Mathematics of Data: From Theory to Computation

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 2: A basic review of probability theory

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

EE-556 (Fall 2015)

lions@epfl

MARIE CURIE ACTIONS
FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION
erc
EPFL

License Information for Mathematics of Data Slides

- This work is released under a Creative Commons License with the following terms:
- **Attribution**
  - The licensor permits others to copy, distribute, display, and perform the work. In return, licensees must give the original authors credit.
- **Non-Commercial**
  - The licensor permits others to copy, distribute, display, and perform the work. In return, licensees may not use the work for commercial purposes – unless they get the licensor's permission.
- **Share Alike**
  - The licensor permits others to distribute derivative works only under a license identical to the one that governs the licensor's work.
- Full Text of the License

- This lecture
  1. Review of probability theory
  2. Learning as an optimization problem
- Next lecture
  1. Basic concepts in convex analysis
  2. Complexity theory review

Recommended reading

- *Probability and Measure*, Patrick Billingsley, Wiley-Interscience, 1995.
- Chapter 7, 8, & 9 in K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, 2012.
- V. N. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Inf. Theory*, vol. 10, no. 5, pp. 988–999, Sep. 1999.
- *Chapter 5 in A. W. van der Vaart, *Asymptotic Statistics*, Cambridge Univ. Press, 1998.

Motivation

Motivation

This lecture reviews basic probability and statistics.

Basic concepts in probability theory

Definition (Sample space)

The sample space Ω of an experiment is the set of all possible outcomes of that experiment.

Example

If the experiment is tossing a coin, the sample set is the set $\{$head, tail$\}$.

Definition (Event)

An event E corresponds to a subset of the sample space; i.e., $E \subseteq \Omega$.

Definition (Probability measure)

Probability measure $P(E)$ maps event E from Ω onto the interval $[0, 1]$ and satisfies the following Kolmogorov axioms:

- $P(E) \geq 0$,
- $P(\Omega) = 1$ and
- $P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i)$, where $E_1, ..., E_n$ are mutually exclusive (i.e. $\bigcap_{i=1}^{n} E_i = \emptyset$). Such events are called *independent*.

Union of non-disjoint events

Definition (Principle of inclusion-exclusion)

The probability of the union of n events is

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \leq i_1 \leq \ldots \leq i_k \leq n} P(E_{i_1} \cap \ldots \cap E_{i_k}),$$

where the second sum is over all subsets of k events.

Union of non-disjoint events

Definition (Principle of inclusion-exclusion)

The probability of the union of n events is

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \leq i_1 \leq \ldots \leq i_k \leq n} P(E_{i_1} \cap \ldots \cap E_{i_k}),$$

where the second sum is over all subsets of k events.

Example

Suppose we throw two dices and ask what is the probability that the outcome is even or larger than 7. Let A and B denote the event of having an even number and the event of getting the number that exceeds 7, respectively. Then, $P(A) = \frac{1}{2}$, $P(B) = \frac{15}{36}$ and $P(A \cap B) = \frac{9}{36}$.

By the inclusion-exclusion principle, $P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{2}{3}$.

The rules of probability

Let A and B denote two events in a sample space Ω, and let $P(B) \neq 0$.

Definition (Marginal probability)

The probability of an event (A) occuring $(P(A))$.

Definition (Joint probability)

$P(A, B)$ is the probability of event A and event B occuring. Symmetry property holds, i.e. $P(A, B) = P(B, A)$.

Definition (Conditional probability)

$P(B|A)$ is the probability that B will occur given that A has occurred.

Rules

- Sum rule: $P(A) = \sum_B P(A, B)$
- Product rule: $P(A, B) = P(B|A)P(A)$.

Bayes' rule

Bayes' rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Constituents:

- $P(A)$, the prior probability, is the probability of A before B is observed.
- $P(A|B)$, the posterior probability, is the probability of A given B, i.e., after B is observed.
- $P(B|A)$ is the probability of observing B given A. As a function of A with B fixed, this is the likelihood.

Random variable

Definition

A real-valued random variable is a **function** that associates a value to the outcome of a randomized experiment $X : \Omega \to \mathbb{R}$.

Example

- Whether a coin flip was heads: a function from $\Omega = \{H, T\}$ to $\{0, 1\}$
- Number of heads in a sequence of n throws: function from $\Omega = \{H, T\}^n$ to $\{0, 1, ..., n\}$.

Discrete random variable

Probability mass function (Pmf)

The probability mass function is the function from values to its probability, $P_X(x) = P(X = x)$ for $x \in \mathcal{X}$ (i.e., a countable subset of the reals) with properties:

- $P_X(x) \geq 0$ for every $x \in \mathcal{X}$,
- $\sum_{x \in \mathcal{X}} P_X(x) = 1$

Example

Discrete distributions:

- Bernoulli distribution - distribution of a binary variable $x \in \{0, 1\}$; single parameter $\mu \in [0, 1]$ represents the probability of $x = 1$:

$$\text{Bern}(x|\mu) = \mu^x (1 - \mu)^{1-x}.$$

- Binomial distribution - probability of observing m occurrences of 1 in a set of N samples from a Bernoulli distribution:

$$\text{Bin}(m|N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{1-m}.$$

- Other important discrete distributions: Categorical, Multinomial, Poisson, Geometric, Negative binomial, etc.

Probability density function (pdf)

- A continuous random variable can have uncountably infinite possible values.

Probability density function (pdf)

The probability density function of a continuous random variable X is an integrable function $p(x)$ satisfying the following:

1. The density is nonnegative: i.e., $p(x) \geq 0$ for any x,
2. Probabilities integrate to 1: i.e., $\int_{-\infty}^{\infty} p(x)dx = 1$,
3. The probability that x belongs to the interval $[a, b]$ is given by the integral of $p(x)$ over that interval: i.e.,

$$P(a \leq X \leq b) = \int_a^b p(x)dx.$$

Basic rules of probability

1. Analog of sum rule: $p(x) = \int p(x, y)dy$
2. Product rule: $p(x, y) = p(y|x)p(x)$.

Expectations and variances

Definition (Expectation (1^{st} moment, mean))

$$\mathbb{E}[X] = \begin{cases} \sum_{x\in\mathcal{X}} xP(X=x) & \text{discrete} \\ \int_{-\infty}^{\infty} xp(x)dx & \text{continuous} \end{cases}$$

Definition (Variance (2^{nd} moment))

$$\mathbb{V}[X] = \begin{cases} \sum_{x\in\mathcal{X}} (x-\mathbb{E}[X])^2 P(X=x) & \text{discrete} \\ \int_{-\infty}^{\infty} (x-\mathbb{E}[X])^2 p(x)dx & \text{continuous} \end{cases}$$

Definition (Conditional expectation and Covariance)

$$\mathbb{E}[X|Y=y] = \sum_{x\in\mathcal{X}} xP(X=x|Y=y)$$

$$\text{cov}[x,y] = \mathbb{E}\Big[\Big(x-\mathbb{E}[X]\Big)\Big(y-\mathbb{E}[Y]\Big)\Big]$$

Probability distributions for continuous variables

Common distributions:

- Uniform
- Normal / Gaussian
- Beta
- Chi-Squared
- Exponential
- Gamma
- Laplace

Normal (Gaussian) Distribution

Gaussian distribution

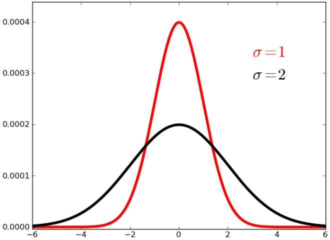
For $\mathbf{x} \in \mathbb{R}^d$, the multivariate Gaussian distribution takes the form

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right),$$

where $\boldsymbol{\mu} \in \mathbb{R}^d$ is the mean, $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ is the covariance matrix and $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$.

- In the case of a single variable

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$



Law of large numbers and central limit theorem

Theorem (Strong Law of Large Numbers)

Let X be a real-valued random variable with the finite first moment $\mathbb{E}[X]$, and let $X_1, X_2, ..., X_n$ be an infinite sequence of independent and identically distributed copies of X. Then the empirical average of this sequence $\bar{X}_n := \frac{1}{n}(X_1 + ... + X_n)$ converges almost surely to $\mathbb{E}[X]$ i.e., $P\Big(lim_{n\to\infty}\bar{X}_n = \mathbb{E}[X]\Big) = 1$.

Theorem (Central Limit Theorem)

Let $X_1, ...X_n$ be a sequence of independent and identically distributed random variables each having mean μ and variance σ^2. Then the distribution of $\frac{X_1 + ... + X_n - n\mu}{\sigma\sqrt{n}}$ tends to the standard normal as $n \to \infty$. That is, for $-\infty < a < \infty$,

$$P\left(\frac{X_1 + ... + X_n - n\mu}{\sigma\sqrt{n}} \leq a\right) \to \frac{1}{2\pi}\int_{-\infty} ae^{-x^2/2}dx$$

as $n \to \infty$.

- Intuitively, the sampling distribution of the mean will be close to Gaussian, if you just take enough independent samples.

Basic statistics

Parametric estimation model

A parametric estimation model consists of the following four elements:

1. A *parameter space*, which is a subset $\mathcal{X}$ of $\mathbb{R}^p$
2. A *parameter* $\mathbf{x}^\natural$, which is an element of the parameter space
3. A class of probability distributions $\mathcal{P}_\mathcal{X} := \{\mathbb{P}_\mathbf{x} : \mathbf{x} \in \mathcal{X}\}$, parametrized by $\mathbf{x} \in \mathcal{X}$
4. A *sample* $\mathbf{b}$, which follows the probability distribution $\mathbf{b} \sim \mathbb{P}_{\mathbf{x}^\natural} \in \mathcal{P}_\mathcal{X}$

Statistical estimation seeks to approximate the value of $\mathbf{x}^\natural$, given $\mathcal{X}$, $\mathcal{P}_\mathcal{X}$, and $\mathbf{b}$.

Definition (Estimator)

An estimator $\hat{\mathbf{x}}$ is a mapping that takes $\mathcal{X}$, $\mathcal{P}_\mathcal{X}$, and $\mathbf{b}$ as inputs, and outputs a value in $\mathbb{R}^p$.

- The output of an estimator depends on the sample, and hence, is random.
- The output of an estimator is not necessarily equal to $\mathbf{x}^\natural$.

Ordinary least-squares estimator

Ordinary least-squares estimator (OLS)

The ordinary least-squares estimator is given by

$$\hat{\mathbf{x}}_{\mathsf{OLS}} \in \arg\min_{\mathbf{x}} \left\{ \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 : \mathbf{x} \in \mathbb{R}^p \right\}.$$

Ordinary least squares estimator: An intuitive model

Gaussian linear model

Let $\mathbf{x}^\natural \in \mathbb{R}^p$. Let $\mathbf{b} := \mathbf{A}\mathbf{x}^\natural + \mathbf{w} \in \mathbb{R}^n$ for some matrix $\mathbf{A} \in \mathbb{R}^{n\times p}$, where $\mathbf{w}$ is a Gaussian vector with zero mean and covariance matrix $\sigma^2 I$.

The probability density function $p_{\mathbf{x}}(\cdot)$ is given by

$$p_{\mathbf{x}}(\mathbf{b}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2\right).$$

Therefore, the maximum likelihood (ML) estimator is defined as

$$\hat{\mathbf{x}}_{\mathsf{ML}} \in \arg\min_{\mathbf{x}} \left\{-\log p_{\mathbf{x}}(\mathbf{b}) = -\frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 : \mathbf{x} \in \mathbb{R}^p\right\},$$

which is equivalent to

$$\hat{\mathbf{x}}_{\mathsf{ML}} \in \arg\min_{\mathbf{x}} \left\{\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 : \mathbf{x} \in \mathbb{R}^p\right\}.$$

OLS is the ML estimator for the Gaussian linear model.

Maximum-likelihood estimator

Recall the general setting.

Parametric estimation model

A parametric estimation model consists of four elements:

1. A *parameter space*, which is a subset $\mathcal{X}$ of $\mathbb{R}^p$,
2. A *parameter* $\mathbf{x}^\natural$, which is an element of the parameter space,
3. A class of probability distributions $\mathcal{P}_{\mathcal{X}} := \{\mathbb{P}_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$, parametrized by $\mathbf{x} \in \mathcal{X}$,
4. A *sample* $\mathbf{b}$, which follows the probability distribution $\mathbb{P}_{\mathbf{x}^\natural} \in \mathcal{P}_{\mathcal{X}}$.

Definition (Maximum-likelihood estimator)

The maximum-likelihood (ML) estimator is given by

$$\hat{\mathbf{x}}_{\mathsf{ML}} \in \arg\min_{\mathbf{x}} \{-\log p_{\mathbf{x}}(\mathbf{y})\},$$

where $p_{\mathbf{x}}(\cdot)$ denotes the probability density function or probability mass function of $\mathbb{P}_{\mathbf{x}}$, for $\mathbf{x} \in \mathcal{X}$.

Logistic regression

Logistic regression [1]

Let $\mathbf{x}^\natural \in \mathbb{R}^p$. Let $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^p$ be given. The sample is given by $\mathbf{b} := (b_1, \ldots, b_n) \in \{-1, 1\}^n$, where each b_i is a Bernoulli random variable satisfying

$$\mathbb{P}\left\{b_i = 1\right\} = 1 - \mathbb{P}\left\{b_i = -1\right\} = \left[1 + \exp\left(-\left\langle \mathbf{a}_i, \mathbf{x}^\natural \right\rangle\right)\right]^{-1},$$

and $b_1, \ldots, b_n$ are independent.

The probability mass function $p_{\mathbf{x}}(\cdot)$ is given by

$$p_{\mathbf{x}}(\mathbf{b}) = \Pi_{i=1}^n \left[1 + \exp\left(-b_i \left\langle \mathbf{a}_i, \mathbf{x} \right\rangle\right)\right]^{-1}.$$

Therefore, the maximum-likelihood estimator is defined as

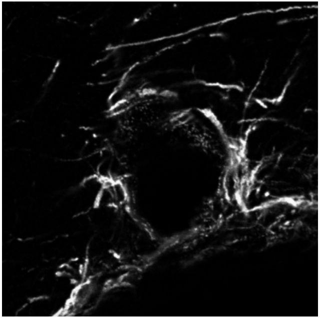
$$\hat{\mathbf{x}}_{\mathsf{ML}} \in \arg\min_{\mathbf{x}} \left\{ -\log p_{\mathbf{x}}(\mathbf{b}) = \sum_{i=1}^n \log\left[1 + \exp\left(-b_i \left\langle \mathbf{a}_i, \mathbf{x} \right\rangle\right)\right] : \mathbf{x} \in \mathbb{R}^p \right\}.$$

- $\hat{\mathbf{x}}_{\mathsf{ML}}$ defines a *linear classifier*. For any new $\mathbf{a}_i$, $i \geq n+1$, we can predict the corresponding b_i by predicting $b_i = 1$ if $\langle \mathbf{a}_i, \hat{\mathbf{x}}_{\mathsf{ML}} \rangle \geq 0$, and $b_i = -1$ otherwise.

ML estimation in photon-limited imaging systems

Statistical model of a photon-limited imaging system [2, 3]

Let $\mathbf{x}^\natural \in \mathbb{R}^p$. Let $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^p$ be given vectors. The sample is given by $\mathbf{b} := (b_1, \ldots, b_n) \in \mathbb{N}^n$, where each b_i is a Poisson random variable with mean $\left\langle \mathbf{a}_i, \mathbf{x}^\natural \right\rangle$ that denotes the number of detected photons, and $b_1, \ldots, b_n$ are independent.



Confocal imaging

In confocal imaging, the vectors $\mathbf{a}_i$ can be used to capture the lens effects, including blur and (spatial) low-pass filtering (due to the numerical aperture of the lens).

ML estimation in photon-limited imaging systems contd.

Statistical model of a photon-limited imaging system [2, 3]

Let $\mathbf{x}^\natural \in \mathbb{R}^p$. Let $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^p$ be given vectors. The sample is given by $\mathbf{b} := (b_1, \ldots, b_n) \in \mathbb{N}^n$, where each b_i is a Poisson random variable with mean $\left\langle \mathbf{a}_i, \mathbf{x}^\natural \right\rangle$ that denotes the number of detected photons, and $b_1, \ldots, b_n$ are independent.

The probability mass function $p_{\mathbf{x}}(\cdot)$ is given by

$$p_{\mathbf{x}}(\mathbf{b}) = \Pi_{i=1}^{n} (b_i!)^{-1} \exp\left(-\langle \mathbf{a}_i, \mathbf{x}\rangle\right) \langle \mathbf{a}_i, \mathbf{x}\rangle^{b_i}.$$

Therefore, the maximum-likelihood estimator is defined as

$$\hat{\mathbf{x}}_{\mathsf{ML}} \in \arg\min_{\mathbf{x}} \left\{ -\log p_{\mathbf{x}}(\mathbf{b}) = \sum_{i=1}^{n} \left[\log(b_i!) + \langle \mathbf{a}_i, \mathbf{x}\rangle - b_i \log\left(\langle \mathbf{a}_i, \mathbf{x}\rangle\right)\right] : \mathbf{x} \in \mathbb{R}^p \right\},$$

which is equivalent to

$$\hat{\mathbf{x}}_{\mathsf{ML}} \in \arg\min_{\mathbf{x}} \left\{ \sum_{i=1}^{n} \left[\langle \mathbf{a}_i, \mathbf{x}\rangle - b_i \log\left(\langle \mathbf{a}_i, \mathbf{x}\rangle\right)\right] : \mathbf{x} \in \mathbb{R}^p \right\}.$$

Regression

Basic regression model

Let $\mathbf{x}^\natural \in \mathbb{R}^p$. Let $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^p$ be given vectors. The sample is given by $\mathbf{b} := (b_1, \ldots, b_n) \in \mathbb{B}^n$ for some set $\mathbb{B}$, where each b_i follows a probability distribution $\mathbb{P}_{\mathbf{x}^\natural, \mathbf{a}_i}$ determined by $\mathbf{x}^\natural$ and $\mathbf{a}_i$, and $b_1, \ldots, b_n$ are independent.

Examples

The statistical models we have discussed are all regression models.

- The *Gaussian linear regression model* is a regression model, where each b_i is a Gaussian random variable with mean $\left\langle \mathbf{a}_i, \mathbf{x}^\natural \right\rangle$ and variance σ^2, for some $\sigma > 0$.
- The *logistic regression model* is a regression model, where each b_i is a Bernoulli random variable with

$$\mathbb{P}\left\{b_i = 1\right\} = 1 - \mathbb{P}\left\{b_i = -1\right\} = \left[1 + \exp\left(-\left\langle \mathbf{a}_i, \mathbf{x}^\natural \right\rangle\right)\right]^{-1}.$$

- The statistical model for photon-limited imaging systems is a *Poisson regression model*, where each b_i is a Poisson random variable with mean $\left\langle \mathbf{a}_i, \mathbf{x}^\natural \right\rangle$.

M-Estimators

Recall that an ML estimator $\hat{\mathbf{x}}_{\mathsf{ML}}$ takes the form

$$\hat{\mathbf{x}}_{\mathsf{ML}} \in \arg\min_{\mathbf{x}} \left\{ L(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^p \right\},$$

where L denotes the negative log-likelihood function. In general, L can be replaced by another suitably designed function.

Definition (M-Estimator)

An M-estimator $\hat{\mathbf{x}}_M$ is an estimator of the form

$$\hat{\mathbf{x}}_M \in \arg\min_{\mathbf{x}} \left\{ f(\mathbf{x}) : \mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^p \right\},$$

for some function f *depending on* the sample space $\mathcal{X}$, class of probability distributions $\mathcal{P}_{\mathcal{X}}$, and sample $\mathbf{b}$.

- The term "M-estimator" denotes "maximum-likelihood-type estimator" [4].

Graphical model learning

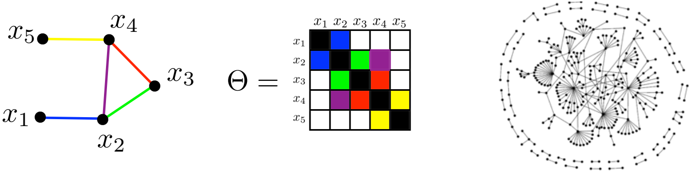
Graphical model selection

Let $\mathbf{\Theta}^{\natural} \in \mathbb{R}^{p \times p}$ be a positive-definite matrix. The sample is given by $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$, which are i.i.d. random vectors with zero mean and covariance matrix ${\mathbf{\Theta}^{\natural}}^{-1}$.

We can consider the M-estimator

$$\widehat{\mathbf{\Theta}}_M \in \arg\min_{\mathbf{\Theta}} \left\{ \mathrm{Tr}\left(\widehat{\mathbf{\Sigma}} \mathbf{\Theta} \right) - \log\det\left(\mathbf{\Theta} \right) : \mathbf{\Theta} \in \mathbb{S}^p_{++} \right\},$$

where $\widehat{\mathbf{\Sigma}}$ is the empirical covariance matrix, i.e., $\widehat{\mathbf{\Sigma}} := (1/n) \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$ [5].



Graphical model learning contd.

Graphical model selection

Let $\boldsymbol{\Theta}^\natural \in \mathbb{R}^{p \times p}$ be a positive-definite matrix. The sample is given by $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$, which are i.i.d. random vectors with zero mean and covariance matrix ${\boldsymbol{\Theta}^\natural}^{-1}$.

The M-estimator becomes the ML estimator when $\mathbf{x}_i$'s are Gaussian random vectors. The probability density function $p_{\boldsymbol{\Theta}}(\cdot)$ is given by

$$
\begin{aligned}
p_{\boldsymbol{\Theta}}(\mathbf{x}_1, \ldots, \mathbf{x}_n) &= \Pi_{i=1}^n \left[(2\pi)^{-p/2} \det\left(\boldsymbol{\Theta}^{-1}\right)^{-1/2} \exp\left(-\frac{1}{2}\mathbf{x}_i^T \boldsymbol{\Theta} \mathbf{x}_i\right)\right] \\
&= (2\pi)^{-np/2} \det(\boldsymbol{\Theta})^{n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^n \left(\mathbf{x}_i^T \boldsymbol{\Theta} \mathbf{x}_i\right)\right]
\end{aligned}
$$

Therefore, the ML estimator is defined as

$$
\hat{\mathbf{x}}_{\mathsf{ML}} \in \arg\min_{\boldsymbol{\Theta}} \left\{ -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log\det(\boldsymbol{\Theta}) + \frac{n}{2}\mathrm{Tr}\left(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}\right) : \boldsymbol{\Theta} \in \mathbb{S}^p_{++} \right\},
$$

which is equivalent to the M-estimator $\widehat{\boldsymbol{\Theta}}_M$.

Checking the fidelity

Given an estimator $\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathcal{X}} \{F(\mathbf{x})\}$, we need to address two key questions:

1. Is the formulation reasonable?
2. What is the role of the data size?

Standard approach to checking the fidelity

Standard approach

1. Specify a performance criterion $\mathcal{L}(\hat{\mathbf{x}}, \mathbf{x}^\natural)$ that should be small if $\hat{\mathbf{x}} = \mathbf{x}^\natural$.
2. Show that $\mathcal{L}$ is actually *small in some sense* when *some condition* is satisfied.

Example

Take the ℓ_2-error $\mathcal{L}(\hat{\mathbf{x}}, \mathbf{x}^\natural) := \left\|\hat{\mathbf{x}} - \mathbf{x}^\natural\right\|_2^2$ as an example. Then we may verify the fidelity via one of the following ways, where ε denotes a small enough number:

1. $\mathbb{E}\left[\mathcal{L}(\hat{\mathbf{x}}, \mathbf{x}^\natural))\right] \leq \varepsilon$ (expected error),
2. $\mathbb{P}\left(\mathcal{L}(\hat{\mathbf{x}}, \mathbf{x}^\natural) \geq \epsilon\right) \leq \delta$ for some δ depending on ϵ (consistency),
3. $\sqrt{n}(\hat{\mathbf{x}} - \mathbf{x}^\natural)$ converges in distribution to $\mathcal{N}(0, \mathbf{I})$ (asymptotic normality),
4. $\sqrt{n}(\hat{\mathbf{x}} - \mathbf{x}^\natural)$ converges in distribution to $\mathcal{N}(0, \mathbf{I})$ in a local neighborhood (local asymptotic normality).

if *some condition* is satisfied. Such conditions typically revolve around the data size.

Approach 1: Expected error

Gaussian linear model

Let $\mathbf{x}^\natural \in \mathbb{R}^p$ and let $\mathbf{A} \in \mathbb{R}^{n \times p}$. The samples are given by $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, where $\mathbf{w}$ is a sample of a Gaussian random vector $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$.

What is the performance of the ML estimator

$$\hat{\mathbf{x}}_{\mathsf{ML}} \in \arg\min_{\mathbf{x} \in \mathbb{R}^p} \left\{ \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 \right\}?$$

Theorem (Performance of the LS estimator [6])

If $\mathbf{A}$ is a matrix of independent and identically distributed (i.i.d.) standard Gaussian distributed entries, and if $n > p + 1$, then

$$\mathbb{E}\left[\left\| \hat{\mathbf{x}}_{ML} - \mathbf{x}^\natural \right\|_2^2 \right] = \frac{p}{n - p - 1}\sigma^2 \to 0 \text{ as } \frac{n}{p} \to \infty.$$

*Approach 2: Consistency

Covariance estimation

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be samples of a sub-Gaussian random vector with zero mean and some unknown positive-definite covariance matrix $\boldsymbol{\Sigma}^\natural \in \mathbb{R}^{p \times p}$. (Sub-Gaussian random variables will be defined in recitation.)

What is the performance of the M-estimator $\widehat{\boldsymbol{\Sigma}} := \widehat{\boldsymbol{\Theta}}^{-1}$, where

$$\widehat{\boldsymbol{\Theta}}_{\mathsf{ML}} \in \arg \min_{\boldsymbol{\Theta} \in \mathbb{S}^p_{++}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[- \log \det \left(\boldsymbol{\Theta} \right) + \mathbf{x}_i^T \boldsymbol{\Theta} \mathbf{x}_i \right] \right\} ?$$

- If $\mathbf{y} = f(\mathbf{x})$, then $\hat{\mathbf{y}}_{\mathsf{ML}} = f(\hat{\mathbf{x}}_{\mathsf{ML}})$. This is called the *functional invariance* property of ML estimators.

Theorem (Performance of the ML estimator [5])

Suppose that the diagonal elements of $\boldsymbol{\Sigma}^\natural$ *are bounded above by* $\kappa > 0$*, and each* $X_i / \sqrt{\left(\Sigma^\natural\right)_{i,i}}$ *is sub-Gaussian with parameter* c*. Then*

$$\mathbb{P}\left(\left\{ \left| \left(\widehat{\boldsymbol{\Sigma}}_{ML} \right)_{i,j} - \left(\boldsymbol{\Sigma}^\natural \right)_{i,j} \right| > t \right\} \right) \leq 4 \exp \left[- \frac{n t^2}{128 \left(1 + 4c^2 \right) \kappa^2} \right] \to 0 \text{ as } n \to \infty$$

for all $t \in \left(0, 8\kappa \left(1 + 4c^2 \right) \right)$.

*Approach 3: Asymptotic normality

Logistic regression

Let $\mathbf{x}^\natural \in \mathbb{R}^p$, and let $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^p$. Let $b_1, \ldots, b_n$ be samples of independent random variables $B_1, \ldots, B_n$. Each random variable B_i takes values in $\{-1, 1\}$ and follows $\mathbb{P}\left(\{B_i = 1\}\right) := \ell_i(\mathbf{x}^\natural) = \left[1 + \exp\left(-\left\langle \mathbf{a}_i, \mathbf{x}^\natural \right\rangle\right)\right]^{-1}$ (i.e., the logistics loss).

What is the performance of the ML estimator

$$\hat{\mathbf{x}}_{\mathsf{ML}} \in \arg\min_{\mathbf{x} \in \mathbb{R}^p} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \ln \left[\mathbb{I}_{\{B_i = 1\}} \ell_i(\mathbf{x}) + \mathbb{I}_{\{B_i = 0\}} \left(1 - \ell_i(\mathbf{x})\right) \right] := -\frac{1}{n} f_n(\mathbf{x}) \right\}?$$

*Approach 3: Asymptotic normality

Theorem (Performance of the ML estimator [7] (*also valid for generalized linear models))

The random variable $\mathbf{J}(\mathbf{x}^\natural)^{-1/2}\left(\hat{\mathbf{x}}_{ML} - \mathbf{x}^\natural\right)$ *converges in distribution to* $\mathcal{N}(\mathbf{0}, \mathbf{I})$ *if* $\lambda_{\min}(\mathbf{J}(\mathbf{x}^\natural)) \to \infty$ *and*

$$\max_{\mathbf{x}\in\mathbb{R}^p}\left\{\left\|\mathbf{J}(\mathbf{x}^\natural)^{-1/2}\mathbf{J}(\mathbf{x})\mathbf{J}(\mathbf{x}^\natural)^{-1/2} - \mathbf{I}\right\|_{2\to 2} : \left\|\mathbf{J}(\mathbf{x}^\natural)^{1/2}\left(\mathbf{x} - \mathbf{x}^\natural\right)\right\|_2 \leq \delta\right\} \to 0 \quad (1)$$

for all $\delta > 0$ *as* $n \to \infty$, *where* $\mathbf{J}(\mathbf{x}) := -\mathbb{E}\left[\nabla^2 f_n(\mathbf{x})\right]$ *is the Fisher information matrix.*

Roughly speaking, assuming that p is fixed, we have the following observations.

1. The technical condition (1) means that $\mathbf{J}(\mathbf{x}) \sim \mathbf{J}(\mathbf{x}^\natural)$ for all $\mathbf{x}$ in a neighborhood $N_{\mathbf{x}^\natural}(\delta)$ of $\mathbf{x}^\natural$, and $N_{\mathbf{x}^\natural}(\delta)$ becomes larger with increasing n.
2. $\left\|\mathbf{J}(\mathbf{x}^\natural)^{-1/2}\left(\hat{\mathbf{x}}_{\mathrm{ML}} - \mathbf{x}^\natural\right)\right\|_2^2 \sim \mathrm{Tr}\left(\mathbf{I}\right) = p$, which means that $\left\|\hat{\mathbf{x}}_{\mathrm{ML}} - \mathbf{x}^\natural\right\|_2^2$ decreases at the rate $\lambda_{\min}(\mathbf{J}(\mathbf{x}^\natural))^{-1} \to 0$ asymptotically.

*Approach 4: Local asymptotic normality

In general, the asymptotic normality does not hold even in the independent identically distributed (i.i.d.) case, but we may have the *local asymptotic normality (LAN)*.

ML estimation with i.i.d. samples

Let $b_1, \ldots, b_n$ be independent samples of a random variable B, whose probability density function is known to be in the set $\{p_{\mathbf{x}}(b) : \mathbf{x} \in \mathcal{X}\}$ with some $\mathcal{X} \subseteq \mathbb{R}^p$.

What is the performance of the ML estimator

$$\hat{\mathbf{x}}_{\mathsf{ML}} \in \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \ln\left[p_{\mathbf{x}}(b_i)\right] \right\}?$$

⋆Approach 4: Local asymptotic normality

Theorem (Performance of the ML estimator (cf. [8, 9] for details))

Under some technical conditions, the random variable $\sqrt{n}\,\mathbf{J}^{-1/2}\left(\hat{\mathbf{x}}_{ML} - \mathbf{x}^\natural\right)$ *converges in distribution to* $\mathcal{N}(\mathbf{0}, \mathbf{I})$*, where* $\mathbf{J}$ *is the Fisher information matrix associated with one sample, i.e.,*

$$\mathbf{J} := -\mathbb{E}\left[\nabla^2_{\mathbf{x}} \ln\left[p_{\mathbf{x}}(B)\right]\right]\Big|_{\mathbf{x}=\mathbf{x}^\natural}.$$

Roughly speaking, assuming that p is fixed, we can observe that

- $\left\| \sqrt{n}\,\mathbf{J}^{-1/2}\left(\hat{\mathbf{x}}_{\mathrm{ML}} - \mathbf{x}^\natural\right)\right\|_2^2 \sim \mathrm{Tr}\left(\mathbf{I}\right) = p$,
- $\left\|\hat{\mathbf{x}}_{\mathrm{ML}} - \mathbf{x}^\natural\right\|_2^2 = \mathcal{O}(1/n)$.

Example: ML estimation for quantum tomography

Problem (Quantum tomography)

A quantum system of q qubits can be characterized by a density operator, i.e., a Hermitian positive semidefinite $\mathbf{X}^\natural \in \mathbb{C}^{p \times p}$ with $p = 2^q$. Let $\{\mathbf{A}_1, \ldots, \mathbf{A}_m\} \subseteq \mathbb{C}^{p \times p}$ be a probability operator-valued measure, i.e., a set of Hermitian positive semidefinite matrices summing to $\mathbf{I}$. Let $b_1, \ldots, b_n$ be samples of independent random variables $B_1, \ldots, B_n$, with probability distribution

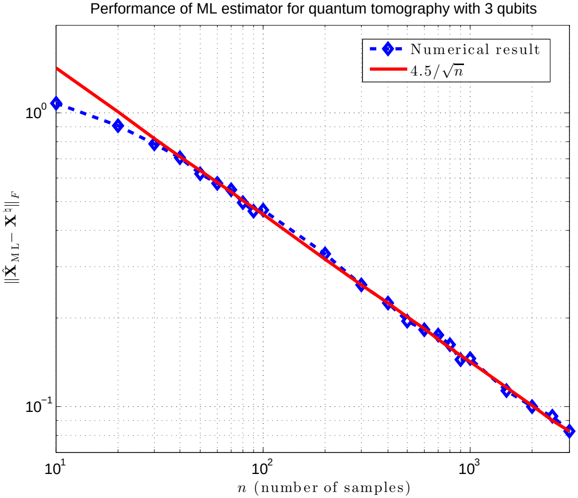
$$\mathbb{P}\left(\{b_i = k\}\right) = \mathrm{Tr}\left(\mathbf{A}_k \mathbf{X}^\natural\right), \quad k = 1, \ldots, m$$

How do we estimate $\mathbf{X}^\natural$ given $\{\mathbf{A}_1, \ldots, \mathbf{A}_m\}$ and $b_1, \ldots, b_n$?

ML approach

$$\hat{\mathbf{X}}_{\mathsf{ML}} \in \arg \min_{\mathbf{X} \in \mathbb{C}^{p \times p}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{m} \mathbb{I}_{\{b_i = k\}} \ln\left[\mathrm{Tr}\left(\mathbf{A}_k \mathbf{X}\right)\right] : \mathbf{X} = \mathbf{X}^H, \mathbf{X} \succeq \mathbf{0} \right\}.$$

Example: ML estimation for quantum tomography



Caveat Emptor

The ML estimator does not always yield the optimal performance. We show a simple yet very powerful example below.

Problem

Let $\mathbf{b}$ *be a sample of a Gaussian random vector* $\mathbf{b} \sim \mathcal{N}(\mathbf{x}^\natural, \mathbf{I})$ *with some* $\mathbf{x}^\natural \in \mathbb{R}^p$*. How do we estimate* $\mathbf{x}^\natural$ *given* $\mathbf{b}$*?*

ML approach

The ML estimator is given by $\hat{\mathbf{x}}_{\mathsf{ML}} := \mathbf{b}$.

James-Stein estimator [10]

The James-Stein estimator is given by

$$\hat{\mathbf{x}}_{\mathsf{JS}} := \left(1 - \frac{p-2}{\|\mathbf{b}\|_2^2}\right)_+ \mathbf{b},$$

for all $p \geq 3$, where $(a)_+ = \max(a, 0)$.

Observation: The James-Stein estimator *shrinks* $\mathbf{b}$ towards the origin.

Caveat Emptor

Theorem (Performance comparison: ML vs. James-Stein [10])

For all $\mathbf{x}^\natural \in \mathbb{R}^p$ *with* $p \geq 3$, *we have*

$$\mathbb{E}\left[\left\|\hat{\mathbf{x}}_{JS} - \mathbf{x}^\natural\right\|_2^2\right] < \mathbb{E}\left[\left\|\hat{\mathbf{x}}_{ML} - \mathbf{x}^\natural\right\|_2^2\right].$$

Performance of the ML estimator is uniformly dominated by the performance of the James-Stein estimator [10].

Important take home message

The ML approach is not always the best.

Caveat Emptor

Theorem (Performance comparison: ML vs. James-Stein [10])

For all $\mathbf{x}^\natural \in \mathbb{R}^p$ *with* $p \geq 3$, *we have*

$$\mathbb{E}\left[\left\|\hat{\mathbf{x}}_{JS} - \mathbf{x}^\natural\right\|_2^2\right] < \mathbb{E}\left[\left\|\hat{\mathbf{x}}_{ML} - \mathbf{x}^\natural\right\|_2^2\right].$$

Performance of the ML estimator is uniformly dominated by the performance of the James-Stein estimator [10].

Important take home message

The ML approach is not always the best.

Remark

The James-Stein estimator inspires the study of *shrinkage estimators* and the use of *oracle inequalities*, which play important roles in contemporary statistics and machine learning [11].

Basic statistical learning

Statistical Learning Model [12]

A statistical learning model consists of the following three elements.

1. A sample of i.i.d. random variables $(\mathbf{a}_i, b_i) \in \mathcal{A} \times \mathcal{B}$, $i = 1, \ldots, n$, following an *unknown* probability distribution $\mathbb{P}$.
2. A class (set) $\mathcal{F}$ of functions $f : \mathcal{A} \to \mathcal{B}$.
3. A loss function $L : \mathcal{B} \times \mathcal{B} \to \mathbb{R}$.

Definition

Let $(\mathbf{a}, b)$ follow the probability distribution $\mathbb{P}$ and be independent of $(\mathbf{a}_1, b_1), \ldots, (\mathbf{a}_n, b_n)$. Then, the risk corresponding to any $f \in \mathcal{F}$ is its expected loss:

$$R(f) := \mathbb{E}_{(\mathbf{a},b)} \left[L(f(\mathbf{a}), b) \right].$$

Statistical learning seeks to find a $f^\star \in \mathcal{F}$ that minimizes the risk, i.e., it solves

$$f^\star \in \arg\min_f \left\{ R(f) : f \in \mathcal{F} \right\}.$$

- Since $\mathbb{P}$ is unknown, the optimization problem above is intractable.

Empirical risk minimization (ERM)

By the law of large numbers, we can expect that for each $f \in \mathcal{F}$,

$$R(f) := \mathbb{E}\left[L(\mathbf{a}, b)\right] \approx \frac{1}{n}\sum_{i=1}^{n} L(f(\mathbf{a}_i), b_i)$$

when n is large enough, with high probability.

Empirical risk minimization (ERM) [12]

We approximate $f^\star$ by minimizing the *empirical average of the loss* instead of the risk. That is, we consider the optimization problem

$$\hat{f}_n \in \arg\min_{f} \left\{ \frac{1}{n}\sum_{i=1}^{n} L(f(\mathbf{a}_i), b_i) : f \in \mathcal{F} \right\}.$$

Least squares revisited

Recall that the LS estimator is given by

$$\hat{\mathbf{x}}_{\mathsf{LS}} \in \arg\min\left\{\|\mathbf{b}-\mathbf{A}\mathbf{x}\|_2^2 : \mathbf{x}\in\mathbb{R}^p\right\} = \arg\min\left\{\frac{1}{n}\sum_{i=1}^{n}\left(b_i-\langle\mathbf{a}_i,\mathbf{x}\rangle\right)^2 : \mathbf{x}\in\mathbb{R}^p\right\},$$

where we define $\mathbf{b} := (b_1,\ldots,b_n)$ and $\mathbf{a}_i$ to be the i-th row of $\mathbf{A}$.

A statistical learning view of least squares

This corresponds to a statistical learning model, for which

- the sample is given by $(\mathbf{a}_i, b_i)\in\mathbb{R}^p\times\mathbb{R}$, $i=1,\ldots,n$,
- the function class $\mathcal{F}$ is given by $\mathcal{F} := \{f_{\mathbf{x}}(\cdot) := \langle\cdot,\mathbf{x}\rangle : \mathbf{x}\in\mathbb{R}^p\}$, and
- the loss function is given by $L(f_{\mathbf{x}}(\mathbf{a}), b) := (b - f_{\mathbf{x}}(\mathbf{a}))^2$.

The corresponding ERM solution is

$$\hat{f}_n(\cdot) := \langle\cdot,\hat{\mathbf{x}}_{\mathsf{LS}}\rangle\,.$$

- Thus the LS estimator also seeks to, given $\mathbf{a}$, minimize the error of predicting the corresponding b by a linear function in terms of the squared error.

References I

[1] M. I. Jordan, "Why the logistic function? a tutorial discussion on probabilities and neural networks," MIT Computational Cognitive Science Report 9503, 1995.

[2] N. Dey, L. Blanc-Feraud, C. Zimmer, P. Roux, Z. Kam, J.-C. Olivo-Marin, and J. Zerubia, "Richardson-Lucy algorithm with total variation regularization for 3D confocal microscope deconvolution," *Microsc. Res. Tech.*, vol. 69, pp. 260–266, 2006.

[3] G. M. P. van Kempen, L. J. van Vliet, P. J. Verveer, and H. T. M. van der Voort, "A quantitative comparison of image restoration methods for confocal microscopy," *J. Microsc.*, vol. 185, pp. 354–365, 1997.

[4] P. J. Huber and E. M. Ronchetti, *Robust Statistics*. Hoboken, NJ: John Wiley & Sons, 2009.

[5] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu, "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," *Electron. J. Stat.*, vol. 5, pp. 935–980, 2011.

[6] S. Oymak, C. Thrampoulidis, and B. Hassibi, "The squared-error of generalized LASSO: A precise analysis," 2013, arXiv:1311.0830v2 [cs.IT].

[7] L. Fahrmeir and H. Kaufmann, "Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models," *Ann. Stat.*, vol. 13, no. 1, pp. 342–368, 1985.

[8] L. Le Cam, *Asymptotic methods in Statistical Decision Theory*. New York, NY: Springer-Verl., 1986.

References II

[9] A. W. van der Vaart, *Asymptotic Statistics*. Cambridge, UK: Cambridge Univ. Press, 1998.

[10] W. James and C. Stein, "Estimation with quadratic loss," in *Proc. 4th Berkeley Symp. Mathematical Statistics and Probability*, vol. 1. Univ. Calif. Press, 1961, pp. 361–379.

[11] E. J. Candès, "Modern statistical estimation via oracle inequalities," *Acta Numer.*, vol. 15, pp. 257–325, May 2006.

[12] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Inf. Theory*, vol. 10, no. 5, pp. 988–999, Sep. 1999.