**High dim time series** **(Always take notes)**

# Unfamiliar knowledge in class

Name: Weihao LI, Netid: 12243473

## Learn Toeplitz matrix

- Trace of a matrix is the sum of its eigenvalues
- The weak norm (or Hilbert-Schmidt norm) of an $n \times n$ matrix $A = [a_{k,j}]$ is defined by

$$|A| = \left( \frac{1}{n} \sum_{k=0}^{n-1} \sum_{j=0}^{n-1} |a_{k,j}|^2 \right)^{1/2}$$
$$= \left( \frac{1}{n} \operatorname{Tr}[A^* A] \right)^{1/2} = \left( \frac{1}{n} \sum_{k=0}^{n-1} \lambda_k \right)^{1/2}$$

  The quantity $\sqrt{n}|A|$ is sometimes called the Frobenius norm. A=$URU^*, \alpha = diag(R)$, $|A|^2 \geq \frac{1}{n} \sum_{k=0}^{n-1} |\alpha_k|^2$, with equality iff $A$ is normal.
- The Hilbert-Schmidt norm is the "weaker" of the two norms since

$$\|A\|^2 = \max_k \lambda_k \geq \frac{1}{n} \sum_{k=0}^{n-1} \lambda_k = |A|^2$$

  A matrix is said to be bounded if it is bounded in both norms.
- $|GH| \leq \|G\||H|$

- **(Definition)** Two sequences of $n \times n$ matrices $\{A_n\}$ and $\{B_n\}$ are said to be asymptotically equivalent if
  1. $A_n$ and $B_n$ are uniformly bounded in strong (and hence in weak) norm:
  $$\|A_n\|, \|B_n\| \leq M < \infty, n = 1, 2, \ldots$$
  2. $A_n - B_n = D_n$ goes to zero in weak norm as $n \to \infty$
  $$\lim_{n\to\infty} |A_n - B_n| = \lim_{n\to\infty} |D_n| = 0$$

  Asymptotic equivalence of the sequences $\{A_n\}$ and $\{B_n\}$ will be abbreviated $A_n \sim B_n$
- Theorem 2.1. Let $\{A_n\}$ and $\{B_n\}$ be sequences of matrices with eigenvalues $\{\alpha_n, i\}$ and $\{\beta_n, i\}$, respectively.

  (1) If $A_n \sim B_n$, then $\lim_{n\to\infty} |A_n| = \lim_{n\to\infty} |B_n|$

  (2) If $A_n \sim B_n$ and $B_n \sim C_n$, then $A_n \sim C_n$ (3) If $A_n \sim B_n$ and $C_n \sim D_n$, then $A_n C_n \sim B_n D_n$ (4) If $A_n \sim B_n$ and $\left\|A_n^{-1}\right\|, \left\|B_n^{-1}\right\| \leq K < \infty$, all $n$, then $A_n^{-1} \sim B_n^{-1}$ (5) If $A_n B_n \sim C_n$ and $\left\|A_n^{-1}\right\| \leq K < \infty$, then $B_n \sim A_n^{-1} C_n$

  (6) If $A_n \sim B_n$, then there are finite constants $m$ and $M$ such that $m \leq \alpha_{n,k}, \beta_{n,k} \leq M, \quad n = 1, 2, \ldots \quad k = 0, 1, \ldots, n-1$

- Lemma 2.4. Given two matrices $A$ and $B$ with eigenvalues $\{\alpha_k\}$ and $\{\beta_k\}$, respectively, then

$$\left|\frac{1}{n}\sum_{k=0}^{n-1}\alpha_k - \frac{1}{n}\sum_{k=0}^{n-1}\beta_k\right| \leq |A-B|$$

  Proof:
  $\sum_{k=0}^{n-1}\alpha_k - \sum_{k=0}^{n-1}\beta_k = \mathrm{Tr}(A) - \mathrm{Tr}(B) = \mathrm{Tr}(D)$. Applying the Cauchy-Schwarz inequality

$$\begin{aligned}|\,\mathrm{Tr}(D)|^2 = \left|\sum_{k=0}^{n-1} d_{k,k}\right|^2 &\leq n\sum_{k=0}^{n-1}|d_{k,k}|^2 \\ &\leq n\sum_{k=0}^{n-1}\sum_{j=0}^{n-1}|d_{k,j}|^2 = n^2|D|^2\end{aligned}$$

- Corollary 2.1. Given two sequences of asymptotically equivalent matrices $\{A_n\}$ and $\{B_n\}$ with eigenvalues $\{\alpha_{n,k}\}$ and $\{\beta_{n,k}\}$, respectively, then

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}(\alpha_{n,k}-\beta_{n,k}) = 0$$

  Proof: Let $D_n = \{d_{k,j}\} = A_n - B_n$. we have $\lim_{n\to\infty}\frac{1}{n}\mathrm{Tr}\,(D_n) = 0$ Dividing by $n^2$, and taking the limit, results in

$$0 \leq \left|\frac{1}{n}\mathrm{Tr}\,(D_n)\right|^2 \leq |D_n|^2 \overset{n\to\infty}{\longrightarrow} 0$$

  Corollary can be interpreted as saying the sample or arithmetic means of the eigenvalues of two matrices are asymptotically equal if the matrices are asymptotically equivalent.

- Corollary 2.2. Given two sequences of asymptotically equivalent Hermitian matrices $\{A_n\}$ and $\{B_n\}$ with eigenvalues $\{\alpha_{n,k}\}$ and $\{\beta_{n,k}\}$ respectively, then

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\left(\alpha_{n,k}^2-\beta_{n,k}^2\right) = 0$$

  Proof:

$$\begin{aligned}|D_n| &\geq ||A_n| - |B_n|| \\ &= \left|\sqrt{\frac{1}{n}\sum_{k=0}^{n-1}\alpha_{n,k}^2} - \sqrt{\frac{1}{n}\sum_{k=0}^{n-1}\beta_{n,k}^2}\right| \overset{n\to\infty}{\longrightarrow} 0\end{aligned}$$

- Theorem 2.2. Let $\{A_n\}$ and $\{B_n\}$ be asymptotically equivalent sequences of matrices with eigenvalues $\{\alpha_{n,k}\}$ and $\{\beta_{n,k}\}$, respectively. Then for any positive integer $s$ the sequences of matrices $\{A_n^s\}$ and $\{B_n^s\}$ are also asymptotically equivalent,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\left(\alpha_{n,k}^s-\beta_{n,k}^s\right) = 0$$

  Proof: Let $A_n = B_n + D_n$ as in the proof of Corollary 2.1 and consider $A_n^s - B_n^s \triangleq \Delta_n$, since the eigenvalues of $A_n^s$ are $\alpha_{n,k}^s$. using binomial expansion: $A_n^s - B_n^s = (B_n + D_n)^s - B_n^s$. The matrix $\Delta_n$ is a sum of several terms each being a product of $D_n$ 's and $B_n$ 's, but containing at least one $D_n$. Then use inequality $|GH| \leq \|G\||H|$, we can get $|\Delta_n| \leq K\,|D_n| \overset{n\to\infty}{\longrightarrow} 0$ which imply

$$\lim_{n\to\infty}\frac{1}{n}\mathrm{Tr}\,(\Delta_n) = 0$$

- Corollary 2.3. Suppose that $\{A_n\}$ and $\{B_n\}$ are asymptotically equivalent sequences of matrices with eigenvalues $\{\alpha_{n,k}\}$ and $\{\beta_{n,k}\}$ respectively, and let $f(x)$ be any polynomial. Then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \left(f\left(\alpha_{n,k}\right) - f\left(\beta_{n,k}\right)\right) = 0$$

- Theorem 2.3. (Weierstrass) If $F(x)$ is a continuous complex function on $[a, b]$, there exists a sequence of polynomials $p_n(x)$ such that

$$\lim_{n\to\infty} p_n(x) = F(x)$$

  uniformly on $[a, b]$, any continuous function defined on a real interval can be approximated arbitrarily closely and uniformly by a polynomial.

- Theorem 2.4. Let $\{A_n\}$ and $\{B_n\}$ be asymptotically equivalent sequences of Hermitian matrices with eigenvalues $\{\alpha_{n,k}\}$ and $\{\beta_{n,k}\}$, respectively. From Theorem 2.1 there exist finite numbers $m$ and $M$ such that

$$m \le \alpha_{n,k}, \beta_{n,k} \le M, \quad n = 1, 2, \ldots \quad k = 0, 1, \ldots, n-1$$

  Let $F(x)$ be an arbitrary function continuous on $[m, M]$. Then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \left(F\left(\alpha_{n,k}\right) - F\left(\beta_{n,k}\right)\right) = 0$$

- Corollary 2.4. Let $\{A_n\}$ and $\{B_n\}$ be asymptotically equivalent sequences of Hermitian matrices with eigenvalues $\{\alpha_{n,k}\}$ and $\{\beta_{n,k}\}$, respectively, such that $\alpha_{n,k}, \beta_{n,k} \ge m > 0$. Then if either limit exists,i.e $\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} F\left(\alpha_{n,k}\right), \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} F\left(\beta_{n,k}\right)$, then

$$\lim_{n\to\infty} \left(\det A_n\right)^{1/n} = \lim_{n\to\infty} \left(\det B_n\right)^{1/n}$$

  Proof: From Theorem 2.4 we have for $F(x) = \ln x$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \ln \alpha_{n,k} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \ln \beta_{n,k}$$

  and hence

$$\lim_{n\to\infty} \exp\left[\frac{1}{n} \ln \prod_{k=0}^{n-1} \alpha_{n,k}\right] = \lim_{n\to\infty} \exp\left[\frac{1}{n} \ln \prod_{k=0}^{n-1} \beta_{n,k}\right]$$

  equivalently

$$\lim_{n\to\infty} \exp\left[\frac{1}{n} \ln \det A_n\right] = \lim_{n\to\infty} \exp\left[\frac{1}{n} \ln \det B_n\right]$$

  Remark: The difficulty with allowing the eigenvalues to approach 0 is that their logarithms are not bounded. Furthermore, the function ln x is not continuous at x = 0, so Theorem 2.4 does not apply.

Circulant Matrices

$$C = \begin{bmatrix} c_0 & c_1 & c_2 & & \cdots & c_{n-1} \\ c_{n-1} & c_0 & c_1 & c_2 & & \vdots \\ & c_{n-1} & c_0 & c_1 & \ddots & \\ \vdots & \ddots & \ddots & \ddots & & c_2 \\ & & & & & c_1 \\ c_1 & \cdots & & c_{n-1} & & c_0 \end{bmatrix}$$

- The eigenvalues $\psi_k$ and the eigenvectors $y^{(k)}$ of $C$ are the solutions of

$$Cy = \psi y$$

or, equivalently, of the $n$ difference equations

$$\sum_{k=0}^{m-1} c_{n-m+k} y_k + \sum_{k=m}^{n-1} c_{k-m} y_k = \psi y_m; m = 0, 1, \ldots, n-1$$

One can solve difference equations as one solves differential equations — by guessing an intuitive solution and then proving that it works. Since the equation is linear with constant coefficients a reasonable guess is $y_k = \rho^k$ (analogous to $y(t) = e^{s\tau}$ in linear time invariant differential equations). We have $\rho^{-n} = 1$, i.e., $\rho$ is one of the $n$ distinct complex $n^{th}$ roots of unity, then we have an eigenvalue $\psi = \sum_{k=0}^{n-1} c_k \rho^k$ with corresponding eigenvector $y = n^{-1/2} \left(1, \rho, \rho^2, \ldots, \rho^{n-1}\right)^T$

- Choosing $\rho_m$ as the complex $n^{\text{th}}$ root of unity, $\rho_m = e^{-2\pi i m/n}$, we have eigenvalue

$$\psi_m = \sum_{k=0}^{n-1} c_k e^{-2\pi i m k/n}$$

and eigenvector $y^{(m)} = \frac{1}{\sqrt{n}} \left(1, e^{-2\pi i m/n}, \cdots, e^{-2\pi i m(n-1)/n}\right)^T$ Thus from the definition of eigenvalues and eigenvectors,

$$Cy^{(m)} = \psi_m y^{(m)}, m = 0, 1, \ldots, n-1$$

- **Definition:** The *discrete Fourier transform* transforms a sequence of $N$ complex numbers $\{\mathbf{x_n}\} := x_0, x_1, \ldots, x_{N-1}$ into another sequence of complex numbers, $\{\mathbf{X_k}\} := X_0, X_1, \ldots, X_{N-1}$, which is defined by

$$\begin{aligned} X_k &= \sum_{n=0}^{N-1} x_n \cdot e^{-\frac{i2\pi}{N} kn} \\ &= \sum_{n=0}^{N-1} x_n \cdot \left[\cos\left(\frac{2\pi}{N} kn\right) - i \cdot \sin\left(\frac{2\pi}{N} kn\right)\right] \end{aligned}$$

The discrete Fourier transform is an invertible, linear transformation. The inverse transform is given by:

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k \cdot e^{i2\pi kn/N}$$

- Theorem 3.1. Every circulant matrix $C$ has eigenvectors $y^{(m)} = \frac{1}{\sqrt{n}} \left(1, e^{-2\pi i m/n}, \cdots, e^{-2\pi i m(n-1)/n}\right)^T, m = 0, 1, \ldots, n-1$, and corresponding eigenvalues

$$\psi_m = \sum_{k=0}^{n-1} c_k e^{-2\pi i m k/n}$$

and can be expressed in the form $C = U\Psi U^*$, where $U$ has the eigenvectors as columns in order and $\Psi$ is diag $(\psi_k)$. In particular all circulant matrices share the same eigenvectors, the same matrix $U$ works for all circulant matrices, and any matrix of the form $C = U\Psi U^*$ is circulant.

- Let $C = \{c_{k-j}\}$ and $B = \{b_{k-j}\}$ be circulant $n \times n$ matrices with eigenvalues

$$\psi_m = \sum_{k=0}^{n-1} c_k e^{-2\pi imk/n}, \quad \beta_m = \sum_{k=0}^{n-1} b_k e^{-2\pi imk/n}$$

respectively. Then

1. $C$ and $B$ commute and
$$CB = BC = U\gamma U^*$$
where $\gamma = \text{diag}\,(\psi_m\beta_m)$, and $CB$ is also a circulant matrix.
2. $C + B$ is a circulant matrix and
$$C + B = U\Omega U^*$$
where $\Omega = \{(\psi_m + \beta_m)\,\delta_{k-m}\}$
3. If $\psi_m \neq 0; m = 0, 1, \ldots, n-1$, then $C$ is nonsingular and
$$C^{-1} = U\Psi^{-1}U^*$$

- We shall see that suitably chosen sequences of circulant matrices asymptotically approximate sequences of Toeplitz matrices and hence results similar to those in Theorem 3.1 will hold asymptotically for sequences of Toeplitz matrices.

Toeplitz Matrices

- Consider the infinite sequence $\{t_k\}$ and define the corresponding sequence of $n \times n$ Toeplitz matrices $T_n = [t_{k-j}; k, j = 0, 1, \ldots, n-1]$. The most general is to assume that the $t_k$ are square summable, i.e., $\sum_{k=-\infty}^{\infty} |t_k|^2 < \infty$
- We will make the stronger assumption that the $t_k$ are absolutely summable, i.e., $\sum_{k=-\infty}^{\infty} |t_k| < \infty$ Why stronger? $\sum_{k=-\infty}^{\infty} |t_k|^2 \leq \left\{\sum_{k=-\infty}^{\infty} |t_k|\right\}^2$
- Absolutely summable make sure Fourier series $f(\lambda)$ exists:

$$f(\lambda) = \sum_{k=-\infty}^{\infty} t_k e^{ik\lambda} = \lim_{n\to\infty} \sum_{k=-n}^{n} t_k e^{ik\lambda}$$

it converges uniformly:

$$\begin{aligned}\left| f(\lambda) - \sum_{k=-n}^{n} t_k e^{ik\lambda} \right| &= \left| \sum_{k=-\infty}^{-n-1} t_k e^{ik\lambda} + \sum_{k=n+1}^{\infty} t_k e^{ik\lambda} \right| \\ &\leq \left| \sum_{k=-\infty}^{-n-1} t_k e^{ik\lambda} \right| + \left| \sum_{k=n+1}^{\infty} t_k e^{ik\lambda} \right| \\ &\leq \sum_{k=-\infty}^{-n-1} |t_k| + \sum_{k=n+1}^{\infty} |t_k|\end{aligned}$$

Thus given $\epsilon$ there is a single $N$, not depending on $\lambda$, such that

$$\left| f(\lambda) - \sum_{k=-n}^{n} t_k e^{ik\lambda} \right| \leq \epsilon, \text{ all } \lambda \in [0, 2\pi], \text{ if } n \geq N$$

- Furthermore, if absolutely summable, then $f(\lambda)$ is Riemann integrable and the $t_k$ can be recovered from $f$ from the ordinary Fourier inversion formula:

$$t_k = \frac{1}{2\pi}\int_0^{2\pi} f(\lambda)e^{-ik\lambda}d\lambda$$

- A sequence of Toeplitz matrices $T_n = [t_{k-j}]$ for which the $t_k$ are absolutely summable is said to be in the Wiener class,. Similarly, a function $f(\lambda)$ defined on $[0, 2\pi]$ is said to be in the Wiener class if it has a Fourier series with absolutely summable Fourier coefficients. It will often be of interest to begin with a function $f$ in the Wiener class and then define the sequence of of $n \times n$ Toeplitz matrices

$$T_n(f) = \left[\frac{1}{2\pi}\int_0^{2\pi} f(\lambda)e^{-i(k-j)\lambda}d\lambda; \quad k, j = 0, 1, \cdots, n-1\right]$$

  The Toeplitz matrix $T_n(f)$ will be Hermitian if and only if $f$ is real.

- Functions f in the Wiener class are bounded since $|f(\lambda)| \leq \sum_{k=-\infty}^{\infty} \left|t_k e^{ik\lambda}\right| \leq \sum_{k=-\infty}^{\infty} |t_k|$ so that $\sup_f, \inf_f$:

$$m_{|f|}, M_{|f|} \leq \sum_{k=-\infty}^{\infty} |t_k|$$

Bounds on Eigenvalues of Toeplitz Matrices

- Lemma 4.1. Let $\tau_{n,k}$ be the eigenvalues of a Toeplitz matrix $T_n(f)$ If $T_n(f)$ is Hermitian, then

$$m_f \leq \tau_{n,k} \leq M_f$$

  Whether or not $T_n(f)$ is Hermitian,

$$\|T_n(f)\| \leq 2M_{|f|}$$

  so that the sequence of Toeplitz matrices $\{T_n(f)\}$ is uniformly bounded over $n$ if the essential supremum of $|f|$ is finite.

-

$$\begin{aligned}
|T_n(f)|^2 &= \frac{1}{n}\sum_{k=0}^{n-1}\sum_{j=0}^{n-1}|t_{k-j}|^2 \\
&= \frac{1}{n}\sum_{k=-(n-1)}^{n-1}(n-|k|)\,|t_k|^2 \\
&= \sum_{k=-(n-1)}^{n-1}(1-|k|/n)\,|t_k|^2
\end{aligned}$$

# 1 Appendix

The knowledge is based on following tutorial: https://ee.stanford.edu/~gray/toeplitz.pdf

https://www.stat.berkeley.edu/~bartlett/courses/153-fall2010/lectures/19.pdf

- If a time series $\{X_t\}$ has autocovariance $\gamma$ satisfying $\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty$, then we define its **spectral density** as

$$f(\nu) = \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i \nu h}$$

for $-\infty < \nu < \infty$

Ways to estimate $f(\nu)$, (1) replace by $\hat{\gamma}(\cdot)$ (2) periodogram

$$\hat{f}(\nu) = \sum_{h=-n+1}^{n-1} \hat{\gamma}(h) e^{-2\pi i \nu h}$$

for $-1/2 \leq \nu \leq 1/2$

- Discrete Fourier transform: For a sequence $(x_1, \ldots, x_n)$, define the discrete Fourier transform (DFT) as $(X(\nu_0), X(\nu_1), \ldots, X(\nu_{n-1}))$, where

$$X(\nu_k) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} x_t e^{-2\pi i \nu_k t}$$

and $\nu_k = k/n$ (for $k = 0, 1, \ldots, n-1$) are called the Fourier frequencies. (Think of $\{\nu_k : k = 0, \ldots, n-1\}$ as the discrete version of the frequency range $\nu \in [0, 1]$.)

Remark: View the DFT as a representation of $x$ in a different basis, the *Fourier basis*

- Orthonormal basis: Suppose that a set $\{\phi_j : j = 0, 1, \ldots, n-1\}$ of $n$ vectors in $\mathbb{C}^n$ are orthonormal:

$$\langle \phi_j, \phi_k \rangle = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{otherwise} \end{cases}$$

Then these $\{\phi_j\}$ span the vector space $\mathbb{C}^n$, and so for any vector $x$, we can write $x$ in terms of this new orthonormal basis: $x = \sum_{j=0}^{n-1} \langle \phi_j, x \rangle \phi_j$

-

$$\left\{ e_j = \frac{1}{\sqrt{n}} \left( e^{2\pi i \nu_j}, e^{2\pi i 2 \nu_j}, \ldots, e^{2\pi i n \nu_j} \right)' : j = 0, \ldots, n-1 \right\}$$

$$\langle e_j, e_k \rangle = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{otherwise} \end{cases}$$

- The vector of discrete Fourier transform coefficients $(X(\nu_0), \ldots, X(\nu_{n-1}))$ is the representation of $x$ in the Fourier basis.

$$x = \sum_{j=0}^{n-1} \langle e_j, x \rangle e_j = \sum_{j=0}^{n-1} X(\nu_j) e_j$$

- An alternative way to represent the DFT

$$\begin{aligned} X(\nu_j) = \langle e_j, x \rangle &= \frac{1}{\sqrt{n}} \sum_{t=1}^{n} e^{-2\pi i t \nu_j} x_t \\ &= \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \cos(2\pi t \nu_j) x_t - i \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \sin(2\pi t \nu_j) x_t \\ &= X_c(\nu_j) - i X_s(\nu_j) \end{aligned}$$

- The periodogram is defined as

$$\begin{aligned} I(\nu_j) &= |X(\nu_j)|^2 \\ &= \frac{1}{n}\left|\sum_{t=1}^{n} e^{-2\pi i t \nu_j} x_t\right|^2 \\ &= X_c^2(\nu_j) + X_s^2(\nu_j) \end{aligned}$$

- Orthonormality of the $e_j$ implies that we can write

$$\begin{aligned} x^* x &= \left(\sum_{j=0}^{n-1} X(\nu_j) e_j\right)^* \left(\sum_{j=0}^{n-1} X(\nu_j) e_j\right) \\ &= \sum_{j=0}^{n-1} |X(\nu_j)|^2 = \sum_{j=0}^{n-1} I(\nu_j) \end{aligned}$$

- **Discrete version of continuous version**

  $I(\nu_j)$ as the discrete version of $f(\nu_j)$

  $(1/n)\sum_{\nu_j} \cdot$ as the discrete version of $\int_\nu \cdot d\nu$

  For $\bar{x} = 0$, we can write this as $\hat{\sigma}_x^2 = \frac{1}{n}\sum_{t=1}^{n} x_t^2 = \frac{1}{n}\sum_{j=0}^{n-1} I(\nu_j)$. This is the discrete analog of the identity $\sigma_x^2 = \gamma_x(0) = \int_{-1/2}^{1/2} f_x(\nu) d\nu$

- Why is the periodogram at a Fourier frequency (that is, $\nu = \nu_j$ ) the same as computing $f(\nu)$ from the sample autocovariance?

$$\begin{aligned} I(\nu_j) &= \frac{1}{n}\left|\sum_{t=1}^{n} e^{-2\pi i t \nu_j} x_t\right|^2 = \frac{1}{n}\left|\sum_{t=1}^{n} e^{-2\pi i t \nu_j}(x_t - \bar{x})\right|^2 \\ &= \frac{1}{n}\left(\sum_{t=1}^{n} e^{-2\pi i t \nu_j}(x_t - \bar{x})\right)\left(\sum_{t=1}^{n} e^{2\pi i t \nu_j}(x_t - \bar{x})\right) \\ &= \frac{1}{n}\sum_{s,t} e^{-2\pi i (s-t)\nu_j}(x_s - \bar{x})(x_t - \bar{x}) = \sum_{h=-n+1}^{n-1} \hat{\gamma}(h) e^{-2\pi i h \nu_j} = \hat{f}(\nu_j) \end{aligned}$$

  Recall $\nu_j \neq 0$ implies $\sum_{t=1}^{n} e^{-2\pi i t \nu_j} = 0$.

- is discrete version good enough to approximate the continuous version, let us see the asymptotic behavior of the periodogram $I(\nu)$.

  **example**: Suppose that $X_1, \ldots, X_n$ are i.i.d. $N\left(0, \sigma^2\right)$ (Gaussian white noise)

$$X_c(\nu_j) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} \cos(2\pi t \nu_j) x_t, \quad X_s(\nu_j) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} \sin(2\pi t \nu_j) x_t$$

  we have that $X_c(\nu_j)$ and $X_s(\nu_j)$ are normal, with

$$\mathbf{E} X_c(\nu_j) = \mathbf{E} X_s(\nu_j) = 0$$

$$\begin{aligned} \operatorname{Var}(X_c(\nu_j)) &= \frac{\sigma^2}{n}\sum_{t=1}^{n} \cos^2(2\pi t \nu_j) \\ &= \frac{\sigma^2}{2n}\sum_{t=1}^{n} (\cos(4\pi t \nu_j) + 1) = \frac{\sigma^2}{2} \end{aligned}$$

  Similarly, $\operatorname{Var}(X_s(\nu_j)) = \sigma^2/2$

$\operatorname{Cov}\left(X_c\left(\nu_j\right), X_s\left(\nu_j\right)\right) = \operatorname{Cov}\left(X_c\left(\nu_j\right), X_c\left(\nu_k\right)\right) = \operatorname{Cov}\left(X_s\left(\nu_j\right), X_s\left(\nu_k\right)\right) = \operatorname{Cov}\left(X_c\left(\nu_j\right), X_s\left(\nu_k\right)\right) = 0$
$f(\nu_j) = \sigma^2$

$$\frac{2}{f(\nu_j)} I\left(\nu_j\right) = \frac{2}{\sigma^2} I\left(\nu_j\right) = \frac{2}{\sigma^2}\left(X_c^2\left(\nu_j\right) + X_s^2\left(\nu_j\right)\right) \sim \chi_2^2$$

So $\mathbb{E} I\left(\hat{\nu_j}\right) = f(\nu_j)$

- Generally, as $n$ increases, $\hat{\nu}^{(n)} \to \nu$ under some condition, $f\left(\hat{\nu}^{(n)}\right) \to f(\nu)$. In this case, we have

$$\frac{2}{f(\nu)} I\left(\hat{\nu}^{(n)}\right) = \frac{2}{f(\nu)}\left(X_c^2\left(\hat{\nu}^{(n)}\right) + X_s^2\left(\hat{\nu}^{(n)}\right)\right) \xrightarrow{\mathrm{d}} \chi_2^2$$

  Thus,

$$\begin{aligned}\mathbf{E} I\left(\hat{\nu}^{(n)}\right) &= \frac{f(\nu)}{2}\mathbf{E}\left(\frac{2}{f(\nu)}\left(X_c^2\left(\hat{\nu}^{(n)}\right) + X_s^2\left(\hat{\nu}^{(n)}\right)\right)\right) \\ &\to \frac{f(\nu)}{2}\mathbf{E}\left(Z_1^2 + Z_2^2\right) = f(\nu)\end{aligned}$$

  Periodogram is asymptotically unbiased.

HANSON-WRIGHT INEQUALITY

http://www-personal.umich.edu/~rudelson/papers/rv-Hanson-Wright.pdf https://sites.ualberta.ca/~omarr/publications/subgaussians.pdf

- Proposition If $X$ is b-subgaussian, and $\mathbb{E}(X) = 0$ and $\mathrm{Var}(X) \leq b^2$

  Sub gaussian decay at least as fast as gaussian.

- For a centered random variable $X$, the following statements are equivalent:

  (1) Laplace transform condition: $\exists b > 0, \quad \forall t \in \mathbb{R}, \quad \mathbb{E}e^{tX} \leq e^{b^2t^2/2}$

  (2) subgaussian tail estimate: $\exists c > 0, \quad \forall \lambda > 0, \quad \mathbb{P}(|X| \geq \lambda) \leq 2e^{-c\lambda^2}$

  (3) $\psi_2$ -condition: $\exists a > 0, \quad \mathbb{E}e^{aX^2} \leq 2$

- Proposition If $X$ is b-subgaussian, then for any $p > 0$ one has

$$\mathbb{E}|X|^p \leq p2^{\frac{p}{2}} b^p \Gamma\left(\frac{p}{2}\right)$$

  Consequently, for $p \geq 1$

$$\|X\|_{L_p} = (\mathbb{E}|X|^p)^{1/p} \leq Cb\sqrt{p}$$

  Conversely, if a centered random variable $X$ satisfies $(\mathbb{E}|X|^p)^{1/p} \leq Cb\sqrt{p}$ for all $p \geq 1$, then $X$ is subgaussian.

  **Proof:**

$$\mathbb{E}|X|^p = \int_0^\infty pt^{p-1}\mathbb{P}(|X| > t)dt \leq \int_0^\infty pt^{p-1} \cdot 2e^{-t^2/2b^2}dt$$

  using the substitution $u = t^2/2b^2$ the last integral is

$$\begin{aligned} &= p\left(2b^2\right)^{\frac{p}{2}} \int_0^\infty u^{\frac{p}{2}-1}e^{-u}du \\ &= p2^{\frac{p}{2}} b^p \Gamma\left(\tfrac{p}{2}\right) \end{aligned}$$

  In particular, using Stirling's formula($\sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n} \leq n! \leq en^{n+\frac{1}{2}}e^{-n}$, $x^{1/x} \leq e^{1/e}$) one gets $(\mathbb{E}|X|^p)^{1/p} \leq Cb\sqrt{p}$ Conversely, suppose $X$ satisfies $(\mathbb{E}|X|^p)^{1/p} \leq Cb\sqrt{p}$ for all $p \geq 1$. Then using the Taylor expansion for the exponential function and Lebesgue's Dominated Convergence Theorem, for any $a > 0$ we have

$$\begin{aligned} \mathbb{E}e^{aX^2} = \sum_{n=0}^\infty \frac{a^n\mathbb{E}\left(|X|^{2n}\right)}{n!} &= 1 + \sum_{n=1}^\infty \frac{a^n\mathbb{E}\left(|X|^{2n}\right)}{n!} \\ &\leq 1 + \sum_{n=1}^\infty \frac{a^n(Cb\sqrt{2n})^{2n}}{n!} = \sum_{n=0}^\infty \frac{a^n(Cb\sqrt{2n})^{2n}}{n!} \end{aligned}$$

  Taking $a$ small enough one gets $\mathbb{E}e^{aX^2} \leq 2$

- Theorem 1.1 (Hanson-Wright inequality). Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector with independent components $X_i$ which satisfy $\mathbb{E}X_i = 0$ and $\|X_i\|_{\psi_2} \leq \mathrm{K}$. Let A be an $n \times n$ matrix. Then, for every $t \geq 0$

$$\mathbb{P}\left\{\left|X^\top AX - \mathbb{E}X^\top AX\right| > t\right\} \leq 2\exp\left[-c\min\left(\frac{t^2}{K^4\|A\|_{\mathrm{F}}^2}, \frac{t}{K^2\|A\|_2}\right)\right]$$

  K is subgaussian parameter, e.g: $\sigma$

-

Marchenko-Pastur Law

http://www.math.wisc.edu/~valko/courses/833/2009f/lec_6_7.pdf

- Let
$$X = (\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_n) \in \mathbb{R}^{p\times n}$$
where $X_{ij}$ are iid, $E(X_{ij}) = 0, E\left(X_{ij}^2\right) = 1$ and $p = p(n)$ Define
$$S_n = \frac{1}{n} X X^T \in \mathbb{R}^{p\times p}$$
and let
$$\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_p$$
denote the eigenvalues of the matrix $S_n$
Define the random spectral measure by
$$\mu_n = \frac{1}{p}\sum_{i=1}^{p} \delta_{\lambda_i}$$

- **(Marchenko-Pastur Law)** Let$S_n, \mu_n$ be as above. Assume that $p/n \overset{n\to\infty}{\longrightarrow} y \in (0, 1]$. Then we have
$$\mu_n(\cdot, \omega) \Rightarrow \mu \quad \text{a.s}$$
where $\mu$ is a deterministic measure whose density is given by
$$\frac{d\mu}{dx} = \frac{1}{2\pi xy}\sqrt{(b-x)(x-a)}1_{(a\leq x\leq b)}$$
Here a and b are functions of y given by
$$a(y) = (1-\sqrt{y})^2, \quad b(y) = (1+\sqrt{y})^2$$

- Let $X_1, \ldots, X_n$ be independent random variables with zero means and finite absolute moments of order $p \geqq 2$. Then
$$\mathrm{E}\,|S_n|^p \leqq C_p n^{p/2-1}\sum_{k=1}^{n}\mathrm{E}\,|X_k|^p$$
https://projecteuclid.org/euclid.aoms/1177697526
Some names: Marcinkiewicz–Zygmund inequality, Khintchine inequality and Rosenthal inequalities

Ljung–Box test

- The Ljung-Box test (named for Greta M. Ljung and George E. P. Box) is a type of statistical test of whether any of a group of autocorrelations of a time series are different from zero. Instead of testing randomness at each distinct lag, it tests the "overall" randomness based on a number of lags, and is therefore a **portmanteau test**.
- A **portmanteau test** is a type of statistical hypothesis test in which the null hypothesis is well specified, but the alternative hypothesis is more loosely specified.
- The Ljung-Box test may be defined as:
  - $H_0$ : The data are independently distributed (i.e. the correlations in the population from which the sample is taken are 0, so that any observed correlations in the data result from randomness of the sampling process).
  - $H_a$ : The data are not independently distributed; they exhibit serial correlation.
- The test statistics is given by :

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k}$$

  where $n$ is the sample size, $\hat{\rho}_k$ is the sample autocorrelation at lag $k$, and $h$ is the number of lags being tested. Under $H_0$ the statistic $Q$ asymptotically follows a $\chi^2_{(h)}$. For significance level $a$, the critical region for rejection of the hypothesis of randomness is: $Q > \chi^2_{1-\alpha,h}$

- $$\alpha(X,Y) = 2 \sup_{(x,y)\in\mathbb{R}^2} |\mathbb{P}(X > x, Y > y) - \mathbb{P}(X > x)\mathbb{P}(Y > y)|$$

- Rosenblatt strong mixing coefficient

  $$\alpha(\mathcal{A},\mathcal{B}) = \sup_{B\in\mathcal{B}} \alpha\left(\mathcal{A}, \mathbb{I}_B\right) = 2\sup\left\{|\operatorname{Cov}\left(\mathbb{I}_A, \mathbb{I}_B\right)| : (A,B) \in \mathcal{A}\times\mathcal{B}\right\}$$

  This coefficient vanishes if and only if the $\sigma$ -fields are independent.

- $|\operatorname{Cov}\left(\mathbb{I}_A, \mathbb{I}_B\right)| \leq \sqrt{\operatorname{Var}\mathbb{I}_A \operatorname{Var}\mathbb{I}_B} \leq 1/4$ it follows that $0 \leq \alpha(\mathcal{A},\mathcal{B}) \leq 1/2$

- Different formulation:

  $$\alpha(\mathcal{A},\mathcal{B}) = \sup\left\{|\operatorname{Cov}\left(\mathbb{I}_A - \mathbb{I}_{A^c}, \mathbb{I}_B\right)| : (A,B) \in \mathcal{A}\times\mathcal{B}\right\}$$

  $$\operatorname{Cov}\left(\mathbb{I}_A - \mathbb{I}_{A^c}, \mathbb{I}_B\right) = \mathbb{E}\left(\left(\mathbb{P}(B \mid \mathcal{A}) - \mathbb{P}(B)\right)\left(\mathbb{I}_A - \mathbb{1}_{A^c}\right)\right)$$

  and consequently, for a fixed $B$, the maximum over $\mathcal{A}$ is reached by the measurable set $A = (\mathbb{P}(B \mid \mathcal{A}) > \mathbb{P}(B))$. Consequently $\alpha(\mathcal{A},\mathcal{B}) = \sup\{\mathbb{E}(|\mathbb{P}(B \mid \mathcal{A}) - \mathbb{P}(B)|) : B \in \mathcal{B}\}$

- In the same way, one can prove that

  $$\alpha(\mathcal{A}, X) = \sup_{x\in\mathbb{R}} \mathbb{E}(|\mathbb{P}(X \leq x \mid \mathcal{A}) - \mathbb{P}(X \leq x)|)$$

- A mixing time series can be viewed as a sequence of random variables for which the past and distant future are asymptotically *independent*.

  The idea is to define mixing coefficients to measure the strength (in different ways) of dependence for the two segments of a time series that are apart from each other in time.

-

**High dim time series** **(Report)**

# Stat308 final project

Name: Weihao LI, Netid: 12243473

**Joint Estimation of Multiple Graphical Models from High Dimensional Time Series**

## 1 Introduction

Graph is a good tool to understand and visualize the dependence relationship. As we all know, under the normality assumption, the graphical model can be represented by precision matrix, which motivates many researchers to propose different methods to estimate the the precision matrix. There are volumes of literature on the precision matrix estimation based on n independent and identical distributed normal random variables. In this paper, researcher focus on the non-identical distributed random variables, which is motivated by brain connectivity networks: since the relationship between different nodes in our brain can change as we grow up, so different age can have different graph structure (precision matrix). Also, this paper consider the temporal dependence for each subjects, that is to say, we have $n$ subjects, each subject is a time series with length $T$.

## 2 Literature review

There are three main methods to estimated a single Gaussian graphical model for high dimension case. Suppose $X_i = (X_{i1}, ..., X_{ip}) \sim N(0, \Sigma = \Omega^{-1})$, if we regress $X_1 \sim X_{-1}$ based on the condition distribution of multivariate normal, we can have $X_1|X_{-1} \sim N(-\Omega_{11}^{-1}\Omega_{1,-1}X_{-1}, \Omega_{11}^{-1})$, then Meinshausen and Bühlmann (2006) apply Lasso to estimate each column of the precision matrix and they show the consistent estimation for sparse graph structure is achieved. Second main method direct estimate the the precision matrix based on the maximum likelihood. Stanford researcher(FHT, 2008) propose to estimate $\Theta$ directly using the $\ell_1$ penalty to encourage the sparsity. But since the computation of FHT method involve the determinant of the large matrix, which is hard to compute for very large $p$, in order to achieve the large scale computation, Tony Cai use the idea of Dantzig method(CLIME) to decompose original problem to p sub-problem, which is computationally easy and we can apply parallel computing to fasten the computation. Also in low dimension, Drton and Perlman use the hypothesis testing procedure to test whether edge between i and j should be included or not, since there are $\frac{p(p-1)}{2}$ testing problem, they use Bonferroni correction to control the error conservatively. Detail of above mention method is in attached appendix.

Above method rely on the assumption that we have i.i.d samples, but this assumption is not so realistic since the model will change over time, the relaxing assumption is that subjects are independent and non-identical distributed. Guo et al consider case when subjects fall into K

different categories with each categories has there own graph structure, they impose additional penalty in graphical lasso to encourage the sparsity. Estimation of multiple Graphical models was also investigated by Zhou (2010), they consider white noise setting with different time has different covariance matrix(distribution evolves over time), they apply the kernel smoothing idea to get a weighted covariance matrix at any time point and then use graphical lasso to get the estimated graph.

# 3 Model and method

In this paper, author consider the time series setting and aim to estimate the conditional independence structure of the time series. Let $\{\boldsymbol{X}^u\}_{u\in[0,1]}$ be a series of $d$ -dimensional random vectors indexed by the label $u$, there is a natural ordering of the subjects , we can view $u$ as normalized age for each subject. For each subject $\boldsymbol{X}^{u_i}$, we have T observations $\boldsymbol{x}_{i1},\ldots,\boldsymbol{x}_{iT}\in\mathbb{R}^d$ with a temporal dependence structure among them. It assumes that $\{\boldsymbol{x}_{it}\}_{t=1}^T$ follows a lag one stationary VAR model:

$$\boldsymbol{x}_{it}=\mathbf{A}\left(u_i\right)\boldsymbol{x}_{i(t-1)}+\boldsymbol{\epsilon}_{it},\ \text{for } i=1,\ldots,n, t=2,\ldots,T$$

and $\boldsymbol{x}_{it}\sim N_d\left\{\mathbf{0},\boldsymbol{\Sigma}\left(u_i\right)\right\}$ for $t=1,\ldots,T$.

Since proposed model is motivated by brain network estimation, we can find the correspondence between statistical model and real problem to help us better understand the model. Suppose we have n people with age ($u$) from 10 to 20, for each person we get his brain image($x_{it}$) every 5 minute, it is intuitive to understand that the brain image at time t depends on that of time t-5, t-10, etc. Each image compose d parts, we want to research on the graphical relation between these d parts. For simplicity, we may assume that each image only depends on the image five minutes ago, we can collect T=200 sequential images for each person.

Given the target time $u_0$, we first estiamte the covariance $\boldsymbol{S}(u_0)$ :

$$\mathbf{S}\left(u_0\right):=\sum_{i=1}^n\omega_i\left(u_0,h\right)\widehat{\boldsymbol{\Sigma}}_i$$

where $\omega_i\left(u_0,h\right)$ is a weight function and $\widehat{\Sigma}_i$ is the sample covariance matrix of $\boldsymbol{x}_{i1},\ldots,\boldsymbol{x}_{iT}$

$$\omega_i\left(u_0,h\right):=\frac{c\left(u_0\right)}{nh}K\left(\frac{u_i-u_0}{h}\right),\quad \widehat{\boldsymbol{\Sigma}}_i:=\frac{1}{T}\sum_{t=1}^T\boldsymbol{x}_{it}\boldsymbol{x}_{it}^\top\in\mathbb{R}^{d\times d}$$

Here $c\left(u_0\right)=2I\left(u_0\in\{0,1\}\right)+I\left\{u_0\in(0,1)\right\}$ is a constant depending on whether $u_0$ is on the boundary or not, and $h$ is the bandwidth parameter. The choices of Kernel function can be Epanechnikov kernel: $K(s)=3\left(1-s^2\right)I(|s|\leq 1)/4$ for example.

After getting the estimated covariance matrix $\boldsymbol{S}(0)$, we put it into CLIME algorithm to get the estimated sparse precision matrix. i.e

$$\widehat{\Omega}\left(u_0\right)=\underset{\mathbf{M}\in\mathbb{R}^{d\times d}}{\operatorname{argmin}}\sum_{jk}\left|\mathbf{M}_{jk}\right|,\quad \text{subject to } \left\|\mathbf{S}\left(u_0\right)\mathbf{M}-\mathbf{I}_d\right\|_{\max}\leq\lambda$$

# 4 Theoretical property

Since estimating the covariance is the crucial step, we can first see whether the estimated covariance matrix is good or not.

Under some smoothness and parameter assumptions , we have

$$\max_{jk}\left|\{\mathbf{S}(u_0)\}_{jk}-\mathbf{\Sigma}(u_0)_{jk}\right|=O_P(\sqrt{\frac{\log d}{\mathrm{Tnh}}})$$

this rate can beat those methods which do not have group effect. Recall from the class, apply the Hanson Wright inequality, we can get

$$\max_{i,j}|\hat{\sigma}_{ij}-\sigma_{ij}|=O_P(99\sqrt{\frac{\log p}{n}})$$

The extra T term in the denominator characterizes the strength one can borrow across different individuals information on parameter estimation. But in my view, the sharper bound is due to increasing sample size(from n to nT).

The author also give convergence rate of i.i.d process for each subject, they compare the rate of the i.i.d case with VAR(1) case, the results is given by:

VAR(1):

$$\|\mathbf{S}(u_0)-\mathbf{\Sigma}(u_0)\|_{\max}=O_P\left[\left\{\frac{\xi\sup_{u\in[0,1]}\|\mathbf{\Sigma}(u)\|_2}{1-\sup_{u\in[0,1]}\|\mathbf{A}(u)\|_2}\sqrt{\frac{\log d}{Tn}}\right\}^{1/2}+n^{-\frac{2}{2+\eta}}\right]$$

where $\xi:=\sup_{u\in[0,1]}\frac{\max_j[\mathbf{\Sigma}(u)]_{jj}}{\min_j[\mathbf{\Sigma}(u)]_{jj}}$
i.i.d

$$\|\mathbf{S}(u_0)-\mathbf{\Sigma}(u_0)\|_{\max}=O_P\left\{\left(\frac{\log d}{Tn}\right)^{1/3}+n^{-\frac{2}{2+\eta}}\right\}$$

The rate for i.i.d case match with the rate in the paper of Zhou (2010) up to a logarithmic factor. The rate of Zhou is $O_P\left(\frac{\sqrt{\log n}}{n^{1/3}}\right)$. Remark: both rate of convergence depends on specific choice of bandwidth parameter $h$.

As we see, the rate in the i.i.d case is sharper because i.i.d assumption is very strong and unrealistic. Interestingly, we notice that the rate of convergence is negative related with the spectrum norm of the transition matrix $\mathbf{A}$. Intuitively, the spectrum norm of $\mathbf{A}$ describe the magnitude of the dependence between $\boldsymbol{x}_{it}$ and $\boldsymbol{x}_{i(t-1)}$, the higher dependence, the less information $\boldsymbol{x}_{i1},\ldots,\boldsymbol{x}_{iT}$ will give(think about the extreme case when $\mathbf{A}=\mathbf{I}$, then it is equivalent to have just one sample for subject $i$; If $\mathbf{A}=\mathbf{0}$, then it is equivalent to have T i.i.d samples for subject $i$). Instead of explaining in a intuitive way, the author give proof for two special cases of $\mathbf{A}$.

- under the case $\mathbf{A} = \text{diag}(\rho_1, \ldots, \rho_d)$

$$\|\mathbf{S}(u_0) - \mathbf{\Sigma}(u_0)\|_{\max} = O_P\left[\left\{\frac{\xi \sup_{u\in[0,1]} \|\mathbf{\Sigma}(u)\|_2}{1 - \max_{j=1,\ldots,d}(|\rho_j|)}\sqrt{\frac{\log d}{Tn}}\right\}^{1/2} + n^{-\frac{2}{2+\eta}}\right]$$

- Under the case $\mathbf{A}_{ij} = \rho I(|i-j| = 1)$

$$\|\mathbf{S}(u_0) - \mathbf{\Sigma}(u_0)\|_{\max} = O_P\left[\left\{\frac{\xi \sup_{u\in[0,1]} \|\mathbf{\Sigma}(u)\|_2}{1 - 2|\rho|\cos\{\pi/(d+1)\})}\sqrt{\frac{\log d}{Tn}}\right\}^{1/2} + n^{-\frac{2}{2+\eta}}\right]$$

For these two cases, $\rho$ is used to measure the temporal dependence between two consecutive pair, it is easy to see the magnitude of the dependence has a negative effect on the convergence rate of covariance matrix.

In order to use the CLIME to estimate the sparse precision matrix, the standard assumption is true precision matrix lying in uniformity class of matrices, i.e

$$\Theta(u_0) := \{\mathbf{\Sigma}(u_0)\}^{-1} \in \mathcal{M}(q = 0, s, M_d) := \left\{\mathbf{M} \in \mathbb{R}^{d\times d} : \mathbf{M} \succ 0, \max_{1\le k\le d}\sum_{j=1}^{d} |M_{jk}|^0 \le s, \|\mathbf{M}\|_1 \le M_d\right\}$$

Since we already get the bound for $\|\mathbf{S}(u_0) - \mathbf{\Sigma}(u_0)\|_{\max}$, according to the theorem of Tong Cai paper, we immediately get the bound : (with a specific choice of $\lambda$)

$$\left\|\widehat{\Theta}(u_0) - \Theta(u_0)\right\|_2 = O_P\left[M_d^2 s\left(\left\{\frac{\xi \sup_{u\in[0,1]} \|\Sigma(u)\|_2}{1 - \sup_{u\in[0,1]} \|\mathbf{A}(u)\|_2}\sqrt{\frac{\log d}{Tn}}\right\}^{1/2} + n^{-\frac{2}{2+\eta}}\right)\right]$$

# 5 Summary and Discussion

The author generalize the previous research with i.i.d samples to the cases where there exists a natural ordering (age) for $n$ subjects, the underlying true convariance matrix change smoothly according to this ordering (age). And for each subject, we observe a VAR(p) process with length $T$. This kind of setting help the author to better deal with the resting state functional magnetic resonance imaging (rs-fMRI) data, where there exist many natural orderings corresponding to measures of health status, demographics, and many other subject-specific covariates.

As far I am concerned, we can also generalize the i.i.d setting with the following reasonable setting: Since it is natural to assume that the covariance matrix at age t depends on the covariance matrix at age (t-1), so instead of using kernel smoothing trick to build connection between covariance matrix at different ages, we can also assume a autoregressive relation between consecutive covariance matrix. That is to say, we can use Matrix autoregressive model (MAR(1)) model. The task of this setting can be: we observe $n$ subjects, each subject is VAR(1) process with length $T$, we are going to estimate the graph structure at time $n+1$. The first step of the estimation is same with this paper: we get the sample covariance matrix $\hat{\Sigma}_1, \ldots, \hat{\Sigma}_n$. Second step: there are two way to predict the convariance matrix at time $n+1$. (Rong et al (2018))

1. Using the idea of VAR(1) model, we can vectorize the covariance matrix and assume $\text{vec}\left(\hat{\mathbf{\Sigma}}_t\right) = \mathbf{A}\,\text{vec}\left(\hat{\mathbf{\Sigma}}_{t-1}\right) + \text{vec}\left(\boldsymbol{E}_t\right)$, we can apply the Dantzig method we learn in the class to estimate the coefficient $\mathbf{A}$. Then use $\hat{\mathbf{A}}\hat{\mathbf{\Sigma}}_n$ as estimate for time $n+1$.

2. Second idea is from Rong Chen paper, they assume $\hat{\mathbf{\Sigma}}_t = \boldsymbol{A}\hat{\mathbf{\Sigma}}_{t-1}\boldsymbol{B}' + \boldsymbol{E}_t$, and we can apply the projection method they use to estimate $\mathbf{A}, \mathbf{B}$, then use $\hat{\mathbf{A}}\hat{\mathbf{\Sigma}}_n\hat{\mathbf{B}}'$ as estimate for time $n+1$.

The final step is to apply CLIME to $\hat{\mathbf{\Sigma}}_{n+1}$ and get $\hat{\mathbf{\Omega}}_{n+1}$ as the estimate for precision matrix at time $n+1$.

# 6 Reference


1. Cai, T., Liu, W., and Luo, X. (2011). A constrained L1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

2. Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

3. Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15.

4. Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

5. Rong Chen, Han Xiao, and Dan Yang. "Autoregressive models for matrix-valued time series." *arXiv preprint arXiv:1812.08916* (2018).

6. Zhou, S., Lafferty, J., and Wasserman, L. (2010). Time varying undirected graphs. *Machine Learning*, 80(2-3):295–319.


# 7 Appendix

**Relation between graph and precision matrix:**
Consider $X = (X_1, \ldots, X_p)^T \sim N(0, \Omega^{-1})$, where $\Omega \in \mathbb{R}^{p\times p}$ is a precision matrix with entries $\Omega_{jk}$. Prove $\Omega_{jk} = 0$ if and only if $X_j$ and $X_k$ are conditional independent given all other variables $\{X_l\}_{l\neq j,k}$
**Proof:**
Reference: "Graphical Models Lauriten 1995 Page 130".
Block matrix inversion:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} E^{-1} & -E^{-1}G \\ -FE^{-1} & D^{-1} + FE^{-1}G \end{pmatrix}$$

where $E = A - BD^{-1}C, F = D^{-1}C$ and $G = BD^{-1}$
Write down the covariance matrix as

$$P = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix}$$

where $P_{11}$ is covariance matrix of $(X_i, X_j)$ and $P_{22}$ is covariance matrix of $\mathbf{V}\backslash\{X_i, X_j\}$
$\mathbf{V}\backslash\{X_i, X_j\}$ means all other variables $\{X_l\}_{l\neq j,i}$
Let $\Omega = P^{-1}$. Similarly, write down $\Omega$ as

$$\Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}$$

By Block matrix inversion formula,

$$\Omega_{11}^{-1} = P_{11} - P_{12}P_{22}^{-1}P_{21}$$

We also know that $P_{11} - P_{12}P_{22}^{-1}P_{21}$ is the covariance matrix of $(X_i, X_j)\,|\mathbf{V}\backslash\{X_i, X_j\}$ from multivariate normal knowledge.
The partial correlation is therefore

$$\rho_{X_iX_j\cdot\mathbf{V}\backslash\{X_i,X_j\}} = \frac{\left[\Omega_{11}^{-1}\right]_{12}}{\sqrt{\left[\Omega_{11}^{-1}\right]_{11}\left[\Omega_{11}^{-1}\right]_{22}}}$$

Inversion formula of 2 by 2 matrix,

$$\begin{pmatrix} \left[\Omega_{11}^{-1}\right]_{11} & \left[\Omega_{11}^{-1}\right]_{12} \\ \left[\Omega_{11}^{-1}\right]_{21} & \left[\Omega_{11}^{-1}\right]_{22} \end{pmatrix} = \Omega_{11}^{-1} = \frac{1}{\det\Omega_{11}} \begin{pmatrix} [\Omega_{11}]_{22} & -[\Omega_{11}]_{12} \\ -[\Omega_{11}]_{21} & [\Omega_{11}]_{11} \end{pmatrix}$$

Therefore,

$$\rho_{X_iX_j,\mathbf{V}\backslash X_i,X_j\}} = \frac{\left[\Omega_{11}^{-1}\right]_{12}}{\sqrt{\left[\Omega_{11}^{-1}\right]_{11}\left[\Omega_{11}^{-1}\right]_{22}}} = \frac{-\frac{1}{\det\Omega_{11}}[\Omega_{11}]_{12}}{\sqrt{\frac{1}{\det\Omega_{11}}[\Omega_{11}]_{22}\frac{1}{\det\Omega_{11}}[\Omega_{11}]_{11}}} = \frac{-[\Omega_{11}]_{12}}{\sqrt{[\Omega_{11}]_{22}[\Omega_{11}]_{11}}}$$

So $\Omega_{jk} = 0$ if and only if $X_j$ and $X_k$ are conditional independent given all other variables.

(Mathias Drton and Michael D. Perlman (2007))
$\Sigma^{-1} = \{\sigma^{ij}\}$, $\rho_{ij \cdot V\setminus\{i,j\}} = \frac{-\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{jj}}}$. Gaussian graphical models can be defined by pairwise conditional independence hypotheses or equivalently by vanishing of partial correlations. Considering the $p(p-1)/2$ testing problems

$$H_{ij} : \rho_{ij \cdot V\setminus\{i,j\}} = 0 \quad \text{vs.} \quad K_{ij} : \rho_{ij \cdot V\setminus\{i,j\}} \neq 0$$

$\pi_{ij}$ be the $p$ -value of the test of hypothesis $H_{ij}$. Then the graph $\hat{G}(\alpha)$ that is selected at level $\alpha$ has the adjacency matrix $\hat{A}(\alpha) = (\hat{a}_{ij}(\alpha)) \in \mathbb{R}^{p\times p}$ with entries

$$\hat{a}_{ij}(\alpha) = \begin{cases} 1, & \text{if } \pi_{ij} \leq \alpha \\ 0, & \text{if } \pi_{ij} > \alpha \end{cases}$$

Since we use have low dimension assumption, we can invert our sample covariance matrix to get sample partial correlation $r_{ij} \cdot V\setminus\{i,j\}$, under the null hypothesis that $\rho_{ij} \cdot V\setminus\{i,j\} = 0$, $\sqrt{n-2} \cdot r_{ij}/\sqrt{1-r_{ij}^2}$ has a t-distribution with $n-2$ degrees of freedom. ($r_{ij}$ is shorthand for $r_{ij} \cdot V\setminus\{i,j\}$)

(Meinshausen and Bühlmann (2006))
$\boldsymbol{X} \in \mathbb{R}^p, \boldsymbol{X_{-1}} \in \mathbb{R}^{p-1}, \Omega_{-1,1} = \Omega_{1,-1}^T \in \mathbb{R}^{(p-1)x1}, \Omega_{-1,-1} \in \mathbb{R}^{(p-1)\times(p-1)}, \Omega^{-1} = \Sigma$, we can rewrite

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} X_1 \\ \boldsymbol{X_{-1}} \end{bmatrix} \sim N\left(0, \begin{pmatrix} \Omega_{11} & \Omega_{1,-1} \\ \Omega_{-1,1} & \Omega_{-1,-1} \end{pmatrix}^{-1}\right)$$

Apply the knowledge of block matrix inversion and conditional distribution of multivariate gaussian $\Longrightarrow X_1|\boldsymbol{X_{-1}} \sim N(-\Omega_{11}^{-1}\Omega_{1,-1}\boldsymbol{X_{-1}}, \Omega_{11}^{-1})$ Then we view $\boldsymbol{X_{-1}}$ as predicator variable in regression, $-\Omega_{11}^{-1}\Omega_{1,-1}$ as $\beta$, $\Omega_{11}^{-1}$ as $\sigma^2$, then we run the Lasso to get $\hat{\beta}, \hat{\sigma}$

(Friedman, J., Hastie, T., and Tibshirani, R. (2008)—graphical lasso)
let $\Theta = \Sigma^{-1}$, sample covariance $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$, $\|\Theta\|_1 = \sum_{jk}|\Theta_{jk}|$

$$\hat{\Theta} = \underset{\Theta\succeq 0}{argmax} \log\det\Theta - \mathrm{tr}(\hat{\Sigma}\Theta) - \rho\|\Theta\|_1$$

There are some issue on the asymmetry of the graphical lasso estimator, which can be found at https://arxiv.org/abs/1111.2667, in that discussion, they also propose some ways to do symmetrization fo the estimator. "A note on the lack of symmetry in the graphical lasso"

(Cai, T., Liu, W., and Luo, X. (2011)——-CLIME)
We want to estimate population precision matrix $\Omega_0$. CLIME estimator: Let $\hat{\Omega}_1$ be the solution of the following optimization problem:

$$\min \|\Omega\|_1 \text{ subject to: } \left|\Sigma_n \Omega - I\right|_\infty \leq \lambda_n, \quad \Omega \in \mathbb{R}^{p\times p}$$

The final CLIME estimator of $\Omega_0$ is obtained by symmetrizing $\hat{\Omega}_1$ as follows. Write $\hat{\Omega}_1 = \left(\hat{\omega}_{ij}^1\right) = \left(\hat{\omega}_1^1, \ldots, \hat{\omega}_p^1\right)$. The CLIME estimator $\hat{\Omega}$ of $\Omega_0$ is defined as

$$\hat{\Omega} = \left(\hat{\omega}_{ij}\right), \quad \text{where} \quad \hat{\omega}_{ij} = \hat{\omega}_{ji} = \hat{\omega}_{ij}^1 I\left\{\left|\hat{\omega}_{ij}^1\right| \leq \left|\hat{\omega}_{ji}^1\right|\right\} + \hat{\omega}_{ji}^1 I\left\{\left|\hat{\omega}_{ij}^1\right| > \left|\hat{\omega}_{ji}^1\right|\right\}$$

parallel computing is possible since $\hat{\boldsymbol{\Omega}}_1 = \hat{\boldsymbol{B}}$, where $\hat{\boldsymbol{B}} := \left(\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_p\right)$ and $\hat{\boldsymbol{\beta}}_i$ are solutions to

$$\min |\boldsymbol{\beta}|_1 \text{ subject to } \left|\boldsymbol{\Sigma}_n \boldsymbol{\beta} - \boldsymbol{e}_i\right|_\infty \leq \lambda_n$$

(Zhou, S., Lafferty, J., and Wasserman, L. (2010))
Let $Z^t \sim N(0, \Sigma(t))$ be independent.

$$\widehat{S}_n(t) = \frac{\sum_s w_{st} Z_s Z_s^T}{\sum_s w_{st}}$$

is a weighted covariance matrix, with weights $w_{st} = K\left(\frac{|s-t|}{h_n}\right)$ given by a symmetric non-negative function kernel over time. Then we plug $\widehat{S}_n(t)$ into graphical lasso algorithm to get the graph of time t. i.e

$$\hat{\Theta}_n(t) = \underset{\Theta \succeq 0}{\operatorname{argmax}} \log\det\Theta - \operatorname{tr}(\widehat{S}_n(t)\Theta) - \rho\|\Theta\|_1$$

How can we do better, how improve in the future, possible tools to solve that problem?

script? In order to represent such models in a way that is easy to visualize and communicate, it is natural to draw a graph with one vertex for each variable and an edge between any two variables that exhibit a desired type of dependence.

When I learn the concentration bound, there are many proof which exploits the properties of the Gaussian distribution, which motivate me to do a review on this fancy distribution. I plan to spend one day on normal distribution in order to be familiar with magic staff.

The knowledge I want to learn involve moment, pdf(Mills ratio this kind of thing), rotation invariance, concentration bound, etc

- (Stein's Lemma) Let $\mathrm{X} \sim \mathrm{n}(\theta, \sigma^2)$, and let $g$ be a differentiable function satisfying $\mathrm{E}\,|g'(X)| < \infty$. Then
$$\mathrm{E}[g(X)(X-\theta)] = \sigma^2 \mathrm{E} g'(X)$$

- (Maximal entropy) Among all distributions on $\mathbb{R}$ with mean $\mu$ variance $\sigma^2$, $N(\mu, \sigma^2)$ has the maximal entropy. Proof. For any distribution $P$ with mean $\mu$ variance $\sigma^2$
$$\begin{aligned} D(P\|\Phi) &= \int p \log p - \int p \log \phi \\ &= \int p \log p - \int p(x)[-\frac{1}{2}\log(2\pi\sigma) - \frac{(x-\mu)^2}{2\sigma^2}]dx \\ &= \int p \log p - \int \phi(x)[-\frac{1}{2}\log(2\pi\sigma) - \frac{(x-\mu)^2}{2\sigma^2}]dx \\ &= \int p \log p - \int \phi \log \phi \\ &= H(\Phi) - H(P) \end{aligned}$$
The conclusion follows the property $D(P\|\Phi) \geq 0$

- **(Mills ratio)**: Let $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ be the density function of a standard normal, we have $\phi'(z) + z\phi(z) = 0$. Then :
$$\phi(z)\left(\frac{1}{z} - \frac{1}{z^3}\right) \leq \mathbb{P}[Z \geq z] \leq \phi(z)\left(\frac{1}{z} - \frac{1}{z^3} + \frac{3}{z^5}\right) \qquad \text{for all } z > 0$$
Proof:
Using the above, we may note first that $\mathbb{P}[Z \geq z] = \int_z^\infty \phi(t)dt$, and substituting $\phi(z) = \frac{-\phi'(z)}{z}$ and using integration by parts, we get
$$\int_z^\infty \phi(t)dt = \int_z^\infty \frac{-\phi'(t)}{t}dt = \left[\frac{-\phi(t)}{t}\right]_z^\infty - \int_z^\infty \frac{\phi(t)}{t^2}dt$$
since $\lim_{t\to\infty} \frac{-\phi(t)}{t} = 0$, we get the top as $\frac{\phi(z)}{z} - \int_z^\infty \frac{\phi(t)}{t^2}dt$, where we may use the same substitution and apply integration by parts again:
$$\begin{aligned} \tfrac{\phi(z)}{z} - \textstyle\int_z^\infty \frac{-\phi'(t)}{t^3}dt &= \tfrac{\phi(z)}{z} + \left[\tfrac{\phi(t)}{t^3}\right]_z^\infty - \textstyle\int_z^\infty \frac{-3\phi(t)}{t^4}dt \\ &= \tfrac{\phi(z)}{z} - \tfrac{\phi(z)}{z^3} + \textstyle\int_z^\infty \frac{3\phi(t)}{t^4}dt \end{aligned}$$
Thus since $\int_z^\infty \frac{3\phi(t)}{t^4}dt > 0$ we get $\phi(z)\left(\frac{1}{z} - \frac{1}{z^3}\right) < \mathbb{P}[Z \geq z]$. Applying the trick again to $\int_z^\infty \frac{3\phi(t)}{t^4}dt$ yields

$$\begin{aligned}\int_z^\infty \frac{-3\phi'(t)}{t^5}dt &= \left[\frac{-3\phi(t)}{t^5}\right]_z^\infty - \int_z^\infty \frac{15\phi(t)}{t^6}dt \\ &= \frac{3\phi(z)}{z^5} - \int_z^\infty \frac{15\phi(t)}{t^6}dt\end{aligned}$$

and since $-\int_z^\infty \frac{15\phi(t)}{t^6}dt < 0$, we get $\mathbb{P}[Z \geq z] < \phi(z)\left(\frac{1}{z} - \frac{1}{z^3} + \frac{3}{z^5}\right)$, which proves the claim.

- **(Maxwell's theorem)**: Let $\boldsymbol{Z} \in \mathbb{R}^n$ be a random vector for which (i) projections into orthogonal subspaces are independent and (ii) the distribution of $\boldsymbol{Z}$ depends only on the length $\|\boldsymbol{Z}\|$. Then $\boldsymbol{Z}$ is normally distributed.
- **(Rotation invariance)**: For standard normal distribution: $Z_i$ are independent standard normal random variables. The joint density of $Z_1, \ldots, Z_n$ is

  $$f(z_1, \ldots, z_n) = (2\pi)^{-n/2} e^{-\left(z_1^2 + \ldots + z_n^2\right)/2} = (2\pi)^{-n/2} e^{-\|\mathbf{z}\|^2/2}$$

  which is rotationally invariant, i.e. invariant under rotations of $n$ -dimensional space, because it only depends on the length of the vector $\mathbf{z} = (z_1, \ldots, z_n)$, and determinant of the Jocabian matrix is **one**.

Stability

- **Theorem** If $X_1, X_2$ are two i. i. d. random variables such that $X_1$ and $(X_1 + X_2)/\sqrt{2}$ have the same distribution, then $X_1$ is normal.
- **Corollary** Suppose $X_1, X_2$ are i. i. *d.* random variables with finite second moments and such that for some scale factor $\kappa$ and some location parameter $\alpha$ the distribution of $X_1 + X_2$ is the same as the distribution of $\kappa(X_1 + \alpha)$. Then $X_1$ is normal.