

Additional Topics

Dr. Edmund Lam

Department of Electrical and Electronic Engineering
The University of Hong Kong

ELEC2844: Probabilistic Systems Analysis
(Second Semester, 2018-19)

<http://www.eee.hku.hk/~elec2844>

1 Linear inverse problems

2 Hypothesis testing

3 Central limit theorem

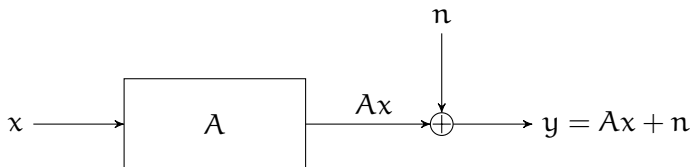
Inverse problem

We have already seen how ML relates to the minimum mean square error (or least square, or ℓ_2 -norm) solution in the context of linear regression. We can generalize even further to linear inverse problems.

We generalize two aspects:

- 1 From regression to inverse problem
- 2 From ML to MAP, with a prior

Inverse problem



Consider x is an input, A is a known linear system, and n is noise. The observation is $y = Ax + n$.

- If $A = I$ (identity matrix), then we are back to regression problem.
- We assume x and n are realizations of random variables X and N respectively, with corresponding PDFs.

Inverse problem

We call this an **inverse problem** because given the output y , we want to derive the input x .

The MAP approach suggests we want to maximize $f_{X|Y}(x|y)$, noting that

$$f_{X|Y}(x|y) \propto f_{Y|X}(y|x) f_X(x)$$

$$\arg \max_x (\log f_{X|Y}(x|y)) = \arg \max_x (\log f_{Y|X}(y|x) + \log f_X(x))$$

Let us now consider specific *choices* of $f_{Y|X}(y|x)$ and $f_X(x)$.

Inverse problem

Suppose we have Gaussian noise, $N \sim \mathcal{N}(0, \sigma^2)$, so

$$f_N(n) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{n^2}{2\sigma^2} \right\}$$

For Y given X , we can compute $n = y - Ax$, so

$$f_{Y|X}(y|x) = f_N(y - Ax) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(y - Ax)^2}{2\sigma^2} \right\}$$

$$\log f_{Y|X}(y|x) = \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(y - Ax)^2}{2\sigma^2}$$

Inverse problem

Suppose we have Gaussian noise, $N \sim \mathcal{N}(0, \sigma^2)$, so

$$f_N(n) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{n^2}{2\sigma^2} \right\}$$

For Y given X , we can compute $n = y - Ax$, so

$$f_{Y|X}(y|x) = f_N(y - Ax) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(y - Ax)^2}{2\sigma^2} \right\}$$

$$\log f_{Y|X}(y|x) = \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(y - Ax)^2}{2\sigma^2}$$

Since x and y are in general vectors, to maximize $\log f_{Y|X}(y|x)$, we need to *minimize*

$$\|y - Ax\|^2$$

Inverse problem

But there's also $f_X(x)$.

- 1 Suppose we have an *uninformative prior*: In this case, $\log f_X(x)$ is a constant and does not affect the maximization. Hence, the optimal MAP (or simply the ML) solution is to

$$\text{minimize } \|y - Ax\|^2 \quad (1)$$

- 2 Suppose X is also Gaussian. Similar calculations as above show that we have an extra term $\|x\|^2$ to minimize. So overall, we should

$$\text{minimize } \|y - Ax\|^2 + \lambda \|x\|^2 \quad (2)$$

where λ balances the relative contribution of the “data fidelity” and “prior” terms.

Inverse problem

The approach

$$\text{minimize} \quad \|y - Ax\|^2 + \lambda \|x\|^2$$

is called Tikhonov regularization and is a well established approach for inverse problems, which can also be motivated without using probability. The solution is the pseudo-inverse:

$$x = (A^T A + \lambda I)^{-1} A^T y \quad (3)$$

- 1 Linear inverse problems
- 2 Hypothesis testing**
- 3 Central limit theorem

Hypothesis testing as an estimation problem

Binary hypothesis testing involves deciding between two “hypotheses”, e.g., $\mathbb{P}(H) = 0.5$ versus $\mathbb{P}(H) = 0.1$ for a particular coin, based on observations of outcomes.

We can look at this as a special case of parameter estimation, where θ takes on two possible values. Both ML and MAP can be used. Nevertheless, we'll introduce the specific terminologies relating to hypothesis testing.

Hypothesis testing terminology

We use H_0 and H_1 to denote the two hypotheses.

H_0 : null hypothesis

H_1 : alternative hypothesis

$X = (X_1, \dots, X_n)$ is a vector of random variables whose distribution depends on the hypothesis. Certain set (or **acceptance region**) of X will lead us to decide on H_0 ; outside that set (or **rejection region**), we will conclude on H_1 .

We are interested to find a rule to determine where the regions are.

Hypothesis testing terminology

We can have two types of errors:

Type I (false rejection): reject H_0 even though it is true

Type II (false acceptance): accept H_0 even though it is false

Type I error means X lies in the rejection region R , even when the reality should follow H_0 . This occurs with probability

$$\alpha(R) = \mathbb{P}(X \in R; H_0) \quad (4)$$

Type II error means X does not lie in the rejection region R , even when the reality should follow H_1 . This occurs with probability

$$\beta(R) = \mathbb{P}(X \notin R; H_1) \quad (5)$$

We often have to tradeoff between the two types of errors.

Hypothesis testing and ML

We define the **likelihood ratio** $L(x)$ for both discrete and continuous r.v.:

$$L(x) = \frac{p_X(x; H_1)}{p_X(x; H_0)} \quad \text{or} \quad L(x) = \frac{f_X(x; H_1)}{f_X(x; H_0)} \quad (6)$$

Hypothesis testing and ML

We define the **likelihood ratio** $L(x)$ for both discrete and continuous r.v.:

$$L(x) = \frac{p_X(x; H_1)}{p_X(x; H_0)} \quad \text{or} \quad L(x) = \frac{f_X(x; H_1)}{f_X(x; H_0)} \quad (6)$$

ML says we pick H_1 if $L(x) > 1$, and H_0 otherwise.

Hypothesis testing and MAP

Let H_0 means $\Theta = \theta_0$ with prior probability $p_{\Theta}(\theta_0)$, and H_1 means $\Theta = \theta_1$ with prior probability $p_{\Theta}(\theta_1)$. The MAP rule for $\Theta = \theta_1$ is

$$p_{\Theta}(\theta_0) p_{X|\Theta}(x|\theta_0) < p_{\Theta}(\theta_1) p_{X|\Theta}(x|\theta_1)$$

Hypothesis testing and MAP

Let H_0 means $\Theta = \theta_0$ with prior probability $p_{\Theta}(\theta_0)$, and H_1 means $\Theta = \theta_1$ with prior probability $p_{\Theta}(\theta_1)$. The MAP rule for $\Theta = \theta_1$ is

$$p_{\Theta}(\theta_0) p_{X|\Theta}(x|\theta_0) < p_{\Theta}(\theta_1) p_{X|\Theta}(x|\theta_1)$$

We define the **likelihood ratio** $L(x)$ for both discrete and continuous r.v.:

$$L(x) = \frac{p_{X|\Theta}(x|\theta_1)}{p_{X|\Theta}(x|\theta_0)} \quad \text{or} \quad L(x) = \frac{f_{X|\Theta}(x|\theta_1)}{f_{X|\Theta}(x|\theta_0)} \quad (7)$$

Hypothesis testing and MAP

Let H_0 means $\Theta = \theta_0$ with prior probability $p_{\Theta}(\theta_0)$, and H_1 means $\Theta = \theta_1$ with prior probability $p_{\Theta}(\theta_1)$. The MAP rule for $\Theta = \theta_1$ is

$$p_{\Theta}(\theta_0) p_{X|\Theta}(x|\theta_0) < p_{\Theta}(\theta_1) p_{X|\Theta}(x|\theta_1)$$

We define the **likelihood ratio** $L(x)$ for both discrete and continuous r.v.:

$$L(x) = \frac{p_{X|\Theta}(x|\theta_1)}{p_{X|\Theta}(x|\theta_0)} \quad \text{or} \quad L(x) = \frac{f_{X|\Theta}(x|\theta_1)}{f_{X|\Theta}(x|\theta_0)} \quad (7)$$

MAP says we pick H_1 if $L(x) > \zeta$, and H_0 otherwise, with

$$\zeta = \frac{p_{\Theta}(\theta_0)}{p_{\Theta}(\theta_1)} \quad (8)$$

Example

We have a die and have two hypotheses about its fairness:

$$H_0(\text{fair die}) : \quad p_X(x; H_0) = \frac{1}{6} \quad x = 1, \dots, 6$$

$$H_1(\text{loaded die}) : \quad p_X(x; H_1) = \begin{cases} \frac{1}{4} & x = 1, 2 \\ \frac{1}{8} & x = 3, 4, 5, 6 \end{cases}$$

Example

We have a die and have two hypotheses about its fairness:

$$H_0(\text{fair die}) : \quad p_X(x; H_0) = \frac{1}{6} \quad x = 1, \dots, 6$$

$$H_1(\text{loaded die}) : \quad p_X(x; H_1) = \begin{cases} \frac{1}{4} & x = 1, 2 \\ \frac{1}{8} & x = 3, 4, 5, 6 \end{cases}$$

$$\textcircled{1} \text{ For } x = 1, 2: L(x) = \frac{1/4}{1/6} = \frac{3}{2}$$

$$\textcircled{2} \text{ For } x = 3, 4, 5, 6: L(x) = \frac{1/8}{1/6} = \frac{3}{4}$$

If we use ML, observing 1 or 2 will lead us to H_1 , and observing 3 to 6 will lead us to H_0 .

Example

If we use “MAP”, we are more flexible to set ζ .

- If we set $\zeta < \frac{3}{4}$: we will reject H_0 for all x .

$$\alpha(\zeta) = 1 \quad \beta(\zeta) = 0$$

- If we set $\zeta > \frac{3}{2}$: we will accept H_0 for all x .

$$\alpha(\zeta) = 0 \quad \beta(\zeta) = 1$$

- If we set $\frac{3}{4} < \zeta < \frac{3}{2}$: we will accept H_0 for $x = 3, 4, 5, 6$ and reject it for $x = 1, 2$.

$$\alpha(\zeta) = \mathbb{P}(X = 1, 2; H_0) = \frac{1}{3}$$

$$\beta(\zeta) = \mathbb{P}(X = 3, 4, 5, 6; H_1) = \frac{1}{2}$$

Tradeoff

Choosing ζ represents a tradeoff in Type I and Type II errors. In fact, using the likelihood ratio as a means of deciding on the acceptance and rejection regions, reducing one type of error leads to an increase in another, given by the **Neyman-Pearson Lemma**.

Consider using the likelihood ratio $L(X)$ for testing, with error probabilities

$$\mathbb{P}(L(X) > \zeta; H_0) = \alpha \quad \text{and} \quad \mathbb{P}(L(X) \leq \zeta; H_1) = \beta$$

Suppose some other test results in a rejection region R achieving a smaller (or equal) false rejection probability $\mathbb{P}(X \in R; H_0) \leq \alpha$, then we must have

$$\mathbb{P}(X \notin R; H_1) \geq \beta$$

with strict inequality $\mathbb{P}(X \notin R; H_1) > \beta$ when $\mathbb{P}(X \in R; H_0) < \alpha$.

- 1 Linear inverse problems
- 2 Hypothesis testing
- 3 Central limit theorem

Ubiquity of Gaussian random variables

A very important mathematical result supporting the general use of Gaussian random variables is the **central limit theorem**.

Let $X_1, X_2, \dots$ be a sequence of n independent, identically distributed random variables with mean μ and variance σ^2 . We define

$$Z_n = \frac{X_1 + \dots + X_n - n\mu}{\sigma\sqrt{n}} \quad (9)$$

The central limit theorem states that Z_n “converges” to $\mathcal{N}(0, 1)$.

Central limit theorem

We can easily show that

$$\mathbb{E}(Z_n) = \frac{\mathbb{E}(X_1 + \cdots + X_n) - n\mu}{\sigma\sqrt{n}} = 0$$

$$\text{var}(Z_n) = \frac{\text{var}(X_1 + \cdots + X_n)}{\sigma^2 n} = \frac{n\sigma^2}{n\sigma^2} = 1$$

Also, the “convergence” is in the technical sense that the CDF of Z_n converges to the standard normal CDF,

$$\lim_{n \rightarrow \infty} \mathbb{P}(Z_n \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-x^2/2} dx$$

and that it is sufficient to show that the moment generating function of Z_n converges to that of a standard normal when n approaches infinity. Before we go into the proof, let us first work out some of the uses.

Central limit theorem

Normal Approximation Based on the Central Limit Theorem

Let $S_n = X_1 + \cdots + X_n$, where the X_i are independent identically distributed random variables with mean μ and variance σ^2 . If n is large, the probability $\mathbb{P}(S_n \leq c)$ can be approximated by treating S_n as if it were normal, according to the following procedure:

- ① Calculate the mean $n\mu$ and the variance $n\sigma^2$ of S_n .
- ② Calculate the normalized value $z = (c - n\mu)/\sigma\sqrt{n}$.
- ③ Use the approximation

$$\mathbb{P}(S_n \leq c) \approx \Phi(z)$$

where $\Phi(z)$ is available from standard normal CDF.

Example

We have 100 packages with independent weights uniformly distributed between 5 and 50 kilograms. What is the (approximate) probability that the total weight exceeds 3000 kilograms?

Example

We have 100 packages with independent weights uniformly distributed between 5 and 50 kilograms. What is the (approximate) probability that the total weight exceeds 3000 kilograms?

ANS: Each package has weight X_i , and $S_{100} = X_1 + \dots + X_{100}$.

$$\textcircled{1} \quad n\mu = 100 \times (5 + 50)/2 = 2750 ; \quad n\sigma^2 = 100 \times (50 - 5)^2/12 = 16875$$

$$\textcircled{2} \quad z = \frac{3000 - 2750}{\sqrt{16875}} = \frac{250}{129.9} = 1.92$$

$$\textcircled{3} \quad \mathbb{P}(S_{100} > 3000) = 1 - \mathbb{P}(S_{100} \leq 3000) \approx 1 - \Phi(1.92) = 0.0274$$

Example

A 3D printer prints out different designs in an amount of time that is uniformly distributed between 1 and 5 hours, independent of each other. Find the (approximate) probability that the number of parts processed within 320 hours, denoted by N_{320} , is at least 100.

Example

A 3D printer prints out different designs in an amount of time that is uniformly distributed between 1 and 5 hours, independent of each other. Find the (approximate) probability that the number of parts processed within 320 hours, denoted by N_{320} , is at least 100.

ANS: Each design takes X_i , and $S_{100} = X_1 + \dots + X_{100}$. Note that the events $\{N_{320} \geq 100\}$ and $\{S_{100} \leq 320\}$ are equivalent.

$$\textcircled{1} \quad n\mu = 100 \times (1 + 5)/2 = 300 ; \quad n\sigma^2 = 100 \times (5 - 1)^2/12 = 400/3$$

$$\textcircled{2} \quad z = \frac{320 - 300}{\sqrt{400/3}} = 1.73$$

$$\textcircled{3} \quad \mathbb{P}(S_{100} \leq 320) \approx \Phi(1.73) = 0.9582$$

Approximation to the Binomial

The Gaussian approximation can also be used for discrete random variables, such as binomial. This is called De Moivre-Laplace approximation.

A binomial random variable S_n with parameters n and p can be viewed as the sum of n independent Bernoulli random variables $X_1, \dots, X_n$, with common parameter p .

$$\mu = \mathbb{E}(X_i) = p$$

$$\sigma = \sqrt{\text{var}(X_i)} = \sqrt{p(1-p)}$$

We want to approximate the event $\{k \leq S_n \leq \ell\}$.

Approximation to the Binomial

For $k \leq S_n \leq \ell$, we have

$$\frac{k - np}{\sqrt{np(1-p)}} \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq \frac{\ell - np}{\sqrt{np(1-p)}}$$

By the central limit theorem, $(S_n - np) / \sqrt{np(1-p)}$ has approximately a standard normal distribution, and we obtain

$$\begin{aligned} \mathbb{P}(k \leq S_n \leq \ell) &= \mathbb{P}\left(\frac{k - np}{\sqrt{np(1-p)}} \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq \frac{\ell - np}{\sqrt{np(1-p)}}\right) \\ &\approx \Phi\left(\frac{\ell + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) \end{aligned}$$

Shifting by $\frac{1}{2}$ for both k and ℓ makes the approximation more accurate.

Approximation to the Binomial

Example: Let $n = 36$ and $p = 0.5$.

Exact calculation vs Gaussian approximation for the event $\{S_n \leq 21\}$:

$$\mathbb{P}(S_n \leq 21) = \sum_{k=0}^{21} \binom{36}{k} (0.5)^{36} = 0.8785$$

$$\mathbb{P}(S_n \leq 21) \approx \Phi \left(\frac{21.5 - np}{\sqrt{np(1-p)}} \right) = \Phi(1.17) = 0.879$$

Exact calculation vs Gaussian approximation for the event $\{S_n = 19\}$:

$$\mathbb{P}(S_n = 19) = \binom{36}{19} (0.5)^{36} = 0.1251$$

$$\mathbb{P}(S_n = 19) \approx \Phi \left(\frac{19.5 - 18}{3} \right) - \Phi \left(\frac{18.5 - 18}{3} \right) = 0.124$$

Central limit theorem

Proof: For simplicity we assume starting with zero-mean random variables, i.e., let $X_1, X_2, \dots$ be a sequence of n independent, identically distributed random variables with mean zero and variance σ^2 , and moment generating function $M_X(s)$. We define

$$Z_n = \frac{X_1 + \dots + X_n}{\sigma\sqrt{n}}$$

with a moment generating function $M_{Z_n}(s)$. We have three steps:

- ① Approximate $M_X(s)$ by Taylor series expansion around $s = 0$
- ② Express $M_{Z_n}(s)$ in terms of $M_X(s)$
- ③ Take limit for $n \rightarrow \infty$ to simplify the expression for $M_{Z_n}(s)$

Central limit theorem

Step I: We want to express $M_X(s)$ in the form

$$M_X(s) = a + bs + cs^2 + \mathcal{O}(s^2)$$

where $\mathcal{O}(s^2)$ means “higher order terms we can ignore”.

Central limit theorem

Step I: We want to express $M_X(s)$ in the form

$$M_X(s) = a + bs + cs^2 + \mathcal{O}(s^2)$$

where $\mathcal{O}(s^2)$ means “higher order terms we can ignore”.

Remember $M_X(s) = \mathbb{E}(e^{sX}) = \int_{-\infty}^{\infty} e^{sx} f_X(x) dx$. Although we do not know the precise PDF of X_i , except it is zero-mean with variance σ^2 , we know

$$a = M_X(0) = 1$$

$$b = \left. \frac{d}{ds} M_X(s) \right|_{s=0} = \mathbb{E}(X) = 0$$

$$c = \frac{1}{2} \cdot \left. \frac{d^2}{ds^2} M_X(s) \right|_{s=0} = \frac{\mathbb{E}(X^2)}{2} = \frac{\sigma^2}{2}$$

Central limit theorem

Step II: We derive an expression for $M_{Z_n}(s)$ in terms of $M_X(s)$:

$$\begin{aligned}M_{Z_n}(s) &= \mathbb{E}[\exp\{sZ_n\}] \\&= \mathbb{E}\left[\exp\left\{\frac{s}{\sigma\sqrt{n}}\sum_{i=1}^n X_i\right\}\right] \\&= \prod_{i=1}^n \mathbb{E}\left[\exp\left\{\frac{s}{\sigma\sqrt{n}}X_i\right\}\right] \\&= \left[M_X\left(\frac{s}{\sigma\sqrt{n}}\right)\right]^n\end{aligned}$$

Central limit theorem

Step III: Take limit for $n \rightarrow \infty$

$$\begin{aligned} M_X(s) &= 1 + \frac{\sigma^2}{2}s^2 + \mathcal{O}(s^2) \\ M_{Z_n}(s) &= \left[1 + \frac{\sigma^2}{2} \left(\frac{s}{\sigma\sqrt{n}} \right)^2 + \mathcal{O} \left(\frac{s}{\sigma\sqrt{n}} \right)^2 \right]^n \\ &= \left[1 + \frac{s^2}{2n} + \mathcal{O} \left(\frac{s^2}{\sigma^2 n} \right) \right]^n \end{aligned}$$

Using the identity $\lim_{n \rightarrow \infty} (1 + \frac{c}{n})^n = e^c$, and ignoring the terms inside $\mathcal{O}(\cdot)$,

$$\lim_{n \rightarrow \infty} M_{Z_n}(s) = \exp \left\{ \frac{s^2}{2} \right\}$$

which is the moment generating function of $\mathcal{N}(0, 1)$.

Conclusions

Pierre-Simon Laplace (1749–1827) once said:

“Les questions les plus importantes de la vie ne sont en effet, pour la plupart, que des problèmes de probabilité.”

Conclusions

Pierre-Simon Laplace (1749–1827) once said:

“Les questions les plus importantes de la vie ne sont en effet, pour la plupart, que des problèmes de probabilité.”

Pierre-Simon Laplace



The most important questions of life are indeed, for the most part, really only problems of probability.

Maybe this is true on May 7, 2019?