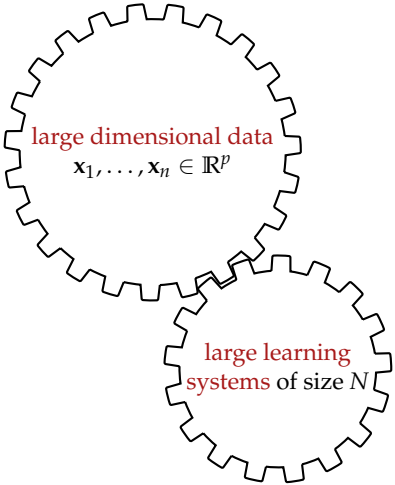
Random Matrix Theory and Its Applications in Large-scale Systems

Zhenyu Liao, Tiebin Mi, Caiming Qiu

Huazhong University of Science and Technology (HUST)
School of Electronic Information and Communications (EIC)

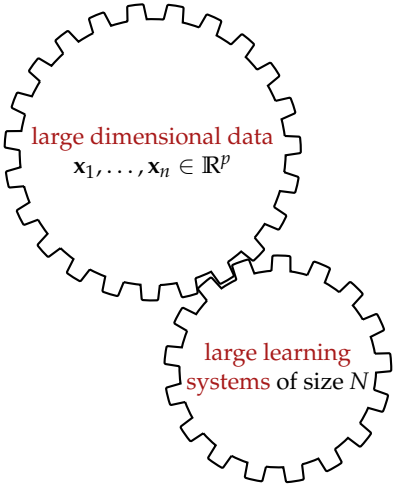
April 21, 2022

Large-scale systems for high-dimensional and massive data



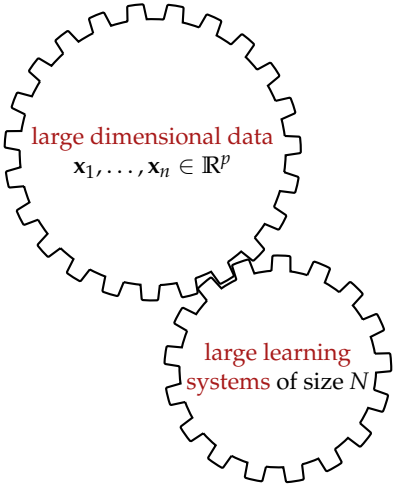
- Big Data era: exploit large n, p, N

Large-scale systems for high-dimensional and massive data



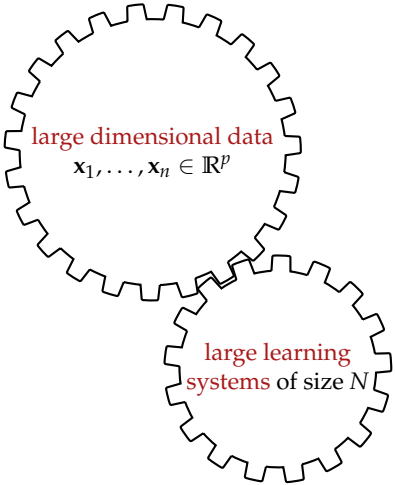
- Big Data era: exploit large n, p, N
- counterintuitive phenomena different from classical asymptotics statistics

Large-scale systems for high-dimensional and massive data



- Big Data era: exploit large n, p, N
- counterintuitive phenomena different from classical asymptotics statistics
- complete change of understanding of many methods

Large-scale systems for high-dimensional and massive data



- Big Data era: exploit large n, p, N
- counterintuitive phenomena different from classical asymptotics statistics
- complete change of understanding of many methods
- **RMT** provides the tools!

Outline

1. Sample covariance matrix for large dimensional data
2. Application of RMT to large-scale telecommunication
3. Application of RMT to large-scale signal processing
4. Application of RMT to large-scale machine learning

Outline

Sample covariance matrix in the large n, p regime

- For $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, estimate population covariance $\mathbf{C} \in \mathbb{R}^{p \times p}$ from n data samples $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$.

Sample covariance matrix in the large n, p regime

- For $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, estimate population covariance $\mathbf{C} \in \mathbb{R}^{p \times p}$ from n data samples $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$.

- Maximum likelihood sample covariance matrix

$$\hat{\mathbf{C}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}} = \frac{1}{n} \mathbf{X}\mathbf{X}^{\mathsf{T}} \in \mathbb{R}^{p \times p},$$

Sample covariance matrix in the large n, p regime

- For $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, estimate population covariance $\mathbf{C} \in \mathbb{R}^{p\times p}$ from n data samples $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p\times n}$.

- Maximum likelihood sample covariance matrix with entry-wise convergence

$$\hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}} = \frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{T}} \in \mathbb{R}^{p\times p}, \quad [\hat{\mathbf{C}}]_{ij} \to [\mathbf{C}]_{ij}$$

  almost surely as $n \to \infty$:

Sample covariance matrix in the large n, p regime

- For $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, estimate population covariance $\mathbf{C} \in \mathbb{R}^{p\times p}$ from n data samples $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p\times n}$.

- Maximum likelihood sample covariance matrix with entry-wise convergence

$$\hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}} = \frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{T}} \in \mathbb{R}^{p\times p}, \quad [\hat{\mathbf{C}}]_{ij} \to [\mathbf{C}]_{ij}$$

  almost surely as $n \to \infty$: optimal for $n \gg p$ (or, for p "small").

Sample covariance matrix in the large n, p regime

- For $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, estimate population covariance $\mathbf{C} \in \mathbb{R}^{p\times p}$ from n data samples $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p\times n}$.

- Maximum likelihood sample covariance matrix with entry-wise convergence

$$\hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}} = \frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{T}} \in \mathbb{R}^{p\times p}, \quad [\hat{\mathbf{C}}]_{ij} \to [\mathbf{C}]_{ij}$$

  almost surely as $n \to \infty$: optimal for $n \gg p$ (or, for p "small").

- In the regime $n \sim p$, conventional wisdom breaks down:
  for $\mathbf{C} = \mathbf{I}_p$ with $n < p$, $\hat{\mathbf{C}}$ has at least $p - n$ zero eigenvalues.

$$\boxed{\|\hat{\mathbf{C}} - \mathbf{C}\| \not\to 0, \quad n, p \to \infty}$$

  $\Rightarrow$ eigenvalue mismatch and not consistent!

Sample covariance matrix in the large n, p regime

- For $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, estimate population covariance $\mathbf{C} \in \mathbb{R}^{p\times p}$ from n data samples $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p\times n}$.

- Maximum likelihood sample covariance matrix with entry-wise convergence

$$\hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}} = \frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{T}} \in \mathbb{R}^{p\times p}, \quad [\hat{\mathbf{C}}]_{ij} \to [\mathbf{C}]_{ij}$$

  almost surely as $n \to \infty$: optimal for $n \gg p$ (or, for p "small").

- In the regime $n \sim p$, conventional wisdom breaks down:
  for $\mathbf{C} = \mathbf{I}_p$ with $n < p$, $\hat{\mathbf{C}}$ has at least $p - n$ zero eigenvalues.

$$\boxed{\|\hat{\mathbf{C}} - \mathbf{C}\| \not\to 0, \quad n, p \to \infty}$$

  $\Rightarrow$ eigenvalue mismatch and not consistent!

- due to $\|\mathbf{A}\|_\infty \le \|\mathbf{A}\| \le p\|\mathbf{A}\|_\infty$ for $\mathbf{A} \in \mathbb{R}^{p\times p}$ and $\|\mathbf{A}\|_\infty \equiv \max_{ij} |\mathbf{A}_{ij}|$.

When is one in the random matrix regime?

What about $n = 100p$?

When is one in the random matrix regime?

What about $n = 100p$? For $\mathbf{C} = \mathbf{I}_p$,

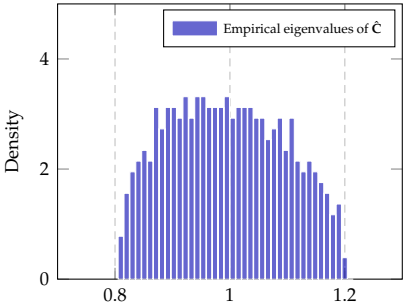


Figure: Eigenvalue distribution of $\hat{\mathbf{C}}$ versus Marčenko-Pastur law, $p = 500$, $n = 50\,000$.

When is one in the random matrix regime?

What about $n = 100p$? For $\mathbf{C} = \mathbf{I}_p$, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$: the Marčenko–Pastur law

$$\mu(dx) = (1 - c^{-1})^+ \delta(x) + \frac{1}{2\pi c x} \sqrt{(x - E_-)^+ (E_+ - x)^+} dx$$

where $E_- = (1 - \sqrt{c})^2$, $E_+ = (1 + \sqrt{c})^2$ and $(x)^+ \equiv \max(x, 0)$.

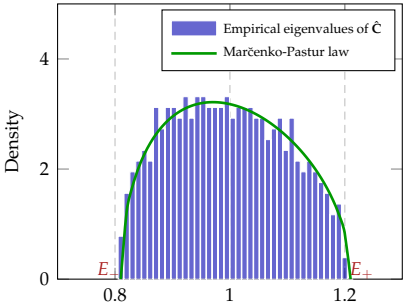


Figure: Eigenvalue distribution of $\hat{\mathbf{C}}$ versus Marčenko-Pastur law, $p = 500$, $n = 50\,000$.

When is one in the random matrix regime?

What about $n = 100p$? For $\mathbf{C} = \mathbf{I}_p$, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$: the Marčenko–Pastur law

$$\mu(dx) = (1 - c^{-1})^+ \delta(x) + \frac{1}{2\pi c x} \sqrt{(x - E_-)^+ (E_+ - x)^+} dx$$

where $E_- = (1 - \sqrt{c})^2$, $E_+ = (1 + \sqrt{c})^2$ and $(x)^+ \equiv \max(x, 0)$. Close match!

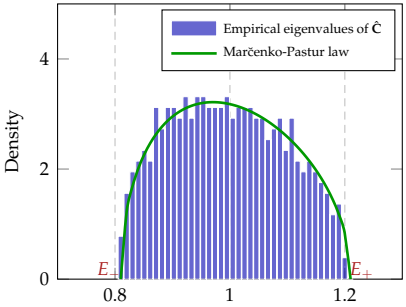


Figure: Eigenvalue distribution of $\hat{\mathbf{C}}$ versus Marčenko-Pastur law, $p = 500$, $n = 50\,000$.

When is one in the random matrix regime?

What about $n = 100p$? For $\mathbf{C} = \mathbf{I}_p$, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$: the Marčenko–Pastur law

$$\mu(dx) = (1 - c^{-1})^+ \delta(x) + \frac{1}{2\pi c x} \sqrt{(x - E_-)^+ (E_+ - x)^+} dx$$

where $E_- = (1 - \sqrt{c})^2$, $E_+ = (1 + \sqrt{c})^2$ and $(x)^+ \equiv \max(x, 0)$. Close match!

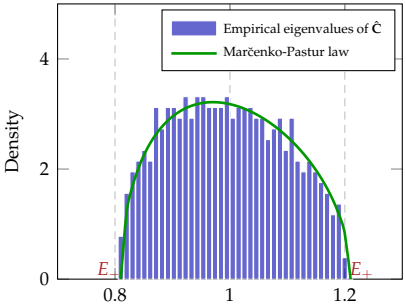


Figure: Eigenvalue distribution of $\hat{\mathbf{C}}$ versus Marčenko-Pastur law, $p = 500$, $n = 50\,000$.

- eigenvalues span on $[E_- = (1 - \sqrt{c})^2, E_+ = (1 + \sqrt{c})^2]$.

When is one in the random matrix regime?

What about $n = 100p$? For $\mathbf{C} = \mathbf{I}_p$, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$: the Marčenko–Pastur law

$$\mu(dx) = (1 - c^{-1})^+ \delta(x) + \frac{1}{2\pi c x} \sqrt{(x - E_-)^+ (E_+ - x)^+} dx$$

where $E_- = (1 - \sqrt{c})^2$, $E_+ = (1 + \sqrt{c})^2$ and $(x)^+ \equiv \max(x, 0)$. Close match!

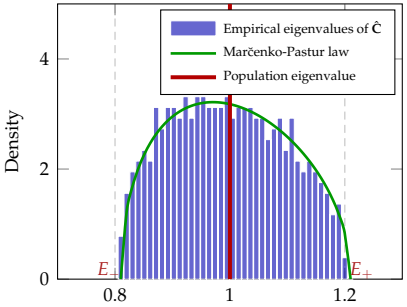


Figure: Eigenvalue distribution of $\hat{\mathbf{C}}$ versus Marčenko-Pastur law, $p = 500$, $n = 50\,000$.

- eigenvalues span on $[E_- = (1 - \sqrt{\mathbf{c}})^2, E_+ = (1 + \sqrt{\mathbf{c}})^2]$.
- for $\mathbf{n = 100p}$, on a range of $\pm\mathbf{2}\sqrt{\mathbf{c}} = \pm\mathbf{0.2}$ around the population eigenvalue $\mathbf{1}$.

When is one in the random matrix regime? Almost always!

What about $n = 100p$? For $\mathbf{C} = \mathbf{I}_p$, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$: the Marčenko–Pastur law

$$\mu(dx) = (1 - c^{-1})^+ \delta(x) + \frac{1}{2\pi c x} \sqrt{(x - E_-)^+ (E_+ - x)^+} dx$$

where $E_- = (1 - \sqrt{c})^2$, $E_+ = (1 + \sqrt{c})^2$ and $(x)^+ \equiv \max(x, 0)$. Close match!

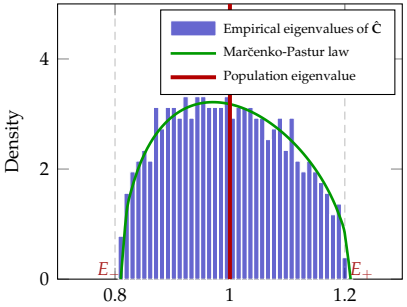


Figure: Eigenvalue distribution of $\hat{\mathbf{C}}$ versus Marčenko-Pastur law, $p = 500$, $n = 50\,000$.

- eigenvalues span on $[E_- = (1 - \sqrt{\mathbf{c}})^2, E_+ = (1 + \sqrt{\mathbf{c}})^2]$.
- for $\mathbf{n = 100p}$, on a range of $\pm\mathbf{2}\sqrt{\mathbf{c}} = \pm\mathbf{0.2}$ around the population eigenvalue $\mathbf{1}$.

Classical large-n asymptotic analysis mostly fails today

- large-n intuition, and many existing popular methods in

Classical large-n asymptotic analysis mostly fails today

- large-n intuition, and many existing popular methods in
  - (nuclear) physics, biology, finance, signal processing, telecommunication, and machine learning

Classical large-n asymptotic analysis mostly fails today

- large-n intuition, and many existing popular methods in
  - (nuclear) physics, biology, finance, signal processing, telecommunication, and machine learning
  - must **fail** even with $n = 100p$!

Classical large-n asymptotic analysis mostly fails today

- large-n intuition, and many existing popular methods in
  - (nuclear) physics, biology, finance, signal processing, telecommunication, and machine learning
  - must **fail** even with $n = 100p$!
- **RMT** appears as a flexible and powerful tool to understand and recreate these methods

Classical large-n asymptotic analysis mostly fails today

- large-n intuition, and many existing popular methods in
  - (nuclear) physics, biology, finance, signal processing, telecommunication, and machine learning
  - must **fail** even with $n = 100p$!
- **RMT** appears as a flexible and powerful tool to understand and recreate these methods
- in essence, "increasing complexity of the system models employed in above fields demand low complexity analysis"

Classical large-n asymptotic analysis mostly fails today

- large-n intuition, and many existing popular methods in
  - (nuclear) physics, biology, finance, signal processing, telecommunication, and machine learning
  - must **fail** even with $n = 100p$!
- **RMT** appears as a flexible and powerful tool to understand and recreate these methods
- in essence, "increasing complexity of the system models employed in above fields demand low complexity analysis"
- in the remainder, provide illustrating examples showing how RMT can be used to assess

Classical large-n asymptotic analysis mostly fails today

- large-n intuition, and many existing popular methods in
  - (nuclear) physics, biology, finance, signal processing, telecommunication, and machine learning
  - must **fail** even with $n = 100p$!
- **RMT** appears as a flexible and powerful tool to understand and recreate these methods
- in essence, "increasing complexity of the system models employed in above fields demand low complexity analysis"
- in the remainder, provide illustrating examples showing how RMT can be used to assess
  - **telecommunication**: code division multiple access (CDMA) technology

Classical large-n asymptotic analysis mostly fails today

- large-n intuition, and many existing popular methods in
  - (nuclear) physics, biology, finance, signal processing, telecommunication, and machine learning
  - must **fail** even with $n = 100p$!
- **RMT** appears as a flexible and powerful tool to understand and recreate these methods
- in essence, "increasing complexity of the system models employed in above fields demand low complexity analysis"
- in the remainder, provide illustrating examples showing how RMT can be used to assess
  - **telecommunication**: code division multiple access (CDMA) technology
  - **signal processing**: generalized likelihood ratio test (GLRT)

Classical large-n asymptotic analysis mostly fails today

- large-n intuition, and many existing popular methods in
  - (nuclear) physics, biology, finance, signal processing, telecommunication, and machine learning
  - must **fail** even with $n = 100p$!
- **RMT** appears as a flexible and powerful tool to understand and recreate these methods
- in essence, "increasing complexity of the system models employed in above fields demand low complexity analysis"
- in the remainder, provide illustrating examples showing how RMT can be used to assess
  - **telecommunication**: code division multiple access (CDMA) technology
  - **signal processing**: generalized likelihood ratio test (GLRT)
  - **machine learning**: kernel spectral clustering

Outline

Application to telecommunication: performance analysis of CDMA via RMT

A quick reminder on code division multiple access (CDMA):

- CDMA in 3G succeeded the time division multiple access (TDMA) technology used in 2G

Application to telecommunication: performance analysis of CDMA via RMT

A quick reminder on code division multiple access (CDMA):

- CDMA in 3G succeeded the time division multiple access (TDMA) technology used in 2G
- with TDMA policy: users are successively allocated an exclusive amount of time to exchange data with the access points (APs)

Application to telecommunication: performance analysis of CDMA via RMT

A quick reminder on code division multiple access (CDMA):

- CDMA in 3G succeeded the time division multiple access (TDMA) technology used in 2G
- with TDMA policy: users are successively allocated an exclusive amount of time to exchange data with the access points (APs)
- major issues: at the same time a very strict maximal number of users could be accepted by a given AP, regardless of the users' requests in terms of quality of service

Application to telecommunication: performance analysis of CDMA via RMT

A quick reminder on code division multiple access (CDMA):

- CDMA in 3G succeeded the time division multiple access (TDMA) technology used in 2G
- with TDMA policy: users are successively allocated an exclusive amount of time to exchange data with the access points (APs)
- major issues: at the same time a very strict maximal number of users could be accepted by a given AP, regardless of the users' requests in terms of quality of service
- CDMA: to increase the max number of users, and to dynamically balancing the quality of service offered to each terminal

Application to telecommunication: performance analysis of CDMA via RMT

A quick reminder on code division multiple access (CDMA):

- CDMA in 3G succeeded the time division multiple access (TDMA) technology used in 2G
- with TDMA policy: users are successively allocated an exclusive amount of time to exchange data with the access points (APs)
- major issues: at the same time a very strict maximal number of users could be accepted by a given AP, regardless of the users' requests in terms of quality of service
- CDMA: to increase the max number of users, and to dynamically balancing the quality of service offered to each terminal
  - each user is allocated a (usually long) **spreading code** that is made roughly orthogonal to the other users' codes

Application to telecommunication: performance analysis of CDMA via RMT

A quick reminder on code division multiple access (CDMA):

- CDMA in 3G succeeded the time division multiple access (TDMA) technology used in 2G
- with TDMA policy: users are successively allocated an exclusive amount of time to exchange data with the access points (APs)
- major issues: at the same time a very strict maximal number of users could be accepted by a given AP, regardless of the users' requests in terms of quality of service
- CDMA: to increase the max number of users, and to dynamically balancing the quality of service offered to each terminal
  - each user is allocated a (usually long) **spreading code** that is made roughly orthogonal to the other users' codes
  - so that all users can simultaneously receive data while experiencing a limited amount of interference from concurrent communications, due to code orthogonality

Application to telecommunication: performance analysis of CDMA via RMT

A quick reminder on code division multiple access (CDMA):

- CDMA in 3G succeeded the time division multiple access (TDMA) technology used in 2G
- with TDMA policy: users are successively allocated an exclusive amount of time to exchange data with the access points (APs)
- major issues: at the same time a very strict maximal number of users could be accepted by a given AP, regardless of the users' requests in terms of quality of service
- CDMA: to increase the max number of users, and to dynamically balancing the quality of service offered to each terminal
  - each user is allocated a (usually long) **spreading code** that is made roughly orthogonal to the other users' codes
  - so that all users can simultaneously receive data while experiencing a limited amount of interference from concurrent communications, due to code orthogonality
  - similarly in the up-link

Application to telecommunication: performance analysis of CDMA via RMT

A quick reminder on code division multiple access (CDMA):

- CDMA in 3G succeeded the time division multiple access (TDMA) technology used in 2G
- with TDMA policy: users are successively allocated an exclusive amount of time to exchange data with the access points (APs)
- major issues: at the same time a very strict maximal number of users could be accepted by a given AP, regardless of the users' requests in terms of quality of service
- CDMA: to increase the max number of users, and to dynamically balancing the quality of service offered to each terminal
  - each user is allocated a (usually long) **spreading code** that is made roughly orthogonal to the other users' codes
  - so that all users can simultaneously receive data while experiencing a limited amount of interference from concurrent communications, due to code orthogonality
  - similarly in the up-link
  - since the spreading codes are rarely fully orthogonal (unless orthogonal codes such as Hadamard codes are used), the more users served by an AP, the more the interference and then the less the quality of service; but at no time is a user rejected for lack of available resource (unless there is an excessive number of users)

Application to telecommunication: performance analysis of CDMA via RMT

A quick reminder on code division multiple access (CDMA):

- CDMA in 3G succeeded the time division multiple access (TDMA) technology used in 2G
- with TDMA policy: users are successively allocated an exclusive amount of time to exchange data with the access points (APs)
- major issues: at the same time a very strict maximal number of users could be accepted by a given AP, regardless of the users' requests in terms of quality of service
- CDMA: to increase the max number of users, and to dynamically balancing the quality of service offered to each terminal
  - each user is allocated a (usually long) **spreading code** that is made roughly orthogonal to the other users' codes
  - so that all users can simultaneously receive data while experiencing a limited amount of interference from concurrent communications, due to code orthogonality
  - similarly in the up-link
  - since the spreading codes are rarely fully orthogonal (unless orthogonal codes such as Hadamard codes are used), the more users served by an AP, the more the interference and then the less the quality of service; but at no time is a user rejected for lack of available resource (unless there is an excessive number of users)
- **Question**: how to evaluate the **capacity** (max achievable transmission data rate) of CDMA network? (which clearly depends on pre-coding strategy)

Application to telecommunication: performance analysis of CDMA via RMT

A quick reminder on code division multiple access (CDMA):

- CDMA in 3G succeeded the time division multiple access (TDMA) technology used in 2G
- with TDMA policy: users are successively allocated an exclusive amount of time to exchange data with the access points (APs)
- major issues: at the same time a very strict maximal number of users could be accepted by a given AP, regardless of the users' requests in terms of quality of service
- CDMA: to increase the max number of users, and to dynamically balancing the quality of service offered to each terminal
  - each user is allocated a (usually long) **spreading code** that is made roughly orthogonal to the other users' codes
  - so that all users can simultaneously receive data while experiencing a limited amount of interference from concurrent communications, due to code orthogonality
  - similarly in the up-link
  - since the spreading codes are rarely fully orthogonal (unless orthogonal codes such as Hadamard codes are used), the more users served by an AP, the more the interference and then the less the quality of service; but at no time is a user rejected for lack of available resource (unless there is an excessive number of users)
- **Question**: how to evaluate the **capacity** (max achievable transmission data rate) of CDMA network? (which clearly depends on pre-coding strategy)
- **Answer**: to use random i.i.d. codes, that is (pseudo-)random CDMA! (that can be shown to outperform TDMA and fully orthogonal CDMA)

For **orthogonal** CDMA, assume:

- frequency flat channel conditions for all users; and
- channel stability over a large number of successive symbol periods;

then the rates achieved in the up-link (from user terminals to APs) are maximal when the orthogonal codes are as long as the number of users n, and we have, for a noise power σ^2, system capacity given by

$$C_{\text{orth}}(\sigma^2) = \frac{1}{n} \log \det \left(\mathbf{I}_n + \frac{1}{\sigma^2} \mathbf{W}\mathbf{G}\mathbf{G}^{\mathsf{H}}\mathbf{W}^{\mathsf{H}} \right), \tag{1}$$

with $\mathbf{W} \in \mathbb{C}^{n \times n}$ having its columns the **orthogonal** CDMA codes (and $\mathbf{W}$ being unitary in this case), and $\mathbf{G} \equiv \text{diag}\{g_i\}_{i=1}^{n}$ the diagonal matrix of the channel **gains** of the users $1, \ldots, n$.

For **orthogonal** CDMA, assume:

- frequency flat channel conditions for all users; and
- channel stability over a large number of successive symbol periods;

then the rates achieved in the up-link (from user terminals to APs) are maximal when the orthogonal codes are as long as the number of users n, and we have, for a noise power σ^2, system capacity given by

$$C_{\text{orth}}(\sigma^2) = \frac{1}{n} \log \det \left(\mathbf{I}_n + \frac{1}{\sigma^2} \mathbf{W} \mathbf{G} \mathbf{G}^{\mathsf{H}} \mathbf{W}^{\mathsf{H}} \right), \tag{1}$$

with $\mathbf{W} \in \mathbb{C}^{n \times n}$ having its columns the **orthogonal** CDMA codes (and $\mathbf{W}$ being unitary in this case), and $\mathbf{G} \equiv \text{diag}\{g_i\}_{i=1}^n$ the diagonal matrix of the channel **gains** of the users $1, \ldots, n$.
It can be easily checked (with Sylvester's identity and $\mathbf{W}$ unitary with $\mathbf{W}^{\mathsf{H}} \mathbf{W} = \mathbf{I}_n$) that

$$C_{\text{orth}}(\sigma^2) = \frac{1}{n} \log \det \left(\mathbf{I}_n + \frac{1}{\sigma^2} \mathbf{G} \mathbf{G}^{\mathsf{H}} \right) = \frac{1}{n} \sum_{i=1}^{n} \log \left(1 + \frac{|g_i|^2}{\sigma^2} \right) = C_{\text{TDMA}}(\sigma^2). \tag{2}$$

This justifies the equivalence between TDMA and **orthogonal** CDMA rate performance.

When it comes to random CDMA, under the same conditions, we have

$$C_{\text{rand}}(\sigma^2) = \frac{1}{n} \log \det \left(\mathbf{I}_n + \frac{1}{\sigma^2} \mathbf{X}\mathbf{G}\mathbf{G}^{\mathsf{H}}\mathbf{X}^{\mathsf{H}} \right), \tag{3}$$

for $\mathbf{X} \in \mathbb{C}^{n \times n}$ with it columns containing the users' random codes.

When it comes to random CDMA, under the same conditions, we have

$$C_{\text{rand}}(\sigma^2) = \frac{1}{n} \log \det \left(\mathbf{I}_n + \frac{1}{\sigma^2} \mathbf{X}\mathbf{G}\mathbf{G}^{\mathsf{H}}\mathbf{X}^{\mathsf{H}} \right), \tag{3}$$

for $\mathbf{X} \in \mathbb{C}^{n \times n}$ with it columns containing the users' random codes.

Question: evaluate $C_{\text{rand}}(\sigma^2)$ for n large, as a function of the gains $\mathbf{G}$ and (the distribution of) the codes $\mathbf{X}$.

- (first?) answered by Shami, Tse, and Verdú in [TV00; VS99];

When it comes to random CDMA, under the same conditions, we have

$$C_{\text{rand}}(\sigma^2) = \frac{1}{n} \log \det \left(\mathbf{I}_n + \frac{1}{\sigma^2} \mathbf{X}\mathbf{G}\mathbf{G}^{\mathsf{H}}\mathbf{X}^{\mathsf{H}} \right), \tag{3}$$

for $\mathbf{X} \in \mathbb{C}^{n \times n}$ with it columns containing the users' random codes.
Question: evaluate $C_{\text{rand}}(\sigma^2)$ for n large, as a function of the gains $\mathbf{G}$ and (the distribution of) the codes $\mathbf{X}$.

- (first?) answered by Shami, Tse, and Verdú in [TV00; VS99];
- however these capacity expressions may not be realistic achievable in practice, due to complicated and nonlinear processing algorithms at the receiving APs;

When it comes to random CDMA, under the same conditions, we have

$$C_{\text{rand}}(\sigma^2) = \frac{1}{n} \log \det \left(\mathbf{I}_n + \frac{1}{\sigma^2} \mathbf{X}\mathbf{G}\mathbf{G}^{\mathsf{H}}\mathbf{X}^{\mathsf{H}} \right), \tag{3}$$

for $\mathbf{X} \in \mathbb{C}^{n \times n}$ with it columns containing the users' random codes.

Question: evaluate $C_{\text{rand}}(\sigma^2)$ for n large, as a function of the gains $\mathbf{G}$ and (the distribution of) the codes $\mathbf{X}$.

- (first?) answered by Shami, Tse, and Verdú in [TV00; VS99];
- however these capacity expressions may not be realistic achievable in practice, due to complicated and nonlinear processing algorithms at the receiving APs;
- if only linear pre-coders and/or decoders (e.g., matched-filter, linear minimum mean square error, LMMSE) are used, optimal solution and performance guarantee are given by

When it comes to random CDMA, under the same conditions, we have

$$C_{\text{rand}}(\sigma^2) = \frac{1}{n} \log \det \left(\mathbf{I}_n + \frac{1}{\sigma^2} \mathbf{X}\mathbf{G}\mathbf{G}^{\mathsf{H}}\mathbf{X}^{\mathsf{H}} \right), \tag{3}$$

for $\mathbf{X} \in \mathbb{C}^{n \times n}$ with it columns containing the users' random codes.

Question: evaluate $C_{\text{rand}}(\sigma^2)$ for n large, as a function of the gains $\mathbf{G}$ and (the distribution of) the codes $\mathbf{X}$.

- (first?) answered by Shami, Tse, and Verdú in [TV00; VS99];
- however these capacity expressions may not be realistic achievable in practice, due to complicated and nonlinear processing algorithms at the receiving APs;
- if only linear pre-coders and/or decoders (e.g., matched-filter, linear minimum mean square error, LMMSE) are used, optimal solution and performance guarantee are given by
  - Tse and Hanly in [TH99] for frequency flat channels;

When it comes to random CDMA, under the same conditions, we have

$$C_{\text{rand}}(\sigma^2) = \frac{1}{n} \log \det \left(\mathbf{I}_n + \frac{1}{\sigma^2} \mathbf{X}\mathbf{G}\mathbf{G}^{\mathsf{H}}\mathbf{X}^{\mathsf{H}} \right), \tag{3}$$

for $\mathbf{X} \in \mathbb{C}^{n \times n}$ with it columns containing the users' random codes.

Question: evaluate $C_{\text{rand}}(\sigma^2)$ for n large, as a function of the gains $\mathbf{G}$ and (the distribution of) the codes $\mathbf{X}$.

- (first?) answered by Shami, Tse, and Verdú in [TV00; VS99];
- however these capacity expressions may not be realistic achievable in practice, due to complicated and nonlinear processing algorithms at the receiving APs;
- if only linear pre-coders and/or decoders (e.g., matched-filter, linear minimum mean square error, LMMSE) are used, optimal solution and performance guarantee are given by
  - Tse and Hanly in [TH99] for frequency flat channels;
  - Evans and Tse in [ET00] for frequency selective channels;

Random versus orthogonal CDMA

When it comes to random CDMA, under the same conditions, we have

$$C_{\text{rand}}(\sigma^2) = \frac{1}{n} \log \det \left(\mathbf{I}_n + \frac{1}{\sigma^2} \mathbf{X}\mathbf{G}\mathbf{G}^{\mathsf{H}}\mathbf{X}^{\mathsf{H}} \right), \tag{3}$$

for $\mathbf{X} \in \mathbb{C}^{n \times n}$ with it columns containing the users' random codes.

Question: evaluate $C_{\text{rand}}(\sigma^2)$ for n large, as a function of the gains $\mathbf{G}$ and (the distribution of) the codes $\mathbf{X}$.

- (first?) answered by Shami, Tse, and Verdú in [TV00; VS99];
- however these capacity expressions may not be realistic achievable in practice, due to complicated and nonlinear processing algorithms at the receiving APs;
- if only linear pre-coders and/or decoders (e.g., matched-filter, linear minimum mean square error, LMMSE) are used, optimal solution and performance guarantee are given by
  - Tse and Hanly in [TH99] for frequency flat channels;
  - Evans and Tse in [ET00] for frequency selective channels;
  - Li and Verdú [LTV04] for reduced-rank LMMSE decoders;

Random versus orthogonal CDMA

When it comes to random CDMA, under the same conditions, we have

$$C_{\text{rand}}(\sigma^2) = \frac{1}{n} \log \det \left(\mathbf{I}_n + \frac{1}{\sigma^2} \mathbf{X}\mathbf{G}\mathbf{G}^{\mathsf{H}}\mathbf{X}^{\mathsf{H}} \right), \tag{3}$$

for $\mathbf{X} \in \mathbb{C}^{n \times n}$ with it columns containing the users' random codes.

Question: evaluate $C_{\text{rand}}(\sigma^2)$ for n large, as a function of the gains $\mathbf{G}$ and (the distribution of) the codes $\mathbf{X}$.

- (first?) answered by Shami, Tse, and Verdú in [TV00; VS99];
- however these capacity expressions may not be realistic achievable in practice, due to complicated and nonlinear processing algorithms at the receiving APs;
- if only linear pre-coders and/or decoders (e.g., matched-filter, linear minimum mean square error, LMMSE) are used, optimal solution and performance guarantee are given by
  - Tse and Hanly in [TH99] for frequency flat channels;
  - Evans and Tse in [ET00] for frequency selective channels;
  - Li and Verdú [LTV04] for reduced-rank LMMSE decoders;
  - Verdú and Shamai in [VS99] for several receivers in frequency flat channels;

When it comes to random CDMA, under the same conditions, we have

$$C_{\text{rand}}(\sigma^2) = \frac{1}{n} \log \det \left(\mathbf{I}_n + \frac{1}{\sigma^2} \mathbf{X}\mathbf{G}\mathbf{G}^{\mathsf{H}}\mathbf{X}^{\mathsf{H}} \right), \tag{3}$$

for $\mathbf{X} \in \mathbb{C}^{n \times n}$ with it columns containing the users' random codes.

Question: evaluate $C_{\text{rand}}(\sigma^2)$ for n large, as a function of the gains $\mathbf{G}$ and (the distribution of) the codes $\mathbf{X}$.

- (first?) answered by Shami, Tse, and Verdú in [TV00; VS99];
- however these capacity expressions may not be realistic achievable in practice, due to complicated and nonlinear processing algorithms at the receiving APs;
- if only linear pre-coders and/or decoders (e.g., matched-filter, linear minimum mean square error, LMMSE) are used, optimal solution and performance guarantee are given by
  - Tse and Hanly in [TH99] for frequency flat channels;
  - Evans and Tse in [ET00] for frequency selective channels;
  - Li and Verdú [LTV04] for reduced-rank LMMSE decoders;
  - Verdú and Shamai in [VS99] for several receivers in frequency flat channels;
  - etc., etc.

Outline

Motivation:

- Shannon made us realize that, in order to achieve high rate of information transfer, increasing the transmission bandwidth is largely preferred over increasing the transmission power

Motivation:

- Shannon made us realize that, in order to achieve high rate of information transfer, increasing the transmission bandwidth is largely preferred over increasing the transmission power
- to ensure high rate communications with a finite power budget, need **frequency multiplexing**

Motivation:

- Shannon made us realize that, in order to achieve high rate of information transfer, increasing the transmission bandwidth is largely preferred over increasing the transmission power
- to ensure high rate communications with a finite power budget, need **frequency multiplexing**
- the idea of **cognitive radio**: to communicate not by exploiting the over-used frequency domain, or by exploiting the over-used space domain, but by exploiting so-called spectrum holes, jointly in time, space, and frequency

Motivation:

- Shannon made us realize that, in order to achieve high rate of information transfer, increasing the transmission bandwidth is largely preferred over increasing the transmission power
- to ensure high rate communications with a finite power budget, need **frequency multiplexing**
- the idea of **cognitive radio**: to communicate not by exploiting the over-used frequency domain, or by exploiting the over-used space domain, but by exploiting so-called spectrum holes, jointly in time, space, and frequency
- **key**: the effectively delivered communication service is largely discontinuous, and telecommunication networks do not operate constantly in **all** frequency bands, at **all** times, and in **all** places at the maximum of their deliverable capacities

Motivation:

- Shannon made us realize that, in order to achieve high rate of information transfer, increasing the transmission bandwidth is largely preferred over increasing the transmission power
- to ensure high rate communications with a finite power budget, need **frequency multiplexing**
- the idea of **cognitive radio**: to communicate not by exploiting the over-used frequency domain, or by exploiting the over-used space domain, but by exploiting so-called spectrum holes, jointly in time, space, and frequency
- **key**: the effectively delivered communication service is largely discontinuous, and telecommunication networks do not operate constantly in **all** frequency bands, at **all** times, and in **all** places at the maximum of their deliverable capacities

Motivation:

- Shannon made us realize that, in order to achieve high rate of information transfer, increasing the transmission bandwidth is largely preferred over increasing the transmission power
- to ensure high rate communications with a finite power budget, need **frequency multiplexing**
- the idea of **cognitive radio**: to communicate not by exploiting the over-used frequency domain, or by exploiting the over-used space domain, but by exploiting so-called spectrum holes, jointly in time, space, and frequency
- **key**: the effectively delivered communication service is largely discontinuous, and telecommunication networks do not operate constantly in **all** frequency bands, at **all** times, and in **all** places at the maximum of their deliverable capacities

As such, a cognitive radio network (also called a *secondary network*)

- can help reuse the resources in a licensed (*first*) network

Motivation:

- Shannon made us realize that, in order to achieve high rate of information transfer, increasing the transmission bandwidth is largely preferred over increasing the transmission power
- to ensure high rate communications with a finite power budget, need **frequency multiplexing**
- the idea of **cognitive radio**: to communicate not by exploiting the over-used frequency domain, or by exploiting the over-used space domain, but by exploiting so-called spectrum holes, jointly in time, space, and frequency
- **key**: the effectively delivered communication service is largely discontinuous, and telecommunication networks do not operate constantly in **all** frequency bands, at **all** times, and in **all** places at the maximum of their deliverable capacities

As such, a cognitive radio network (also called a *secondary network*)

- can help reuse the resources in a licensed (*first*) network
- but require constant **awareness** of the operations taking place in the licensed networks

Motivation:

- Shannon made us realize that, in order to achieve high rate of information transfer, increasing the transmission bandwidth is largely preferred over increasing the transmission power
- to ensure high rate communications with a finite power budget, need **frequency multiplexing**
- the idea of **cognitive radio**: to communicate not by exploiting the over-used frequency domain, or by exploiting the over-used space domain, but by exploiting so-called spectrum holes, jointly in time, space, and frequency
- **key**: the effectively delivered communication service is largely discontinuous, and telecommunication networks do not operate constantly in **all** frequency bands, at **all** times, and in **all** places at the maximum of their deliverable capacities

As such, a cognitive radio network (also called a *secondary network*)

- can help reuse the resources in a licensed (*first*) network
- but require constant **awareness** of the operations taking place in the licensed networks
- for example, via **signal sensing/detection**

Hypothesis testing in a signal-plus-noise model for cognitive radios

System model: let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$ be a matrix with i.i.d. columns $\mathbf{x}_i \in \mathbb{R}^p$ received by a large array of p sensors, the decision problem is formulated as the following binary hypothesis test:

$$\mathbf{X} = \begin{cases} \sigma \mathbf{Z}, & \mathcal{H}_0 \\ \mathbf{a}\mathbf{s}^{\mathsf{T}} + \sigma \mathbf{Z}, & \mathcal{H}_1 \end{cases}$$

where $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{p \times n}$ with $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, $\mathbf{a} \in \mathbb{R}^p$ deterministic with unit norm $\|\mathbf{a}\| = 1$, signal $\mathbf{s} = [s_1, \ldots, s_n]^{\mathsf{T}} \in \mathbb{R}^n$ with s_i i.i.d. random scalars, and $\sigma > 0$. We also denote $c = p/n \in (0, \infty)$.

Hypothesis testing in a signal-plus-noise model for cognitive radios

System model: let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$ be a matrix with i.i.d. columns $\mathbf{x}_i \in \mathbb{R}^p$ received by a large array of p sensors, the decision problem is formulated as the following binary hypothesis test:

$$\mathbf{X} = \begin{cases} \sigma \mathbf{Z}, & \mathcal{H}_0 \\ \mathbf{a}\mathbf{s}^\mathsf{T} + \sigma \mathbf{Z}, & \mathcal{H}_1 \end{cases}$$

where $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{p \times n}$ with $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, $\mathbf{a} \in \mathbb{R}^p$ deterministic with unit norm $\|\mathbf{a}\| = 1$, signal $\mathbf{s} = [s_1, \ldots, s_n]^\mathsf{T} \in \mathbb{R}^n$ with s_i i.i.d. random scalars, and $\sigma > 0$. We also denote $c = p/n \in (0, \infty)$.

- This model describes the observation of either pure Gaussian **noise** $\sigma \mathbf{z}_i$ with zero mean and covariance $\sigma^2 \mathbf{I}_p$ or of a deterministic **information** vector $\mathbf{a}$ possibly modulated by a scalar (random) **signal** s_i (which could simply be ± 1) added to the noise.

System model: let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p\times n}$ be a matrix with i.i.d. columns $\mathbf{x}_i \in \mathbb{R}^p$ received by a large array of p sensors, the decision problem is formulated as the following binary hypothesis test:

$$\mathbf{X} = \begin{cases} \sigma\mathbf{Z}, & \mathcal{H}_0 \\ \mathbf{a}\mathbf{s}^\mathsf{T} + \sigma\mathbf{Z}, & \mathcal{H}_1 \end{cases}$$

where $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{p\times n}$ with $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, $\mathbf{a} \in \mathbb{R}^p$ deterministic with unit norm $\|\mathbf{a}\| = 1$, signal $\mathbf{s} = [s_1, \ldots, s_n]^\mathsf{T} \in \mathbb{R}^n$ with s_i i.i.d. random scalars, and $\sigma > 0$. We also denote $c = p/n \in (0, \infty)$.

- This model describes the observation of either pure Gaussian **noise** $\sigma\mathbf{z}_i$ with zero mean and covariance $\sigma^2\mathbf{I}_p$ or of a deterministic **information** vector $\mathbf{a}$ possibly modulated by a scalar (random) **signal** s_i (which could simply be ± 1) added to the noise.
- If the parameters $\mathbf{a}$, σ as well as the statistics of s_i are known, a mere Neyman-Pearson test allows one to discriminate between $\mathcal{H}_0$ and $\mathcal{H}_1$ with optimal detection probability: decide on the genuine hypothesis according to the ratio of posterior probabilities

$$\frac{\mathbb{P}(\mathbf{X} \mid \mathcal{H}_1)}{\mathbb{P}(\mathbf{X} \mid \mathcal{H}_0)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \alpha \tag{4}$$

for some $\alpha > 0$ controlling the desired Type I and Type II error rates (that is, the probability of false positives and of false negatives).

However,

- in practice, we do not know σ, nor the information vector $\mathbf{a} \in \mathbb{R}^p$ (to be recovered)

However,

- in practice, we do not know σ, nor the information vector $\mathbf{a} \in \mathbb{R}^p$ (to be recovered)
- in the most generic scenario where $\mathbf{a}$ is fully unknown, and if $s_i \sim \mathcal{N}(0,1)$, instead of the maximum likelihood test in (4), one may resort to a **generalized likelihood ratio test** (GLRT) defined as

$$\frac{\sup_{\sigma,\mathbf{a}} \mathbb{P}(\mathbf{X} \mid \sigma, \mathbf{a}, \mathcal{H}_1)}{\sup_{\sigma,\mathbf{a}} \mathbb{P}(\mathbf{X} \mid \sigma, \mathcal{H}_0)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \alpha.$$

Hypothesis testing in a signal-plus-noise model via GLRT

However,

- in practice, we do not know σ, nor the information vector $\mathbf{a} \in \mathbb{R}^p$ (to be recovered)
- in the most generic scenario where $\mathbf{a}$ is fully unknown, and if $s_i \sim \mathcal{N}(0,1)$, instead of the maximum likelihood test in (4), one may resort to a **generalized likelihood ratio test** (GLRT) defined as

$$\frac{\sup_{\sigma,\mathbf{a}} \mathbb{P}(\mathbf{X} \mid \sigma, \mathbf{a}, \mathcal{H}_1)}{\sup_{\sigma,\mathbf{a}} \mathbb{P}(\mathbf{X} \mid \sigma, \mathcal{H}_0)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \alpha.$$

- under both a Gaussian noise and signal s_i assumption, the GLRT has an explicit expression as a monotonous increasing function of $\|\mathbf{X}\mathbf{X}^\mathsf{T}\| / \operatorname{tr}(\mathbf{X}\mathbf{X}^\mathsf{T})$. so the test is equivalent to

$$T_p \equiv \frac{\left\| \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} \right\|}{\frac{1}{p}\operatorname{tr}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}\right)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} f(\alpha),$$

for some known monotonously increasing function f (the normalizations $1/p$ and $1/n$ so that both the numerator and denominator are of order $O(1)$ as $n, p \to \infty$).

Hypothesis testing in a signal-plus-noise model via GLRT

However,

- in practice, we do not know σ, nor the information vector $\mathbf{a} \in \mathbb{R}^p$ (to be recovered)
- in the most generic scenario where $\mathbf{a}$ is fully unknown, and if $s_i \sim \mathcal{N}(0,1)$, instead of the maximum likelihood test in (4), one may resort to a **generalized likelihood ratio test** (GLRT) defined as

$$\frac{\sup_{\sigma,\mathbf{a}} \mathbb{P}(\mathbf{X} \mid \sigma, \mathbf{a}, \mathcal{H}_1)}{\sup_{\sigma,\mathbf{a}} \mathbb{P}(\mathbf{X} \mid \sigma, \mathcal{H}_0)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \alpha.$$

- under both a Gaussian noise and signal s_i assumption, the GLRT has an explicit expression as a monotonous increasing function of $\|\mathbf{X}\mathbf{X}^\mathsf{T}\| / \operatorname{tr}(\mathbf{X}\mathbf{X}^\mathsf{T})$. so the test is equivalent to

$$T_p \equiv \frac{\left\| \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} \right\|}{\frac{1}{p}\operatorname{tr}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}\right)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} f(\alpha),$$

  for some known monotonously increasing function f (the normalizations $1/p$ and $1/n$ so that both the numerator and denominator are of order $O(1)$ as $n, p \to \infty$).
- to evaluate the power of GLRT above, we need to assess the max and mean eigenvalues of SCM $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$

Hypothesis testing in a signal-plus-noise model via GLRT

As a consequence, in order to set a maximum false alarm rate (or false positive, or Type I error) of $r > 0$ in the limit of large n, p, one must choose a threshold $f(\alpha)$ for T_p such that

$$\mathbb{P}(T_p \leq f(\alpha)) = r,$$

that is, such that

$$\mu_{\mathrm{TW}_1}((-\infty, A_p]) = r, \quad A_p = (f(\alpha) - (1+\sqrt{c})^2)(1+\sqrt{c})^{-\frac{4}{3}} c^{\frac{1}{6}} n^{\frac{2}{3}} \tag{5}$$

with μ_{TW_1} the Tracy-Widom distribution.

As a consequence, in order to set a maximum false alarm rate (or false positive, or Type I error) of $r > 0$ in the limit of large n, p, one must choose a threshold $f(\alpha)$ for T_p such that

$$\mathbb{P}(T_p \leq f(\alpha)) = r,$$

that is, such that

$$\mu_{\mathrm{TW}_1}((-\infty, A_p]) = r, \quad A_p = (f(\alpha) - (1+\sqrt{c})^2)(1+\sqrt{c})^{-\frac{4}{3}} c^{\frac{1}{6}} n^{\frac{2}{3}} \tag{5}$$

with μ_{TW_1} the Tracy-Widom distribution.

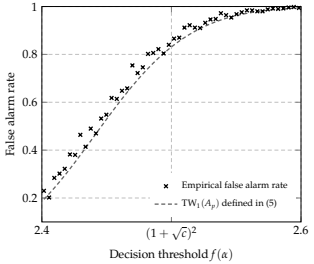


Figure: Comparison between empirical false alarm rates and $\mathrm{TW}_1(A_p)$ for A_p of the form in (5), as a function of the threshold $f(\alpha) \in [(1+\sqrt{c})^2 - 5n^{-2/3}, (1+\sqrt{c})^2 + 5n^{-2/3}]$, for $p = 256$, $n = 1\,024$ and $\sigma = 1$. Results obtained from 500 runs.

Outline

"Curse of dimensionality": loss of relevance of Euclidean distance

- Binary Gaussian mixture classification $\mathbf{x} \in \mathbb{R}^p$:

$$\mathcal{C}_1 : \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_1), \text{ versus } \mathcal{C}_2 : \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_2);$$

"Curse of dimensionality": loss of relevance of Euclidean distance

- Binary Gaussian mixture classification $\mathbf{x} \in \mathbb{R}^p$:

$$\mathcal{C}_1 : \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_1), \text{ versus } \mathcal{C}_2 : \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_2);$$

- Neyman-Pearson test: classification is possible only when [CLM18]

$$\boxed{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| \geq C_{\boldsymbol{\mu}}, \text{ or } \|\mathbf{C}_1 - \mathbf{C}_2\| \geq C_{\mathbf{C}} \cdot p^{-1/2}}$$

for some constants $C_{\boldsymbol{\mu}}, C_{\mathbf{C}} > 0$.

[1] Romain Couillet, Zhenyu Liao, and Xiaoyi Mai. "Classification asymptotics in the random matrix regime". In: *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE. 2018, pp. 1875–1879

"Curse of dimensionality": loss of relevance of Euclidean distance

- Binary Gaussian mixture classification $\mathbf{x} \in \mathbb{R}^p$:

$$\mathcal{C}_1 : \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_1), \text{ versus } \mathcal{C}_2 : \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_2);$$

- Neyman-Pearson test: classification is possible only when [CLM18]

$$\boxed{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| \geq C_{\boldsymbol{\mu}}, \text{ or } \|\mathbf{C}_1 - \mathbf{C}_2\| \geq C_{\mathbf{C}} \cdot p^{-1/2}}$$

  for some constants $C_{\boldsymbol{\mu}}, C_{\mathbf{C}} > 0$.

- In this non-trivial setting, for $\mathbf{x}_i \in \mathcal{C}_a, \mathbf{x}_j \in \mathcal{C}_b$:

$$\boxed{\max_{1 \leq i \neq j \leq n} \left\{ \frac{1}{p} \|\mathbf{x}_i - \mathbf{x}_j\|^2 - \frac{2}{p} \operatorname{tr} \mathbf{C}^\circ \right\} \xrightarrow{a.s.} 0}$$

  as $n, p \to \infty$ (i.e., $n \sim p$), for $\mathbf{C}^\circ \equiv \frac{1}{2}(\mathbf{C}_1 + \mathbf{C}_2)$, regardless of the classes $\mathcal{C}_a, \mathcal{C}_b$! (In fact even for $n = p^m$.)

[1] Romain Couillet, Zhenyu Liao, and Xiaoyi Mai. "Classification asymptotics in the random matrix regime". In: *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE. 2018, pp. 1875–1879

Loss of relevance of Euclidean distance in large dimensions: visual representation

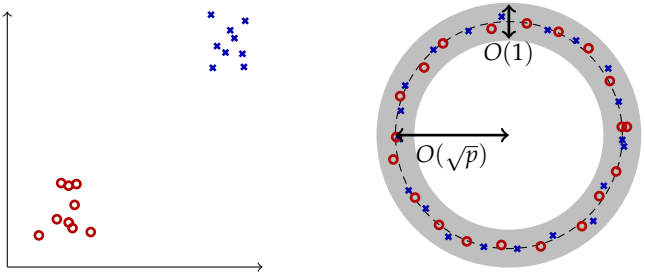


Figure: Visual representation of classification in (**left**) small and (**right**) large dimensions.

Loss of relevance of Euclidean distance in large dimensions: visual representation

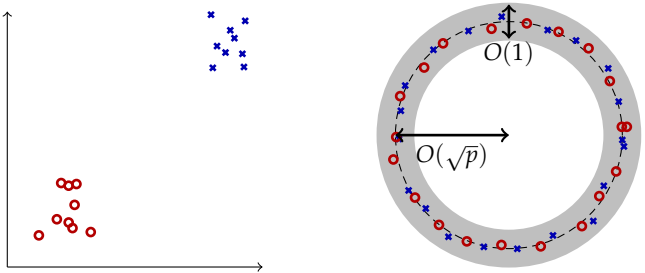
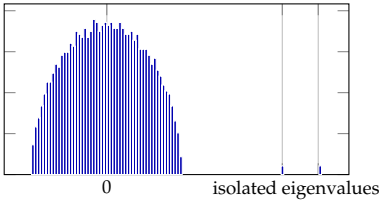


Figure: Visual representation of classification in (**left**) small and (**right**) large dimensions.

$\Rightarrow$ Direct consequence to various distance-based machine learning methods (e.g., kernel spectral clustering)!

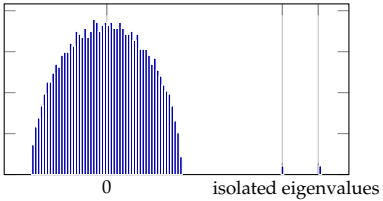
Reminder on kernel spectral clustering

Two-step classification of n data points based on distance kernel matrix $\mathbf{K} \equiv \{f(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)\}_{i,j=1}^n$:

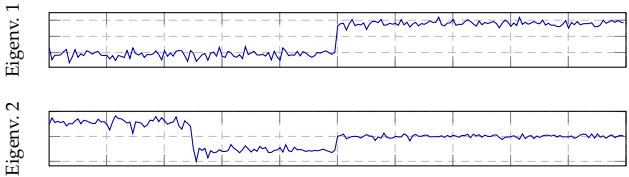


Reminder on kernel spectral clustering

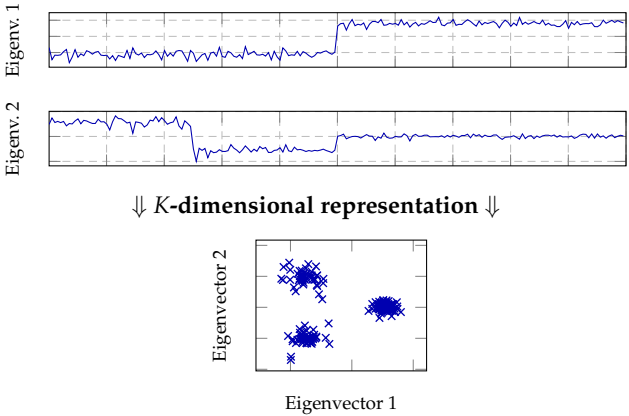
Two-step classification of n data points based on distance kernel matrix $\mathbf{K} \equiv \{f(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)\}_{i,j=1}^n$:



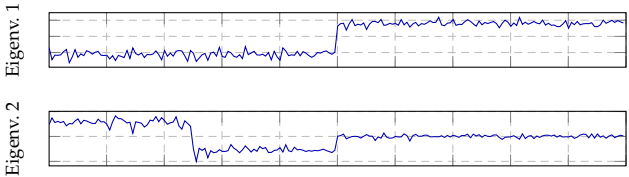
Reminder on kernel spectral clustering



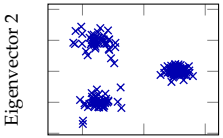
Reminder on kernel spectral clustering



Reminder on kernel spectral clustering



$\Downarrow$ K-**dimensional representation** $\Downarrow$



$\Downarrow$

EM or k-means clustering.

(Three classes/clusters in this example.)

Visualization of kernel matrices for large dimensional Gaussian data

Objective: "cluster" Gaussian data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbf{R}^p$ into $\mathcal{C}_1$ or $\mathcal{C}_2$.

Visualization of kernel matrices for large dimensional Gaussian data

Objective: "cluster" Gaussian data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbf{R}^p$ into $\mathcal{C}_1$ or $\mathcal{C}_2$.

Consider Gaussian kernel matrix $\mathbf{K}_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2p)$ and the second top eigenvectors $\mathbf{v}_2$ for small (**left**) and large (**right**) dimensional data.

Visualization of kernel matrices for large dimensional Gaussian data

Objective: "cluster" Gaussian data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbf{R}^p$ into $\mathcal{C}_1$ or $\mathcal{C}_2$.

Consider Gaussian kernel matrix $\mathbf{K}_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2p)$ and the second top eigenvectors $\mathbf{v}_2$ for small (**left**) and large (**right**) dimensional data.

(a) $p = 5, n = 500$ (b) $p = 250, n = 500$

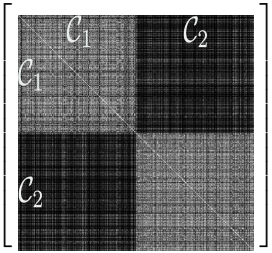
$\mathbf{K} =$ 

Figure: Kernel matrices $\mathbf{K}$ and the second top eigenvectors $\mathbf{v}_2$ for small (left, $p = 5, n = 500$) and large (right, $p = 250, n = 500$) dimensional data.

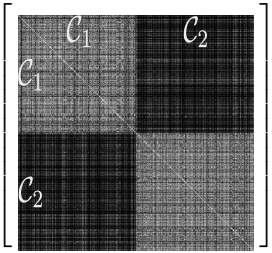
Visualization of kernel matrices for large dimensional Gaussian data

Objective: "cluster" Gaussian data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbf{R}^p$ into $\mathcal{C}_1$ or $\mathcal{C}_2$.

Consider Gaussian kernel matrix $\mathbf{K}_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2p)$ and the second top eigenvectors $\mathbf{v}_2$ for small (**left**) and large (**right**) dimensional data.

(a) $p = 5, n = 500$

$\mathbf{K} =$



(b) $p = 250, n = 500$

$\mathbf{K} =$

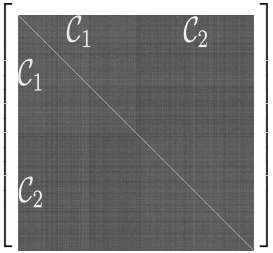


Figure: Kernel matrices $\mathbf{K}$ and the second top eigenvectors $\mathbf{v}_2$ for small (left, $p = 5, n = 500$) and large (right, $p = 250, n = 500$) dimensional data.

Visualization of kernel matrices for large dimensional Gaussian data

Objective: "cluster" Gaussian data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbf{R}^p$ into $\mathcal{C}_1$ or $\mathcal{C}_2$.

Consider Gaussian kernel matrix $\mathbf{K}_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2p)$ and the second top eigenvectors $\mathbf{v}_2$ for small (**left**) and large (**right**) dimensional data.

(a) $p = 5, n = 500$ (b) $p = 250, n = 500$

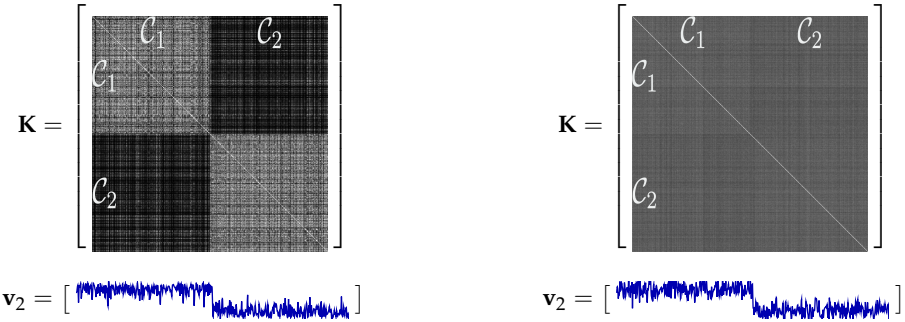
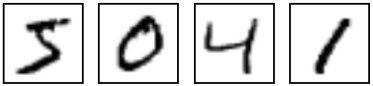


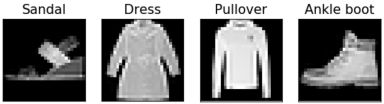
Figure: Kernel matrices $\mathbf{K}$ and the second top eigenvectors $\mathbf{v}_2$ for small (left, $p = 5, n = 500$) and large (right, $p = 250, n = 500$) dimensional data.

Kernel matrices for large dimensional real-world data

(a) MNIST

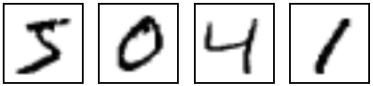


(b) Fashion-MNIST

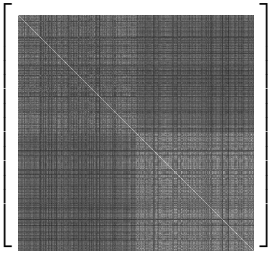


Kernel matrices for large dimensional real-world data

(a) MNIST

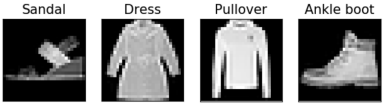


$\mathbf{K} =$

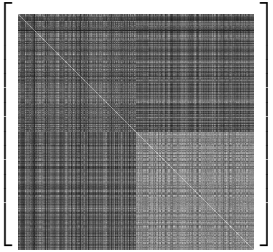


$\mathbf{v}_2 =$

(b) Fashion-MNIST



$\mathbf{K} =$



$\mathbf{v}_2 =$

A spectral viewpoint of large kernel matrices in large dimensions

- "local" linearization of *nonlinear* kernel matrices in large dimensions, e.g., Gaussian kernel matrix $\mathbf{K}_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2p)$ with $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{I}_p$ (e.g., $\mathcal{C}_1 : \mathbf{x}_i = \boldsymbol{\mu}_1 + \mathbf{z}_i$ versus $\mathcal{C}_2 : \mathbf{x}_j = \boldsymbol{\mu}_2 + \mathbf{z}_j$) so that

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2/p \xrightarrow{a.s.} 2, \text{ and } \mathbf{K} = \exp\left(-\frac{2}{2}\right)\left(\mathbf{1}_n\mathbf{1}_n^\mathsf{T} + \frac{1}{p}\mathbf{Z}^\mathsf{T}\mathbf{Z}\right) + g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)\frac{1}{p}\mathbf{j}\mathbf{j}^\mathsf{T} + * + o_{\|\cdot\|}(1)$$

with Gaussian matrix $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{p\times n}$ and $\mathbf{j} = [\mathbf{1}_{n/2}; -\mathbf{1}_{n/2}]$, the class-information vector

A spectral viewpoint of large kernel matrices in large dimensions

- "local" linearization of *nonlinear* kernel matrices in large dimensions, e.g., Gaussian kernel matrix $\mathbf{K}_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2p)$ with $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{I}_p$ (e.g., $\mathcal{C}_1 : \mathbf{x}_i = \boldsymbol{\mu}_1 + \mathbf{z}_i$ versus $\mathcal{C}_2 : \mathbf{x}_j = \boldsymbol{\mu}_2 + \mathbf{z}_j$) so that

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2/p \xrightarrow{a.s.} 2, \text{ and } \mathbf{K} = \exp\left(-\frac{2}{2}\right)\left(\mathbf{1}_n\mathbf{1}_n^\mathsf{T} + \frac{1}{p}\mathbf{Z}^\mathsf{T}\mathbf{Z}\right) + g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)\frac{1}{p}\mathbf{j}\mathbf{j}^\mathsf{T} + * + o_{\|\cdot\|}(1)$$

  with Gaussian matrix $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{p\times n}$ and $\mathbf{j} = [\mathbf{1}_{n/2}; -\mathbf{1}_{n/2}]$, the class-information vector
- accumulated effect of small "hidden" statistical information ($\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|$ in this case)

A spectral viewpoint of large kernel matrices in large dimensions

- "local" linearization of *nonlinear* kernel matrices in large dimensions, e.g., Gaussian kernel matrix $\mathbf{K}_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2p)$ with $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{I}_p$ (e.g., $\mathcal{C}_1 : \mathbf{x}_i = \boldsymbol{\mu}_1 + \mathbf{z}_i$ versus $\mathcal{C}_2 : \mathbf{x}_j = \boldsymbol{\mu}_2 + \mathbf{z}_j$) so that

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2/p \xrightarrow{a.s.} 2, \text{ and } \mathbf{K} = \exp\left(-\frac{2}{2}\right)\left(\mathbf{1}_n\mathbf{1}_n^\mathsf{T} + \frac{1}{p}\mathbf{Z}^\mathsf{T}\mathbf{Z}\right) + g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)\frac{1}{p}\mathbf{j}\mathbf{j}^\mathsf{T} + * + o_{\|\cdot\|}(1)$$

  with Gaussian matrix $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{p\times n}$ and $\mathbf{j} = [\mathbf{1}_{n/2}; -\mathbf{1}_{n/2}]$, the class-information vector
- accumulated effect of small "hidden" statistical information ($\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|$ in this case)

Therefore

- **entry-wise**:

$$\mathbf{K}_{ij} = \exp(-1)\Big(1 + \underbrace{\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j}_{O(p^{-1/2})}\Big) \pm \underbrace{\frac{1}{p}g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)}_{O(p^{-1})} + *, \text{ so that } \frac{1}{p}g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|) \ll \frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j,$$

- "local" linearization of *nonlinear* kernel matrices in large dimensions, e.g., Gaussian kernel matrix $\mathbf{K}_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2p)$ with $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{I}_p$ (e.g., $\mathcal{C}_1 : \mathbf{x}_i = \boldsymbol{\mu}_1 + \mathbf{z}_i$ versus $\mathcal{C}_2 : \mathbf{x}_j = \boldsymbol{\mu}_2 + \mathbf{z}_j$) so that

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2/p \xrightarrow{a.s.} 2, \text{ and } \mathbf{K} = \exp\left(-\frac{2}{2}\right)\left(\mathbf{1}_n\mathbf{1}_n^\mathsf{T} + \frac{1}{p}\mathbf{Z}^\mathsf{T}\mathbf{Z}\right) + g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)\frac{1}{p}\mathbf{j}\mathbf{j}^\mathsf{T} + * + o_{\|\cdot\|}(1)$$

  with Gaussian matrix $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{p\times n}$ and $\mathbf{j} = [\mathbf{1}_{n/2}; -\mathbf{1}_{n/2}]$, the class-information vector
- accumulated effect of small "hidden" statistical information ($\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|$ in this case)

Therefore

- **entry-wise**:

$$\mathbf{K}_{ij} = \exp(-1)\Big(1 + \underbrace{\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j}_{O(p^{-1/2})}\Big) \pm \underbrace{\frac{1}{p}g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)}_{O(p^{-1})} + *, \text{ so that } \frac{1}{p}g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|) \ll \frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j,$$

- **spectrum-wise**: (i) $\|\mathbf{K} - \exp(-1)\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\| \not\to 0$; (ii) $\|\frac{1}{p}\mathbf{Z}^\mathsf{T}\mathbf{Z}\| = O(1)$ and $\|g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)\frac{1}{p}\mathbf{j}\mathbf{j}^\mathsf{T}\| = O(1)$!

A spectral viewpoint of large kernel matrices in large dimensions

- "local" linearization of *nonlinear* kernel matrices in large dimensions, e.g., Gaussian kernel matrix $\mathbf{K}_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2p)$ with $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{I}_p$ (e.g., $\mathcal{C}_1 : \mathbf{x}_i = \boldsymbol{\mu}_1 + \mathbf{z}_i$ versus $\mathcal{C}_2 : \mathbf{x}_j = \boldsymbol{\mu}_2 + \mathbf{z}_j$) so that

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2/p \xrightarrow{a.s.} 2, \text{ and } \mathbf{K} = \exp\left(-\frac{2}{2}\right)\left(\mathbf{1}_n\mathbf{1}_n^\mathsf{T} + \frac{1}{p}\mathbf{Z}^\mathsf{T}\mathbf{Z}\right) + g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)\frac{1}{p}\mathbf{j}\mathbf{j}^\mathsf{T} + * + o_{\|\cdot\|}(1)$$

  with Gaussian matrix $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{p\times n}$ and $\mathbf{j} = [\mathbf{1}_{n/2}; -\mathbf{1}_{n/2}]$, the class-information vector
- accumulated effect of small "hidden" statistical information ($\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|$ in this case)

Therefore

- **entry-wise**:

$$\mathbf{K}_{ij} = \exp(-1)\Big(1 + \underbrace{\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j}_{O(p^{-1/2})}\Big) \pm \underbrace{\frac{1}{p}g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)}_{O(p^{-1})} + *, \text{ so that } \frac{1}{p}g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|) \ll \frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j,$$

- **spectrum-wise**: (i) $\|\mathbf{K} - \exp(-1)\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\| \not\to 0$; (ii) $\|\frac{1}{p}\mathbf{Z}^\mathsf{T}\mathbf{Z}\| = O(1)$ and $\|g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)\frac{1}{p}\mathbf{j}\mathbf{j}^\mathsf{T}\| = O(1)$!
- **Same** phenomenon as the sample covariance example: $[\hat{\mathbf{C}} - \mathbf{C}]_{ij} \to 0 \not\Rightarrow \|\hat{\mathbf{C}} - \mathbf{C}\| \to 0$!

A spectral viewpoint of large kernel matrices in large dimensions

- "local" linearization of *nonlinear* kernel matrices in large dimensions, e.g., Gaussian kernel matrix $\mathbf{K}_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2p)$ with $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{I}_p$ (e.g., $\mathcal{C}_1 : \mathbf{x}_i = \boldsymbol{\mu}_1 + \mathbf{z}_i$ versus $\mathcal{C}_2 : \mathbf{x}_j = \boldsymbol{\mu}_2 + \mathbf{z}_j$) so that

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2/p \xrightarrow{a.s.} 2, \text{ and } \mathbf{K} = \exp\left(-\frac{2}{2}\right)\left(\mathbf{1}_n\mathbf{1}_n^\mathsf{T} + \frac{1}{p}\mathbf{Z}^\mathsf{T}\mathbf{Z}\right) + g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)\frac{1}{p}\mathbf{j}\mathbf{j}^\mathsf{T} + * + o_{\|\cdot\|}(1)$$

  with Gaussian matrix $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{p\times n}$ and $\mathbf{j} = [\mathbf{1}_{n/2}; -\mathbf{1}_{n/2}]$, the class-information vector
- accumulated effect of small "hidden" statistical information ($\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|$ in this case)

Therefore

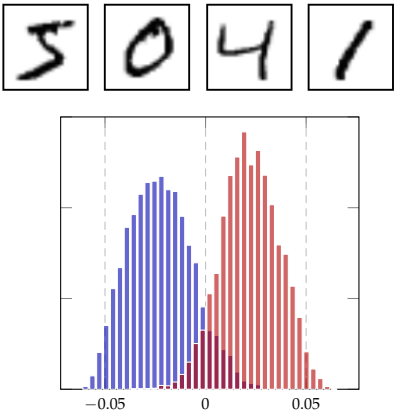
- **entry-wise**:

$$\mathbf{K}_{ij} = \exp(-1)\Big(1 + \underbrace{\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j}_{O(p^{-1/2})}\Big) \pm \underbrace{\frac{1}{p}g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)}_{O(p^{-1})} + *, \text{ so that } \frac{1}{p}g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|) \ll \frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j,$$

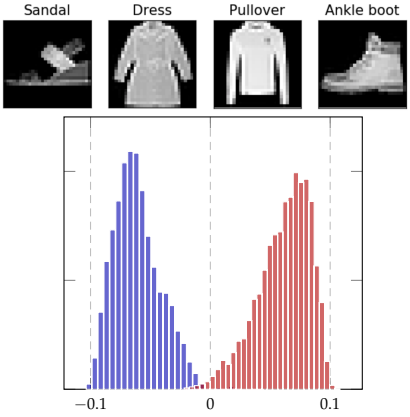
- **spectrum-wise**: (i) $\|\mathbf{K} - \exp(-1)\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\| \not\to 0$; (ii) $\|\frac{1}{p}\mathbf{Z}^\mathsf{T}\mathbf{Z}\| = O(1)$ and $\|g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)\frac{1}{p}\mathbf{j}\mathbf{j}^\mathsf{T}\| = O(1)$!
- **Same** phenomenon as the sample covariance example: $[\hat{\mathbf{C}} - \mathbf{C}]_{ij} \to 0 \not\Rightarrow \|\hat{\mathbf{C}} - \mathbf{C}\| \to 0$!

$\Rightarrow$ With **RMT**, we understand kernel spectral clustering for large dimensional data!

Numerical results



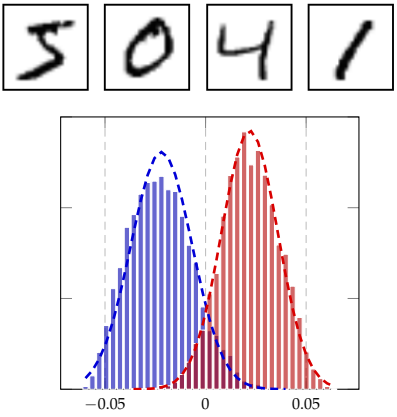
(a) MNIST



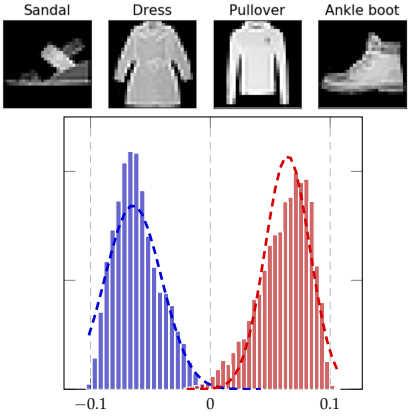
(b) Fashion-MNIST

Figure: Empirical histogram of LS-SVM soft output versus RMT prediction, $n = 2\,048$, $p = 784$, $\gamma = 1$ with Gaussian kernel, for MINST (**left**, 7 versus 9) and Fashion-MNIST (**right**, 8 versus 9) data. Results averaged over 30 runs.

Numerical results



(a) MNIST



(b) Fashion-MNIST

Figure: Empirical histogram of LS-SVM soft output versus RMT prediction, $n = 2\,048$, $p = 784$, $\gamma = 1$ with Gaussian kernel, for MINST (**left**, 7 versus 9) and Fashion-MNIST (**right**, 8 versus 9) data. Results averaged over 30 runs.

Thank you! And some more information

- Find more information in the upcoming book "Random Matrix Methods for Machine Learning" with Cambridge University Press (Online ISBN 9781009128490)

Thank you! And some more information

- Find more information in the upcoming book "Random Matrix Methods for Machine Learning" with Cambridge University Press (Online ISBN 9781009128490)
- with online book draft `https://zhenyu-liao.github.io/pdf/RMT4ML.pdf`

Thank you! And some more information

- Find more information in the upcoming book "Random Matrix Methods for Machine Learning" with Cambridge University Press (Online ISBN 9781009128490)
- with online book draft `https://zhenyu-liao.github.io/pdf/RMT4ML.pdf`
- with online code `https://github.com/Zhenyu-LIAO/RMT4ML`!

Thank you! And some more information

- Find more information in the upcoming book "Random Matrix Methods for Machine Learning" with Cambridge University Press (Online ISBN 9781009128490)
- with online book draft `https://zhenyu-liao.github.io/pdf/RMT4ML.pdf`
- with online code `https://github.com/Zhenyu-LIAO/RMT4ML`!
- and exercise solution `https://zhenyu-liao.github.io/pdf/RMT4ML_solution.pdf`

Thank you! And some more information

- Find more information in the upcoming book "Random Matrix Methods for Machine Learning" with Cambridge University Press (Online ISBN 9781009128490)
- with online book draft `https://zhenyu-liao.github.io/pdf/RMT4ML.pdf`
- with online code `https://github.com/Zhenyu-LIAO/RMT4ML`!
- and exercise solution `https://zhenyu-liao.github.io/pdf/RMT4ML_solution.pdf`

Thank you! Q & A?