Estimating covariance matrices of intermediate size

Martin T. Wells

Department of Statistics and Data Science
Cornell University

Joint work with Didier Chételat

The Blessing of Dimensionality: Theory and Applications
July 17, 2024

Financial portfolio optimization

- You have \$100,000 to invest in the $p = 200$ largest stocks on the NASDAQ.
- You have about 2 years of mostly uncorrelated, identically distributed daily returns, about $n = 500$ observations.

- How much money should you put in each stock to get, say, a 5% return on investment with as little risk as possible?

- Researchers in portfolio management know that the mean-variance portfolio optimization suggests the optimal weights $w_1, ..., w_{200}$ should solve the problem

$$\begin{array}{rl} \text{Minimize} & w^t \Sigma w \\ \text{such that} & w \geq 0, \\ & w^t 1 = 100{,}000, \\ & w^t \mu = 5{,}000. \end{array}$$

where μ and Σ are respectively the mean vector and covariance matrix of the asset returns.

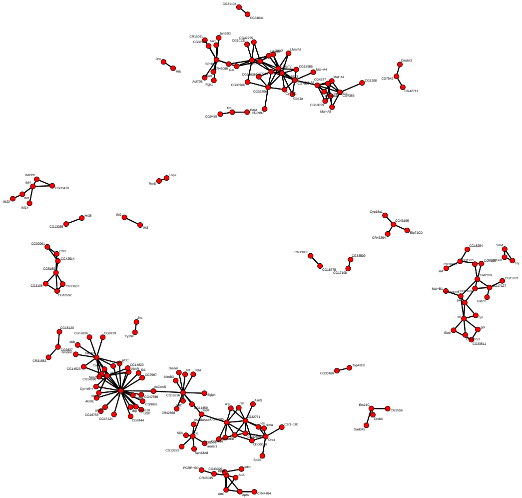
Financial portfolio optimization

- Given μ and Σ, this is quadratic programming (QP).
- Estimating μ is easy, even for many assets.
- However, when the number of assets p ($p = 200$) is large, estimating Σ quickly becomes unrealistic.
  $\Rightarrow \quad p(p+1)/2 = 20{,}100$ free parameters to estimate from $n = 500$ observations!
- Problem: how can we estimate Σ for large p?

Regulatory network with RNA-seq: *Drosophila* immune response

- After a microbial infection, Drosophila launch rapid and efficient immune responses that are crucial to survival.
- Generated a full transcriptional profile of gene expression dynamics in *Drosophila melanogaster* after immune challenge, we injected adult male flies with commercial lipopolysaccharide.
- Flies were sampled for a total of 21 time points throughout the course of five days, which includes an uninfected uninjected sample as control at time zero, and 20 time points after infection.
- After normalization, only genes with more than 5 counts in at least 2 samples were kept, leaving $12,657$ genes for further analysis.

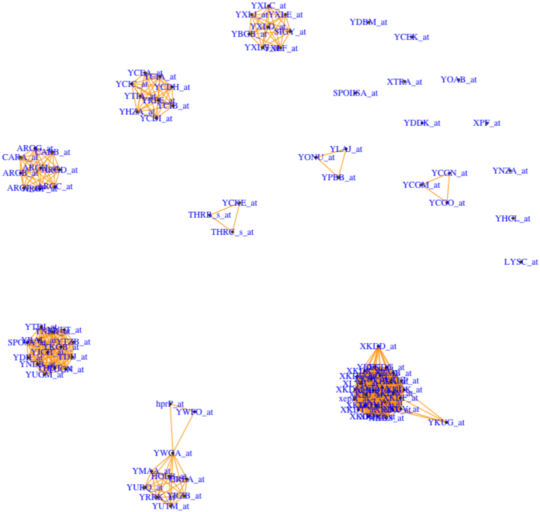
Drosophila immune response



Typical 'omics regression – Riboflavin data

- Y_i is the logarithm of riboflavin (vitamin B12) production rate, $i = 1, \ldots, 71$.
- X_k are normalized expression levels of $4,088$ genes.
- Not only $p \gg n$, but also out of 8,353,828 pairs of genes, there are 70,349 with correlation coefficient greater than 0.8 (in absolute value).
- Other examples include, Quantitative Trait Loci (**QTL**): Thousands, or even 10s of thousands, of genetic markers (predictors). The number of subjects is much smaller ($n < 1000$) or larger ($n > 500,000$) .

Riboflavin data



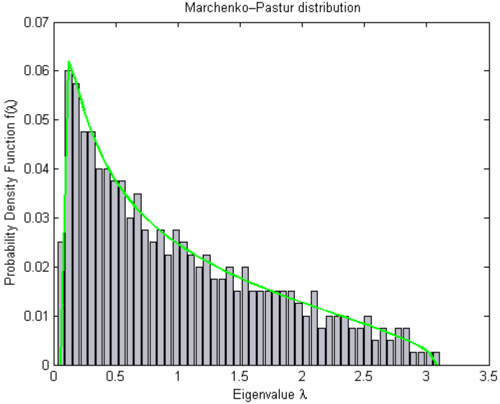
- The natural covariance estimator is the **sample covariance matrix** $S_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^t$. Many nice theoretical properties (MVUE and MLE for normal data, etc. in nice "classical" settings).

- Theoretical justification: say the underlying data is normal and that $\Sigma_p = I_p$, so that $S_n \sim \mathrm{W}_p(n, I_p/n)$ (Wishart distribution). If the eigenvalues of S_n were consistent, we would have $\lambda_1(S_n), \ldots, \lambda_p(S_n) \to 1$.

Covariance estimation

- Also need covariance matrix or its inverse (the precision matrix) estimates in
  - principal components analysis (PCA)
  - linear discriminant analysis (LDA)
  - graphical models
  - factor analysis
  - multidimensional scaling
  - canonical correlation analysis
  - multivariate analysis of variance (MANOVA)
  - multivariate regression analysis
  - discriminant analysis
  - tests of equality of covariance matrices
  - random graph theory ...
- However, the behavior of the sample covariance is **terrible in practice** when p is large! It is know that S_n doesn't estimate the true eigenvalues of Σ well when p is large.

Marchenko and Pastur (1967) showed in that when both $p, n \to \infty$ such that $p/n \to c \in (0, \infty)$, the eigenvalues converged instead to the "Marchenko-Pastur distribution" with parameter c.



Marchenko-Pastur density:

$$f_{\mathrm{MP}(c)}(\lambda) = \frac{\sqrt{(c_+ - \lambda)(\lambda - c_-)}}{2\pi c \lambda}$$
for $\lambda \in [c_-, c_+]$,

where $c_\pm = (1 \pm \sqrt{c})^2$.

High-dimensional covariance estimation

- So S_n is not consistent when $p \approx n$, even in the simplest case possible (normal data, $\Sigma_p = I_p$.)
- But we know its asymptotic behavior $\Rightarrow$ so we can try to "correct" S_n in order to improve its behavior!
- This has motivated much research in the past decades to find covariance estimators that behave well in the **high-dimensional regime** where $p/n \to (0, \infty)$.
- For the most part this line of research has been quite successful! See Johnstone (2006, 2008, ...) and Ledoit and Wolf (2012).

High-dimensional covariance estimation

- So S_n is not consistent when $p \approx n$, even in the simplest case possible (normal data, $\Sigma_p = I_p$.)
- But we know its asymptotic behavior $\Rightarrow$ so we can try to "correct" S_n in order to improve its behavior!
- This has motivated much research in the past decades to find covariance estimators that behave well in the **high-dimensional regime** where $p/n \to (0, \infty)$.
- For the most part this line of research has been quite successful! See Johnstone (2006, 2008, ...) and Ledoit and Wolf (2012).

High-dimensional covariance estimation

Question: What happens when $p \to \infty$ but $p/n \to 0$?

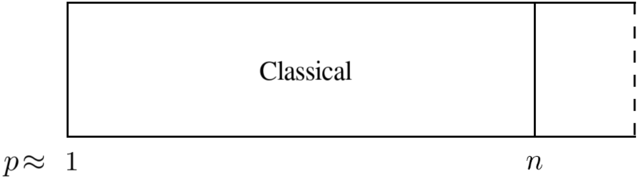
- We know that when $p/n \to c \in (0, \infty)$, the eigenvalues of a $S_n \sim \mathrm{W}_p(n, I_p/n)$ tend to a Marchenko-Pastur distribution.
- As $c \to 0$, this distribution concentrates all probability mass at $\lambda = 1$ (the right value).
- This suggests that the eigenvalues will be **consistent**, just like when p is fixed!

- Binary view: behavior of sample covariance matrix is classical when $p \to \infty$ until $p \approx n$, where there is a **regime change**.

Question: What happens when $p \to \infty$ but $p/n \to 0$?

- We know that when $p/n \to c \in (0, \infty)$, the eigenvalues of a $S_n \sim W_p(n, I_p/n)$ tend to a Marchenko-Pastur distribution.
- As $c \to 0$, this distribution concentrates all probability mass at $\lambda = 1$ (the right value).
- This suggests that the eigenvalues will be **consistent**, just like when p is fixed!

- Binary view: behavior of sample covariance matrix is classical when $p \to \infty$ until $p \approx n$, where there is a **regime change**.



Intermediate regimes

- A real symmetric matrix follows the Gaussian orthogonal ensemble GOE(p) distribution if X_{kl}, $k \leq l$ are all independent, with diagonal elements $X_{kk} \sim \mathrm{N}(0,2)$ and off-diagonal elements $X_{kl} \sim \mathrm{N}(0,1)$.

$$\mathrm{GOE}(p) = \begin{bmatrix} \mathrm{N}(0,2) & \mathrm{N}(0,1) & \dots & \mathrm{N}(0,1) \\ \mathrm{N}(0,1) & \mathrm{N}(0,2) & \dots & \mathrm{N}(0,1) \\ & & \ddots & \\ \mathrm{N}(0,1) & \mathrm{N}(0,1) & \dots & \mathrm{N}(0,2) \end{bmatrix} \quad \text{(symmetric)}$$

- When p is fixed, multivariate central limit theorem shows that

$$\sqrt{n}\left(\mathrm{W}_p(n, I_p/n) - I_p\right) \Rightarrow \mathbf{GOE(p)}$$

as $n \to \infty$. That is, the sample covariance matrix is "asymptotically normal."

Intermediate regimes

- A real symmetric matrix follows the Gaussian orthogonal ensemble GOE(p) distribution if X_{kl}, $k \leq l$ are all independent, with diagonal elements $X_{kk} \sim \mathrm{N}(0,2)$ and off-diagonal elements $X_{kl} \sim \mathrm{N}(0,1)$.

$$\mathrm{GOE}(p) = \begin{bmatrix} \mathrm{N}(0,2) & \mathrm{N}(0,1) & \dots & \mathrm{N}(0,1) \\ \mathrm{N}(0,1) & \mathrm{N}(0,2) & \dots & \mathrm{N}(0,1) \\ & & \ddots & \\ \mathrm{N}(0,1) & \mathrm{N}(0,1) & \dots & \mathrm{N}(0,2) \end{bmatrix} \quad \text{(symmetric)}$$

- When p is fixed, multivariate central limit theorem shows that

$$\sqrt{n}\big(\mathrm{W}_p(n, I_p/n) - I_p\big) \Rightarrow \mathbf{GOE(p)}$$

as $n \to \infty$. That is, the sample covariance matrix is "asymptotically normal."

Intermediate regimes?

- Does this still hold when $p \to \infty$ but $p/n \to 0$?

- One needs to be careful! Bubeck and Ganguly (2015) showed that as $n, p \to \infty$,

$$d_{\mathrm{TV}}\Big(\sqrt{n}(\mathrm{W}_p(n, I_p/n) - I_p), \mathrm{GOE}(p)\Big) \to 0$$

iff $p^3/n \to 0$.

(d_{TV} is the total variation distance $\int |f_X - f_Y| dx$.)

- What is going on when $\sqrt[3]{n} \lesssim p \lesssim n$?

Intermediate regimes?

- Does this still hold when $p \to \infty$ but $p/n \to 0$?

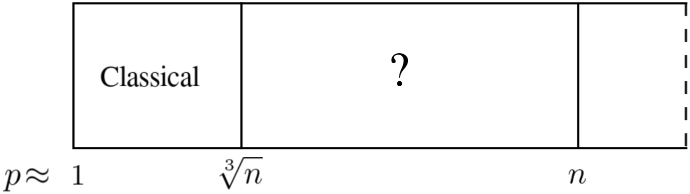
- One needs to be careful! Bubeck and Ganguly (2015) showed that as $n, p \to \infty$,

$$d_{\mathrm{TV}}\Big(\sqrt{n}(\mathrm{W}_p(n, I_p/n) - I_p), \mathrm{GOE}(p)\Big) \to 0$$

  iff $p^3/n \to 0$.

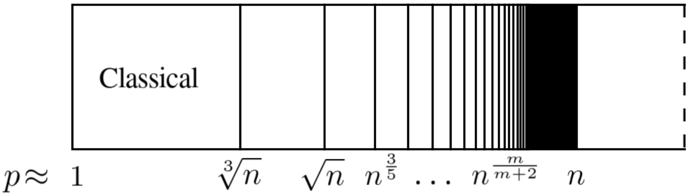
  (d_{TV} is the total variation distance $\int |f_X - f_Y| dx$.)

- What is going on when $\sqrt[3]{n} \lesssim p \lesssim n$?



Intermediate regimes!

- There may be an infinite countable number of “intermediate“ regimes.
- SPOILER ALERT We show that regime changes occur exactly at $p \approx n^{\frac{m}{m+2}}$ for $m = 1, 2, 3, \ldots$.



The $\mathcal{F}$-conjugate

- The space of $p \times p$ symmetric matrices $\mathbb{S}_p(\mathbb{R})$ can be assimilated to $\mathbb{R}^{p(p+1)/2}$ by mapping a symmetric matrix to its upper triangle. By integration over $\mathbb{S}_p(\mathbb{R})$, we mean integration with respect to the pullback Lebesgue measure under this isomorphism, that is,

$$\int_{\mathbb{S}_p(\mathbb{R})} f(X)\, dX = \int_{\mathbb{R}^{p(p+1)/2}} f(X) \prod_{i \leq j}^{p} dX_{ij}.$$

- For an integrable function f on $\mathbb{S}_p(\mathbb{R})$, the Fourier transform with kernel $\exp\{-i\,\mathrm{tr}(XT)\}$, normalized to be an L^2-isometry, satisfies

$$\mathcal{F}\{f\}(T) = 2^{-\frac{p}{2}} \pi^{-\frac{p(p+1)}{4}} \int_{\mathbb{S}_p(\mathbb{R})} e^{-i\,\mathrm{tr}(XT)} f(X) dX.$$

The $\mathcal{F}$-conjugate

- Properties of $\psi = \mathcal{F}\{f^{1/2}\}^2$.
  1. ψ is a function $\mathbb{R} \to \mathbb{C}$.
  2. By Parseval's theorem, modulus is itself a density!

$$\int |\psi(t)|dt = \int |\psi^{\frac{1}{2}}(t)|^2 dt = \int |f^{\frac{1}{2}}(x)|^2 dx = \int |f(x)|dx = 1.$$

  $\approx$ wavefunction in quantum mechanics...

  3. By the Plancherel theorem, we can express the Hellinger distance in terms of G-transforms!

$$\mathrm{H}^2(f_1, f_2) = \int |f_1^{\frac{1}{2}} - f_2^{\frac{1}{2}}|^2 dx = \int |\psi_1^{\frac{1}{2}} - \psi_2^{\frac{1}{2}}|^2 dt.$$

  (Note that $H^2(f_1, f_2) \to 0$ iff $d_{\mathrm{TV}}(f_1, f_2) \to 0$.)

The $\mathcal{F}$-conjugate

- Properties of $\psi = \mathcal{F}\{f^{1/2}\}^2$.
  1. ψ is a function $\mathbb{R} \to \mathbb{C}$.
  2. By Parseval's theorem, modulus is itself a density!

$$\int |\psi(t)| dt = \int |\psi^{\frac{1}{2}}(t)|^2 dt = \int |f^{\frac{1}{2}}(x)|^2 dx = \int |f(x)| dx = 1.$$

     $\approx$ wavefunction in quantum mechanics...
  3. By the Plancherel theorem, we can express the Hellinger distance in terms of G-transforms!

$$\mathrm{H}^2(f_1, f_2) = \int |f_1^{\frac{1}{2}} - f_2^{\frac{1}{2}}|^2 dx = \int |\psi_1^{\frac{1}{2}} - \psi_2^{\frac{1}{2}}|^2 dt.$$

     (Note that $H^2(f_1, f_2) \to 0$ iff $d_{\mathrm{TV}}(f_1, f_2) \to 0$.)

The $\mathcal{F}$-conjugate

Definition

The $\mathcal{F}$-conjugate of an absolutely continuous distribution F with density f on an Euclidean space is the distribution F^* with density $|\mathcal{F}\{f^{1/2}\}|^2$.

- Can't really see intermediate regimes in the density of $\mathrm{NW}(n,p) = \sqrt{n}(\mathrm{W}_p(n, I_p/n) - I_p)$. We will see them in the characteristic function!
- But it is difficult to relate total variation distance to characteristic functions.
- The $\mathcal{F}$-conjugate is a **middle ground**, similar to characteristic functions, but lets you control the Hellinger distance.

The $\mathcal{F}$-conjugates

- The $\mathcal{F}$-conjugate density, $f_{\text{GOE}^*}(T)$, of the GOE is given by

$$\left|\mathcal{F}\left\{\frac{\exp\{-\frac{1}{8}\sum_{i,j=1}^{p}X_{ij}^2\}}{2^{p(p+3)/8}\pi^{p(p+1)/8}}\right\}\right|^2(T) = \frac{2^{p(3p+1)/4}}{\pi^{p(p+1)/4}}\exp\left\{-4\sum_{i,j=1}^{p}T_{ij}^2\right\},$$

- When $n \geq p-2$, the $\mathcal{F}$-conjugate of a normalized Wishart distribution, $\sqrt{n}[\mathrm{W}_p(n, I_p/n) - I_p]$, has a density on $\mathbb{S}_p(\mathbb{R})$ given by

$$f_{\text{NW}^*}(T) = \frac{2^{\frac{p(n+2p)}{2}}}{\pi^{\frac{p(p+1)}{2}}n^{\frac{p(p+1)}{4}}}\frac{\Gamma_p^2\left(\frac{n+p+1}{4}\right)}{\Gamma_p\left(\frac{n}{2}\right)}\left|I_p + \frac{16\,T^2}{n}\right|^{-\frac{n+p+1}{4}}.$$

- When $p = 1$, this is the $t_{n/2}/\sqrt{8}$ distribution, so it would be natural to interpret this distribution as the parametrization of some generalization of the t-distribution to real-valued symmetric matrices.

The $\mathcal{F}$-conjugate

Definition (Symmetric matrix variate t-distribution)

We say a real symmetric $p \times p$ matrix T has the symmetric matrix variate t-distribution with $\nu \geq p/2 - 1$ degrees of freedom and $p \times p$ positive-definite scale matrix Ω, denoted Sym-$t_\nu(\Omega)$, if it has density

$$f_{T_n(\Omega)}(T) = \frac{2^{p(\nu-1)}\Gamma_p^2\left(\frac{\nu+(p+1)/2}{2}\right)}{\pi^{\frac{p(p+1)}{2}}\nu^{\frac{p(p+1)}{4}}\Gamma_p(\nu)}|\Omega|^{-\frac{p+1}{4}}\left|I_p + \frac{T\Omega^{-1}T}{\nu}\right|^{-\frac{\nu+(p+1)/2}{2}}.$$

- With this definition, the $\mathcal{F}$-conjugate of the normalized Wishart distribution is the Sym-$t_{n/2}(I_p/8)$ distribution. The fact that this is indeed a density with $n = \nu/2$ degrees of freedom.
- Most of the subsequent results are a consequence of the asymptotic behavior of the Sym-t distribution.

The $\mathcal{F}$-conjugate

- By the definition of the Kullback-Leibler divergence,

$$\mathrm{d}_{\mathrm{KL}}\Big(\mathrm{GOE}(p)^* \;\|\; \sqrt{n}\big[W_p(n, I_p/n) - I_p\big]^*\Big)$$
$$= \mathsf{E}\left[\log C_{\mathrm{GOE}/4} - 4\,\mathrm{tr}\,T^2 - \log C_t + \frac{n+p+1}{4}\log\left|I_p + \frac{16\,T^2}{n}\right|\right]$$

 where $C_{\mathrm{GOE}/4}$ and C_t are the normalization constants of the $\mathrm{GOE}(p)/4$ and the Sym-$t_{n/2}(I_p/8)$ distributions, respectively.

- Therefore we need to examine the asymptotic properties of $\log C_{\mathrm{GOE}} - \log C_t$ and $\mathsf{E}\left[-4\,\mathrm{tr}\,T^2 + \frac{n+p+1}{4}\log\left|I_p + \frac{16\,T^2}{n}\right|\right]$ as functions of n and p as both $\to \infty$.

Asymptotic behavior of the normalization constants

- We start by studying the asymptotic behavior of its normalization constants $C_{\mathrm{GOE}/4}$ and C_t .

Lemma (Behavior of the normalization constants)

For every $K \in \mathbb{N}$, there exist constants a_k, b_k such that

$$\log C_t = \log C_{GOE/4} + \sum_{k=1}^{K} a_k \frac{p^{k+2}}{n^k} + \sum_{k=1}^{K} b_k \frac{p^{k+1}}{n^k} + O\Big(\frac{p^{K+3}}{n^{K+1}}\Big)$$

as n, $p \to \infty$, where the symbol $C_{GOE/4}$ stands for the normalization constant of $GOE(p)/4$ distribution.

Asymptotic behavior of the moments

- For any integer L and any real x, we have the inequality $\left|-\frac{1}{2}\log(1+x^2) - \sum_{l=1}^{L}(-1)^l x^{2l}/2l\right| \leq x^{2L+2}/(2L+2)$.

- Thus we have the bound

$$\left|\mathrm{E}\left[-\frac{n+p+1}{4}\log\left|I_p+\frac{X^2}{n}\right|\right]-\mathrm{E}\left[\frac{n+p+1}{2}\sum_{k=1}^{K+1}(-1)^k\frac{\operatorname{tr}X^{2k}}{2kn^k}\right]\right|$$

$$\leq \frac{n+p+1}{2}\frac{\mathrm{E}\left[\operatorname{tr}X^{2(K+2)}\right]}{2(K+2)n^{K+2}} \qquad = O\left(\frac{p^{K+3}}{n^{K+1}}\right)$$

as $n, p \to \infty$, since $\mathrm{E}\left[\operatorname{tr}X^{2(K+2)}\right] = O(p^{K+3})$ as $p \to \infty$.

- Next, we study the empirical moments of the Sym-$t_{n/2}(I_p/8)$ distribution. For any integer partition $\kappa = (\kappa_1, \ldots, \kappa_q)$ in decreasing order $\kappa_1 \geq \cdots \geq \kappa_q > 0$, define its associated power sum polynomial to be

$$r_\kappa(Z) = \prod_{i=1}^{q} \operatorname{tr} Z^{\kappa_i}.$$

The norm of the partition κ is $|\kappa| = \kappa_1 + \cdots + \kappa_q > 0$, which should not be confused with its length $q(\kappa) = q$ (number of elements).

Lemma (Behavior of the the moments)

Let $T \sim \text{Sym-}t_{n/2}(I_p/8)$. *Then for any* $k \in \mathbb{N}$, *whenever* n *is large enough so that* $n \geq p + 8k + 6$, *the* $2k^{th}$ *moment of* T *can be written*

$$\mathrm{E}\left[\operatorname{tr} T^{2k}\right] = \frac{(-1)^k}{n^k} \sum_{|\kappa| \leq 2k} b_\kappa(n, m, p)\, \mathrm{E}\left[r_\kappa(Y^{-1})\right]$$

for a $Y^{-1} \sim W_p^{-1}(n, I_p/n)$ *and some polynomials* b_κ *in* n, m, p, *indexed by integer partitions* κ, *whose degrees satisfy* $\deg b_\kappa \leq 2k + 1 - q(\kappa)$. *The sums are taken over all partitions of the integers* κ *satisfying* $|\kappa| \leq 2k$, *including the empty partition.*

Asymptotic behavior of the moments

The next step is to compute expected power sum polynomials of an inverse Wishart. (Letac and Massam, 2004).

- For any integer partition κ, there exist coefficients $c_{\kappa,\lambda}$ (which depend solely on κ and λ) such that

$$r_\kappa(Y^{-1}) = \sum_{|\lambda|=|\kappa|} c_{\kappa,\lambda} C_\lambda(Y^{-1}),$$

  for C_λ the so-called zonal polynomials.

- The expected zonal polynomials for $Y^{-1} \sim \mathrm{W}_p^{-1}(n, I_p/n)$ are

$$\begin{aligned} \mathrm{E}\big[C_\lambda(Y^{-1})\big] &= \frac{n^{|\lambda|}}{2^{|\lambda|}\prod_{i=1}^{q(\lambda)}\frac{m-i+1}{2}} C_\lambda(I_p) \\ &= \frac{2^{|\lambda|}|\lambda|!\prod_{i<j}^{q(\lambda)}(2\lambda_i - 2\lambda_j - i + j)}{\prod_{i=1}^{q(\lambda)}(2\lambda_i + q(\lambda) - i)!} n^{|\lambda|} \prod_{i=1}^{q(\lambda)} \prod_{l=0}^{\lambda_i - 1} \frac{p + (1 - i + 2l)}{m - (1 - i + 2l)}. \end{aligned}$$

- From this, we can exactly compute $\mathrm{E}\big[r_\kappa(Y^{-1})\big]$ and thus $\mathrm{E}\big[\mathrm{tr}\, T^{2k}\big]$, as a function of p and n.

K-L distance for $\mathcal{F}$-conjugates

Theorem (K-L distance for $\mathcal{F}$-conjugates)

For any $K \in \mathbb{N}$, there exists constants a_k, b_k such that we have an asymptotic expansion

$$\mathrm{d}_{\mathrm{KL}}\Big(\mathrm{GOE}(p)^* \,\Big\|\, \sqrt{n}\big[W_p(n, I_p/n) - I_p\big]^*\Big) = \sum_{k=1}^{K} a_k \frac{p^{k+2}}{n^k} + \sum_{\substack{k=1 \\ k \text{ even}}}^{K} b_k \sqrt{\frac{p^{k+2}}{n^k}} + O\left(\frac{p^{K+3}}{n^{K+1}}\right)$$

as $p, n \to \infty$.

Note that $K = 1$ is the result of Bubeck and Ganguly (2015).

- The middle-scale regimes of higher degree correspond to previously unknown behavior. This raises the question whether we can find approximations of the normalized Wishart density for such middle-scale regimes when $K > 0$?

Definition (Middle-scale densities)

For $n \geq 3p - 3$ and any $K \in \mathbb{N}$, we define F_K as the distribution on the space of real symmetric matrices with density $f_K(X)$

$$\propto \left| \mathrm{E}\left[\exp\left\{ \frac{i \operatorname{tr}(XZ)}{\sqrt{8}} - \frac{n}{4} \sum_{k=3}^{2K_1} \frac{i^k}{k} \operatorname{tr}\left(\frac{\sqrt{2}Z}{\sqrt{n}}\right)^k + \frac{p+1}{4} \sum_{k=1}^{2K_2} \frac{i^k}{k} \operatorname{tr}\left(\frac{\sqrt{2}Z}{\sqrt{n}}\right)^k \right\} \right] \right|^2$$

for $Z \sim \mathrm{GOE}(p)$ and $K_1 = K + 1 + \mathbb{1}[K\,\mathrm{odd}]$, $K_2 = K + \mathbb{1}[K\,\mathrm{even}]$.

Convergence of middle-scale densities

- Since $\mathrm{E}\left[\exp\{i \operatorname{tr}(XZ)/\sqrt{8}\}\right]^2 = \exp\{-\operatorname{tr} X^2/4\}$, f_0 is the Gaussian orthogonal ensemble density. The theorem below provides as a special case an independent proof of the classical GOE approximation when $p^3/n \to 0$.

Theorem (Middle-scale densities)

For any $K \in \mathbb{N}$, the distribution F_K is well-defined whenever $n \geq 3p - 3$. Moreover, the total variation distance between the normalized Wishart distribution $\sqrt{n}[W_p(n, I_p/n) - I_p]$ and F_K satisfies

$$\mathrm{d}_{\mathrm{TV}}\left(\sqrt{n}\left[W_p(n, I_p/n) - I_p\right],\, F_K\right) \to 0$$

as $n \to \infty$ with $p^{K+3}/n^{K+1} \to 0$.

Why is this useful?

- Since intermediate regimes for $\mathrm{W}_p(n, \mathbf{I_p}/n)$ exist, they will also exist for $S_n \sim \mathrm{W}_p(n, \mathbf{\Sigma_p}/n)$ in general.
- In most recent research on covariance estimation, when p large, folks automatically use tools (e.g. Ledoit-Wolf) that correct S_n for high-dimensional regime ($p \approx n$) asymptotics.
- One should likely derive estimators for intermediate regimes ($p \approx n^{\frac{m+2}{m}}$) instead. One would get **better estimation/inference/prediction** when p is large but not as large as n, e.g. introductory example ($p = 200$, $n = 500$).

High-dimensional covariance estimation

- Asymptotic equivalence (via LeCam theory) gives interesting consequences about the "difficulty" of estimation:

  1. Optimal **minimax rates** of the two asymptotically equivalent estimation problems must be the **same**.

  2. If we put a prior on the parameter, **Bayes risks** of the two estimation problems must asymptotically be the **same**.

- One can **transfer computation** of optimal minimax rate of a complicated problem to a simpler setting.

High-dimensional covariance estimation

- Asymptotic equivalence (via LeCam theory) gives interesting consequences about the "difficulty" of estimation:

  1. Optimal **minimax rates** of the two asymptotically equivalent estimation problems must be the **same**.
  2. If we put a prior on the parameter, **Bayes risks** of the two estimation problems must asymptotically be the **same**.

- One can **transfer computation** of optimal minimax rate of a complicated problem to a simpler setting.

High-dimensional covariance estimation

- By using our result that
$d_{\mathrm{TV}}\Big(\sqrt{n}(\mathrm{W}_p(n, I_p/n) - I_p), \mathrm{GOE}(p)\Big) \to 0$ when $p^3/n \to 0$,
if follows that

$$\begin{aligned} \text{Estimating } \Sigma_p \text{ from a sample} \quad & X_1, ..., X_n \sim \mathrm{N}_p(0, \Sigma_p) \\ \text{and estimating } \Sigma_p \text{ from a single} \quad & Y \sim \log \Sigma_p + \frac{1}{n}\mathrm{GOE}(p) \end{aligned}$$

are **asymptotically equivalent** as $p^3/n \to 0$ (under some conditions on the eigen-structure of Σ_p).

- The second problem is much simpler to study since the covariance problem is reduced to estimating a **mean** from a normal distribution using a sample of size 1!

High-dimensional covariance estimation

- Could we prove analogue for the **other regimes**? That is for

$$\text{Estimating } \Sigma_p \text{ from a sample} \quad X_1, ..., X_n \sim \mathrm{N}_p(0, \Sigma_p),$$
$$\text{and estimating } \Sigma_p \text{ from a single} \quad Y \sim \log \Sigma_p + \frac{1}{n} G,$$

where $G \sim f_m$ are asymptotically equivalent as $p^{m+2}/n^m \to 0$, under growth conditions on Σ_p.

- Proving that some estimator is optimal would be then **much easier**, since computing the optimal minimax rate would be reduced to finding the analogue for the second estimation problem.

High-dimensional covariance estimation

- Could we prove analogue for the **other regimes**? That is for

$$\text{Estimating } \Sigma_p \text{ from a sample} \qquad X_1, ..., X_n \sim \mathrm{N}_p(0, \Sigma_p),$$
$$\text{and estimating } \Sigma_p \text{ from a single} \qquad Y \sim \log \Sigma_p + \frac{1}{n} G,$$

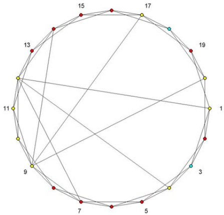
  where $G \sim f_m$ are asymptotically equivalent as $p^{m+2}/n^m \to 0$, under growth conditions on Σ_p.

- Proving that some estimator is optimal would be then **much easier**, since computing the optimal minimax rate would be reduced to finding the analogue for the second estimation problem.

Random graph theory connections

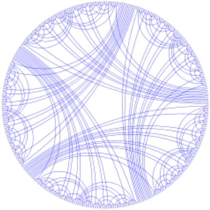
Beyond covariance estimation: random graph theory and networks.

- For vertices $\{1, \ldots, p\}$, **Erdős-Rényi random graph** with probability q, $G(p, q)$, is generated by drawing an $p(p+1)/2$ sample $U_{12}, \ldots, U_{p-1,p}$ $\sim \mathrm{U}(0,1)$ and connecting i and j if $U_{ij} > q$.



Random graph theory connections

- For vertices $\{1, ..., p\}$, **random geometric graph** on $\mathbb{S}^{n-1}$ with probability q, $G(p, q, n)$, is generated by drawing an p sample $U_1, ..., U_p \overset{\text{i.i.d.}}{\sim}$ $\text{Unif}(\mathbb{S}^{n-1})$ and connecting vertices i and j if $\langle U_i, U_j \rangle \geq t$, where t is such that $\text{P}[\langle U_1, U_2 \rangle \geq t] = q$.



Random graph theory connections

- It turns out the GOE(p) and the $W_p(n, I_p/n)$ distributions are closely related to these models.
- Let $X \sim \text{GOE}(p)$, and let Φ be the standard normal cdf. The thresholded matrix

$$A_{ij} = \begin{cases} \mathbb{1}\left[X_{ij} > \Phi^{-1}(q)\right] & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

is the **adjacency matrix** of an Erdős-Rényi graph $G(p, q)$.

- Let $Y \sim W_p(n, I_p/n)$, and let Ψ_n be the cdf of the r.v. $\langle U_1, U_1 \rangle$ for U_1, U_2 i.i.d. Unif($\mathbb{S}^{n-1}$). The thresholded matrix

$$B_{ij} = \begin{cases} \mathbb{1}\left[\frac{Y_{ij}}{\sqrt{Y_{ii}Y_{jj}}} > \Psi_n^{-1}(q)\right] & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

is the **adjacency matrix** of a random geometric graph $G(p, q, n)$.

Random graph theory connections

- It turns out the $\mathrm{GOE}(p)$ and the $\mathrm{W}_p(n, I_p/n)$ distributions are closely related to these models.
- Let $X \sim \mathrm{GOE}(p)$, and let Φ be the standard normal cdf. The thresholded matrix

$$A_{ij} = \begin{cases} \mathbb{1}\left[X_{ij} > \Phi^{-1}(q)\right] & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

  is the **adjacency matrix** of an Erdős-Rényi graph $\mathrm{G}(p, q)$.
- Let $Y \sim W_p(n, I_p/n)$, and let Ψ_n be the cdf of the r.v. $\langle U_1, U_1 \rangle$ for U_1, U_2 i.i.d. $\mathrm{Unif}(\mathbb{S}^{n-1})$. The thresholded matrix

$$B_{ij} = \begin{cases} \mathbb{1}\left[\frac{Y_{ij}}{\sqrt{Y_{ii} Y_{jj}}} > \Psi_n^{-1}(q)\right] & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

  is the **adjacency matrix** of a random geometric graph $\mathrm{G}(p, q, n)$.

Random graph theory connections

- It turns out the GOE(p) and the $W_p(n, I_p/n)$ distributions are closely related to these models.
- Let $X \sim \text{GOE}(p)$, and let Φ be the standard normal cdf. The thresholded matrix

$$A_{ij} = \begin{cases} \mathbb{1}\left[X_{ij} > \Phi^{-1}(q)\right] & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

  is the **adjacency matrix** of an Erdős-Rényi graph $G(p, q)$.
- Let $Y \sim W_p(n, I_p/n)$, and let Ψ_n be the cdf of the r.v. $\langle U_1, U_1 \rangle$ for U_1, U_2 i.i.d. Unif($\mathbb{S}^{n-1}$). The thresholded matrix

$$B_{ij} = \begin{cases} \mathbb{1}\left[\frac{Y_{ij}}{\sqrt{Y_{ii}Y_{jj}}} > \Psi_n^{-1}(q)\right] & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

  is the **adjacency matrix** of a random geometric graph $G(p, q, n)$.

Random graph theory connections

- Relationship Erdős-Rényi graph $\leftrightarrow$ random geometric graph is analogue to relationship GOE(p) $\leftrightarrow$ normalized Wishart!

- For example, known that the random geometric graph $G(p, q, n)$ is approximated by a Erdős-Rényi random graph $G(p, q)$ as $p, n \to \infty$ if and only if $p^3/n \to 0$.

- This also suggests we could find distinct random graph models $G_1(p, q), G_2(p, q), \ldots$ such that

$$G(p, q, n) \approx G_m(p, q) \text{ when } p^{m+2}/n^m \to 0,$$

with $G_1(p, q)$ the Erdős-Rényi random graph.

Other work on high-dimensional estimation

- Applied work with colleagues at Cornell's Ithaca Campus and Medical College in 'omics (i.e. gene expression, metabolomics, microbiome) data, in the cross-section and time course, using Bayes and penalized approaches.
- Applying quantile regression based graphical models to understand dynamic gene networks. (work with Haim Bar and James Booth)
- Portfolio selection problems using 1000s of ETFs combined with financial news text.
- A mixture-model of beta distributions framework introduced to identify significant correlations when p is large. *betaMix* relies on theorems in random matrix theory and convex geometry. (work with Haim Bar)
- Improved (shrinkage) estimation of a mean, covariance, precision, and discriminant function in the settings where $p >> n$. This work involves working with the singular Wishart distribution.

Thank you for listening!

References

- Bubeck and Ganguly (2015). Entropic CLT and phase transition in high-dimensional Wishart matrices. Preprint arXiv:1509:03258 [math.PR].
- Bubeck, Ding, Eldan and Rácz (2014). Testing for high-dimensional geometry in random graphs. Preprint arXiv:1411.5713 [math.ST].
- Chételat and Wells (2020). The middle-scale asymptotics of Wishart matrices. Annals of Statistics, 47 2639–2670.
- Jiang and Li (2013). Approximation of rectangular beta-Laguerre ensembles and large deviations. Journal of Theoretical Probability, 1-44.
- Johnstone (2006). High Dimensional Statistical Inference and Random Matrices. Proc ICM Vol 1, 307-–333.
- Johnstone (2008). Multivariate Analysis and Jacobi Ensembles: Largest eigenvalue,Tracy-Widom Limits and Rates of Convergence, Ann. Statist. 36, 2638-2716.
- Ledoit and Wolf (2012) Nonlinear shrinkage estimation of large-dimensional covariance matrices. Annals of Statistics 40, 1024–1060.
- Letac and Massam (2004). All invariant moments of the Wishart distribution. Scandinavian Journal of Statistics, 31(2), 295-318.
- Marchenko and Pastur (1967) The distribution of eigenvalues in certain sets of random matrices. Mat. Sb., 72, 507-536.

The $\mathcal{F}$-conjugate

Theorem (Kullback-Leibler inequality for $\mathcal{F}$-conjugates)

Let F be a distribution on $\mathbb{S}_p(\mathbb{R})$ with density f, and let $\psi \in L^2(\mathbb{S}_p(\mathbb{R}))$. Then the L^2-distance between $\mathcal{F}\{f^{1/2}\}$ and ψ satisfies $\mathrm{d}^2_{L^2}\big(\mathcal{F}\{f^{1/2}\}, \psi\big)$ is less or equal to

$$\Big[\|\psi\|^2_{L^2} - 1\Big] + E\Big[\Re \log \frac{\mathcal{F}\{f^{1/2}\}^2(T)}{\psi^2(T)}\Big] + 2\|\psi\|_{L^2} E\Big[\Big|\Im \log \frac{\mathcal{F}\{f^{1/2}\}^2(T)}{\psi^2(T)}\Big|\Big]^{1/2}$$

for $T \sim F^*$, where $\log$ stands for the principal branch of the complex logarithm and F^* the $\mathcal{F}$-conjugate of F.