Maximum likelihood estimation

Nuno Vasconcelos
UCSD

Maximum likelihood

- parameter estimation in three steps:
  - 1) choose a parametric model for probabilities
    to make this clear we denote the vector of parameters by Θ

$$P_X(x;\Theta)$$

    note that this means that Θ is NOT a random variable
  - 2) assemble $\mathcal{D} = \{x_1, ..., x_n\}$ of examples drawn independently
  - 3) select the parameters that maximize the probability of the data

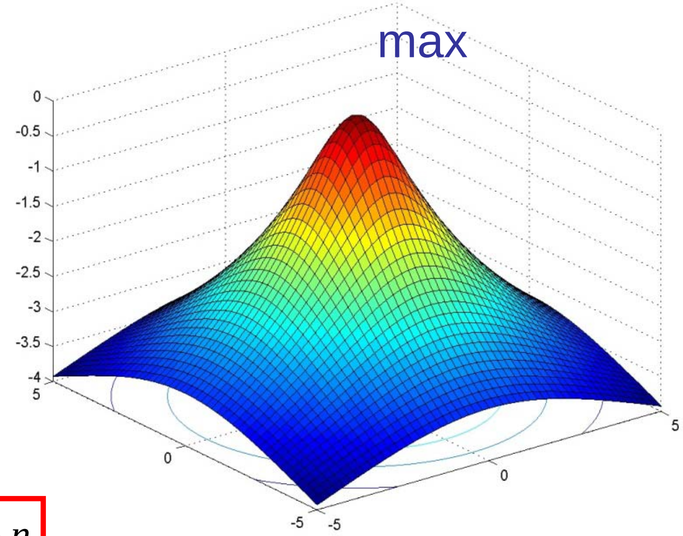
$$\begin{aligned}\Theta^* &= \arg\max_{\Theta} P_X(D;\Theta) \\ &= \arg\max_{\Theta} \log P_X(D;\Theta)\end{aligned}$$

- $P_X(\mathcal{D};\Theta)$ is the likelihood of parameter Θ with respect to the data

Maximum likelihood

- in summary, given a sample, we need to solve

$$\Theta^* = \arg\max_{\Theta} P_X(D;\Theta)$$



- the solutions are the parameters such that

$$\nabla_{\Theta} P_X(x;\Theta) = 0$$

$$\theta^t \nabla_{\Theta}^{2} P_X(x;\theta)\theta \le 0, \quad \forall \theta \in \Re^n$$

- note that you always have to check the second-order condition!

Maximum likelihood

- we solved the Gaussian case

$$f(T) = \frac{1}{\sigma_T\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{T-\bar{T}}{\sigma_T}\right)^2}$$

- given a sample $\{T_1, \ldots, T_N\}$ of independent points
- the log-likelihood is

$$\Lambda = \ln L = -\frac{N}{2}\ln(2\pi) - N\ln\sigma_T - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{T_i - \bar{T}}{\sigma_T}\right)^2$$

- the ML estimates of the mean and variance are

$$\bar{T} = \frac{1}{N}\sum_{i=1}^{N} T_i \qquad \hat{\sigma}_T^2 = \frac{1}{N}\sum_{i=1}^{N}(T_i - \bar{T})^2$$

Estimators

- when we talk about estimators, it is important to keep in mind that
  - an estimate is a number
  - an estimator is a random variable

$$\hat{\theta} = f(X_1, \ldots, X_n)$$

- an estimate is the value of the estimator for a given sample.
- if $\mathcal{D} = \{x_1, \ldots, x_n\}$, when we say $\hat{\mu} = \frac{1}{n} \sum_j x_j$

  what we mean is $\hat{\mu} = f(X_1, \ldots, X_n)\Big|_{X_1 = x_1, \ldots, X_n = x_n}$ with

$$f(X_1, \ldots, X_n) = \frac{1}{n} \sum_j X_j$$

← the X_i are random variables

Bias and variance

- we know how to produce estimators (by ML)
- how do we evaluate an estimator?
- Q_1: is the expected value equal to the true value?
- this is measured by the bias
  - if

$$\hat{\theta} = f(X_1, \ldots, X_n)$$

  then

$$Bias\left(\hat{\theta}\right) = E_{X_1, \ldots, X_n}\left[f(X_1, \ldots, X_n) - \theta \right]$$

  - an estimator that has bias will usually not converge to the perfect estimate θ, no matter how large the sample is
  - e.g. if θ is negative and the estimator is $f(X_1, \ldots, X_n) = \frac{1}{n}\sum_j X_j^2$ the bias is clearly non-zero

Bias and variance

- the estimators is said to be biased
  - this means that it is not expressive enough to approximate the true value arbitrarily well
  - this will be clearer when we talk about density estimation
- Q_2: assuming that the estimator converges to the true value, how many sample points do we need?
  - this can be measured by the variance

$$Var\left(\hat{\theta}\right) = E_{X_1,\ldots,X_n}\left\{\left(f(X_1,\ldots,X_n) - E_{X_1,\ldots,X_n}\left[f(X_1,\ldots,X_n)\right]\right)^2\right\}$$

  - the variance usually decreases as one collects more training examples

Example

- ML estimator for the mean of a Gaussian $N(\mu,\sigma^2)$

$$Bias(\hat{\mu}) = E_{X_1,\dots,X_n}[\hat{\mu} - \mu] = E_{X_1,\dots,X_n}[\hat{\mu}] - \mu$$

$$= E_{X_1,\dots,X_n}\left[\frac{1}{n}\sum_i X_i\right] - \mu$$

$$= \frac{1}{n}\sum_i E_{X_1,\dots,X_n}[X_i] - \mu$$

$$= \frac{1}{n}\sum_i E_{X_i}[X_i] - \mu$$

$$= \mu - \mu = 0$$

- the estimator is unbiased

Example

- variance of ML estimator for mean of a Gaussian $N(\mu,\sigma^2)$

$$Var(\hat{\mu}) = E_{X_1,\dots,X_n}\left\{\left(\hat{\mu} - E_{X_1,\dots,X_n}[\hat{\mu}]\right)^2\right\} = E_{X_1,\dots,X_n}\left\{(\hat{\mu}-\mu)^2\right\}$$

$$= E_{X_1,\dots,X_n}\left\{\left(\frac{1}{n}\sum_i X_i - \mu\right)^2\right\}$$

$$= \frac{1}{n^2} E_{X_1,\dots,X_n}\left\{\left(\sum_i (X_i - \mu)\right)^2\right\}$$

$$= \frac{1}{n^2} E_{X_1,\dots,X_n}\left\{\sum_{ij} (X_i - \mu)(X_j - \mu)\right\}$$

Example

- ML estimator for the mean of a Gaussian $N(\mu,\sigma^2)$

$$Var(\hat{\mu}) = \frac{1}{n^2}\sum_{ij} E_{X_i,X_j}\left[(X_i-\mu)(X_j-\mu)\right]$$

$$= \frac{1}{n^2}\sum_{ij}\sigma_{ij}$$

- and since X_i, X_j are independent, $\sigma_{ij} = 0, \quad \forall i \neq j$

$$Var(\hat{\mu}) = \frac{1}{n^2}\sum_{i}\sigma_i^2 = \frac{\sigma^2}{n}$$

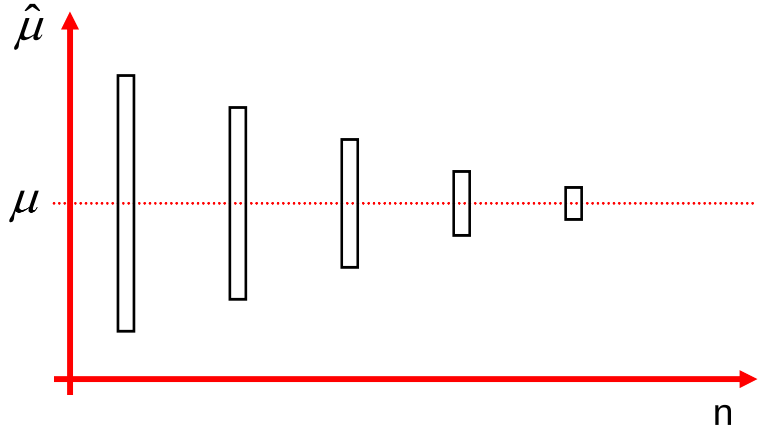
- the variance goes to zero as n increases!

Example

- in summary, for ML estimator for the mean of a Gaussian $N(\mu,\sigma^2)$

$$E[\hat{\mu}] = \mu \qquad Var(\hat{\mu}) = \frac{\sigma^2}{n}$$

- this means that if I have a large sample, the value of the estimate will be close to the true value with high probability



Example

- is this always true?
- ML estimator for the variance of a Gaussian $N(\mu,\sigma^2)$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_i \left(X_i - \hat{\mu}\right)^2 = \frac{1}{n}\sum_i \left(X_i^2 - 2X_i\hat{\mu} + \hat{\mu}^2\right)$$

$$= \frac{1}{n}\sum_i X_i^2 - \hat{\mu}^2$$

- the expected value is

$$E_{X_1,\ldots,X_n}\left[\hat{\sigma}^2\right] = \frac{1}{n}\sum_i E_{X_1,\ldots,X_n}\left[X_i^2\right] - E_{X_1,\ldots,X_n}\left[\hat{\mu}^2\right]$$

$$= \frac{1}{n}\sum_i E_{X_i}\left[X_i^2\right] - E_{X_1,\ldots,X_n}\left[\hat{\mu}^2\right] = E_X\left[X^2\right] - E_{X_1,\ldots,X_n}\left[\hat{\mu}^2\right]$$

Example

- using

$$E_{X_1,\ldots,X_n}\left[\hat{\mu}^2\right] = E_{X_1,\ldots,X_n}\left[\frac{1}{n^2}\sum_{ij} X_i X_j\right] = \frac{1}{n^2}\sum_{ij} E_{X_i,X_j}\left[X_i X_j\right]$$

$$= \frac{1}{n^2}\sum_{i} E_{X_i}\left[X_i^2\right] + \frac{1}{n^2}\sum_{i,j\neq i} E_{X_i,X_j}\left[X_i X_j\right]$$

$$= \frac{1}{n} E_X\left[X^2\right] + \frac{1}{n^2}\sum_{i,j\neq i} E_{X_i}\left[X_i\right] E_{X_j}\left[X_j\right]$$

$$= \frac{1}{n} E_X\left[X^2\right] + \frac{1}{n^2}\sum_{i} E_{X_i}\left[X_i\right]\sum_{j\neq i} E_{X_j}\left[X_j\right]$$

$$= \frac{1}{n} E_X\left[X^2\right] + \frac{1}{n^2}\sum_{i} E_{X_i}\left[X_i\right](n-1) E_X\left[X\right]$$

Example

- using

$$E_{X_1,\ldots,X_n}\left[\hat{\mu}^2\right]=\frac{1}{n}E_X\left[X^2\right]+\frac{1}{n^2}\sum_i E_{X_i}\left[X_i\right](n-1)E_X\left[X\right]$$

$$=\frac{1}{n}E_X\left[X^2\right]+\frac{(n-1)}{n}\left(E_X\left[X\right]\right)^2$$

$$=\frac{1}{n}E_X\left[X^2\right]+\frac{(n-1)}{n}\mu^2$$

- we get

$$E_{X_1,\ldots,X_n}\left[\hat{\sigma}^2\right]=E_X\left[X^2\right]-E_{X_1,\ldots,X_n}\left[\hat{\mu}^2\right]$$

$$=\frac{n-1}{n}E_X\left[X^2\right]-\frac{n-1}{n}\mu^2=\left(1-\frac{1}{n}\right)\sigma^2$$

Example

- in summary

$$E_{X_1,\ldots,X_n}\left[\hat{\sigma}^2\right] = \left(1 - \frac{1}{n}\right)\sigma^2$$

- the estimator is biased
- Q: do we care?
  - clearly

$$\lim_{n\to\infty} E_{X_1,\ldots,X_n}\left[\hat{\sigma}^2\right] = \sigma^2$$

  - so, for large samples it is (for all practical purposes) unbiased
  - what about small samples? the variance is likely to be large to start with, a little bit of bias is not going to make much difference
  - so, in practice, it is fine

Important note

- since the estimator is a random variable
  - we can never say that an estimate obtained with more samples is "better" than an estimate from less samples.
  - e.g., if

$$\mu_1 = \frac{1}{100}\sum_{i=1}^{100} X_i \qquad \mu_2 = \frac{1}{10{,}000}\sum_{i=1}^{10{,}000} X_i$$

    we measure and obtain

$$\hat{\mu}_1 = 10.5 \qquad \hat{\mu}_2 = 10.3$$

    is 10.3 a better estimate of μ than 10.5?

  - we can never know, all we know is that

$$\mu_1 = N\left(\mu, \sigma^2 / 100\right) \qquad \mu_2 = N\left(\mu, \sigma^2 / 10{,}000\right)$$

Important note

- and we can use this to compute

$$P(|\mu_2 - \mu| < |\mu_1 - \mu|)$$

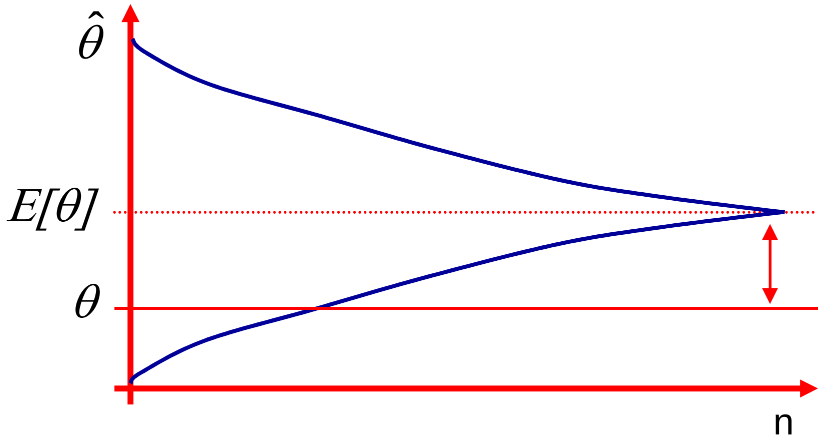
- but there is always a probability that the estimate produced by μ_1 is better than that produced by μ_2
- even though μ_2 has much smaller variance
- all that we can hope for, is to make the estimator better in a probabilistic sense
- this means making

$$P_{\hat{\Theta}}(\theta)$$

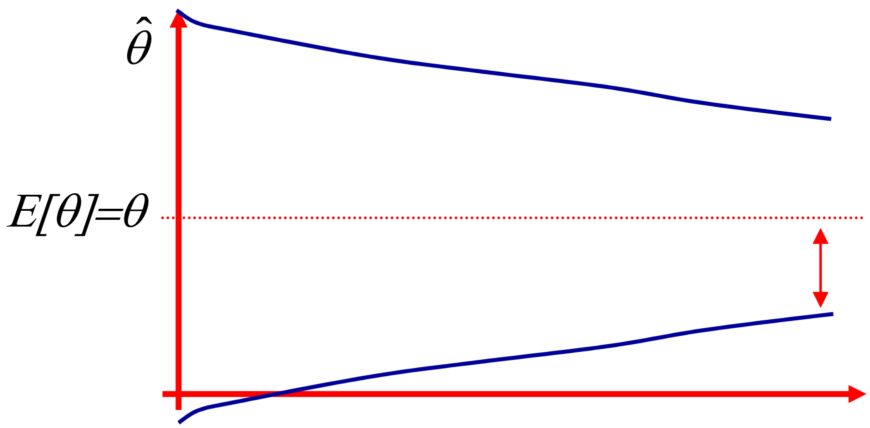
  as concentrated as possible around the true value
- in this sense, emphasizing bias or variance can be wrong

Bias and variance

- we really care about the conjunction of the two factors
  - working hard to decrease variance if bias is large is useless



  - working hard to decrease bias if variance is large is useless



Mean squared error

- one possibility to account for both bias and variance is to minimize the mean squared error
  - if $$\hat{\theta} = f(X_1, \ldots, X_n)$$
  - then $$MSE(\hat{\theta}) = E_{X_1,\ldots,X_n}\left[\{f(X_1, \ldots, X_n) - \theta\}^2\right]$$
- the connection to bias and variance follows from

$$MSE(\hat{\theta}) = E\left[\left\{\hat{\Theta} - E[\hat{\Theta}] + E[\hat{\Theta}] - \theta\right\}^2\right]$$

$$= E\left[\left\{\hat{\Theta} - E[\hat{\Theta}]\right\}^2\right] + 2E\left[\{\hat{\Theta} - E[\hat{\Theta}]\}\{E[\hat{\Theta}] - \theta\}\right]$$

$$+ E\left[\left\{E[\hat{\Theta}] - \theta\right\}^2\right]$$

Mean squared error

- $$MSE(\hat{\theta}) = E\left[\left\{\hat{\Theta} - E[\hat{\Theta}] + E[\hat{\Theta}] - \theta\right\}^2\right]$$
  $$= E\left[\left\{\hat{\Theta} - E[\hat{\Theta}]\right\}^2\right] + 2E\left(\left\{\hat{\Theta} - E[\hat{\Theta}]\right\}\left\{E[\hat{\Theta}] - \theta\right\}\right)$$
  $$+ E\left[\left\{E[\hat{\Theta}] - \theta\right\}^2\right]$$
  $$= \mathrm{var}(\hat{\Theta}) + 2E\left(\left\{\hat{\Theta} - E[\hat{\Theta}]\right\}\right)\left\{E[\hat{\Theta}] - \theta\right\} + \left\{E[\hat{\Theta}] - \theta\right\}^2$$
  $$= \mathrm{var}(\hat{\Theta}) + 2\left\{E[\hat{\Theta}] - E[\hat{\Theta}]\right\}\left\{E[\hat{\Theta}] - \theta\right\} + Bias^2(\hat{\Theta})$$
  - and

$$\boxed{MSE(\hat{\Theta}) = \mathrm{var}(\hat{\Theta}) + Bias^2(\hat{\Theta})}$$

Bias variance trade-off

- in general, the MSE estimator has non-zero bias and variance
- we can only reduce bias at the cost of increased variance and vice-versa
  - suppose we are not happy with the 1/n decay of the variance of

$$\hat{\mu} = \frac{1}{n}\sum_i X_i$$

  - one possibility is to use

$$\hat{\hat{\mu}} = \frac{\alpha}{n}\sum_i X_i = \alpha\hat{\mu}$$

  - this has

$$E\left[\hat{\hat{\mu}}\right] = \alpha\mu \qquad Bias\left[\hat{\hat{\mu}}\right] = (1-\alpha)\mu \qquad \text{var}\left[\hat{\hat{\mu}}\right] = \frac{\alpha^2\sigma^2}{n}$$

Bias variance trade-off

- this has

$$Bias\left[\hat{\hat{\mu}}\right] = (1-\alpha)\mu$$

$$\mathrm{var}\left[\hat{\hat{\mu}}\right] = \frac{\alpha^2\sigma^2}{n}$$

- by choosing $\alpha < 1$ we can decrease the variance, but the bias will no longer be zero
- what value of α minimizes the MSE?

$$MSE\left[\hat{\hat{\mu}}\right] = \mathrm{var}\left[\hat{\hat{\mu}}\right] + Bias^2\left[\hat{\hat{\mu}}\right]$$

$$= \frac{\alpha^2\sigma^2}{n} + (1-\alpha)^2\mu^2$$

- and

$$\frac{\partial MSE\left[\hat{\hat{\mu}}\right]}{\partial\alpha} = 2\alpha\frac{\sigma^2}{n} - 2(1-\alpha)\mu^2$$

Bias variance trade-off

– from which

$$\frac{\partial MSE\left[\hat{\hat{\mu}}\right]}{\partial \alpha} = 0 \Leftrightarrow \alpha \frac{\sigma^2}{n} + \alpha\mu^2 = \mu^2$$

$$\Leftrightarrow \alpha\left(\frac{\sigma^2}{n} + \mu^2\right) = \mu^2 \Leftrightarrow \alpha = \frac{\mu^2}{\left(\frac{\sigma^2}{n} + \mu^2\right)}$$

– and the MSE estimator of μ is

$$\hat{\hat{\mu}} = \frac{\mu^2}{\sigma^2 + n\mu^2} \sum_i X_i$$

- one can immediately detect a problem
  - the optimal estimator depends on the quantity that we are trying to estimate!
  - the estimator is unrealizable

Estimators

- unrealizable solutions are a common source of problems for the MSE estimator
- one alternative is to
  - constrain the estimator to be in a class (e.g. unbiased)
  - find, among all solutions in the class, that of least MSE
- many ideas on how to do this
  - BLUE: best linear unbiased estimator
  - MVUE: minimum variance unbiased
  - check the parameter estimation literature
- why is the ML estimator so popular?
  - many of these alternatives are frequently unrealizable
  - the ML solution typically makes intuitive sense
  - connections to Bayesian estimation (we will talk about this later)

Estimators

- consider BLUE estimator for the population mean

$$\mu_{BLUE} = \sum_i w_i X_i$$

  – what are the weights w_i such that

$$E[\mu_{BLUE}] = E[X] = \mu$$

$$\text{var}[\mu_{BLUE}] = MSE[X] \quad \text{is minimal?}$$

  – the answer is

$$\mu_{BLUE} = \frac{1}{n}\sum_i X_i$$

  – note that this holds independently of whether X is Gaussian
  – but, for Gaussian X, it is the same as ML!
  – “when there is an easy realizable solution ML gets it”

Any questions?