Statistical Estimation in Algebraically Structured Models

Pan Jing Bin

Final Year Project Midterm Presentation

Outline

Introduction to Statistical Estimation

General setting: Let P be an **unknown** probability distribution on a sample space Ω. Let $X_1, X_2, \cdots, X_n$ be samples that are drawn independently and randomly from Ω according to the probability distribution P.

Introduction to Statistical Estimation

General setting: Let P be an **unknown** probability distribution on a sample space Ω. Let $X_1, X_2, \cdots, X_n$ be samples that are drawn independently and randomly from Ω according to the probability distribution P.

Problem: How to **accurately** recover the probability distribution P from the samples $X_1, X_2 \cdots, X_n$?

Introduction to Statistical Estimation

General setting: Let P be an **unknown** probability distribution on a sample space Ω. Let $X_1, X_2, \cdots, X_n$ be samples that are drawn independently and randomly from Ω according to the probability distribution P.

Problem: How to **accurately** recover the probability distribution P from the samples $X_1, X_2 \cdots, X_n$?

Unrestricted hypothesis space: The space of all probability distributions on Ω is **huge!**

Introduction to Statistical Estimation

General setting: Let P be an **unknown** probability distribution on a sample space Ω. Let $X_1, X_2, \cdots, X_n$ be samples that are drawn independently and randomly from Ω according to the probability distribution P.

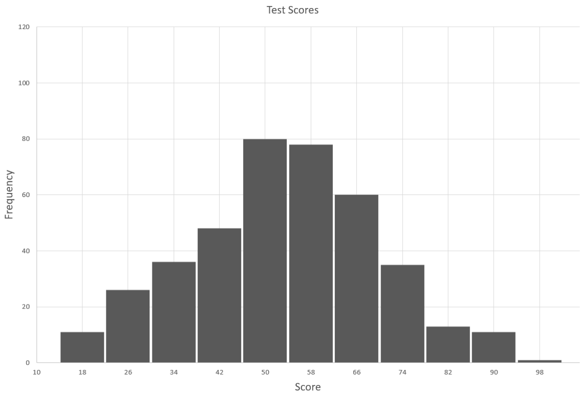
Problem: How to **accurately** recover the probability distribution P from the samples $X_1, X_2 \cdots, X_n$?

Unrestricted hypothesis space: The space of all probability distributions on Ω is **huge!**

Restricted hypothesis space: In real world statistical estimation, we already have a rough idea of what the probability distribution P should look like.

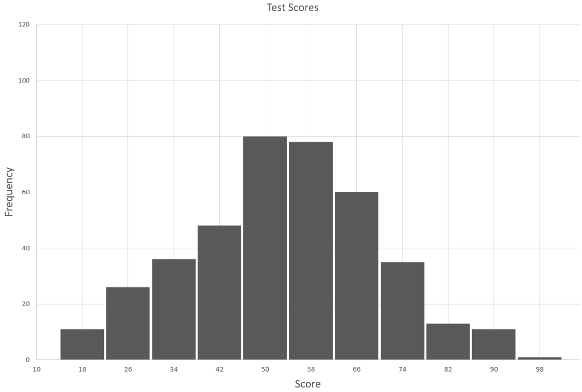
Toy Example

Final exam for a large course ($\sim$ 400 students):



Toy Example

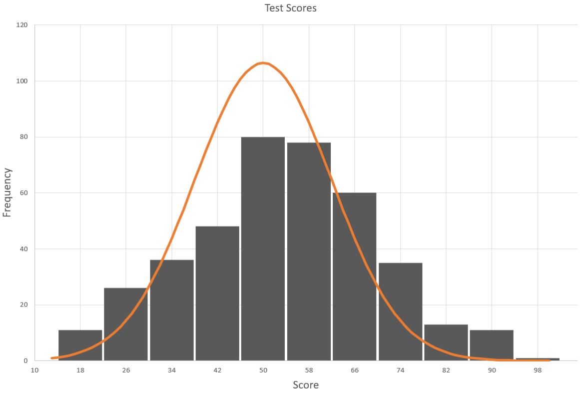
Final exam for a large course ($\sim$ 400 students):



The distribution is expected to be approximately **normal.**

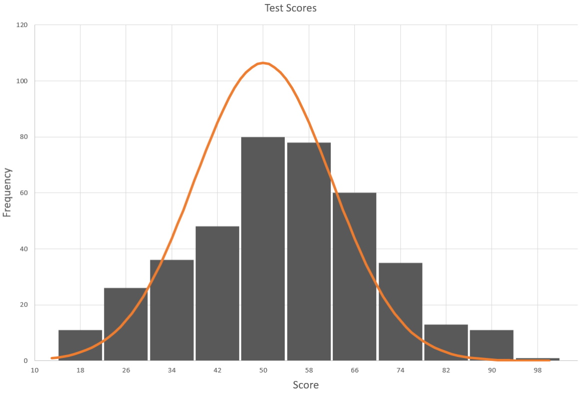
Toy Example

The normal distribution is completely determined by its **mean** and **variance**.



Toy Example

The normal distribution is completely determined by its **mean** and **variance**.



Finding the "best fit" curve on 400 samples is now simply an optimization problem in two parameters.

Algebraically Structured Model

General setting: Let $\theta \in \mathbb{R}^d$ be an **unknown** vector. Consider two types of corruptions on θ:

$$P_\theta \sim G\theta + \xi \tag{1}$$

Algebraically Structured Model

General setting: Let $\theta \in \mathbb{R}^d$ be an **unknown** vector. Consider two types of corruptions on θ:

$$P_\theta \sim G\theta + \xi \tag{1}$$

1. Additive Gaussian noise:

$$\xi \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_d).$$

Used to model many small, independent sources of randomness.

Algebraically Structured Model

General setting: Let $\theta \in \mathbb{R}^d$ be an **unknown** vector. Consider two types of corruptions on θ:

$$P_\theta \sim G\theta + \xi \tag{1}$$

1. Additive Gaussian noise:

   $$\xi \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_d).$$

   Used to model many small, independent sources of randomness.

2. Random rotation: G is drawn **uniformly** at random from a compact subgroup $\mathcal{G}$ of the orthogonal group $O(d)$ given by

   $$O(d) := \left\{ \boldsymbol{A} \in \mathrm{Mat}_{d\times d}(\mathbb{R}) \ : \ \boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I}_d = \boldsymbol{A}^T\boldsymbol{A} \right\}.$$

Algebraically Structured Model

General setting: Let $\theta \in \mathbb{R}^d$ be an **unknown** vector. Consider two types of corruptions on θ:

$$P_\theta \sim G\theta + \xi \qquad (1)$$

1. Additive Gaussian noise:
$$\xi \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_d).$$
Used to model many small, independent sources of randomness.

2. Random rotation: G is drawn **uniformly** at random from a compact subgroup $\mathcal{G}$ of the orthogonal group $O(d)$ given by
$$O(d) := \left\{ \boldsymbol{A} \in \mathrm{Mat}_{d\times d}(\mathbb{R}) \;:\; \boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I}_d = \boldsymbol{A}^T\boldsymbol{A} \right\}.$$
Motivation: Image reconstruction (computer vision), molecule structure determination (physics, chemistry), etc.

Algebraically Structured Model

General setting: Let $\theta \in \mathbb{R}^d$ be an **unknown** vector. Consider two types of corruptions on θ:

$$P_\theta \sim G\theta + \xi \tag{1}$$

1. Additive Gaussian noise:
$$\xi \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_d).$$
Used to model many small, independent sources of randomness.

2. Random rotation: G is drawn **uniformly** at random from a compact subgroup $\mathcal{G}$ of the orthogonal group $O(d)$ given by
$$O(d) := \left\{ \boldsymbol{A} \in \mathrm{Mat}_{d\times d}(\mathbb{R}) \ : \ \boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I}_d = \boldsymbol{A}^T\boldsymbol{A} \right\}.$$
Motivation: Image reconstruction (computer vision), molecule structure determination (physics, chemistry), etc.

Given independent samples $Y_1, \cdots, Y_n$ drawn from $\mathbb{R}^d$ according to (1), we want to recover the vector θ. This setup is known as an **algebraically structured model**.

Algebraically Structured Model

Problem Setup:

$$Y_i = G\theta + \xi$$

where $\theta \in \mathbb{R}^d, \quad G \in \mathcal{G} \subseteq O(d)$ and $\xi \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_d)$.

Additional assumption: The observations Y_i are **noisy** (i.e. $\|\theta\|^2 / \sigma^2$ is low) and precise estimates are **difficult**.

Algebraically Structured Model

Problem Setup:

$$Y_i = G\theta + \xi$$

where $\theta \in \mathbb{R}^d, \quad G \in \mathcal{G} \subseteq O(d)$ and $\xi \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_d)$.

Additional assumption: The observations Y_i are **noisy** (i.e. $\|\theta\|^2 / \sigma^2$ is low) and precise estimates are **difficult**.

But the **noise level** σ can be **decreased** with improvements in technology.

Algebraically Structured Model

Problem Setup:

$$Y_i = G\theta + \xi$$

where $\theta \in \mathbb{R}^d, \quad G \in \mathcal{G} \subseteq O(d)$ and $\xi \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_d)$.

Additional assumption: The observations Y_i are **noisy** (i.e. $\|\theta\|^2 / \sigma^2$ is low) and precise estimates are **difficult**.

But the **noise level** σ can be **decreased** with improvements in technology.

Problem of interest: How does the **performance** (e.g. rate of convergence) of the algorithm depend on σ?

Algebraically Structured Model

Problem Setup:

$$Y_i = G\theta + \xi$$

where $\theta \in \mathbb{R}^d$, $\quad G \in \mathcal{G} \subseteq O(d)$ and $\xi \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_d)$.

Additional assumption: The observations Y_i are **noisy** (i.e. $\|\theta\|^2 / \sigma^2$ is low) and precise estimates are **difficult**.

But the **noise level** σ can be **decreased** with improvements in technology.

Problem of interest: How does the **performance** (e.g. rate of convergence) of the algorithm depend on σ?

To build an **abstract framework**, we first need some new mathematical tools.

Kullback-Leibler Divergence

Each vector $\theta \in \mathbb{R}^d$ defines a corresponding **probability distribution** P_θ according to

$$G\theta + \xi.$$

If two vectors θ and ϕ define similar probability distributions, then we expect algorithms to have a hard time distinguishing between them (and vice versa).

Kullback-Leibler Divergence

Each vector $\theta \in \mathbb{R}^d$ defines a corresponding **probability distribution** P_θ according to

$$G\theta + \xi.$$

If two vectors θ and ϕ define similar probability distributions, then we expect algorithms to have a hard time distinguishing between them (and vice versa).

Question: How to mathematically quantify "the level of similarity" between two probability distributions?

Kullback-Leibler Divergence

Each vector $\theta \in \mathbb{R}^d$ defines a corresponding **probability distribution** P_θ according to

$$G\theta + \xi.$$

If two vectors θ and ϕ define similar probability distributions, then we expect algorithms to have a hard time distinguishing between them (and vice versa).

Question: How to mathematically quantify "the level of similarity" between two probability distributions?

Answer: Statistical divergences. The **Kullback-Leibler Divergence** between two probability distributions P_θ and P_ϕ (with densities f_θ and f_ϕ respectively) is defined to be

$$D(P_\theta \parallel P_\phi) := \int_{\mathbb{R}^d} f_\theta(x) \log \frac{f_\theta(x)}{f_\phi(x)}\, dx.$$

Kullback-Leibler Divergence

$$D(P_\theta \parallel P_\phi) := \int_{\mathbb{R}^d} f_\theta(x) \log \frac{f_\theta(x)}{f_\phi(x)} \, dx.$$

Kullback-Leibler Divergence

$$D(P_\theta \parallel P_\phi) := \int_{\mathbb{R}^d} f_\theta(x) \log \frac{f_\theta(x)}{f_\phi(x)} \, dx.$$

Two preliminary observations:

1. If $P_\theta = P_\phi$, then $D(P_\theta \parallel P_\phi) = 0$;

Kullback-Leibler Divergence

$$D(P_\theta \parallel P_\phi) := \int_{\mathbb{R}^d} f_\theta(x) \log \frac{f_\theta(x)}{f_\phi(x)} \, dx.$$

Two preliminary observations:

1. If $P_\theta = P_\phi$, then $D(P_\theta \parallel P_\phi) = 0$;
2. Otherwise, $D(P_\theta \parallel P_\phi) > 0$;

Kullback-Leibler Divergence

$$D(P_\theta \parallel P_\phi) := \int_{\mathbb{R}^d} f_\theta(x) \log \frac{f_\theta(x)}{f_\phi(x)} \, dx.$$

Two preliminary observations:

1. If $P_\theta = P_\phi$, then $D(P_\theta \parallel P_\phi) = 0$;
2. Otherwise, $D(P_\theta \parallel P_\phi) > 0$;

The distribution P_θ represents the **true** distribution and P_ϕ represents another probability distribution. In general, the **larger** the KL-divergence, the **easier** it is to distinguish between the two distributions.

Kullback-Leibler Divergence

$$D(P_\theta \parallel P_\phi) := \int_{\mathbb{R}^d} f_\theta(x) \log \frac{f_\theta(x)}{f_\phi(x)} \, dx.$$

Two preliminary observations:

1. If $P_\theta = P_\phi$, then $D(P_\theta \parallel P_\phi) = 0$;
2. Otherwise, $D(P_\theta \parallel P_\phi) > 0$;

The distribution P_θ represents the **true** distribution and P_ϕ represents another probability distribution. In general, the **larger** the KL-divergence, the **easier** it is to distinguish between the two distributions.

The KL-divergence has deep connections in many different areas (e.g. information theory, machine learning etc).

Kullback-Leibler Divergence

$$D(P_\theta \parallel P_\phi) := \int_{\mathbb{R}^d} f_\theta(x) \log \frac{f_\theta(x)}{f_\phi(x)}\, dx.$$

Two preliminary observations:

1. If $P_\theta = P_\phi$, then $D(P_\theta \parallel P_\phi) = 0$;
2. Otherwise, $D(P_\theta \parallel P_\phi) > 0$;

The distribution P_θ represents the **true** distribution and P_ϕ represents another probability distribution. In general, the **larger** the KL-divergence, the **easier** it is to distinguish between the two distributions.

The KL-divergence has deep connections in many different areas (e.g. information theory, machine learning etc). Many powerful **passages**

$$\text{Performance of estimators} \longleftrightarrow D(P_\theta \parallel P_\phi)$$

have already been established.

Kullback-Leibler Divergence

$$D(P_\theta \parallel P_\phi) := \int_{\mathbb{R}^d} f_\theta(x) \log \frac{f_\theta(x)}{f_\phi(x)} \, dx.$$

Two preliminary observations:

1. If $P_\theta = P_\phi$, then $D(P_\theta \parallel P_\phi) = 0$;
2. Otherwise, $D(P_\theta \parallel P_\phi) > 0$;

The distribution P_θ represents the **true** distribution and P_ϕ represents another probability distribution. In general, the **larger** the KL-divergence, the **easier** it is to distinguish between the two distributions.

The KL-divergence has deep connections in many different areas (e.g. information theory, machine learning etc). Many powerful **passages**

$$\text{Performance of estimators} \longleftrightarrow D(P_\theta \parallel P_\phi)$$

have already been established.

Gauging the performance of estimators essentially boils down to controlling the quantity $D(P_\theta \parallel P_\phi)$.

Summary Statistics

Instead of looking at the **raw data** of test results,

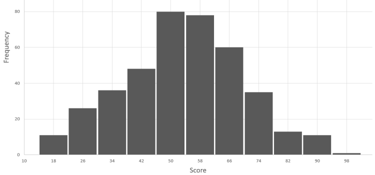
54	59	61	27	44	58	24	48
42	46	63	75	57	25	84	30
62	31	39	49	43	58	59	45
15	63	69	63	46	48	45	24
17	36	41	69	66	52	40	47
25	48	29	62	42	84	73	46
42	45	57	47	41	24	1	42
52	62	50	72	45	84	69	44
13	36	44	60	59	59	31	28
34	84	46	22	28	45	96	59
33	36	65	45	65	40	80	38
73	60	39	53	78	73	65	39
28	82	64	55	44	96	26	63
54	70	49	37	38	69	36	27
37	47	20	60	64	31	63	52

Summary Statistics

Instead of looking at the **raw data** of test results,

54	59	61	27	44	58	24	48
42	46	63	75	57	25	84	30
62	31	39	49	43	58	59	45
15	63	69	63	46	48	45	24
17	36	41	69	66	52	40	47
25	48	29	62	42	84	73	46
42	45	57	47	41	24	1	42
52	62	50	72	45	84	69	44
13	36	44	60	59	59	31	28
34	84	46	22	28	45	96	59
33	36	65	45	65	40	80	38
73	60	39	53	78	73	65	39
28	82	64	55	44	96	26	63
54	70	49	37	38	69	36	27
37	47	20	60	64	31	63	52

we look at **simplified representations** such as



as well as **summary statistics** such as

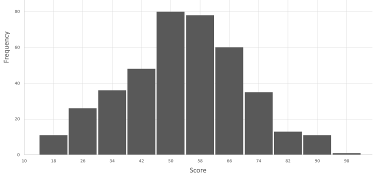
$$\mathbb{E}[X], \quad \mathrm{Var}(X)$$

Summary Statistics

Instead of looking at the **raw data** of test results,

54	59	61	27	44	58	24	48
42	46	63	75	57	25	84	30
62	31	39	49	43	58	59	45
15	63	69	63	46	48	45	24
17	36	41	69	66	52	40	47
25	48	29	62	42	84	73	46
42	45	57	47	41	24	1	42
52	62	50	72	45	84	69	44
13	36	44	60	59	59	31	28
34	84	46	22	28	45	96	59
33	36	65	45	65	40	80	38
73	60	39	53	78	73	65	39
28	82	64	55	44	96	26	63
54	70	49	37	38	69	36	27
37	47	20	60	64	31	63	52

we look at **simplified representations** such as



as well as **summary statistics** such as

$$\mathbb{E}[X], \quad \mathrm{Var}(X), \quad \mathbb{E}[X^3], \quad \mathbb{E}[X^4], \quad \cdots$$

Summary Statistics in Multiple Dimensions

For a (real-valued) random variable X, we have:

$$\mathbb{E}[X], \quad \mathrm{Var}(X), \quad \mathbb{E}[X^3], \quad \mathbb{E}[X^4], \quad \cdots$$

Summary Statistics in Multiple Dimensions

For a (real-valued) random variable X, we have:

$$\mathbb{E}[X], \quad \mathrm{Var}(X), \quad \mathbb{E}[X^3], \quad \mathbb{E}[X^4], \quad \cdots$$

For a random vector $\boldsymbol{X} = (X_1, \cdots, X_n)$ with $\mathrm{Cov}(X_i, X_j) = \sigma_{ij}$, we have

$$\boldsymbol{\mu} = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix}$$

Summary Statistics in Multiple Dimensions

For a (real-valued) random variable X, we have:

$$\mathbb{E}[X], \quad \mathrm{Var}(X), \quad \mathbb{E}[X^3], \quad \mathbb{E}[X^4], \quad \cdots$$

For a random vector $\boldsymbol{X} = (X_1, \cdots, X_n)$ with $\mathrm{Cov}(X_i, X_j) = \sigma_{ij}$, we have

$$\boldsymbol{\mu} = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix}, \quad \cdots \ ?$$

How to **generalise** the higher moments $\mathbb{E}[X^k]$ to the multivariate setting?

Moment Tensors

$$\boldsymbol{\mu} = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix}$$

Moment Tensors

$$\boldsymbol{\mu} = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix}$$

The mean vector is a $n \times 1$ vector and the covariance matrix is a $n \times n$ matrix. Hence we expect the mth moment to be a $\underbrace{n \times n \times \cdots \times n}_{m \text{ times}}$ array.

Moment Tensors

$$\boldsymbol{\mu} = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix}$$

The mean vector is a $n \times 1$ vector and the covariance matrix is a $n \times n$ matrix. Hence we expect the mth moment to be a $\underbrace{n \times n \times \cdots \times n}_{m \text{ times}}$ array.

Recall that

$$\sigma_{ij} = \mathrm{Cov}(X_i, X_j) = \mathbb{E}\big[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])\big].$$

Moment Tensors

$$\boldsymbol{\mu} = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix}$$

The mean vector is a $n \times 1$ vector and the covariance matrix is a $n \times n$ matrix. Hence we expect the mth moment to be a $\underbrace{n \times n \times \cdots \times n}_{m \text{ times}}$ array.

Recall that

$$\sigma_{ij} = \mathrm{Cov}(X_i, X_j) = \mathbb{E}\big[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])\big].$$

Let $Y_i = X_i - \mathbb{E}[X_i]$. The covariance matrix can be written in the form

$$\mathbb{E}\left[\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \begin{pmatrix} Y_1, \cdots, Y_n \end{pmatrix}\right]$$

Moment Tensors

$$\boldsymbol{\mu} = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix}$$

The mean vector is a $n \times 1$ vector and the covariance matrix is a $n \times n$ matrix. Hence we expect the mth moment to be a $\underbrace{n \times n \times \cdots \times n}_{m \text{ times}}$ array.

Recall that

$$\sigma_{ij} = \mathrm{Cov}(X_i, X_j) = \mathbb{E}\big[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])\big].$$

Let $Y_i = X_i - \mathbb{E}[X_i]$. The covariance matrix can be written in the form

$$\mathbb{E}\left[\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \left(Y_1, \cdots, Y_n\right)\right] = \mathbb{E}[\boldsymbol{Y} \otimes \boldsymbol{Y}]$$

which is a 2-**tensor**.

Moment Tensor

For a random vector $\boldsymbol{X} = (X_1, \cdots, X_m)$, the mth moment should be an ***m*-tensor**

$$\mathbb{E}\big[\underbrace{\boldsymbol{X} \otimes \boldsymbol{X} \otimes \cdots \otimes \boldsymbol{X}}_{m \text{ times}}\big] = \mathbb{E}\big[(\boldsymbol{X})^{\otimes m}\big].$$

The $(i_1, i_2, \cdots, i_m)$-entry is given by

$$\mathbb{E}\big[X_{i_1} X_{i_2} \cdots X_{i_m}\big].$$

Moment Tensor

For a random vector $\boldsymbol{X} = (X_1, \cdots, X_m)$, the mth moment should be an ***m*-tensor**

$$\mathbb{E}\Big[\underbrace{\boldsymbol{X} \otimes \boldsymbol{X} \otimes \cdots \otimes \boldsymbol{X}}_{m \text{ times}}\Big] = \mathbb{E}\big[(\boldsymbol{X})^{\otimes m}\big].$$

The $(i_1, i_2, \cdots, i_m)$-entry is given by

$$\mathbb{E}\big[X_{i_1} X_{i_2} \cdots X_{i_m}\big].$$

Returning to our setting, the ***m*th moment tensor** between two vectors $\theta, \phi \in \mathbb{R}^d$ is defined to be

$$\Delta_m := \mathbb{E}\big[(G\theta)^{\otimes m} - (G\phi)^{\otimes m}\big].$$

Universal Upper and Lower Bounds

Our goal is to establish upper and lower bounds on $D(P_\theta \parallel P_\phi)$.

Universal Upper and Lower Bounds

Our goal is to establish upper and lower bounds on $D(P_\theta \parallel P_\phi)$.

Bandeira-Rigollet-Weed (2017)

Let $\theta, \phi \in \mathbb{R}^d$ be vectors satisfying some technical conditions. For any $k \geq 1$, there exist constants $\underline{C}$ and $\overline{C}$ such that

$$\underline{C} \sum_{m=1}^{\infty} \frac{\|\Delta_m\|^2}{(\sqrt{3}\sigma)^{2m} m!} \leq D(P_\theta \parallel P_\phi)$$

Universal Upper and Lower Bounds

Our goal is to establish upper and lower bounds on $D(P_\theta \parallel P_\phi)$.

Bandeira-Rigollet-Weed (2017)

Let $\theta, \phi \in \mathbb{R}^d$ be vectors satisfying some technical conditions. For any $k \geq 1$, there exist constants $\underline{C}$ and $\overline{C}$ such that

$$\underline{C} \sum_{m=1}^{\infty} \frac{\|\Delta_m\|^2}{(\sqrt{3}\sigma)^{2m} m!} \leq D(P_\theta \parallel P_\phi) \leq \overline{C} \sum_{m=1}^{\infty} \frac{\|\Delta_m\|^2}{\sigma^{2m} m!}.$$

We have both an upper bound and lower bound on the KL divergence in terms of the **same quantity**.

Universal Upper and Lower Bounds

Our goal is to establish upper and lower bounds on $D(P_\theta \parallel P_\phi)$.

Bandeira-Rigollet-Weed (2017)

Let $\theta, \phi \in \mathbb{R}^d$ be vectors satisfying some technical conditions. For any $k \geq 1$, there exist constants $\underline{C}$ and $\overline{C}$ such that

$$\underline{C} \sum_{m=1}^{\infty} \frac{\|\Delta_m\|^2}{(\sqrt{3}\sigma)^{2m} m!} \leq D(P_\theta \parallel P_\phi) \leq \overline{C} \sum_{m=1}^{\infty} \frac{\|\Delta_m\|^2}{\sigma^{2m} m!}.$$

We have both an upper bound and lower bound on the KL divergence in terms of the **same quantity**.

This allows us to extend the **passageway:**

$$\text{Performance of Estimators} \longleftrightarrow D(P_\theta \parallel P_\phi)$$

Universal Upper and Lower Bounds

Our goal is to establish upper and lower bounds on $D(P_\theta \parallel P_\phi)$.

Bandeira-Rigollet-Weed (2017)

Let $\theta, \phi \in \mathbb{R}^d$ be vectors satisfying some technical conditions. For any $k \geq 1$, there exist constants $\underline{C}$ and $\overline{C}$ such that

$$\underline{C} \sum_{m=1}^{\infty} \frac{\|\Delta_m\|^2}{(\sqrt{3}\sigma)^{2m} m!} \leq D(P_\theta \parallel P_\phi) \leq \overline{C} \sum_{m=1}^{\infty} \frac{\|\Delta_m\|^2}{\sigma^{2m} m!}.$$

We have both an upper bound and lower bound on the KL divergence in terms of the **same quantity**.

This allows us to extend the **passageway:**

$$\text{Performance of Estimators} \longleftrightarrow D(P_\theta \parallel P_\phi) \longleftrightarrow \left\{ \|\Delta_m\| \ : \ m \in \mathbb{Z}_{\geq 1} \right\}.$$

Multi-reference Alignment

If we are able to understand how the family of moment tensors

$$\Delta_m = \mathbb{E}\left[(G\theta)^{\otimes m} - (G\phi)^{\otimes m}\right]$$

varies with θ and ϕ, it should give us a better understanding of the **fundamental difficulty** of solving the algebraically structured model.

Multi-reference Alignment

If we are able to understand how the family of moment tensors

$$\Delta_m = \mathbb{E}\left[(G\theta)^{\otimes m} - (G\phi)^{\otimes m}\right]$$

varies with θ and ϕ, it should give us a better understanding of the **fundamental difficulty** of solving the algebraically structured model.

However, the action of the orthogonal group $O(d)$ on $\mathbb{R}^d$ is **highly complicated**.

Multi-reference Alignment

If we are able to understand how the family of moment tensors

$$\Delta_m = \mathbb{E}\left[(G\theta)^{\otimes m} - (G\phi)^{\otimes m}\right]$$

varies with θ and ϕ, it should give us a better understanding of the **fundamental difficulty** of solving the algebraically structured model.

However, the action of the orthogonal group $O(d)$ on $\mathbb{R}^d$ is **highly complicated**.

1. Direct integration of $O(d)$:

Multi-reference Alignment

If we are able to understand how the family of moment tensors

$$\Delta_m = \mathbb{E}\left[(G\theta)^{\otimes m} - (G\phi)^{\otimes m}\right]$$

varies with θ and ϕ, it should give us a better understanding of the **fundamental difficulty** of solving the algebraically structured model.

However, the action of the orthogonal group $O(d)$ on $\mathbb{R}^d$ is **highly complicated**.

1. Direct integration of $O(d)$: Involves $\dfrac{d^2 - d}{2}$ parameters.

Multi-reference Alignment

If we are able to understand how the family of moment tensors

$$\Delta_m = \mathbb{E}\left[(G\theta)^{\otimes m} - (G\phi)^{\otimes m}\right]$$

varies with θ and ϕ, it should give us a better understanding of the **fundamental difficulty** of solving the algebraically structured model.

However, the action of the orthogonal group $O(d)$ on $\mathbb{R}^d$ is **highly complicated**.

1. Direct integration of $O(d)$: Involves $\dfrac{d^2 - d}{2}$ parameters.
2. Subgroups of $O(d)$:

Multi-reference Alignment

If we are able to understand how the family of moment tensors

$$\Delta_m = \mathbb{E}\left[(G\theta)^{\otimes m} - (G\phi)^{\otimes m}\right]$$

varies with θ and ϕ, it should give us a better understanding of the **fundamental difficulty** of solving the algebraically structured model.

However, the action of the orthogonal group $O(d)$ on $\mathbb{R}^d$ is **highly complicated**.

1. Direct integration of $O(d)$: Involves $\dfrac{d^2 - d}{2}$ parameters.
2. Subgroups of $O(d)$: Every finite group is a subgroup of $O(d)$ for some $d \in \mathbb{Z}_{\geq 1}$.

Multi-reference Alignment

If we are able to understand how the family of moment tensors

$$\Delta_m = \mathbb{E}\left[(G\theta)^{\otimes m} - (G\phi)^{\otimes m}\right]$$

varies with θ and ϕ, it should give us a better understanding of the **fundamental difficulty** of solving the algebraically structured model.

However, the action of the orthogonal group $O(d)$ on $\mathbb{R}^d$ is **highly complicated**.

1. Direct integration of $O(d)$: Involves $\dfrac{d^2 - d}{2}$ parameters.
2. Subgroups of $O(d)$: Every finite group is a subgroup of $O(d)$ for some $d \in \mathbb{Z}_{\geq 1}$.

Further simplifications are needed to make the problem **tractable**.

Multi-reference Alignment

Define

$$\mathcal{G} := \left\{ R_\ell \;:\; 0 \leq \ell \leq d-1 \right\}.$$

where

$$R_\ell\big((\theta_1, \cdots, \theta_d)\big) := \big(\theta_{1+\ell}, \theta_{2+\ell}, \cdots, \theta_{d+\ell}\big).$$

This setup is known as the **Multi-reference Alignment** model.

Multi-reference Alignment

Define

$$\mathcal{G} := \left\{ R_\ell \ : \ 0 \leq \ell \leq d-1 \right\}.$$

where

$$R_\ell\big((\theta_1, \cdots, \theta_d)\big) := \big(\theta_{1+\ell}, \theta_{2+\ell}, \cdots, \theta_{d+\ell}\big).$$

This setup is known as the **Multi-reference Alignment** model.

Greatly simplified but still **mathematically interesting**.

Multi-reference Alignment

Define

$$\mathcal{G} := \{R_\ell \ : \ 0 \le \ell \le d-1\}.$$

where

$$R_\ell\big((\theta_1, \cdots, \theta_d)\big) := \big(\theta_{1+\ell}, \theta_{2+\ell}, \cdots, \theta_{d+\ell}\big).$$

This setup is known as the **Multi-reference Alignment** model.

Greatly simplified but still **mathematically interesting**.

The **Discrete Fourier Transform** $\hat{\theta}$ of a vector $\theta \in \mathbb{R}^d$ is given by

$$\hat{\theta}_j := \frac{1}{\sqrt{d}} \sum_{k=1}^{d} e^{\frac{2\pi ijk}{d}} \theta_k, \qquad -\lfloor d/2 \rfloor \le j \le \lfloor d/2 \rfloor.$$

The Fourier Domain

By passing through the passage

$$\theta \xleftrightarrow{\text{DFT}} \hat{\theta},$$

The Fourier Domain

By passing through the passage

$$\theta \xleftrightarrow{\text{DFT}} \hat{\theta},$$

we obtain **explicit formulas** for the moment tensors:

$$\mathbb{E}\left[(\widehat{G\theta})\right]_i = \begin{cases} \hat{\theta}_0 & \text{if } i = 0, \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathbb{E}\left[(\widehat{G\theta})^{\otimes 2}\right]_{ij} = \begin{cases} |\hat{\theta}_i|^2 & \text{if } i + j = 0, \\ 0 & \text{otherwise.} \end{cases}$$

$$\vdots$$

$$\mathbb{E}\left[(\widehat{G\theta})^{\otimes m}\right]_{i_1 \cdots i_m} = \begin{cases} \hat{\theta}_{i_1}\hat{\theta}_{i_2} \cdots \hat{\theta}_{i_m} & \text{if } i_1 + \cdots + i_m = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Moment Matching

$$\underline{C}\sum_{m=1}^{\infty}\frac{\|\Delta_m\|^2}{(\sqrt{3}\sigma)^{2m}m!} \leq D(P_\theta \parallel P_\phi) \leq \overline{C}\sum_{m=1}^{\infty}\frac{\|\Delta_m\|^2}{\sigma^{2m}m!}.$$

Key Idea: If the first $k-1$ moments match, then $D(P_\theta \parallel P_\phi)$ is of order $O(\sigma^{-2k})$.

Establish upper bounds: Construct two vectors θ and ϕ such that the first k moments **cancel out**.

Moment Matching

$$\underline{C}\sum_{m=1}^{\infty}\frac{\|\Delta_m\|^2}{(\sqrt{3}\sigma)^{2m}m!} \leq D(P_\theta \parallel P_\phi) \leq \overline{C}\sum_{m=1}^{\infty}\frac{\|\Delta_m\|^2}{\sigma^{2m}m!}.$$

Key Idea: If the first $k-1$ moments match, then $D(P_\theta \parallel P_\phi)$ is of order $O(\sigma^{-2k})$.

Establish upper bounds: Construct two vectors θ and ϕ such that the first k moments **cancel out**.

Establish lower bounds: Show that no such cancellation is possible.

Moment Matching

$$\text{DC:} \qquad \mathbb{E}\left[(\widehat{G\theta})\right]_i = \begin{cases} \hat{\theta}_0 & \text{if } i = 0 \\ 0 & \text{otherwise.} \end{cases}$$

$$\text{Power spectrum:} \qquad \mathbb{E}\left[(\widehat{G\theta})^{\otimes 2}\right]_{ij} = \begin{cases} |\hat{\theta}_i|^2 & \text{if } i + j = 0 \\ 0 & \text{otherwise.} \end{cases}$$

$$\text{Bispectrum:} \qquad \mathbb{E}\left[(\widehat{G\theta})^{\otimes 3}\right]_{ijk} = \begin{cases} \hat{\theta}_i\hat{\theta}_j\hat{\theta}_k & \text{if } i + j + k = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Most of the tensor entries **vanishes**.

Moment Matching

$$\text{DC:} \qquad \mathbb{E}\left[(\widehat{G\theta})\right]_i = \begin{cases} \hat{\theta}_0 & \text{if } i = 0 \\ 0 & \text{otherwise.} \end{cases}$$

$$\text{Power spectrum:} \qquad \mathbb{E}\left[(\widehat{G\theta})^{\otimes 2}\right]_{ij} = \begin{cases} |\hat{\theta}_i|^2 & \text{if } i + j = 0 \\ 0 & \text{otherwise.} \end{cases}$$

$$\text{Bispectrum:} \qquad \mathbb{E}\left[(\widehat{G\theta})^{\otimes 3}\right]_{ijk} = \begin{cases} \hat{\theta}_i\hat{\theta}_j\hat{\theta}_k & \text{if } i + j + k = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Most of the tensor entries **vanishes**.The remaining non-vanishing terms are special quantities in **signal processing**.

Moment Matching

$$\text{DC:} \qquad \mathbb{E}\left[(\widehat{G\theta})\right]_i = \begin{cases} \hat{\theta}_0 & \text{if } i = 0 \\ 0 & \text{otherwise.} \end{cases}$$

$$\text{Power spectrum:} \qquad \mathbb{E}\left[(\widehat{G\theta})^{\otimes 2}\right]_{ij} = \begin{cases} |\hat{\theta}_i|^2 & \text{if } i + j = 0 \\ 0 & \text{otherwise.} \end{cases}$$

$$\text{Bispectrum:} \qquad \mathbb{E}\left[(\widehat{G\theta})^{\otimes 3}\right]_{ijk} = \begin{cases} \hat{\theta}_i \hat{\theta}_j \hat{\theta}_k & \text{if } i + j + k = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Most of the tensor entries **vanishes**.The remaining non-vanishing terms are special quantities in **signal processing**.

We extend the passageway by one more step:

$$\begin{array}{c}\text{Performance of} \\ \text{Estimators}\end{array} \longleftrightarrow D(P_\theta \parallel P_\phi) \longleftrightarrow \left\{\|\Delta_m\|\right\}_{m=1}^{\infty} \longleftrightarrow \begin{array}{c}\text{Support of} \\ \hat{\theta} \text{ and } \hat{\phi}\end{array}$$

The End

Thank you for your attention.