Parameter Estimation

Ben Gravell

benjamin.gravell@utdallas.edu
The Erik Jonsson School of Engineering and Computer Science
The University of Texas at Dallas
800 W. Campbell Rd.
Richardson, TX 75080

Outline

Parameter estimation

In many applications:

- Distribution of a random variable X is unknown or too complicated to compute
- Only need some parameter θ that characterizes the distribution

Goal: Obtain a good approximation of parameter θ based only on observations of X.

Estimator

An **estimator** $\hat{\Theta}$ is a function of the data $\{X_i\}$ that approximates θ, but is not an explicit function of θ.

Judging parameter estimators

How do we judge the quality of an estimator?

Consistency

An estimator $\hat{\Theta}_n$ computed from n samples is **consistent** if

$$\lim_{n \to \infty} P\left[\ |\hat{\Theta}_n - \theta| > \varepsilon\ \right] = 0 \tag{1}$$

for any positive tolerance $\varepsilon > 0$.

Consistency means "we can guarantee arbitrarily accurate estimates if we use an arbitrarily large amount of data"

Judging parameter estimators

What we really want:

Confidence bound

An estimator $\hat{\Theta}_n$ is **ε-accurate with $1-\delta$ confidence** if

$$P\left[\ |\hat{\Theta}_n - \theta| > \varepsilon\ \right] \leq \delta \tag{2}$$

- This is like soft consistency w/ finite data
- Consistency allows us to take ε and δ as small as we like (so long as we can pay for it with infinite data $n \to \infty$)
- Quantifying n
  - Can be done exactly in certain special cases
    - e.g. estimating the mean of a Gaussian
  - Can be done conservatively using concentration inequalities in more general cases
    - e.g. estimating the mean of any distribution w/ finite variance

Judging parameter estimators

Confidence interval

Consider an estimator $\hat{\Theta}_n$. Fix the number of samples n and fix a failure probability δ. The $1-\delta$ **confidence interval** is the smallest accuracy tolerance ε such that

$$P\left[\ |\hat{\Theta}_n - \theta| > \varepsilon\ \right] \leq \delta \tag{3}$$

i.e. the estimator $\hat{\Theta}_n$ is ε-accurate with $1-\delta$ confidence.

Basically the same as the confidence criterion where we fixed ε and sought n, but here we fix n and seek ε

Judging parameter estimators

Many classical results use two proxies for the ε-δ criterion:

- Bias
  - "systematic errors"
  - "location"
- Variance
  - "random errors"
  - "spread"

Bias

The **bias** of an estimator $\hat{\Theta}$ is

$$|\mathbb{E}[\hat{\Theta}] - \theta|. \tag{4}$$

The estimator is **unbiased** if

$$\mathbb{E}[\hat{\Theta}] = \theta. \tag{5}$$

Variance

The **variance** of an estimator $\hat{\Theta}$ is

$$\mathbb{E}[(\hat{\Theta} - \theta)^2]. \tag{6}$$

The estimator is **minimum variance** if

$$\hat{\Theta} = \underset{\Theta}{\text{argmin}}\ \mathbb{E}[(\Theta - \theta)^2]. \tag{7}$$

Judging parameter estimators

Sometimes bias can be eliminated without affecting the variance

- We will see an example of such a correction

Sometimes bias can only be reduced at the expense of higher variance

- In machine learning this is a well-studied phenomenon called the **bias-variance tradeoff**

Sample average estimator of a RV

The **sample average estimator** of a random variable X given N observations $\{X_i\}_{i=1}^{N}$ is

$$\hat{\mu}_X(n) := \frac{1}{N}\sum_{i=1}^{N} X_i$$

Sample average estimator of a function of a RV

The **sample average estimator** of a function g of a random variable X given N observations $\{X_i\}_{i=1}^{N}$ is

$$\hat{\mu}_{g(X)}(n) := \frac{1}{N}\sum_{i=1}^{N} g(X_i)$$

It's easy to show that the sample average is **unbiased**:

$$
\begin{aligned}
\mathbb{E}\left[\hat{\mu}_X(n)\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] && \text{(def. of } \hat{\mu}_X(n)\text{)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i\right] && \text{(linearity of } \mathbb{E}[\cdot]\text{)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu_X && \text{(def. of } \mu_X\text{)} \\
&= \frac{1}{n} \cdot n \cdot \mu_X && (8) \\
&= \mu_X && (9)
\end{aligned}
$$

The **variance** of the sample average is not much harder to find:

$$
\begin{aligned}
\sigma^2_{\hat{\mu}}(n) &:= \mathbb{E}\left[\left(\hat{\mu}_X(n) - \mathbb{E}\left[\hat{\mu}_X(n)\right]\right)^2\right] && \text{(def. of } \sigma^2_{\hat{\mu}}(n)\text{)} \\
&= \mathbb{E}\left[\left(\hat{\mu}_X(n) - \mu_X\right)^2\right] && \text{(since } \hat{\mu} \text{ unbiased)} \\
&= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)\right)^2\right] && \text{(def. of } \hat{\mu}\text{)} \\
&= \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}(X_i - \mu_X)^2\right] + \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}(X_i - \mu_X)(X_j - \mu_X)\right] \\
& && \text{(expand squared sum)} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[(X_i - \mu_X)^2\right] + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i\neq j}^{n}\mathbb{E}\left[(X_i - \mu_X)(X_j - \mu_X)\right] \\
& && \text{(linearity of } \mathbb{E}[\cdot]\text{)} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma_X^2 + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i\neq j}^{n} 0 && \text{(def. of } \sigma_X^2\text{, uncorrelation of } X_i\text{)} \\
&= \sigma_X^2 / n && (10)
\end{aligned}
$$

We can get a **confidence bound** by using the Chebyshev inequality:

$$P\left[|\hat{\mu}_X(n) - \mu_X| \geq \varepsilon\right] \leq \frac{\sigma^2_{\hat{\mu}}(n)}{\varepsilon^2} = \frac{1}{n} \cdot \frac{\sigma_X^2}{\varepsilon^2} \tag{11}$$

Taking $n \to \infty$ reveals that the **sample average is consistent**:

$$\lim_{n\to\infty} P\left[|\hat{\mu}_X(n) - \mu_X| \geq \varepsilon\right] = \lim_{n\to\infty} \frac{1}{n} \cdot \frac{\sigma_X^2}{\varepsilon^2} = 0 \tag{12}$$

Remark: If we knew the form of the distribution e.g. Gaussian we could get an exact confidence bound using the standard normal CDF.

Remark: This confidence bound involves the true variance σ_X^2, which is typically unknown. If X is Gaussian and σ_X^2 is replaced by a sample variance estimate, an exact confidence bound can still be obtained using the **student T-distribution** CDF - see Ch. 6.3 of [1].

So far we estimated the mean - what about estimating the variance?

If we **knew the true mean μ** we could create the variance estimator

$$\hat{\sigma}_X^2(n) = \frac{1}{n} \sum_{i=0}^{n} (X_i - \mu)^2 \tag{13}$$

But of course we **don't know the true mean μ!**

Natural idea: just use the sample mean in place of the true mean:

$$\hat{\sigma}_X^2(n) = \frac{1}{n} \sum_{i=0}^{n} (X_i - \hat{\mu})^2 \tag{14}$$

But there is an issue with this...

Homework P4-1

Compute the expectation of the sample variance estimator

$$\hat{\sigma}_X^2(n) = \frac{1}{n}\sum_{i=0}^{n}(X_i - \hat{\mu}_X(n))^2 \tag{15}$$

where

$$\hat{\mu}_X(n) = \frac{1}{n}\sum_{i=0}^{n} X_i \tag{16}$$

1. Is this sample variance estimator $\hat{\sigma}_X^2(n)$ biased?
2. If so, how much is the bias?
3. How does the bias change with the number of samples n?
4. What correction needs to be made to $\hat{\sigma}_X^2(n)$ in order to make the estimator unbiased?

Maximum likelihood estimation

Maximum likelihood estimation provides a principled way to design estimators based on optimization.

Likelihood

The **likelihood** function $L(\theta)$ of the random variables $\{X_i\}_{i=1}^n$ for outcome $\{x_i\}_{i=1}^n$ under parameter θ is the parametric joint pdf

$$L(\theta) = f_{\{X_i\}_{i=1}^n}(\{x_i\}_{i=1}^n; \theta). \tag{17}$$

As a special case, if $\{X_i\}_{i=1}^n$ are i.i.d. random variables then

$$L(\theta) = \prod_{i=1}^n f_X(x_i; \theta) \tag{18}$$

Maximum likelihood estimation

Maximum likelihood estimate

The **maximum likelihood estimate** for outcome $\{x_i\}_{i=1}^n$ is the parameter $\theta^*(\{x_i\}_{i=1}^n)$ that maximizes the likelihood, i.e.

$$\theta^*(\{x_i\}_{i=1}^n) = \underset{\theta}{\text{argmax}}\ L(\theta) \tag{19}$$

The **maximum likelihood estimator** is the random variable

$$\hat{\theta} = \theta^*(\{X_i\}_{i=1}^n) \tag{20}$$

We start by assuming the *form* of the distribution is Gaussian with variance σ^2. We are estimating the mean, so the parameter is $\theta = \mu$

The likelihood is

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right) \tag{21}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(\sum_{i=1}^{n} -\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right) \tag{22}$$

Since the log function is monotonic increasing, the argmax of $L(\mu)$ is the same as the argmax of $\log L(\mu)$. The log is easier to work with.

$$\log L(\mu) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \tag{23}$$

MLE for mean of a Gaussian

To maximize the log likelihood we find the stationary point

$$0 = \left.\frac{\partial \log L(\mu)}{\partial \mu}\right|_{\mu^*} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu^*) \tag{24}$$

which implies the MLE is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{25}$$

which happens to be the sample mean.

Homework P4-2:
Derive the expression for the maximum likelihood estimator of the mean and variance of a Gaussian. Is the MLE variance biased?
Hint: Use the log-likelihood

$$\log L(\mu, \sigma) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{26}$$

Least-squares estimation

Suppose we wish to estimate a vector parameter which is exposed through the **linear observation model**

$$Y = H\theta + N \tag{27}$$

- Y is an **observation vector**
- H is a known constant **observation matrix**
- θ is an unknown constant **parameter vector**
- N is a **random observation noise vector**

The observation Y is directly measured, but the noise N is not.

Least-squares estimation

Define the **residual**

$$E = Y - H\theta \tag{28}$$

which measures the error between the observation and its expected value.

A natural idea is to choose a parameter estimate that minimizes an objective function $v(\theta)$ which increases with the size of the residual.

$$\hat{\theta} = \underset{\theta}{\text{argmin}}\ v(\theta) \tag{29}$$

In particular, choose $v(\theta)$ as the squared norm of the residual:

$$v(\theta) = \|E\|^2 = (Y - H\theta)^\intercal (Y - H\theta) \tag{30}$$

Least-squares estimation

For the next step we need some basic knowledge from optimization and matrix calculus.

Since $v(\theta)$ is a quadratic form, we can compute the minimizer in closed-form by finding the **stationary point** where the gradient of the objective vanishes:

$$0 = \left.\frac{\partial v(\theta)}{\partial \theta}\right|_{\hat{\theta}} = 2(H^\intercal H)\hat{\theta} - 2H^\intercal Y \tag{31}$$

Rearranging yields the so-called **normal equation**

$$(H^\intercal H)\hat{\theta} = H^\intercal Y \tag{32}$$

If $H^\intercal H$ is invertible, we obtain the **least-squares estimate (LSE)**

$$\hat{\theta} = (H^\intercal H)^{-1} H^\intercal Y \tag{33}$$

Remark: If N is white Gaussian noise, i.e. $N \sim \mathcal{N}(0, I)$, then one can show the LSE is unbiased, minimum variance, and maximum likelihood.

Homework P4-3: We are given the following data:

$$\begin{bmatrix} 6.2 \\ 7.8 \\ 2.2 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \\ 1 \end{bmatrix} \theta + \begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix} \tag{34}$$

where n_i are random variables. Find a least-squares estimate for θ.

Asymptotics

Asymptotics

In this section we see major results from classical statistics

Claims are **asymptotic**; they only hold as the amount of data $\to \infty$

Claims are all about **convergence** of some kind

Contrast with finite-sample results c.f. [2]

Weak law of large numbers (WLLN)

Weak law of large numbers

Let X_i be an infinite sequence of i.i.d. random variables with a finite, common true mean μ and variance σ^2. Consider the sample mean

$$\hat{\mu}(n) = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{35}$$

Then for any fixed positive tolerance $\varepsilon > 0$ we have

$$\lim_{n \to \infty} \mathbb{P}\left[|\hat{\mu}(n) - \mu| < \varepsilon \right] = 1 \tag{36}$$

i.e. the sample mean **converges in probability** to the true mean.

Proof: We already proved that the sample mean is consistent, which is the same thing as the WLLN.

Strong law of large numbers

Let X_i be an infinite sequence of i.i.d. random variables with a finite, common true mean μ and variance σ^2. Consider the sample mean

$$\hat{\mu}(n) = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{37}$$

Then we have

$$\mathbb{P}\left[\lim_{n \to \infty} \hat{\mu}(n) = \mu\right] = 1 \tag{38}$$

i.e. the sample mean **converges almost surely** to the true mean.

Proof: More involved than the WLLN. Also SLLN implies WLLN.

Notice the difference between weak and strong laws:

1. WLLN: Sequence of success probabilities approaches one
2. SLLN: Sequence of sample means approaches the true mean

Central limit theorem

Let X_i be an infinite sequence of independent random variables with cdf's F_{X_i}, finite means μ_i and finite variances σ_i^2.

Define the variance sum s_n^2 and normalized random variable Z_n

$$s_n^2 = \sum_{i=1}^{n} \sigma_i^2, \quad Z_n = \sum_{i=1}^{n} (X_i - \mu_i)/s_n \tag{39}$$

Suppose there exists $\varepsilon > 0$ and for all n sufficiently large that

$$\sigma_i < \varepsilon s_n, \quad i = 1, \ldots, n \tag{40}$$

Then

$$\lim_{n \to \infty} F_{Z_n}(z) = \Phi(z) \tag{41}$$

i.e. Z_n **converges in distribution** to a standard normal.

Homework: Central limit theorem

Homework P4-4: Let $\{X_i\}_{i=1}^n$ be a sequence of n i.i.d. random variables. Compute the approximate probability

$$\mathbb{P}[a \leq S \leq b] \tag{42}$$

of the sum

$$S(n) = \sum_{i=1}^{n} X_i \tag{43}$$

using the central limit theorem.
For concreteness, assume the X_i are uniform random variables on the unit interval $[0, 1]$, $n = 100$, $a = 45$, and $b = 52.5$.

[1] John Woods and Henry Stark.
Probability, Statistics, and Random Processes for Engineers.
Pearson Higher Ed, 4 edition, 2011.

[2] Martin J Wainwright.
High-dimensional statistics: A non-asymptotic viewpoint, volume 48.
Cambridge University Press, 2019.
https://people.eecs.berkeley.edu/~wainwrig/BibPapers/Wai19.pdf.