Probability and Stochastic Processes II: Estimation

Zhenyu Liao

School of Electronic Information and Communications, HUST

May, 30, 2024

Outline

Estimating the parameters of a distribution

- A **parametric model** is a family of probability distributions that can be described by a finite number of parameters[1]
  - the family of normal/Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, with parameters μ and $\sigma^2 \geq 0$; and
  - the family of Bernoulli distribution $\mathrm{Bern}(p)$, with parameter p; and
  - the family of Gamma distribution $\mathrm{Gamma}(\alpha, \beta)$, with parameters α and β.
- PDF/PMF $\{f(x|\theta) \colon \theta \in \Omega\}$ for general **parameter model**, with **parameters** $\theta \in \mathbb{R}^k$, $\Omega \subset \mathbb{R}^k$ the **parameter space**
- Example: Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, with $\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$, $\Omega = \mathbb{R} \times \mathbb{R}_+$, and

$$f(x|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{1}$$

- Question: given observations $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f(x|\theta)$, how can we estimate the **unknown** parameters θ and possibly quantify the quality of the proposed estimate?

[1] If the number of parameters **increases** with the sample size, the "double asymptotic" regime in RMT.

Method of moments

- if θ is a single number, a simple idea to estimate θ is to "MATCH" the theoretical mean of $X \sim f(x|\theta)$ equals to the sample mean $\bar{X} = \frac{1}{n} \sum_{i=1} X_i$

Poisson distribution

The Poisson distribution with parameter $\lambda > 0$ (denoted Poisson(λ)) is a discrete distribution over the non-negative integers $\{0, 1, \ldots\}$ having PMF

$$f(x|\lambda) = \frac{e^{-\lambda} \lambda^x}{x!}. \tag{2}$$

- if $X \sim \text{Poisson}(\lambda)$, we have $\mathbb{E}[X] = \lambda$, so a simple estimate of λ as

$$\hat{\lambda} = \bar{X} = \frac{1}{n} \sum_{i=1} X_i. \tag{3}$$

Method of moments

- if θ is a single number, a simple idea to estimate θ is to "MATCH" the theoretical mean of $X \sim f(x|\theta)$ equals to the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1} X_i$

Exponential distribution

The exponential distribution with parameter $\lambda > 0$ (denoted $\text{Exp}(\lambda)$) is a continuous distribution over $\mathbb{R}_+$ having PDF

$$f(x|\lambda) = \lambda e^{-\lambda x}. \tag{4}$$

- if $X \sim \text{Exp}(\lambda)$, we have $\mathbb{E}[X] = \frac{1}{\lambda}$, so a simple estimate of λ as

$$\hat{\lambda} = \frac{1}{\bar{X}} = \frac{1}{\frac{1}{n}\sum_{i=1} X_i}. \tag{5}$$

Method of moments

- more generally, for $X \sim f(x|\theta)$ where θ contains k unknown parameters, the **method of moments estimator** proposes to consider the first k **moments** of the distribution of X,

$$\mu_1 = \mathbb{E}[X], \quad \mu_2 = \mathbb{E}[X^2], \quad \ldots, \quad \mu_k = \mathbb{E}[X^k]. \tag{6}$$

- leading to the following empirical estimates

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^n X_i, \quad \hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^n X_i^2, \quad \ldots, \quad \hat{\mu}_k = \frac{1}{n}\sum_{i=1}^n X_i^k. \tag{7}$$

Method of moments: Gaussian distribution

Method of moments: Gaussian distribution

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{E}[X] = \mu$ and $\mathbb{E}[X^2] = \mu^2 + \sigma^2$. With the method of moments estimator, we write the empirical estimates

$$\hat{\mu} = \hat{\mu}_1 = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad \hat{\mu}^2 + \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2. \tag{8}$$

Solving for the parameter estimates $\hat{\mu}$ and $\hat{\sigma}^2$, we get

$$\hat{\mu} = \bar{X}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2. \tag{9}$$

- Question: what can we say about these MoM estimators?
- Answer: characterization via the **mean-squared-error (MSE)**

Bias, variance, and mean-squared-error

- any estimator $\hat{\theta} \equiv \hat{\theta}(X_1, \ldots, X_n)$ is a statistics – randomness from the data $X_1, \ldots, X_n$
- for $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f(x|\theta)$, measure the **quality** of the estimator $\hat{\theta}$ as
  - **bias** of $\hat{\theta}$ as $\mathbb{E}[\hat{\theta}] - \theta$, the expectation taken with respect to the randomness in $X_1, \ldots, X_n$
  - the **standard error** of $\hat{\theta}$ is the standard deviation $\sqrt{\mathrm{Var}[\hat{\theta}]}$
  - the mean-squared-error (MSE) of $\hat{\theta}$ given by $\mathbb{E}[(\hat{\theta} - \theta)^2]$
- Note that

$$\mathbb{E}[(\hat{\theta} - \theta)^2] = \mathrm{Var}[\hat{\theta}] + (\mathbb{E}[\hat{\theta}] - \theta)^2. \tag{10}$$

- This is the **bias-variance decomposition** of MSE:

$$\mathrm{MSE} = \mathrm{Variance} + \mathrm{Bias}^2. \tag{11}$$

MSE of MoM for Poisson distribution

Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$, the MoM estimator of λ is

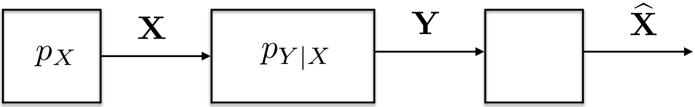
$$\hat{\lambda} = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{12}$$

The bias-variance decomposition of MSE of $\hat{\lambda}$ can be derived as

- bias $\mathbb{E}[\hat{\lambda}] - \lambda = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i] - \lambda = 0$: **unbiased!**
- variance $\text{Var}[\hat{\lambda}] = \text{Var}[\bar{X}] = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}[X_i] = \frac{\lambda}{n}$: **of vanishing variance (order $O(n^{-1})$)!**
- So $\text{MSE}[\hat{\lambda}] = 0 + \frac{\lambda}{n} = \frac{\lambda}{n}$.

MMSE Estimation

- We observe some data y, which we assume to be produced as the realization of some RV Y.
- We have that Y is generated as a random transformation $X \mapsto Y$ of another RV X.
- The random transformation is described by a conditional PDF $p_{Y|X}$.
- X is distributed according to some known PDF p_X (i.e., the **statistical modeling**).
- Goal: find an estimator $\widehat{X} = g(Y)$ such that $\mathbb{E}[\|X - \widehat{X}\|^2]$ is minimized.



Reminder on vector spaces

Definition

A vector space V over $\mathbb{R}$ is a set of elements called *vectors* such that

1. For all $\mathbf{v}, \mathbf{v}' \in V$, $\mathbf{v} + \mathbf{v}' \in V$.
2. $\exists\, \mathbf{0} \in V$ such that $\mathbf{v} + \mathbf{0} = \mathbf{v}$ for all $\mathbf{v} \in V$.
3. For all $\mathbf{v} \in V$ there exists an opposite element $-\mathbf{v} \in V$ such that $\mathbf{v} + (-\mathbf{v}) = \mathbf{0}$.
4. $x\mathbf{v} \in V$ for all $\mathbf{v} \in V$ and $x \in \mathbb{R}$.
5. $0\mathbf{v} = \mathbf{0}$ for all $\mathbf{v} \in V$.
6. $1\mathbf{v} = \mathbf{v}$ for all $\mathbf{v} \in V$.

- This implies that V is closed with respect to linear combinations with coefficients in $\mathbb{R}$.

Reminder on norms and normed vector spaces

Definition

A norm is a function $\|\cdot\| : V \to \mathbb{R}_+$ that satisfies the following properties:

1. $\|\mathbf{v}\| = 0$ if and only if $\mathbf{v} = \mathbf{0}$.
2. $\|\mathbf{v} + \mathbf{u}\| \leq \|\mathbf{v}\| + \|\mathbf{u}\|$ (triangle inequality).
3. $\|x\mathbf{v}\| = |x| \cdot \|\mathbf{v}\|$ for all $\mathbf{v} \in V$ and $x \in \mathbb{R}$.

And a normed vector space is a vector space V with a norm $\|\cdot\|$.

Notice: a norm is a "distance" function.

- For example, one can check that the norm defined as

$$\|\mathbf{v}\|_2 = \sqrt{\sum_{i=1}^{n} v_i^2}$$

  where $V = \mathbb{R}^n$ is the standard Euclidean n-dimensional vector space over $\mathbb{R}$, defines a distance in the usual sense (length of the vector joining two points in $\mathbb{R}^n$).
- Let $\mathbf{v}, \mathbf{u} \in \mathbb{R}^n$, then

$$\|\mathbf{v} - \mathbf{u}\|_2 = \sqrt{\sum_{i=1}^{n} (v_i - u_i)^2}$$

  is the Euclidean distance between the points (vectors) $\mathbf{v}$ and $\mathbf{u}$.

Reminder on inner product

Definition

Given a vector space V over $\mathbb{R}$, an inner product is a function $\langle \cdot, \cdot \rangle \colon V \times V \to \mathbb{R}$ with the following properties:

1. $\langle \mathbf{v}, \mathbf{u} \rangle = \langle \mathbf{u}, \mathbf{v} \rangle$ (symmetry).
2. $\langle x\mathbf{v}, \mathbf{u} \rangle = x \langle \mathbf{v}, \mathbf{u} \rangle$, for all $\mathbf{v}, \mathbf{u} \in V$ and $x \in \mathbb{R}$ (scaling).
3. $\langle \mathbf{v}_1 + \mathbf{v}_2, \mathbf{u} \rangle = \langle \mathbf{v}_1, \mathbf{u} \rangle + \langle \mathbf{v}_2, \mathbf{u} \rangle$ (linearity).
4. $\langle \mathbf{v}, \mathbf{v} \rangle \geq 0$, with equality if and only if $\mathbf{v} = \mathbf{0}$.

A vector space with an inner product is called inner product space.

Theorem (Cauchy-Schwarz inequality)

$$\langle \mathbf{v}, \mathbf{u} \rangle^2 \leq \langle \mathbf{v}, \mathbf{v} \rangle \langle \mathbf{u}, \mathbf{u} \rangle$$

with equality if and only if $a\mathbf{v} = b\mathbf{u}$, *with* $a, b \in \mathbb{R}$ *not both zero.* □

Theorem (2-norm)

Let V be an inner product space. Then, the following is a norm (called 2-norm, or standard Euclidean norm):

$$\|\mathbf{v}\|_2 = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$$

Least Squares approximation

- Let be $\mathbf{x}$ a point (vector) in some vector space V over $\mathbb{R}$ and let $\mathbf{y}_1, \ldots, \mathbf{y}_m$ be a given collection of vectors:
  we wish to find the "best" approximation of $\mathbf{x}$ by a linear combination of the vectors $\{\mathbf{y}_i\}$.
- We have to give a rigorous meaning to the term "best": if V is an inner product space, we shall consider the minimum distance approximation, that is, we look for

$$\widehat{\mathbf{x}} = \sum_{i=1}^{m} a_i \mathbf{y}_i$$

  such that

$$\|\mathbf{x} - \widehat{\mathbf{x}}\|_2^2 = \langle \mathbf{x} - \widehat{\mathbf{x}}, \mathbf{x} - \widehat{\mathbf{x}} \rangle$$

  is minimum.
- This approximation is called (linear) "Least-Squares" (some people call it "linear regression").

LS Solution

- A brute-force approach: we can write, for $\mathbf{a} \in \mathbb{R}^m$,

$$\begin{aligned}
\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 &= \|\mathbf{x}\|_2^2 - 2\langle \mathbf{x}, \hat{\mathbf{x}} \rangle + \|\hat{\mathbf{x}}\|_2^2 \\
&= \|\mathbf{x}\|_2^2 - 2\sum_{i=1}^m \langle \mathbf{x}, \mathbf{y}_i \rangle a_i + \sum_{i=1}^m \sum_{j=1}^m a_i \langle \mathbf{y}_i, \mathbf{y}_j \rangle a_j \\
&= \|\mathbf{x}\|_2^2 - 2\mathbf{r}_{xy}^\mathsf{T} \mathbf{a} + \mathbf{a}^\mathsf{T} \mathbf{G}_y \mathbf{a}
\end{aligned}$$

where we define the "cross-correlation vector"

$$\mathbf{r}_{xy} = \left[\langle \mathbf{x}, \mathbf{y}_1 \rangle, \ldots, \langle \mathbf{x}, \mathbf{y}_m \rangle \right]^\mathsf{T}$$

and the matrix of inner products (Gram matrix)

$$\mathbf{G}_y = \begin{bmatrix} \langle \mathbf{y}_1, \mathbf{y}_1 \rangle & \langle \mathbf{y}_1, \mathbf{y}_2 \rangle & \cdots & \langle \mathbf{y}_1, \mathbf{y}_m \rangle \\ \langle \mathbf{y}_2, \mathbf{y}_1 \rangle & \langle \mathbf{y}_2, \mathbf{y}_2 \rangle & & \vdots \\ \vdots & & & \\ \langle \mathbf{y}_m, \mathbf{y}_1 \rangle & \langle \mathbf{y}_m, \mathbf{y}_2 \rangle & \cdots & \langle \mathbf{y}_m, \mathbf{y}_m \rangle \end{bmatrix}$$

Notice: this is true independent of the "dimension" of the vector space V!

- Notice that $\mathbf{G}_y \in \mathbb{R}^{m \times m}$ is symmetric and positive semi-definite (WHY?).
- Taking the gradient of the distance function with respect to $\mathbf{a}$, we obtain the equation

$$\mathbf{G}_y \mathbf{a} = \mathbf{r}_{xy}$$

- Assuming for simplicity that $\mathbf{G}_y$ is invertible (otherwise, we can eliminate some linearly dependent $\mathbf{y}_i$ and obtain the same subspace), we obtain $\mathbf{a} = \mathbf{G}_y^{-1} \mathbf{r}_{xy}$.
- This leads to the solution $\hat{\mathbf{x}} = \begin{bmatrix} \mathbf{y}_1 & \ldots & \mathbf{y}_m \end{bmatrix} \mathbf{a}$, is this the minimal $\|\mathbf{x} - \hat{\mathbf{x}}\|$? If yes, WHY?
- OBSERVATION: notice that the solution $\hat{\mathbf{x}}$ satisfies the following orthogonality condition:

$$\langle \mathbf{x} - \hat{\mathbf{x}}, \mathbf{y}_i \rangle = 0, \quad \forall \ i = 1, \ldots, m$$

How to prove this?

Zero-mean Finite Covariance RVs

- The space of zero-mean finite covariance RVs forms a vector space.
- Inner product:

$$\langle X, Y \rangle = \mathbb{E}[XY]$$

- Induced 2-norm:

$$\|X\|_2 = \sqrt{\mathbb{E}[|X|^2]}$$

- In this vector space, distance is expressed by the MSE

$$\|X - Y\|_2^2 = \mathbb{E}[|X - Y|^2]$$

Generalization to Random Vectors

- For zero-mean finite covariance random vectors, we can combine the standard inner product in $\mathbb{R}^n$ with what defined before:

$$\langle \mathbf{X}, \mathbf{Y} \rangle = \mathbb{E}[\mathbf{X}^\mathsf{T}\mathbf{Y}] = \sum_{i=1}^{n} \mathbb{E}[X_i Y_i]$$

- The induced 2-norm is given by

$$\sqrt{(\mathbf{X}, \mathbf{X})} = \sqrt{\mathbb{E}\left[\mathbf{X}^\mathsf{T}\mathbf{X}\right]} = \sqrt{\operatorname{tr}\left(\mathbb{E}\left[\mathbf{X}\mathbf{X}^\mathsf{T}\right]\right)} = \sqrt{\operatorname{tr}(\boldsymbol{\Sigma}_x)}$$

- Then, the MSE for the vector case is given by

$$\mathsf{MSE} = \mathbb{E}\left[\|\mathbf{X} - \mathbf{Y}\|^2\right] = \sum_{i=1}^{n} \mathbb{E}[|X_i - Y_i|^2] = \operatorname{tr}\left(\operatorname{Cov}(\mathbf{X} - \mathbf{Y})\right)$$

A remark about notation

- Unfortunately, the same symbol $\|\cdot\|_2$ takes on different meanings depending on the inner product space it is referred to.
- In our case, for all $\omega \in \Omega$, $\mathbf{X}(\omega)$ is an element of $\mathbb{R}^n$, but when defining the vector space V of finite-dimensional random vectors with mean zero and finite per-component variance, we need to be careful!
- We shall use

$$\|\mathbf{X}\|^2 = \sum_{i=1}^{n} |X_i|^2$$

  to denote the standard squared 2-norm in $\mathbb{R}^n$. Since $\mathbf{X}$ is a random vector, $\|\mathbf{X}\|^2$ is a random variable.
- Instead, we use

$$\|\mathbf{X}\|_2^2 = \mathbb{E}[\|\mathbf{X}\|^2]$$

  to denote the squared norm in V. This is a non-random quantity (expectation).

Linear Minimum Mean-Square Error estimation

- We have two jointly distributed random vectors $\mathbf{X} \in \mathbb{R}^n$ and $\mathbf{Y} \in \mathbb{R}^m$.
- We observe $\mathbf{Y}$ and we with to "guess" the value of $\mathbf{X}$ by some estimator $\widehat{\mathbf{X}} = g(\mathbf{Y})$ in order to minimize the Mean-Square-Error sense:

$$\mathsf{MSE} = \mathbb{E}\left[\|\mathbf{X} - \widehat{\mathbf{X}}\|^2\right]$$

- For now, we seek an estimator $\widehat{\mathbf{X}}$ in the form of a **linear** function of the observation $\mathbf{Y}$, that is,

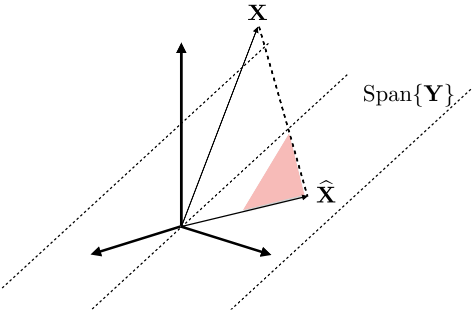
$$\widehat{\mathbf{X}} = \mathbf{A}\mathbf{Y}$$

Orthogonality principle

- The approximation error $\mathbf{X} - \widehat{\mathbf{X}}$ must be orthogonal with respect to the space of linear functions of $\mathbf{Y}$.
- This means that for any matrix $\mathbf{B} \in \mathbb{C}^{n \times m}$ is must be:

$$\mathbb{E}[(\mathbf{X} - \widehat{\mathbf{X}})^\mathsf{T} \mathbf{B} \mathbf{Y}] = 0$$

for all linear functions $\mathbf{BY}$ of the observation.



- The orthogonality principle yields the condition

$$\langle \mathbf{X} - \widehat{\mathbf{X}}, \mathbf{B}\mathbf{Y} \rangle = \mathbb{E}\left[(\mathbf{X} - \widehat{\mathbf{X}})^\mathsf{T} \mathbf{B}\mathbf{Y} \right] = \operatorname{tr}\left(\mathbb{E}\left[\mathbf{B}\mathbf{Y}(\mathbf{X} - \widehat{\mathbf{X}})^\mathsf{T} \right] \right) = 0$$

  for all $\mathbf{B} \in \mathbb{R}^{n \times m}$.
- In turns, by replacing $\widehat{\mathbf{X}} = \mathbf{A}\mathbf{Y}$, we find the condition that, for all $\mathbf{B}$, it must be

$$\operatorname{tr}\left(\mathbf{B} \left(\mathbb{E}\left[\mathbf{Y}\mathbf{X}^\mathsf{T} \right] - \mathbb{E}\left[\mathbf{Y}\mathbf{Y}^\mathsf{T} \right] \mathbf{A}^\mathsf{T} \right) \right) = 0$$

- This yields the equation

$$\mathbf{A}\mathbb{E}\left[\mathbf{Y}\mathbf{Y}^\mathsf{T} \right] = \mathbb{E}[\mathbf{X}\mathbf{Y}^\mathsf{T}]$$

- Solving for $\mathbf{A}$ (under the assumption that the covariance $\mathbb{E}\left[\mathbf{Y}\mathbf{Y}^\mathsf{T}\right]$ is strictly positive definite), we find:

$$\mathbf{A}\mathbb{E}\left[\mathbf{Y}\mathbf{Y}^\mathsf{T}\right] = \mathbb{E}\left[\mathbf{X}\mathbf{Y}^\mathsf{T}\right] \ \Rightarrow \ \mathbf{A} = \mathbb{E}\left[\mathbf{X}\mathbf{Y}^\mathsf{T}\right]\left(\mathbb{E}\left[\mathbf{Y}\mathbf{Y}^\mathsf{T}\right]\right)^{-1}$$

- In the general case of non-zero mean vectors, we define the centralized RVs $\mathbf{X}_0 = \mathbf{X} - \mathbf{m}_x$ and $\mathbf{Y}_0 = \mathbf{Y} - \mathbf{m}_y$, and notice that $\widehat{\mathbf{X}}$ is the LMMSE estimator for $\mathbf{X}$ if and only if $\widehat{\mathbf{X}}_0 = \widehat{\mathbf{X}} - \mathbf{m}_x$ is the LMMSE estimator for $\mathbf{X}_0$:

$$\mathbb{E}\left[\|\mathbf{X} - \widehat{\mathbf{X}}\|^2\right] = \mathbb{E}\left[\|\mathbf{X}_0 - \underbrace{(\widehat{\mathbf{X}} - \mathbf{m}_x)}_{\widehat{\mathbf{X}}_0}\|^2\right]$$

- Furthermore, $\widehat{\mathbf{X}}_0$ must be a (linear) function of $\widehat{\mathbf{Y}}_0$, since $\mathbf{m}_y$ is just an (arbitrary) constant.

- Letting

$$\begin{aligned}\boldsymbol{\Sigma}_{xy} &= \mathrm{Cov}(\mathbf{X},\mathbf{Y}) = \mathbb{E}[(\mathbf{X}-\mathbf{m}_x)(\mathbf{Y}-\mathbf{m}_y)^\mathsf{T}] \\ \boldsymbol{\Sigma}_{y} &= \mathrm{Cov}(\mathbf{Y}) = \mathbb{E}[(\mathbf{Y}-\mathbf{m}_y)(\mathbf{Y}-\mathbf{m}_y)^\mathsf{T}]\end{aligned}$$

we obtain

$$\widehat{\mathbf{X}}_0 = \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_y^{-1}\mathbf{Y}_0$$

and for the non-zero mean case

$$\widehat{\mathbf{X}} = \mathbf{m}_x + \widehat{\mathbf{X}}_0 = \mathbf{m}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_y^{-1}\left(\mathbf{Y}-\mathbf{m}_y\right)$$

MMSE Covariance Matrix

- The MMSE covariance matrix is given by

$$\mathrm{Cov}(\mathbf{X} - \widehat{\mathbf{X}}) = \boldsymbol{\Sigma}_x - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{xy}^{\mathsf{T}}$$

- The resulting MMSE, is given by $\mathbb{E}[\|\mathbf{X} - \widehat{\mathbf{X}}\|^2] = \mathrm{tr}(\mathrm{Cov}(\mathbf{X} - \widehat{\mathbf{X}}))$.
- Notice: The estimation error vector $\mathbf{X} - \widehat{\mathbf{X}}$ is uncorrelated with any linear function of the observation vector $\mathbf{Y}$.

MMSE estimator: the general case

- With the same setting as before, we now seek an estimator $\hat{\mathbf{X}} = g^*(\mathbf{Y})$, in the space of all (measurable, so **not** necessarily linear) functions of the observation $\mathbf{Y}$.

Theorem

The MMSE estimator of $\mathbf{X}$ *given* $\mathbf{Y}$ *is the conditional mean*

$$\hat{\mathbf{X}} = g^*(\mathbf{Y}) = \mathbb{E}[\mathbf{X}|\mathbf{Y}]$$

Proof

We use the orthogonality principle: the optimal estimator $\widehat{\mathbf{X}}$ must satisfy

$$\mathbb{E}\left[(\mathbf{X}-\widehat{\mathbf{X}})^{\mathsf{T}} g(\mathbf{Y})\right]=0, \quad \text{for all functions } g$$

Letting $\widehat{\mathbf{X}}=\mathbb{E}[\mathbf{X}|\mathbf{Y}]$ and using the iterated expectation theorem[2], we find:

$$\begin{aligned}
\mathbb{E}\left[(\mathbf{X}-\mathbb{E}[\mathbf{X}|\mathbf{Y}])^{\mathsf{T}} g(\mathbf{Y})\right] &= \mathbb{E}\left[\mathbb{E}\left[(\mathbf{X}-\mathbb{E}[\mathbf{X}|\mathbf{Y}])^{\mathsf{T}} g(\mathbf{Y})|\mathbf{Y}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbf{X}^{\mathsf{T}} g(\mathbf{Y})|\mathbf{Y}\right]-\mathbb{E}[\mathbf{X}|\mathbf{Y}]^{\mathsf{T}} g(\mathbf{Y})\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbf{X}|\mathbf{Y}\right]^{\mathsf{T}} g(\mathbf{Y})-\mathbb{E}[\mathbf{X}|\mathbf{Y}]^{\mathsf{T}} g(\mathbf{Y})\right] \\
&= 0
\end{aligned}$$

[2] $\mathbb{E}[f(X,Y)]=\mathbb{E}[\mathbb{E}[f(X,Y)|Y]]$.

Reminder on Conditional Gaussian distribution

- Consider a random vector with $n + m$ components, denoted for simplicity by $(\mathbf{X}, \mathbf{Y})$.
- A very important problem in statistics is to find the conditional distribution of a group of components given the other group. Without loss of generality, we are interested in the conditional distribution of $\mathbf{X}$ given $\mathbf{Y}$.
- In particular, suppose that $(\mathbf{X}, \mathbf{Y}) \sim \mathcal{N}(\mathbf{m}, \mathbf{\Sigma})$, with

$$\mathbf{m} = \begin{bmatrix} \mathbf{m}_x \\ \mathbf{m}_y \end{bmatrix}, \quad \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_x & \mathbf{\Sigma}_{xy} \\ \mathbf{\Sigma}_{yx} & \mathbf{\Sigma}_y \end{bmatrix}$$

with $\mathbf{m}_x = \mathbb{E}[\mathbf{X}]$, $\mathbf{m}_y = \mathbb{E}[\mathbf{Y}]$, $\mathbf{\Sigma}_x = \operatorname{cov}(\mathbf{X})$, $\mathbf{\Sigma}_y = \operatorname{cov}(\mathbf{Y})$ and

$$\mathbf{\Sigma}_{xy} = \operatorname{cov}(\mathbf{X}, \mathbf{Y}) = \mathbb{E}\left[(\mathbf{X} - \mathbf{m}_x)(\mathbf{Y} - \mathbf{m}_y)^\mathsf{T}\right]$$

with $\mathbf{\Sigma}_{yx} = \mathbf{\Sigma}_{xy}^\mathsf{T}$.

Reminder on Conditional Gaussian distribution

With the notation defined before,

$$f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^n \det(\mathbf{\Sigma}_{x|y})}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{m}_{x|y})^\mathsf{T} \mathbf{\Sigma}_{x|y}^{-1} (\mathbf{x} - \mathbf{m}_{x|y}) \right)$$

where the conditional mean value is given by

$$\mathbf{m}_{x|y} = \mathbb{E}[\mathbf{X}|\mathbf{Y} = \mathbf{y}] = \mathbf{m}_x + \mathbf{\Sigma}_{xy}\mathbf{\Sigma}_y^{-1}\left(\mathbf{y} - \mathbf{m}_y\right)$$

and the conditional covariance matrix is given by

$$\mathbf{\Sigma}_{x|y} = \mathbb{E}[(\mathbf{X} - \mathbf{m}_{x|y})(\mathbf{X} - \mathbf{m}_{x|y})^\mathsf{T} | \mathbf{Y} = \mathbf{y}] = \mathbf{\Sigma}_x - \mathbf{\Sigma}_{xy}\mathbf{\Sigma}_y^{-1}\mathbf{\Sigma}_{yx}$$

Notice: given jointly Gaussian $\mathbf{X}, \mathbf{Y}$, $\mathbf{X}$ given $\mathbf{Y}$ is Gaussian, with conditional mean affine function of $\mathbf{Y}$ and conditional covariance constant with $\mathbf{Y}$.

MMSE estimation for Gaussian vectors

- If $\mathbf{X}, \mathbf{Y}$ are jointly Gaussian, then the linear MMSE estimator and the optimal MMSE estimator coincide.
- In order to see this, recall

$$f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma}_{x|y})}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{m}_{x|y})^\mathsf{T} \boldsymbol{\Sigma}_{x|y}^{-1} (\mathbf{x} - \mathbf{m}_{x|y})\right)$$

where the conditional mean value is given by

$$\mathbf{m}_{x|y} = \mathbb{E}[\mathbf{X}|\mathbf{Y} = \mathbf{y}] = \mathbf{m}_x + \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_y^{-1} \left(\mathbf{y} - \mathbf{m}_y\right)$$

and the conditional covariance matrix is given by

$$\boldsymbol{\Sigma}_{x|y} = \mathbb{E}[(\mathbf{X} - \mathbf{m}_{x|y})(\mathbf{X} - \mathbf{m}_{x|y})^\mathsf{T} | \mathbf{Y} = \mathbf{y}] = \boldsymbol{\Sigma}_x - \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_y^{-1} \boldsymbol{\Sigma}_{yx}$$

- Hence, in the Gaussian case, the (general) MMSE estimator of $\mathbf{X}$ given $\mathbf{Y}$ coincides with the LMMSE estimator (Wiener filter):

$$\widehat{\mathbf{X}} = \mathbb{E}[\mathbf{X}|\mathbf{Y}] = \mathbf{m}_x + \mathbf{\Sigma}_{xy}\mathbf{\Sigma}_y^{-1}\left(\mathbf{Y} - \mathbf{m}_y\right)$$

- MMSE decomposition:

$$\mathbf{X} = \widehat{\mathbf{X}} + (\mathbf{X} - \widehat{\mathbf{X}}) = \widehat{\mathbf{X}} + \mathbf{V}$$

  where the MMSE estimator $\widehat{\mathbf{X}}$ and the estimation error vector $\mathbf{V}$ are uncorrelated, and therefore independent (in the Gaussian case), where we have

$$\widehat{\mathbf{X}} \sim \mathcal{N}(\mathbf{m}_x, \mathbf{\Sigma}_{xy}\mathbf{\Sigma}_y^{-1}\mathbf{\Sigma}_{yx}), \quad \mathbf{V} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{x|y})$$

Application to proper Gaussian random vectors

- If $\mathbf{X}$ and $\mathbf{Y}$ are proper jointly Gaussian, i.e.,

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim \mathcal{CN}\left(\begin{bmatrix} \mathbf{m}_x \\ \mathbf{m}_y \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma}_x & \mathbf{\Sigma}_{xy} \\ \mathbf{\Sigma}_{yx} & \mathbf{\Sigma}_y \end{bmatrix} \right)$$

  where

$$\mathbf{\Sigma}_x = \mathbb{E}[(\mathbf{X} - \mathbf{m}_x)(\mathbf{X} - \mathbf{m}_x)^{\mathsf{H}}], \quad \mathbf{\Sigma}_y = \mathbb{E}[(\mathbf{Y} - \mathbf{m}_y)(\mathbf{Y} - \mathbf{m}_y)^{\mathsf{H}}]$$

$$\mathbf{\Sigma}_{xy} = \mathbb{E}[(\mathbf{X} - \mathbf{m}_x)(\mathbf{Y} - \mathbf{m}_y)^{\mathsf{H}}]$$

  we define the MSE as

$$\mathsf{MSE} = \mathbb{E}[\|\mathbf{X} - \widehat{\mathbf{X}}\|^2] = \mathbb{E}[(\mathbf{X} - \widehat{\mathbf{X}})^{\mathsf{H}}(\mathbf{X} - \widehat{\mathbf{X}})]$$

- Result: all the derivations and results found before are still valid when replacing "transpose" with "Hermitian transpose".

Gaussian signal in Gaussian noise

- Often we need to estimate a signal observed through a linear transformation $\mathbf{H}$ in additive noise:

$$\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathbf{Z}$$

  where $\mathbf{X} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}_x)$ and $\mathbf{Z} = \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}_z)$.

- In this case, we have

$$\widehat{\mathbf{X}} = \mathbf{\Sigma}_x \mathbf{H}^{\mathsf{H}} \left(\mathbf{H}\mathbf{\Sigma}_x \mathbf{H}^{\mathsf{H}} + \mathbf{\Sigma}_z \right)^{-1} \mathbf{Y}$$

  with estimation error covariance

$$\mathbf{\Sigma}_{x|y} = \mathbf{\Sigma}_x - \mathbf{\Sigma}_x \mathbf{H}^{\mathsf{H}} \left(\mathbf{H}\mathbf{\Sigma}_x \mathbf{H}^{\mathsf{H}} + \mathbf{\Sigma}_z \right)^{-1} \mathbf{H}\mathbf{\Sigma}_x$$

Example: MMSE Multi-user Detection

- A Gaussian Multiple Access Channel can be represented as

$$\mathbf{Y} = \sum_{k=1}^{K} \sqrt{P_k}\mathbf{s}_k X_k + \mathbf{Z} = \mathbf{S}\mathbf{P}^{1/2}\mathbf{X} + \mathbf{Z}$$

where $\mathbf{s}_k = (s_{1,k}, \ldots, s_{N,k})^{\mathsf{T}}$ is the vector formed by the samples of user k waveform, P_k is the received power of user k, X_k are information symbols from a unit energy signal constellation (e.g., QAM), and $\mathbf{Z} \sim \mathcal{CN}(\mathbf{0}, N_0\mathbf{I})$.

- A linear detector for user k consists of a projection of $\mathbf{Y}$ onto a unit vector $\mathbf{u}_k$, forming the scalar observation $\widehat{X}_k = \mathbf{u}_k^{\mathsf{H}}\mathbf{Y}$.
- We define the Signal to Interference plus Noise Ratio (**SINR**) as

$$\text{SINR}_k = \frac{\left|\mathbf{u}_k^{\mathsf{H}}\mathbf{s}_k\right|^2 P_k}{N_0 + \sum_{j \neq k} \left|\mathbf{u}_k^{\mathsf{H}}\mathbf{s}_j\right|^2 P_j}$$

- It can be shown that the SINR is maximized over all linear detectors by choosing

$$\mathbf{u}_k = \alpha_k \left(N_0 \mathbf{I} + \sum_{j=1}^{K} P_j \mathbf{s}_j \mathbf{s}_j^{\mathsf{H}} \right)^{-1} \mathbf{s}_k$$

  where α_k is a normalization constant in order to have $\|\mathbf{u}_k\| = 1$.
- Notice that this SINR-maximizing detector is **proportional** to the MMSE estimator of X_k given $\mathbf{Y}$.
- The resulting maximum SINR can be compactly written as

$$\mathrm{SINR}_k = P_k \mathbf{s}_k^{\mathsf{H}} \left(N_0 \mathbf{I} + \sum_{j \neq k} P_j \mathbf{s}_j \mathbf{s}_j^{\mathsf{H}} \right)^{-1} \mathbf{s}_k.$$

Thank you! Q & A?

Exercises

Method of moments: Gamma distribution

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{Gamma}(\alpha, \beta)$, derive the corresponding MoM estimators $\hat{\alpha}, \hat{\beta}$ for the parameters α and β, and **try** to derive the bias-variance decomposition of their MSE.

Binary Signal in Gaussian noise

Consider X taking values in $\mathcal{X} = \{+1, -1\}$ with equal probability, and the observation

$$Y = hX + Z$$

where $h \in \mathbb{R}_+$ and $Z \sim \mathcal{N}(0, \sigma^2)$. Show that

- the linear MMSE estimator is given by $\widehat{X}_{\text{lin}} = \frac{h}{h^2+\sigma^2} Y$; and
- the optimal MMSE estimator is

$$\widehat{X}_{\text{opt}} = \tanh\left(\frac{hY}{\sigma^2}\right).$$

Method of moments: Gamma distribution

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{Gamma}(\alpha, \beta)$, derive the corresponding MoM estimators $\hat{\alpha}, \hat{\beta}$ for the parameters α and β, and **try** to derive the bias-variance decomposition of their MSE.

- For $X \text{Gamma} \sim (\alpha, \beta)$, we have

$$\mathbb{E}[X] = \frac{\alpha}{\beta}, \quad \mathbb{E}[X^2] = \frac{\alpha^2 + \alpha}{\beta^2}. \tag{13}$$

- This leads to the MoM estimators as

$$\hat{\alpha} =, \quad \hat{\beta} =, \tag{14}$$

with corresponding bias and variance given by

$$\mathbb{E}[\hat{\alpha}] - \alpha =, \quad \mathbb{E}[\hat{\alpha}] - \alpha =, \quad \text{Var}[\hat{\alpha}] =, \quad \text{Var}[\hat{\beta}] = \tag{15}$$

so that MSE as

$$\mathbb{E}[(\hat{\alpha} - \alpha)^2] =, \quad \mathbb{E}[(\hat{\beta} - \beta)^2] =, \tag{16}$$

Exercises

Binary Signal in Gaussian noise

Consider X taking values in $\mathcal{X} = \{+1, -1\}$ with equal probability, and the observation

$$Y = hX + Z$$

where $h \in \mathbb{R}_+$ and $Z \sim \mathcal{N}(0, \sigma^2)$. Show that

- the linear MMSE estimator is given by $\widehat{X}_{\text{lin}} = \frac{h}{h^2+\sigma^2} Y$; and
- the optimal MMSE estimator is

$$\widehat{X}_{\text{opt}} = \tanh\left(\frac{hY}{\sigma^2}\right).$$

- for LMMSE, consider $\widehat{X}_{\text{lin}} = aY$, and it suffices to determine $\alpha \in \mathbb{R}$ that minimizes $\mathbb{E}[(X - \hat{X})^2] = \mathbb{E}[X^2 - 2X\hat{X} + \hat{X}^2] = \mathbb{E}[X^2] - 2a\mathbb{E}[X(hX+Z)] + a^2\mathbb{E}[(hX+Z)^2] = 1 - 2a(h+0) + a^2(h^2+\sigma^2)$.
- This leads to $a = \frac{h}{h^2+\sigma^2}$ and thus the conclusion.
- To derive the optimal MMSE estimator, we use the conclusion that $\widehat{X}_{\text{opt}} = \mathbb{E}[X|Y]$.
- For given Y, we have that $X = \frac{Y-Z}{h}$, for $Z \sim \mathcal{N}(0, \sigma^2)$,