leq k^*|\mathbf{X}_n)} \times \frac{C_0}{C_1}$$

Consistency result: Under $H_0 : k \leq k^*$

$$B_{10}(\mathbf{X}_n) = O_p(T_{n,H_0,k_0,k^*})$$

Under H_1

$$B_{10}(\mathbf{X}_n)^{-1} = O_p(T_{n,H_1,k_0,k^*})$$

Then is the method comparison between (An et al., 2014; Cheng et al., 2017) and it's method. The mainly concern also is the consistency.

> *It is curious about the problem that how the test work, which means significance under such bayes factor, such ratio?*

Quote from the arthor.

> *If H_0 is true, $B_{10}(Y)$ decreases at rate $O_p(e^{-c_0(k_2-k_1)})$ for some constant $c_2 > 0$. On the other hand, if H_1 is true, $B_{10}(Y)^{-1}$ decreases exponentially with $n(k_2 - k_1)\beta_{min}^2$.*

So there should be some threshold for the ratio about "decline the null-hypothesis" or the "accept the alter-hypothesis" like 1 or 0.95 something. Back to the form of the Bayes factor : $B_{10}(X_n) = \frac{p(X_n|H_1)}{p(X_n|H_0)}$,if H_0 is true, then the denominator is larger than the numerator, bayes factor is closer to 0. If H_1 is true, then numerator $p(X_n|H_1)$ should be larger than the denominator. Then the inverse of the Bayes factor should closer to 0. As [Bayes factor] (`https://en.wikipedia.org/wiki/Bayes_factor`) give a slightly discussion about the explaination of the bayes factor. It seems no trivial maps from frequency 5% level significance with test under Bayes factor.

Because as it says, (An et al., 2014; Cheng et al., 2017)'s result is that the statistic they constructed is asymptotically under Normal distribution, then it comes to the normal 95% hypothesis-test-framework. Need notice, the asymptotically achieve for $n \wedge p \to \infty$ rather than $n \to \infty$.

Q : How the bayes factor linked with the Cholesky decomposition?

A: Insider the analytic form of Bayes factor: $p(X_n|\Omega_n, k)$ *and* $\pi(\Omega_n|k)$

3 Two-Sample Bandwidth Test

Problem:

$$X_1, \ldots, X_{n_1} | \Omega_{1n_1} \overset{i.i.d.}{\sim} N_p\left(0, \Omega_{1n_1}^{-1}\right)$$
$$Y_1, \ldots, Y_{n_2} | \Omega_{2n_2} \overset{i.i.d}{\sim} N_p\left(0, \Omega_{2n_2}^{-1}\right)$$

Interesting question: Test of equality between two bandwidths k_1 and k_2.
Hypothesis:

$$H_0 : k_1 = k_2, H_1 : k_1 \neq k_2$$

Bayes factor:

$$B_{10}\left(\mathbf{X}_{n_1}, \mathbf{Y}_{n_2}\right) = \frac{p\left(\mathbf{X}_{n_1}, \mathbf{Y}_{n_2} | H_1\right)}{p\left(\mathbf{X}_{n_1}, \mathbf{Y}_{n_2} | H_0\right)}$$

How to investigate the posterior/probability under the H_1 $p(X_{n1}, Y_{n_2})$?

To answer this question, should deep in P14 the construction and the specifying of the prior distribution inside $a_{1,j}^{(k_0)}$ and $a_{1,j}^{(k_1)}$ for given k_0 and k_1. Then the Bayes factor can be written as analytic form under the prior given

$$B_{10}\left(\mathbf{X}_{n_1}, \mathbf{Y}_{n_2}\right) = \frac{\sum_{k_1 \neq k_2} \pi\left(k_1 | \mathbf{X}_{n_1}\right) \pi\left(k_2 | \mathbf{Y}_{n_2}\right)}{\sum_{k_1 = k_2} \pi\left(k_1 | \mathbf{X}_{n_1}\right) \pi\left(k_2 | \mathbf{Y}_{n_2}\right)} \times R_n^{-1}$$

Use marginal posterior distribution to deal with the alternative hypothesis $k_1 \neq k_2$. And arthor gives a theorem about the consistency of such test.

Here start review about (An et al., 2014) . Summary about introduction. (Will insert to begining of the final ediction of review.)
"When the data are multivariate Gaussian, the precision matrix can be used to infer the conditional dependence structure of random variables."
Link between Precision matrix and conditional dependency structure (Problem: Only for Gaussian?)
"When the variables of interest have a natural order, it is often assumed that two random variable are not partially correlated when the distance between the is large enough." It is interesting to do such assumption, or just intuitively idea. There are things are done by this sentence, one is the "prior" of the importantce of variable has order and is already known. Then, the "most interesting" with "less interesting" variable are less correlated. However, this may not true in some situation, and may cause misleading in prior assumption.

That is, another problem rised here, how to decide the order of the variables, for banded structure, the covariance/precision matrix may differ for $X_1, X_2, ...X_n$ from $X_{i_1}, ..., X_{i_n}$ because the banded covariance structure showed the dependency from a variable to others is only on its neighbor. How to choose a satisfied order of indeces I ? How about the precision matrices? (Under Gaussian is same? How about other distributions?)

model-specification-notations

$X_i (i = 1, \ldots, n)$ observation collected from the ith subject.
$X_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}} \in \mathbb{R}^p (i = 1, \ldots, n)$ p-dimensional vector which are *independent* and *normally distributed* with mean 0 and covariance matrix Σ.

> *Notice: the collected data X is assume as multivariate Normal, and without a respond Y, so it is NOT a regression problem.*

MCD(modified cholesky decomposition) of Σ is denoted by $\Sigma = LDL^{\mathrm{T}}$, correspond formula in precision matrix is $\Omega = T^{\top} D^{-1} T$. Estimating T, D is equivelant to estimate Ω and Σ.

Similarly we have the autoregressive property in such notation as

$$X_{ij} = \sum_{q=1}^{j-1} (-t_{jq}) X_{iq} + \varepsilon_{ij}$$

for banded structure with bandwidth$=K_0$, we have

$$X_{ij} = \sum_{q=(j-K_0)_1} (-t_{jq}) X_{iq} + \varepsilon_{ij}$$

Hypothesis:

$$H_0 : K \leqslant k-1 \text{ vesus } H_1 : K \geqslant k$$

k is prespecified positive number smaller than $n-4$. Define $\mathcal{X} = (X_1, \ldots, X_n)^{\mathrm{T}}$ By fitting the autoregression equation, the estimator for $t_j^{(k)}$ as $\hat{t}_j^{(k)} = -\left(\mathcal{X}_j^{(k)\tau}\mathcal{X}_j^{(k)}\right)^{-1}\mathcal{X}_j^{(k)\mathrm{T}}\mathcal{X}_j$. Estimation for d_j as $\hat{d}_j^{(k)} = \left\|\mathcal{X}_j + \mathcal{X}_j^{(k)}\hat{t}_j^{(k)}\right\|^2 / \left[n - \{j - (j-k)_1\}\right]$.

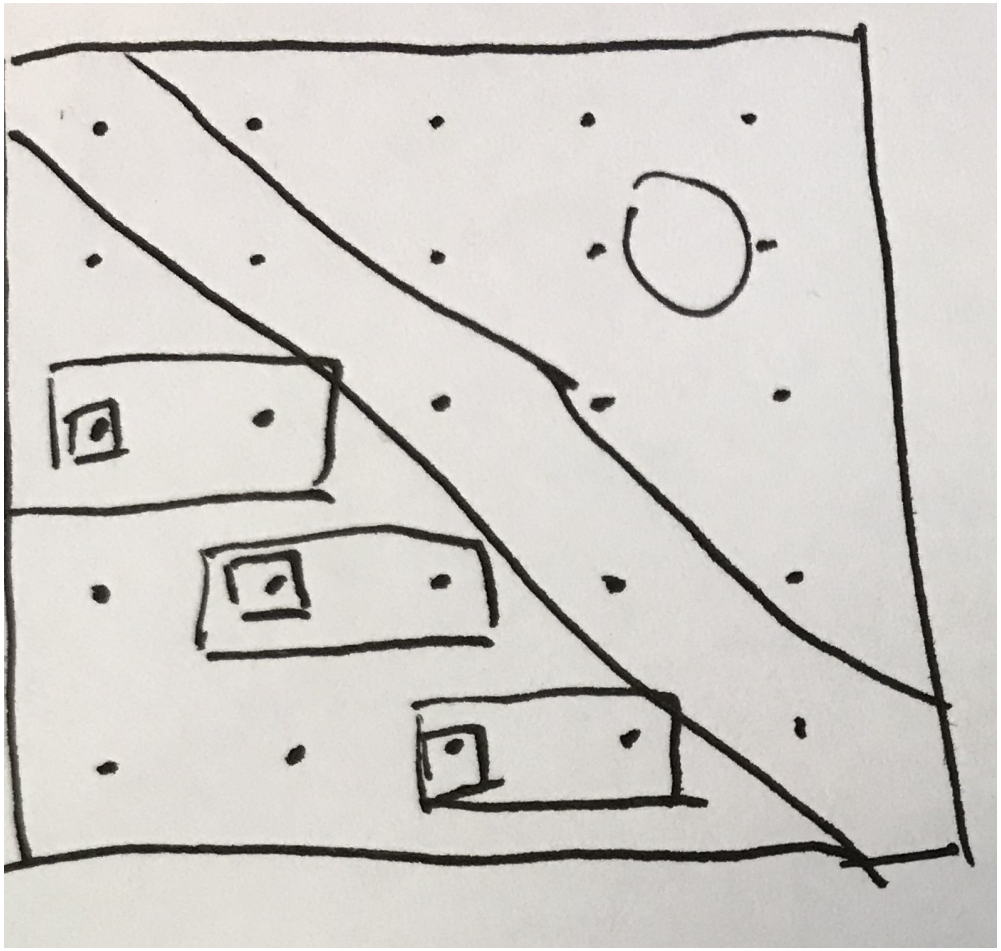
Surprise: finally an example that directly get the estimation by the mcd.

Idea: can we do the autoregression between variables to obtain the sparse covariance structure? such as $X_j = \sum_{i=1, i\neq j}^{n} \beta_i X_i$, *test the significant to 0 for* β *or regularized method to obtain the dependence structure for* $X_1, ..., X_n \sim N(0, \Sigma)$. *Then we change the order of the variables we can obtain a banded structure. Under such model, the bandwidth may change, so can we develope a more specific test for* k_i *?*

When H_0 is true, then $t_{j,j-k} = 0$ for every $j > k$. Then the corresponding estimators should be small. Follow this idea, the test statistic is constructed as

$$L = \sum_{j=k+1}^{p} \hat{t}_{j,j-k}^{(k)2}$$

The distribution, or approximate distribution of the test statistic is needed for an accurate testing procedure.
From the construction, the statistic consider the difference between "k-banded" with "$(k-1)$-banded", that is, showed in the figure below:



here is p=5,k=2 structure matrix.
The $t_j^{(k)}$ is the rectangle box of vector, the statistics L is the sum square of small square entry. It is the "outer-wall" of k-banded with (k-1)-banded.

Result by (Wu & Pourahmadi, 2003): Conditonal on $\mathcal{X}_j^{(k)}$,
$\hat{t}_{j,j-k}^{(k)} \sim N(t_{j,j-k}, \Delta_j \hat{d}_j^{(k)})$ where

$$\Delta_j = \left\{ \mathcal{X}_{j-k}^{\mathrm{T}} \mathcal{X}_{j-k} - \mathcal{X}_{j-k}^{\mathrm{T}} \mathcal{X}_j^{(k-1)} \left(\mathcal{X}_j^{(k-1)\mathrm{T}} \mathcal{X}_j^{(k-1)} \right)^{-1} \mathcal{X}_j^{(k-1)\mathrm{T}} \mathcal{X}_{j-k} \right\}^{-1}.$$

We have $\hat{t}_{j,j-k}^{(k)}$ is Normal, then the $\left(\Delta_j \hat{d}_j^{(k)} \right)^{-1/2} \left(\hat{t}_{j,j-k}^{(k)} - t_{j,j-k} \right) \sim t_{n-k}$.

Then $\left(\hat{t}_{j,j-k}^{(k)} - t_{j,j-k} \right)^2 / \left(\Delta_j \hat{d}_j^{(k)} \right) \sim F_{1,n-k}$.

Take l_j defined as $\hat{t}_{j,j-k}^{(k)2} / \left(\Delta_j \hat{d}_j^{(k)} \right)$, modified L as $L_c = \sum_{j=k+1}^{p} l_j$

Even more, the variance may different for each j, from the $\hat{t}_j^{(k)}$ may different. Then need modified again by standardize it. We have $l_j \sim F_{1,n-k}$ with $E(l_j) = (n-k)/(n-k-2)$ and $var(l_j) = 2(n-k)^2(n-k-1)/\left\{(n-k-2)^2(n-k-4)\right\}$

$$L_f = (p-k)^{-1/2} \sum_{j=k+1}^{p} \frac{l_j - E(l_j)}{\text{var}(l_j)^{1/2}}$$

Then the asymptotic null distribution of L_f as $n, p \to \infty$ is standard Normal.

After this, the arthor showed the power of the test.

Consider the hypothesis test below:

$$H_{0k} : K \leqslant k-1 \quad \text{versus} \quad H_{1k} : K \geqslant k, \quad 1 \leqslant k \leqslant M$$

Then we can estimate bandwidth by $K_0 = \{H_{0k} \text{ is false}\}$.

Algorithm1:

1. Specify significance level α, upper bound of k: M
2. k=0, stop, output $\hat{K} = k$,otherwise, compute the test-statistics L_f , denoted by $L_f^{(k)}$. Let $p_k = 2\{1 - \phi(|L_f^{(k)}|)\}$.
3. Test: if $p_k > \alpha_k$, then do not reject H_{0k} , and update $k = k - 1$, back to Step 2. If $p_k \leqslant \alpha_k$, reject H_{0k} and stop, report the bandwidth as $\hat{K} = k$

Notice: This is a multiple comparison procedure, whould control each α_k in order to control overall significance level α. Refer Efron (2010) for multiple comparison of hypothesis tests. Use Bonferroni procedure sets $\alpha_k = \alpha/M$ to control familywise error so that it is no larger than α. Bonferroni procedure may be too conservative when number of tests is relative large.

Holm (1979) proposed following procedure:
Algorithm2:

1. Initialize α.
2. For k=1,. . . ,M compute test statistics $L_f^{(k)}$. $p_k = 2\{1 - \phi(|L_f^{(k)}|)\}$
3. Sort p_k as $p_{(1)} \leqslant \cdots \leqslant p_{(M)}$ with correpond $H_{0j_i}, \cdots, H_{0j_M}$. Reject H_{0j_k} if $p_{(j)} \leqslant \alpha/(M - j + 1)$ for all $j = 1, ..., k$.
4. No H_0 is rejected: output $\hat{K} = 0$,implies diagonal matrix. Otherwise, $\hat{K} = max\{k : H_{0k}$ is rejected$\}$.

Now start for review Cheng et al. (2017)
Data involve: genetic regulatory networks: gene expression data, medical imaging, risk management and web search problems.
Data example: levels of p genes denotes as $(Y_1, ..., Y_p)$, a concentration network can be described by an undirected graph in which the p vertices represent the p genes and an edge connects gene i and gene j iff the partial correlation ρ_{ij} between Y_i and Y_j.
Such model can describe by an undirected graph is which the pvertices represent the p genes and an edge connects gene i and gene j if and only if the partial correlation ρ_{ij} between Y_i and Y_j is non-zero.

So in this description of undirected graph, the model is Markov random field ?

Edge connected Y_i and Y_j iff partial correlation ρ_{ij} is non-zero. If $(Y_1, \ldots, Y_p)$ are jointly normally distributedm, estimating the structure of undirected graph is equivalent to recovering the support of precision matrix ($\Omega = \Sigma^{-1}$)

> *Question: Just Gaussian graphical model for undirected model works? why only undirect. (Lee & Lin, 2018) is work on DAG, and (An et al., 2014) is work on multi-variable Normal without any specification such as graphical model.*

Normal property for this, see Lauritzen (1996) .

Introduction part is interesting to collect wide range of papers. (Bickel & Levina, 2008) is very important, have to read it all. (Rothman, Levina, & Zhu, 2010) is also important, seems first use Cholesky factor to model such covariance matrices. (But it seems too late? But we should trust Biometrika). But seems I omit it for now.

Notation and Model specification

$\mathbf{Y} = (Y_1, ..., Y_p)'$: p-dimension vector sith mean μ and covariance matrix $\Sigma_y = (\sigma_{y,ij})_{p\times p} = \Omega^{-1} = (\omega_{ij})_{p\times p}$
$\mathcal{B}_\gamma$ is a class of banded matrices with bandwidth γ. A matrix $\mathbf{A} = (a_{ij})$ is said to belong to $\mathcal{B}_\gamma$ if its elements (a_{ij}) are satisfies $a_{ij} = 0$ if $|i - j| > \gamma$.
(The full-scale of bandwidth, then the cholesky factor bandwidth should be $\frac{\gamma}{2}$)
Hypothesis test problem:

$$H_{\gamma,0} : \Omega \in \mathcal{B}_\gamma \text{ vs. } H_{\gamma,1} : \Omega \notin \mathcal{B}_\gamma$$

Remark: Bickel & Lindner (2012) showed: If Ω belongs to $\mathcal{B}_\gamma$, there exists $G_y \in B_{c\gamma}$ and Σ_y can be approximated by G_y in the sense of $||\Sigma_y - G_y|| \leq (1/m)((M-m)/(M+m))^c$ (for every $c \in N_0$). $||\cdot||$ is spectral norm. M and m are largest and smallest eigenvalues of Ω. An efficient method of testing the banded structure of Σ_y due to Qiu & Chen (2012) .

Test Statistic:

$B_\gamma(\Omega) = (\omega_{ij} I\{|i-j| \leq \gamma\})$ is a banded version of Ω with bandwidth γ. Then the difference between Ω and $B_\gamma(\Omega)$ should be small if null hypothesis is true.

The measurement of the difference between two matrices used in this paper is matrix's Frobenius norm $\text{tr}\,(\Omega - B_\gamma(\Omega))^2$. Because it is easier to analysis.

Under the null hypothesis, then we have

$\text{tr}\,(\Omega - B_\gamma(\Omega))^2 = \sum_{|i-j|>\gamma} \omega_{ij}^2 = 0$

Let

$$Y_i = \alpha_i + Y'_{-i}\beta_i + \epsilon_i$$

β_i is a $p-1$ vector $(\beta_{1,i}, \ldots, \beta_{i-1,i}, \beta_{i+1,i}, \ldots, \beta_{p,i})^{\mathrm{T}}$, which satisfies

$$\boldsymbol{\beta}_i = -\frac{\Omega_{-i,i}}{\omega_{ii}}, \quad \text{and} \quad \operatorname{Cov}(\epsilon_i, \epsilon_j) = \frac{\omega_{ij}}{\omega_{ii}\omega_{jj}}$$

$\operatorname{tr}(\Omega - B_\gamma(\Omega))^2 = 0$ implies $\sum_{|i-j|>\gamma} \sigma^2_{\epsilon,ij} = 0$.
$\Sigma_\epsilon = (\operatorname{Cov}(\epsilon_i, \epsilon_j))_{p\times p} = (\sigma_{\epsilon,ij})_{p\times p}$.

Under such notation, Σ_ϵ belons to $\mathcal{B}_\gamma$ if and only if Ω belongs to $\mathcal{B}_\gamma$.
An unbiased estimator of $\sigma^2_{\epsilon,ij}$:

$$\tilde{\sigma}^2_{\epsilon,ij} = \left\{ \frac{1}{P_n^2} \sum_{kl}^{*} (\epsilon_{ki}\epsilon_{kj})(\epsilon_{li}\epsilon_{lj}) - 2\frac{1}{P_n^3} \sum_{kls}^{*} (\epsilon_{ki}\epsilon_{kj})(\epsilon_{li}\epsilon_{sj}) + \frac{1}{P_n^4} \sum_{klst}^{*} (\epsilon_{ki}\epsilon_{lj})(\epsilon_{si}\epsilon \right.$$

location invariant statistic :

$$T_{n\gamma} := \sum_{|i-j|>\gamma} \tilde{\sigma}^2_{\epsilon,ij}$$

$T_{n\gamma}$ is cannot be observed due to ϵ is unobserved. Denote the residuals as

$$\hat{\epsilon}_{ki} = Y_{ki} - \hat{\alpha}_i - Y'_{k,-i}\hat{\boldsymbol{\beta}}_i = Y_{ki} - \overline{Y}_i - \left(\boldsymbol{Y}'_{k,-i} - \overline{\boldsymbol{Y}}'_{-i} \right) \hat{\boldsymbol{\beta}}_i$$

use estimator instead of real error term in $\tilde{\sigma}^2_{\epsilon,ij}$.

then obtain a U statistic as

$$\hat{\sigma}^2_{\epsilon,ij} = \frac{1}{P_n^2} \sum_{kl}^{*} \left(\hat{\epsilon}_{ki}\hat{\epsilon}_{kj}\right)\left(\hat{\epsilon}_{li}\hat{\epsilon}_{jj}\right) - 2\frac{1}{P_n^3} \sum_{kls}^{*} \left(\hat{\epsilon}_{ki}\hat{\epsilon}_{kj}\right)\left(\hat{\epsilon}_{li}\hat{\epsilon}_{sj}\right) \tag{2}$$

$$+ \frac{1}{P_n^4} \sum_{klst}^{*} \left(\hat{\epsilon}_{ki}\epsilon_{jj}\right)\left(\hat{\epsilon}_{si}\hat{\epsilon}_{tj}\right) \tag{3}$$

Then alternative for $T_{n\gamma}$ as $T'_{n\gamma} = \sum_{|i-j|>\gamma} \hat{\sigma}^2_{\epsilon,ij}$.
Then $(T_{n\gamma} - T'_{n\gamma})/\sqrt{Var(T_{n\gamma})} \to 0$ in probability under certain conditions.
Following is the asymptotically theory and converge for such statistic and the assumption for the asymptotically proportion.

An, B., Guo, J., & Liu, Y. (2014). Hypothesis testing for band size detection of high-dimensional banded precision matrices. *Biometrika*, *101*(2), 477–483.
Banerjee, S., & Ghosal, S. (2015). Bayesian structure learning in graphical models. *Journal of Multivariate Analysis*, *136*, 147–162.
Bickel, P. J., & Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics*, *36*(1), 199–227.
Bickel, P., & Lindner, M. (2012). Approximating the Inverse of Banded Matrices by Banded Matrices with Applications to Probability and Statistics. *Theory of Probability & Its Applications*, *56*(1), 1–20.
Cheng, G., Zhang, Z., & Zhang, B. (2017). Test for bandedness of high-dimensional precision matrices, 1–20.
Efron, B. (2010). *Large-Scale Inference by Bradley Efron*. Cambridge University Press.
Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics. Theory and Applications*, *6*(2), 65–70.
Johnstone, I. M., manuscript, A. L. U., & 2004. (n.d.). Sparse principal components analysis. *Pdfs.semanticscholar.org*.

Lauritzen, S. L. (1996). Graphical models.
Lee, K., & Lee, J. (2017). Estimating Large Precision Matrices via Modified Cholesky Decomposition. *arXiv.org*. Retrieved from `http://arxiv.org/abs/1707.01143v1`
Lee, K., & Lin, L. (2018). Bayesian Test and Selection for Bandwidth of High-dimensional Banded Precision Matrices. *arXiv.org*. Retrieved from `http://arxiv.org/abs/1804.08650v1`
Pan, J., & Pan, Y. (2017). jmcm: An RPackage for Joint Mean-Covariance Modeling of Longitudinal Data. *Journal of Statistical Software*, *82*(9), 1–29.
Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, *87*(2), 425–435.
Qiu, Y., & Chen, S. X. (2012). Test for bandedness of high-dimensional covariance matrices and bandwidth estimation. *The Annals of Statistics*, *40*(3), 1285–1314.

Rothman, A. J., Levina, E., & Zhu, J. (2010). A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika*, *97*(3), 539–550.
Wu, W. B., & Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, *90*(4), 831–844.