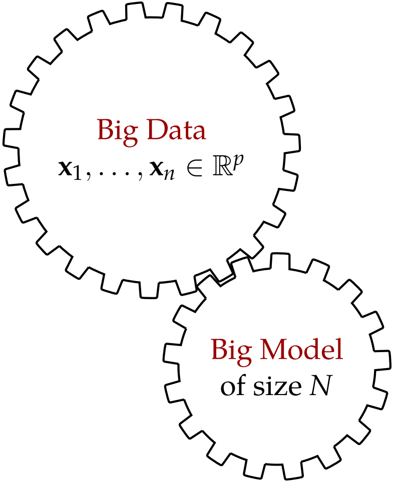
Probability and Stochastic Process II: Random Matrix Theory and Applications Lecture 1: Introduction

Zhenyu Liao, Tiebin Mi, Caiming Qiu

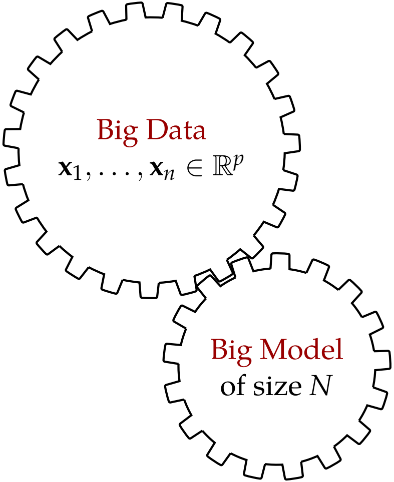
School of Electronic Information and Communications (EIC)
Huazhong University of Science and Technology (HUST)

March 1, 2023

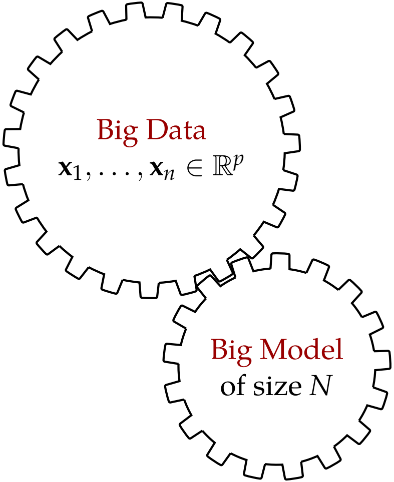
Outline



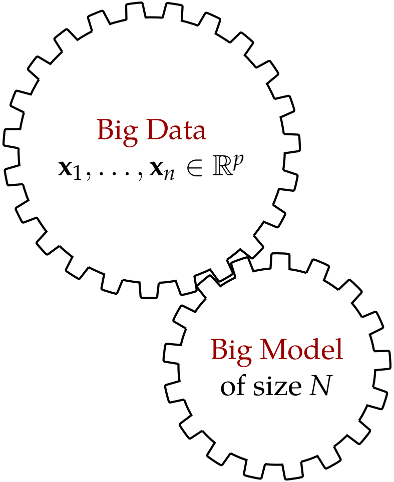
- **Big Data era**: exploit large n, p, N
- counterintuitive phenomena different from classical asymptotics statistics
- complete change of understanding of many methods in statistics, machine learning, signal processing, and wireless communications
- **Random Matrix Theory (RMT)** provides the tools!



- **Big Data era**: exploit large n, p, N
- counterintuitive phenomena different from classical asymptotics statistics
- complete change of understanding of many methods in statistics, machine learning, signal processing, and wireless communications
- **Random Matrix Theory (RMT)** provides the tools!



- **Big Data era**: exploit large n, p, N
- counterintuitive phenomena different from classical asymptotics statistics
- complete change of understanding of many methods in statistics, machine learning, signal processing, and wireless communications
- Random Matrix Theory (RMT) provides the tools!



- **Big Data era**: exploit large n, p, N
- counterintuitive phenomena different from classical asymptotics statistics
- complete change of understanding of many methods in statistics, machine learning, signal processing, and wireless communications
- **Random Matrix Theory (RMT)** provides the tools!

Outline

- **Problem**: estimate covariance $\mathbf{C} \in \mathbb{R}^{p \times p}$ from n data samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$,

- Maximum likelihood sample covariance matrix with entry-wise convergence

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

  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:

$$\boxed{\|\hat{\mathbf{C}} - \mathbf{C}\| \not\to 0, \quad n, p \to \infty} \Rightarrow \text{eigenvalue mismatch and not consistent!}$$

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

» **Problem**: estimate covariance $\mathbf{C} \in \mathbb{R}^{p \times p}$ from n data samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$,

» Maximum likelihood sample covariance matrix with entry-wise convergence

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

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:

$$\boxed{\|\hat{\mathbf{C}} - \mathbf{C}\| \not\to 0, \quad n, p \to \infty} \Rightarrow \text{eigenvalue mismatch and not consistent!}$$

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

» **Problem**: estimate covariance $\mathbf{C} \in \mathbb{R}^{p \times p}$ from n data samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$,

» Maximum likelihood sample covariance matrix with entry-wise convergence

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

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:

$$\boxed{\|\hat{\mathbf{C}} - \mathbf{C}\| \not\to 0, \quad n, p \to \infty} \Rightarrow \text{eigenvalue mismatch and not consistent!}$$

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

» **Problem**: estimate covariance $\mathbf{C} \in \mathbb{R}^{p\times p}$ from n data samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$,

» Maximum likelihood sample covariance matrix with entry-wise convergence

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

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:

$$\boxed{\|\hat{\mathbf{C}} - \mathbf{C}\| \not\to 0, \quad n, p \to \infty} \Rightarrow \text{eigenvalue mismatch and not consistent!}$$

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

- **Problem**: estimate covariance $\mathbf{C} \in \mathbb{R}^{p \times p}$ from n data samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$,
- Maximum likelihood sample covariance matrix with entry-wise convergence

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

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:

$$\boxed{\|\hat{\mathbf{C}} - \mathbf{C}\| \not\to 0, \quad n, p \to \infty} \Rightarrow \text{ eigenvalue mismatch and not consistent!}$$

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

» **Problem**: estimate covariance $\mathbf{C} \in \mathbb{R}^{p \times p}$ from n data samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$,

» Maximum likelihood sample covariance matrix with entry-wise convergence

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

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:

$$\boxed{\|\hat{\mathbf{C}} - \mathbf{C}\| \not\to 0, \quad n, p \to \infty} \Rightarrow \text{ eigenvalue mismatch and not consistent!}$$

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

What about $n = 100p$? For $\mathbf{C} = \mathbf{I}_p$, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$: MP 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 $\pm 2\sqrt{\mathbf{c}} = \pm \mathbf{0.2}$ around the population eigenvalue $\mathbf{1}$.

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

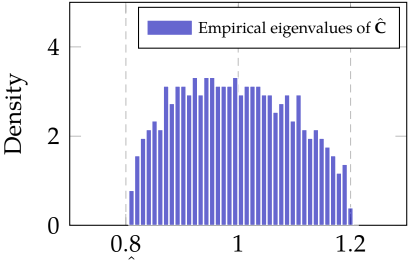


Figure: Eigenvalue distribution of $\hat{\mathbf{C}}$ versus Marčenko-Pastur law, $p = 500$, $n = 50\,000$.

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

» for $\mathbf{n = 100p}$, on a range of $\pm 2\sqrt{c} = \pm \mathbf{0.2}$ around the population eigenvalue $\mathbf{1}$.

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

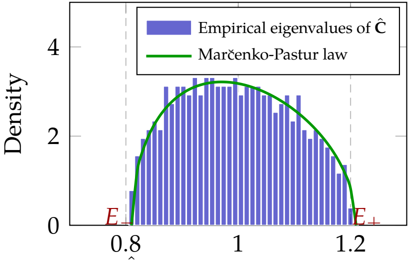


Figure: Eigenvalue distribution of $\hat{\mathbf{C}}$ versus Marčenko-Pastur law, $p = 500$, $n = 50\,000$.

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

» for $\mathbf{n} = \mathbf{100p}$, on a range of $\pm 2\sqrt{c} = \pm 0.2$ around the population eigenvalue $\mathbf{1}$.

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

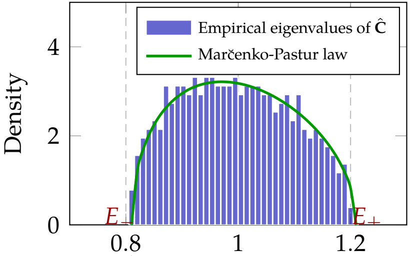


Figure: Eigenvalue distribution of $\hat{\mathbf{C}}$ versus Marčenko-Pastur law, $p = 500$, $n = 50\,000$.

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

» for $\mathbf{n} = \mathbf{100p}$, on a range of $\pm 2\sqrt{\mathbf{c}} = \pm \mathbf{0.2}$ around the population eigenvalue $\mathbf{1}$.

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

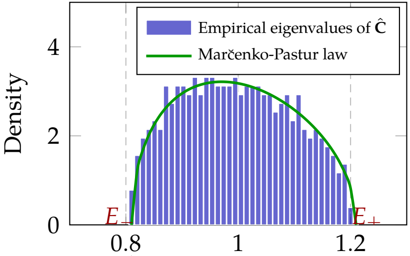


Figure: Eigenvalue distribution of $\hat{\mathbf{C}}$ versus Marčenko-Pastur law, $p = 500$, $n = 50\,000$.

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

» for $\mathbf{n = 100p}$, on a range of $\pm 2\sqrt{\mathbf{c}} = \pm \mathbf{0.2}$ around the population eigenvalue $\mathbf{1}$.

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

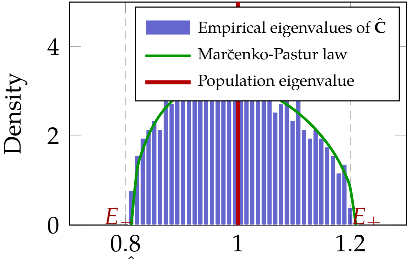


Figure: Eigenvalue distribution of $\hat{\mathbf{C}}$ versus Marčenko-Pastur law, $p = 500$, $n = 50\,000$.

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

Classical large-n asymptotic analysis mostly fails today

- large-n intuition, and many existing popular methods in biology, finance, signal processing, telecommunication, and machine learning, must **fail** even with $n = 100p$!
- **RMT** 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, how RMT can be applied to assess
  - **telecommunication**: code division multiple access (CDMA) technology
  - **signal processing**: generalized likelihood ratio test (GLRT)
  - **machine learning**: principle component analysis (PCA), kernel spectral clustering

Classical large-n asymptotic analysis mostly fails today

- large-n intuition, and many existing popular methods in biology, finance, signal processing, telecommunication, and machine learning, must **fail** even with $n = 100p$!
- **RMT** 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, how RMT can be applied to assess
  - **telecommunication**: code division multiple access (CDMA) technology
  - **signal processing**: generalized likelihood ratio test (GLRT)
  - **machine learning**: principle component analysis (PCA), kernel spectral clustering

Classical large-n asymptotic analysis mostly fails today

- large-n intuition, and many existing popular methods in biology, finance, signal processing, telecommunication, and machine learning, must **fail** even with $n = 100p$!
- **RMT** 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, how RMT can be applied to assess
  - **telecommunication**: code division multiple access (CDMA) technology
  - **signal processing**: generalized likelihood ratio test (GLRT)
  - **machine learning**: principle component analysis (PCA), kernel spectral clustering

Classical large-n asymptotic analysis mostly fails today

- large-n intuition, and many existing popular methods in biology, finance, signal processing, telecommunication, and machine learning, must **fail** even with $n = 100p$!
- **RMT** 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, how RMT can be applied to assess
  - **telecommunication**: code division multiple access (CDMA) technology
  - **signal processing**: generalized likelihood ratio test (GLRT)
  - **machine learning**: principle component analysis (PCA), kernel spectral clustering

Classical large-n asymptotic analysis mostly fails today

- large-n intuition, and many existing popular methods in biology, finance, signal processing, telecommunication, and machine learning, must **fail** even with $n = 100p$!
- **RMT** 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, how RMT can be applied to assess
  - **telecommunication**: code division multiple access (CDMA) technology
  - **signal processing**: generalized likelihood ratio test (GLRT)
  - **machine learning**: principle component analysis (PCA), kernel spectral clustering

Classical large-n asymptotic analysis mostly fails today

- large-n intuition, and many existing popular methods in biology, finance, signal processing, telecommunication, and machine learning, must **fail** even with $n = 100p$!
- **RMT** 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, how RMT can be applied to assess
  - **telecommunication**: code division multiple access (CDMA) technology
  - **signal processing**: generalized likelihood ratio test (GLRT)
  - **machine learning**: principle component analysis (PCA), kernel spectral clustering

Classical large-n asymptotic analysis mostly fails today

- large-n intuition, and many existing popular methods in biology, finance, signal processing, telecommunication, and machine learning, must **fail** even with $n = 100p$!
- **RMT** 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, how RMT can be applied to assess
  - **telecommunication**: code division multiple access (CDMA) technology
  - **signal processing**: generalized likelihood ratio test (GLRT)
  - **machine learning**: principle component analysis (PCA), kernel spectral clustering

Outline

Sample Covariance

RMT for Telecom

RMT for SP

RMT for ML

Application to telecom: performance analysis of CDMA via RMT

- » CDMA in 3G succeeded the TDMA tech in 2G, for which users are successively allocated an exclusive amount of time to exchange data with the APs
- » **major issue**: 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 (long) **spreading code** 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
  - codes not fully orthogonal, more users served, more interference and then less quality of service; but at no time is a user rejected for lack of available resource
- » **Question**: how to evaluate the **capacity** (max achievable transmission data rate) of CDMA network? (which clearly depends on pre-coding strategy)

Application to telecom: performance analysis of CDMA via RMT

» CDMA in 3G succeeded the TDMA tech in 2G, for which users are successively allocated an exclusive amount of time to exchange data with the APs

» **major issue**: 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 (long) **spreading code** 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
- codes not fully orthogonal, more users served, more interference and then less quality of service; but at no time is a user rejected for lack of available resource

» **Question**: how to evaluate the **capacity** (max achievable transmission data rate) of CDMA network? (which clearly depends on pre-coding strategy)

Application to telecom: performance analysis of CDMA via RMT

- CDMA in 3G succeeded the TDMA tech in 2G, for which users are successively allocated an exclusive amount of time to exchange data with the APs
- **major issue**: 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 (long) **spreading code** 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
  - codes not fully orthogonal, more users served, more interference and then less quality of service; but at no time is a user rejected for lack of available resource
- **Question**: how to evaluate the **capacity** (max achievable transmission data rate) of CDMA network? (which clearly depends on pre-coding strategy)

Application to telecom: performance analysis of CDMA via RMT

- CDMA in 3G succeeded the TDMA tech in 2G, for which users are successively allocated an exclusive amount of time to exchange data with the APs
- **major issue**: 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 (long) **spreading code** 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
  - codes not fully orthogonal, more users served, more interference and then less quality of service; but at no time is a user rejected for lack of available resource
- **Question**: how to evaluate the **capacity** (max achievable transmission data rate) of CDMA network? (which clearly depends on pre-coding strategy)

Application to telecom: performance analysis of CDMA via RMT

- CDMA in 3G succeeded the TDMA tech in 2G, for which users are successively allocated an exclusive amount of time to exchange data with the APs
- **major issue**: 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 (long) **spreading code** 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
  - codes not fully orthogonal, more users served, more interference and then less quality of service; but at no time is a user rejected for lack of available resource
- **Question**: how to evaluate the **capacity** (max achievable transmission data rate) of CDMA network? (which clearly depends on pre-coding strategy)

Application to telecom: performance analysis of CDMA via RMT

- CDMA in 3G succeeded the TDMA tech in 2G, for which users are successively allocated an exclusive amount of time to exchange data with the APs
- **major issue**: 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 (long) **spreading code** 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
  - codes not fully orthogonal, more users served, more interference and then less quality of service; but at no time is a user rejected for lack of available resource
- Question: how to evaluate the **capacity** (max achievable transmission data rate) of CDMA network? (which clearly depends on pre-coding strategy)

Application to telecom: performance analysis of CDMA via RMT

- CDMA in 3G succeeded the TDMA tech in 2G, for which users are successively allocated an exclusive amount of time to exchange data with the APs
- **major issue**: 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 (long) **spreading code** 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
  - codes not fully orthogonal, more users served, more interference and then less quality of service; but at no time is a user rejected for lack of available resource
- **Question**: how to evaluate the **capacity** (max achievable transmission data rate) of CDMA network? (which clearly depends on pre-coding strategy)

Orthogonal CDMA versus TDMA

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 are maximal when the orthogonal codes are as long as the number of users n, with system capacity given by

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

with noise power σ^2, $\mathbf{W} \in \mathbb{C}^{n \times n}$ the **orthogonal** CDMA codes ($\mathbf{W}$ unitary), and $\mathbf{G} \equiv \operatorname{diag}\{g_i\}_{i=1}^n$ represents channel **gains** of the users. Note that

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

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

Orthogonal CDMA versus TDMA

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 are maximal when the orthogonal codes are as long as the number of users n, with system capacity given by

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

with noise power σ^2, $\mathbf{W} \in \mathbb{C}^{n \times n}$ the **orthogonal** CDMA codes ($\mathbf{W}$ unitary), and $\mathbf{G} \equiv \mathrm{diag}\{g_i\}_{i=1}^n$ represents channel **gains** of the users. Note that

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

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

Orthogonal CDMA versus TDMA

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 are maximal when the orthogonal codes are as long as the number of users n, with system capacity given by

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

with noise power σ^2, $\mathbf{W} \in \mathbb{C}^{n\times n}$ the **orthogonal** CDMA codes ($\mathbf{W}$ unitary), and $\mathbf{G} \equiv \operatorname{diag}\{g_i\}_{i=1}^n$ represents channel **gains** of the users. Note that

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

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

Orthogonal CDMA versus TDMA

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 are maximal when the orthogonal codes are as long as the number of users n, with system capacity given by

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

with noise power σ^2, $\mathbf{W} \in \mathbb{C}^{n\times n}$ the **orthogonal** CDMA codes ($\mathbf{W}$ unitary), and $\mathbf{G} \equiv \mathrm{diag}\{g_i\}_{i=1}^n$ represents channel **gains** of the users. Note that

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

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

Orthogonal CDMA versus TDMA

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 are maximal when the orthogonal codes are as long as the number of users n, with system capacity given by

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

with noise power σ^2, $\mathbf{W} \in \mathbb{C}^{n\times n}$ the **orthogonal** CDMA codes ($\mathbf{W}$ unitary), and $\mathbf{G} \equiv \operatorname{diag}\{g_i\}_{i=1}^n$ represents channel **gains** of the users. Note that

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

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

Random versus orthogonal CDMA

When it comes to (pseudo-)random CDMA with (random i.i.d. codes), under the same conditions, we have

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

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

Question: C_{rand} as a function of gains $\mathbf{G}$ and (distribution of) codes $\mathbf{X}$?

» (first?) answered by Shami, Tse, and Verdú in [5, 6];

» however capacity expressions not realistically achievable in practice, due to complicated and nonlinear processing algorithms;

» if only linear pre-coders and/or decoders are used, optimal solution:

- Tse and Hanly in [4] for frequency flat channels;
- Evans and Tse in [2] for frequency selective channels;
- Li and Verdú [3] for reduced-rank LMMSE decoders, etc.

Random versus orthogonal CDMA

When it comes to (pseudo-)random CDMA with (random i.i.d. codes), under the same conditions, we have

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

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

Question: C_{rand} as a function of gains $\mathbf{G}$ and (distribution of) codes $\mathbf{X}$?

- (first?) answered by Shami, Tse, and Verdú in [5, 6];
- however capacity expressions not realistically achievable in practice, due to complicated and nonlinear processing algorithms;
- if only linear pre-coders and/or decoders are used, optimal solution:
  - Tse and Hanly in [4] for frequency flat channels;
  - Evans and Tse in [2] for frequency selective channels;
  - Li and Verdú [3] for reduced-rank LMMSE decoders, etc.

Random versus orthogonal CDMA

When it comes to (pseudo-)random CDMA with (random i.i.d. codes), under the same conditions, we have

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

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

Question: C_{rand} as a function of gains $\mathbf{G}$ and (distribution of) codes $\mathbf{X}$?

- (first?) answered by Shami, Tse, and Verdú in [5, 6];
- however capacity expressions not realistically achievable in practice, due to complicated and nonlinear processing algorithms;
- if only linear pre-coders and/or decoders are used, optimal solution:
  - Tse and Hanly in [4] for frequency flat channels;
  - Evans and Tse in [2] for frequency selective channels;
  - Li and Verdú [3] for reduced-rank LMMSE decoders, etc.

Random versus orthogonal CDMA

When it comes to (pseudo-)random CDMA with (random i.i.d. codes), under the same conditions, we have

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

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

Question: $C_{\rm rand}$ as a function of gains $\mathbf{G}$ and (distribution of) codes $\mathbf{X}$?

» (first?) answered by Shami, Tse, and Verdú in [5, 6];

» however capacity expressions not realistically achievable in practice, due to complicated and nonlinear processing algorithms;

» if only linear pre-coders and/or decoders are used, optimal solution:

- Tse and Hanly in [4] for frequency flat channels;
- Evans and Tse in [2] for frequency selective channels;
- Li and Verdú [3] for reduced-rank LMMSE decoders, etc.

Random versus orthogonal CDMA

When it comes to (pseudo-)random CDMA with (random i.i.d. codes), under the same conditions, we have

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

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

Question: $C_{\rm rand}$ as a function of gains $\mathbf{G}$ and (distribution of) codes $\mathbf{X}$?

» (first?) answered by Shami, Tse, and Verdú in [5, 6];

» however capacity expressions not realistically achievable in practice, due to complicated and nonlinear processing algorithms;

» if only linear pre-coders and/or decoders are used, optimal solution:
  - Tse and Hanly in [4] for frequency flat channels;
  - Evans and Tse in [2] for frequency selective channels;
  - Li and Verdú [3] for reduced-rank LMMSE decoders, etc.

Outline

Signal sensing using multi-dimensional sensor arrays

Motivation:

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

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**

Signal sensing using multi-dimensional sensor arrays

Motivation:

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

» high rate communications with finite power budget, need **frequency multiplexing**

» **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

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**

Signal sensing using multi-dimensional sensor arrays

Motivation:

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

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**

Signal sensing using multi-dimensional sensor arrays

Motivation:

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

» high rate communications with finite power budget, need **frequency multiplexing**

» **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

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**

Signal sensing using multi-dimensional sensor arrays

Motivation:

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

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**

Signal sensing using multi-dimensional sensor arrays

Motivation:

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

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**

Signal sensing using multi-dimensional sensor arrays

Motivation:

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

» high rate communications with finite power budget, need **frequency multiplexing**

» **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

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 $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$ with i.i.d. columns $\mathbf{x}_i \in \mathbb{R}^p$ received by array of p sensors, signal decision 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 $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{p \times n}$, $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, $\mathbf{a} \in \mathbb{R}^p$ deterministic of 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, and $\sigma > 0$. Denote $c = p/n > 0$.

» observation of either zero-mean Gaussian **noise** $\sigma \mathbf{z}_i$ of power σ^2, or deterministic **information** vector $\mathbf{a}$ modulated by an added scalar (random) **signal** s_i (e.g., ± 1).

» If $\mathbf{a}$, σ, and statistics of s_i are known, the decision-optimal Neyman-Pearson () test:

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

for some $\alpha > 0$ controlling the Type I and II error rates.

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

System model: let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$ with i.i.d. columns $\mathbf{x}_i \in \mathbb{R}^p$ received by array of p sensors, signal decision 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 $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{p \times n}$, $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, $\mathbf{a} \in \mathbb{R}^p$ deterministic of 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, and $\sigma > 0$. Denote $c = p/n > 0$.

- observation of either zero-mean Gaussian **noise** $\sigma \mathbf{z}_i$ of power σ^2, or deterministic **information** vector $\mathbf{a}$ modulated by an added scalar (random) **signal** s_i (e.g., ± 1).
- If $\mathbf{a}$, σ, and statistics of s_i are known, the decision-optimal Neyman-Pearson () test:

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

for some $\alpha > 0$ controlling the Type I and II error rates.

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

System model: let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$ with i.i.d. columns $\mathbf{x}_i \in \mathbb{R}^p$ received by array of p sensors, signal decision 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 $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{p \times n}$, $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, $\mathbf{a} \in \mathbb{R}^p$ deterministic of 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, and $\sigma > 0$. Denote $c = p/n > 0$.

- observation of either zero-mean Gaussian **noise** $\sigma \mathbf{z}_i$ of power σ^2, or deterministic **information** vector $\mathbf{a}$ modulated by an added scalar (random) **signal** s_i (e.g., ± 1).
- If $\mathbf{a}$, σ, and statistics of s_i are known, the decision-optimal Neyman-Pearson () test:

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

for some $\alpha > 0$ controlling the Type I and II error rates.

Hypothesis testing 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 case of $\mathbf{a}$ fully unknown, one may resort to a **generalized likelihood ratio test** (GLRT) defined as

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

- Gaussian noise and signal s_i, GLRT has an explicit expression as a monotonous increasing function of $\|\mathbf{X}\mathbf{X}^\mathsf{T}\| / \operatorname{tr}(\mathbf{X}\mathbf{X}^\mathsf{T})$, test equivalent to, for some known f,

$$T_p \equiv \frac{\|\mathbf{X}\mathbf{X}^\mathsf{T}\|}{\operatorname{tr}(\mathbf{X}\mathbf{X}^\mathsf{T})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} f(\alpha).$$

- 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 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 case of $\mathbf{a}$ fully unknown, one may resort to a **generalized likelihood ratio test** (GLRT) defined as

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

- Gaussian noise and signal s_i, GLRT has an explicit expression as a monotonous increasing function of $\|\mathbf{X}\mathbf{X}^\mathsf{T}\| / \operatorname{tr}(\mathbf{X}\mathbf{X}^\mathsf{T})$, test equivalent to, for some known f,

$$T_p \equiv \frac{\|\mathbf{X}\mathbf{X}^\mathsf{T}\|}{\operatorname{tr}(\mathbf{X}\mathbf{X}^\mathsf{T})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} f(\alpha).$$

- 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 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 case of $\mathbf{a}$ fully unknown, one may resort to a **generalized likelihood ratio test** (GLRT) defined as

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

- Gaussian noise and signal s_i, GLRT has an explicit expression as a monotonous increasing function of $\|\mathbf{X}\mathbf{X}^\mathsf{T}\|/\operatorname{tr}(\mathbf{X}\mathbf{X}^\mathsf{T})$, test equivalent to, for some known f,

$$T_p \equiv \frac{\|\mathbf{X}\mathbf{X}^\mathsf{T}\|}{\operatorname{tr}(\mathbf{X}\mathbf{X}^\mathsf{T})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} f(\alpha).$$

- 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 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 case of $\mathbf{a}$ fully unknown, one may resort to a **generalized likelihood ratio test** (GLRT) defined as

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

- Gaussian noise and signal s_i, GLRT has an explicit expression as a monotonous increasing function of $\|\mathbf{X}\mathbf{X}^\mathsf{T}\| / \operatorname{tr}(\mathbf{X}\mathbf{X}^\mathsf{T})$, test equivalent to, for some known f,

$$T_p \equiv \frac{\|\mathbf{X}\mathbf{X}^\mathsf{T}\|}{\operatorname{tr}(\mathbf{X}\mathbf{X}^\mathsf{T})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} f(\alpha).$$

- 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 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 case of $\mathbf{a}$ fully unknown, one may resort to a **generalized likelihood ratio test** (GLRT) defined as

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

» Gaussian noise and signal s_i, GLRT has an explicit expression as a monotonous increasing function of $\|\mathbf{X}\mathbf{X}^\mathsf{T}\| / \operatorname{tr}(\mathbf{X}\mathbf{X}^\mathsf{T})$, test equivalent to, for some known f,

$$T_p \equiv \frac{\|\mathbf{X}\mathbf{X}^\mathsf{T}\|}{\operatorname{tr}(\mathbf{X}\mathbf{X}^\mathsf{T})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} f(\alpha).$$

» 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

To set a maximum false alarm rate (or Type I error) of $r > 0$ for large n, p, according to RMT, one must choose a threshold $f(\alpha)$ for T_p:

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

with μ_{TW_1} the Tracy-Widom distribution in RMT.

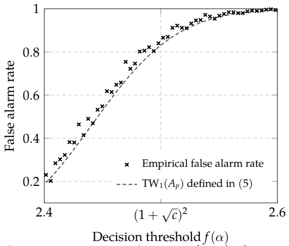


Figure: Comparison between empirical false alarm rates and $\mathrm{TW}_1(A_p)$ for A_p of the form in (5), as a function of the threshold $f(\alpha) \in [(1+\sqrt{c})^2 - 5n^{-2/3}, (1+\sqrt{c})^2 + 5n^{-2/3}]$, for $p = 256$, $n = 1\,024$ and $\sigma = 1$.

Outline

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

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

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

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

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

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

» 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 $\mathcal{C}_a, \mathcal{C}_b$!

[a] Couillet, Liao and Mai "Classification asymptotics in the random matrix regime" (2018).

"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[a]

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

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 $\mathcal{C}_a, \mathcal{C}_b$!

[a] Couillet, Liao and Mai "Classification asymptotics in the random matrix regime" (2018).

"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[a]

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

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 $\mathcal{C}_a, \mathcal{C}_b$!

[a] Couillet, Liao and Mai "Classification asymptotics in the random matrix regime" (2018).

Loss of relevance of Euclidean distance: visual representation

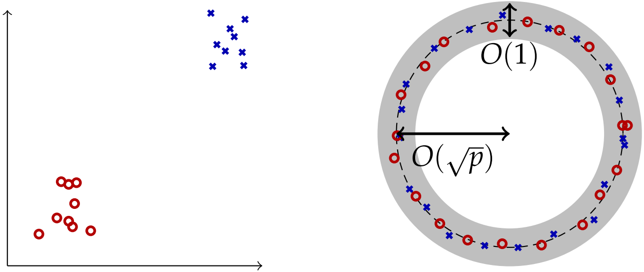


Figure: Visual representation of classification in (**left**) small and (**right**) large dimensions.

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

Loss of relevance of Euclidean distance: visual representation

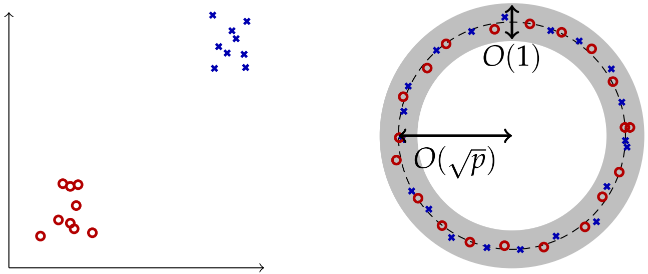
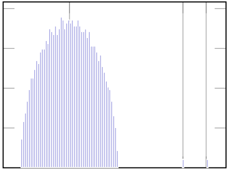


Figure: Visual representation of classification in (**left**) small and (**right**) large dimensions.

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

Reminder on kernel spectral clustering

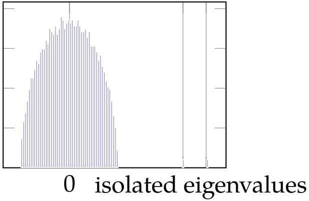
Two-step classification of n data points with distance kernel $\mathbf{K} \equiv \{f(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)\}_{i,j=1}^n$:



⇓ **Top eigenvectors** ⇓

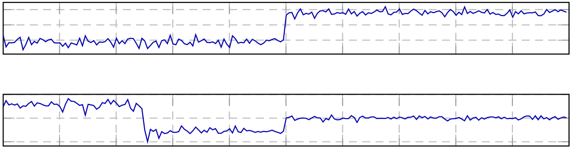
Reminder on kernel spectral clustering

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



$\Downarrow$ **Top eigenvectors** $\Downarrow$

Reminder on kernel spectral clustering



⇓ K-dimensional representation ⇓

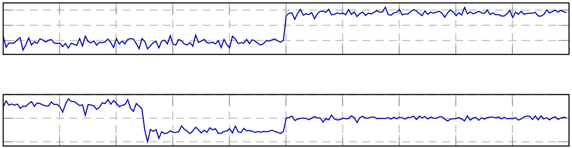
Eig. 2

Eig. 1

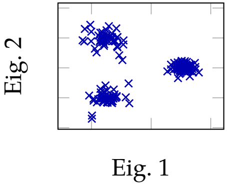
⇓

EM or k-means clustering

Reminder on kernel spectral clustering



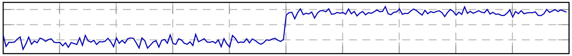
$\Downarrow$ **K-dimensional representation** $\Downarrow$

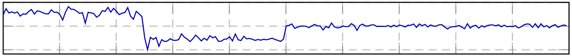


$\Downarrow$

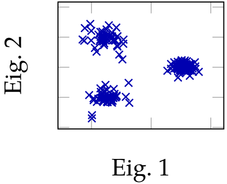
EM or k-means clustering

Reminder on kernel spectral clustering





⇓ K-dimensional representation ⇓

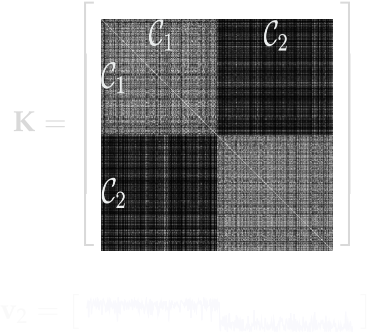


⇓

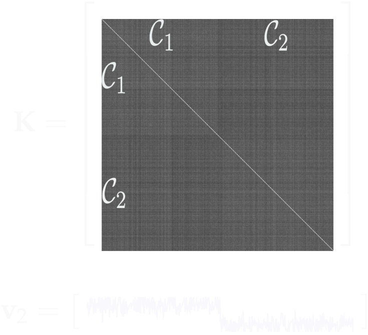
EM or k-means clustering

Cluster Gaussian data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbf{R}^p$ into $\mathcal{C}_1$ or $\mathcal{C}_2$, with second top eigenvectors $\mathbf{v}_2$ of heat kernel $\mathbf{K}_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2p)$, small and large dimensional data.

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

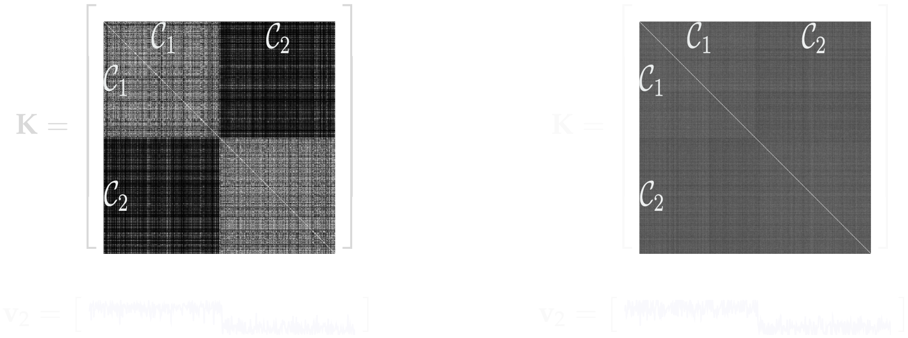


(b) $p = 250, n = 500$



Cluster Gaussian data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbf{R}^p$ into $\mathcal{C}_1$ or $\mathcal{C}_2$, with second top eigenvectors $\mathbf{v}_2$ of heat kernel $\mathbf{K}_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2p)$, small and large dimensional data.

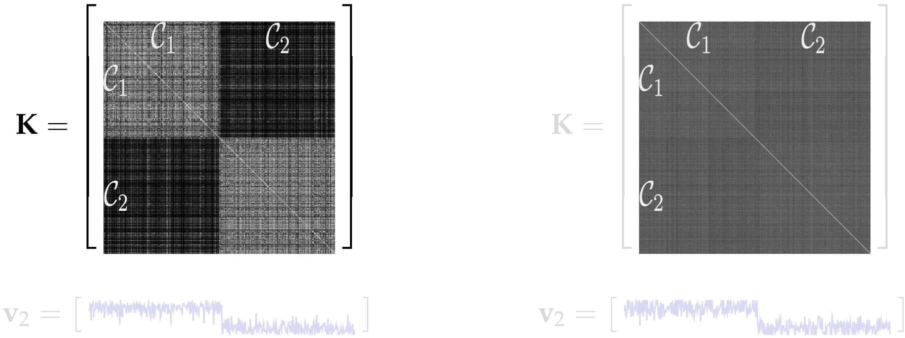
(a) $p = 5, n = 500$ (b) $p = 250, n = 500$



Cluster Gaussian data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbf{R}^p$ into $\mathcal{C}_1$ or $\mathcal{C}_2$, with second top eigenvectors $\mathbf{v}_2$ of heat kernel $\mathbf{K}_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2p)$, small and large dimensional data.

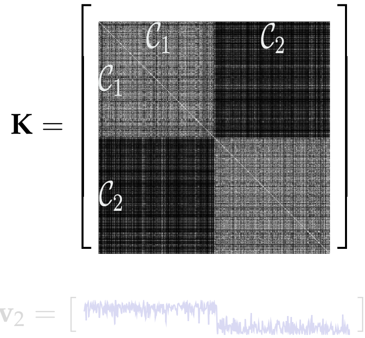
(a) $p = 5, n = 500$

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

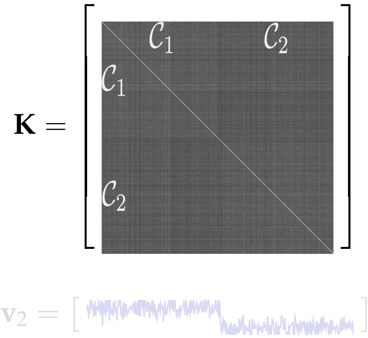


Cluster Gaussian data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbf{R}^p$ into $\mathcal{C}_1$ or $\mathcal{C}_2$, with second top eigenvectors $\mathbf{v}_2$ of heat kernel $\mathbf{K}_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2p)$, small and large dimensional data.

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

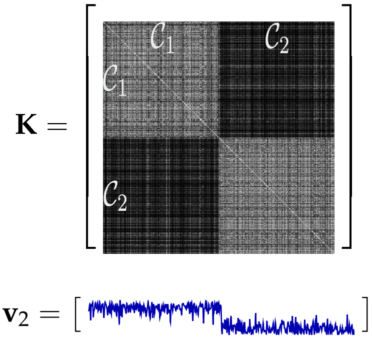


(b) $p = 250, n = 500$

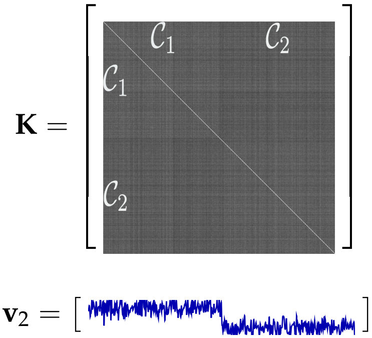


Cluster Gaussian data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbf{R}^p$ into $\mathcal{C}_1$ or $\mathcal{C}_2$, with second top eigenvectors $\mathbf{v}_2$ of heat kernel $\mathbf{K}_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2p)$, small and large dimensional data.

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



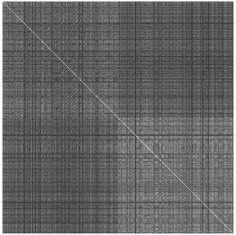
(b) $p = 250$, $n = 500$



Kernel matrices for large dimensional real-world data

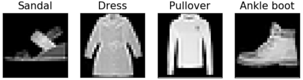
(a) MNIST

$\mathbf{K} =$

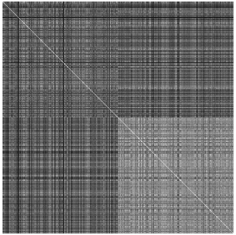


$\mathbf{v}_2 =$

(b) Fashion-MNIST



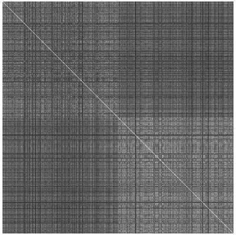
$\mathbf{K} =$

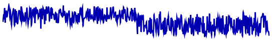


$\mathbf{v}_2 =$

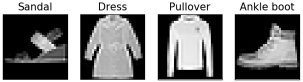
Kernel matrices for large dimensional real-world data

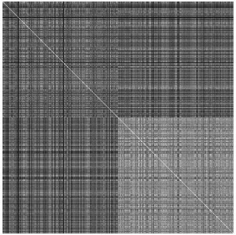
(a) MNIST

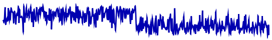
$\mathbf{K} =$ 

$\mathbf{v}_2 =$ 

(b) Fashion-MNIST



$\mathbf{K} =$ 

$\mathbf{v}_2 =$ 

A RMT viewpoint of large kernel matrices

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

» accumulated effect of small "hidden" statistical information ($\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|$ in this case)

A RMT viewpoint of large kernel matrices

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

» accumulated effect of small "hidden" statistical information ($\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|$ in this case)

A RMT viewpoint of large kernel matrices

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**:

- $\|\mathbf{K} - \exp(-1)\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\| \not\to 0$;
- $\|\frac{1}{p}\mathbf{Z}^\mathsf{T}\mathbf{Z}\| = O(1)$ 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!

A RMT viewpoint of large kernel matrices

Therefore

» **entry-wise**:

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

» **spectrum-wise**:

- $\|\mathbf{K} - \exp(-1)\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\| \not\to 0$;
- $\|\frac{1}{p}\mathbf{Z}^\mathsf{T}\mathbf{Z}\| = O(1)$ 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!

A RMT viewpoint of large kernel matrices

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**:

- $\|\mathbf{K} - \exp(-1)\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\| \not\to 0$;
- $\|\frac{1}{p}\mathbf{Z}^\mathsf{T}\mathbf{Z}\| = O(1)$ 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!

A RMT viewpoint of large kernel matrices

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**:

- $\|\mathbf{K} - \exp(-1)\mathbf{1}_n\mathbf{1}_n^{\mathsf{T}}\| \not\to 0$;
- $\|\frac{1}{p}\mathbf{Z}^{\mathsf{T}}\mathbf{Z}\| = O(1)$ 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!

A RMT viewpoint of large kernel matrices

Therefore

» **entry-wise**:

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

» **spectrum-wise**:

- $\|\mathbf{K} - \exp(-1)\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\| \not\to 0$;
- $\|\frac{1}{p}\mathbf{Z}^\mathsf{T}\mathbf{Z}\| = O(1)$ 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!

A RMT viewpoint of large kernel matrices

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**:
  - $\|\mathbf{K} - \exp(-1)\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\| \not\to 0$;
  - $\|\frac{1}{p}\mathbf{Z}^\mathsf{T}\mathbf{Z}\| = O(1)$ 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!

A RMT viewpoint of large kernel matrices

Therefore

» **entry-wise**:

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

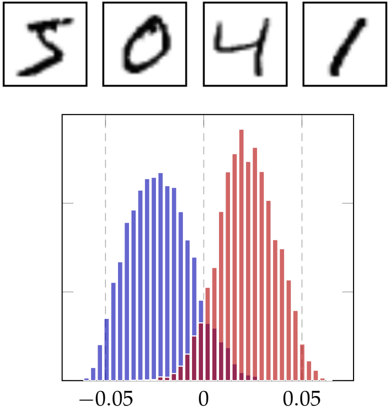
» **spectrum-wise**:

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

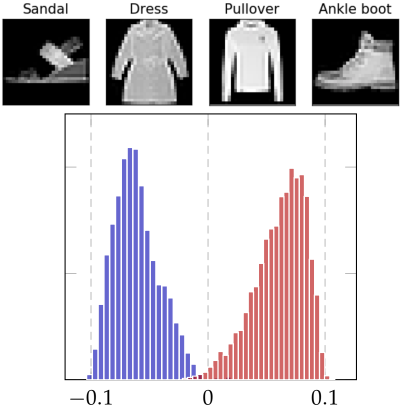
» **Same** phenomenon as the sample covariance example: $[\hat{\mathbf{C}} - \mathbf{C}]_{ij} \to 0 \not\Rightarrow \|\hat{\mathbf{C}} - \mathbf{C}\| \to 0$!

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

Some more numerical results

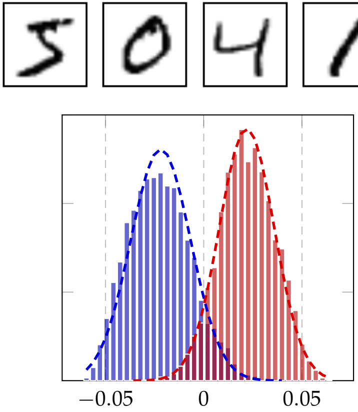


(a) MNIST

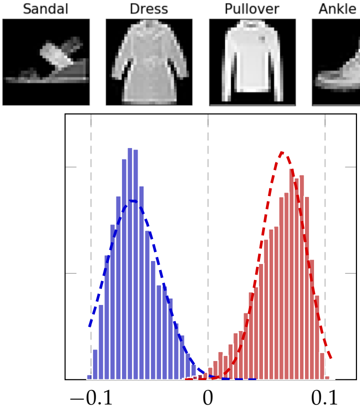


(b) Fashion-MNIST

Some more numerical results



(a) MNIST



(b) Fashion-MNIST

Thank you! And some more information

- Find more information in the monograph "Random Matrix Methods for Machine Learning" with Cambridge University Press
- 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?

Thank you! And some more information

- Find more information in the monograph "Random Matrix Methods for Machine Learning" with Cambridge University Press
- 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?

Thank you! And some more information

- Find more information in the monograph "Random Matrix Methods for Machine Learning" with Cambridge University Press
- 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?

Thank you! And some more information

- Find more information in the monograph "Random Matrix Methods for Machine Learning" with Cambridge University Press
- 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?

Thank you! And some more information

- Find more information in the monograph "Random Matrix Methods for Machine Learning" with Cambridge University Press
- 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?

[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.

[2] J. Evans and D.N.C. Tse. "Large system performance of linear multiuser receivers in multipath fading channels". In: *IEEE Transactions on Information Theory* 46.6 (2000), 2059–2078. DOI: 10.1109/18.868478.

[3] Linbo Li, Antonia M. Tulino and Sergio Verdú. "Design of Reduced-Rank MMSE Multiuser Detectors Using Random Matrix Methods". In: *IEEE Transactions on Information Theory* 50.6 (2004), 986–1008. ISSN: 0018-9448. DOI: 10.1109/tit.2004.828076.

[4] D.N.C. Tse and S.V. Hanly. "Linear multiuser receivers: effective interference, effective bandwidth and user capacity". In: *IEEE Transactions on Information Theory* 45.2 (1999), 641–657. ISSN: 0018-9448. DOI: 10.1109/18.749008.

[5] D.N.C. Tse and S. Verdú. "Optimum asymptotic multiuser efficiency of randomly spread CDMA". In: *IEEE Transactions on Information Theory* 46.7 (2000), 2718–2722.