Lecture 9

# MAXIMUM-LIKELIHOOD ESTIMATION AND FISHER INFORMATION

Harrison H. Barrett
University of Arizona

# OUTLINE

- Estimation as mapping
- Performance metrics: Bias and variance, MSE and EMSE
- Fisher information and Cramér-Rao bounds
- Maximum-likelihood estimation: Why and how?
- Textbook examples (univariate and bivariate)
- A multivariate toolkit
- The EM algorithm

Next lecture: Applications in astronomy

# References

Barrett and Myers, Foundations of Image Science

For this lecture:

Chap. 13, Statistical Decision Theory

For background:

Appendix C, Probability

Chap. 8, Stochastic Properties of Objects and Images

## Notation and terminology

The $M \times 1$ vector $\mathrm{g}$ describes the random data.

The PDF on $\mathrm{g}$ is characterized by the $K \times 1$ parameter vector $\boldsymbol{\theta}$ and denoted $\mathrm{pr}(\mathrm{g}|\boldsymbol{\theta})$. This PDF describes the *sampling distribution* of $\mathrm{g}$ for a given $\boldsymbol{\theta}$.

Once a data vector is measured, $\mathrm{pr}(\mathrm{g}|\boldsymbol{\theta})$ can be regarded as a function of $\boldsymbol{\theta}$, sometimes called the *likelihood*:

$$L(\boldsymbol{\theta}|\mathrm{g}) = \mathrm{pr}(\mathrm{g}|\boldsymbol{\theta})$$

Note that $L(\boldsymbol{\theta}|\mathrm{g})$ is *not* a PDF on $\boldsymbol{\theta}$.

An *estimate* of the parameter is denoted $\hat{\boldsymbol{\theta}}$; in most cases the estimate is a deterministic function of the data, so we can also write it as $\hat{\boldsymbol{\theta}}(\mathrm{g})$. Since $\mathrm{g}$ is random (even for a given $\boldsymbol{\theta}$), so is $\hat{\boldsymbol{\theta}}(\mathrm{g})$.

## Estimation as mapping

$$\boldsymbol{\theta} \longrightarrow \mathbf{g} \longrightarrow \hat{\boldsymbol{\theta}}$$

$$\mathrm{pr}(\boldsymbol{\theta}) \longrightarrow \mathrm{pr}(\mathbf{g}|\boldsymbol{\theta}) \longrightarrow \mathrm{pr}(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta})$$

## Performance metrics

For a given true value of $\theta$, there are two basic kinds of error:

- Random error

  Also called precision, estimation variance, noise, jitter, ...

- Systematic error

  Also called accuracy, estimation bias, calibration error, ...

Both kinds can be quantified by the conditional distribution of the estimates:

$$\mathrm{pr}(\hat{\theta}|\theta)$$

## Bias of the estimate of a scalar parameter

Consider a scalar parameter $\theta$ and an estimate $\hat{\theta}(\mathrm{g})$. The conditional mean of the estimate can be computed from the likelihood:

$$\overline{\hat{\theta}} = \left\langle \hat{\theta}(\mathrm{g}) \right\rangle_{\mathrm{g}|\theta} = \int d^M g \; \mathsf{pr}(\mathrm{g}|\theta)\, \hat{\theta}(\mathrm{g})\,, \tag{13.274}$$

where the subscript $\mathrm{g}|\theta$ indicates that the average is over randomness in the data when the parameter has fixed value equal to $\theta$.

If we know the conditional PDF of $\hat{\theta}$ itself, we can write

$$\overline{\hat{\theta}} = \int d\hat{\theta} \; \mathsf{pr}(\hat{\theta}|\theta)\, \hat{\theta}\,, \tag{13.275}$$

We denote the conditional bias by $b(\theta)$, where

$$b(\theta) = \overline{\hat{\theta}} - \theta\,. \tag{13.276}$$

The overline indicates an average over all realizations of the data, given the true value $\theta$. An *unbiased* estimate is one for which $b(\theta) = 0$ for all $\theta$.

## Estimability

A parameter is said to be be *estimable* or *identifiable* with respect to some data set if there exists an unbiased estimator of it for all true values of the parameter.

Examples:

- Stopped clock
- Integral of grey levels over ROI
- Pixel value

One way of dealing with the issue of estimability is to average the bias over all possible true values of the parameter, defining an average bias by

$$\bar{\bar{\hat{\theta}}} = \int d\theta \ \mathrm{pr}(\theta) \int d^M g \ \mathrm{pr}(\mathbf{g}|\theta)\, \hat{\theta}(\mathbf{g})\,, \qquad (13.277)$$

## Variance and MSE of the estimate of a scalar parameter

Bias specifies accuracy or systematic error. Variance specifies precision or random error. It is defined by:

$$\mathrm{Var}(\hat{\theta}) = \sigma_{\hat{\theta}}^2 = \left\langle |\hat{\theta}(\mathbf{g}) - \overline{\hat{\theta}}|^2 \right\rangle_{\mathbf{g}|\theta} = \int d^M g \; \mathrm{pr}(\mathbf{g}|\theta) \, |\hat{\theta} - \overline{\hat{\theta}}|^2 \, . \qquad (13.279)$$

The variance describes the fluctuations of the estimate about the mean of the estimate, *not the true value*.

Fluctuations around the true value are described by the *mean-square error*:

$$\mathrm{MSE}(\theta) = \left\langle |\hat{\theta} - \theta|^2 \right\rangle_{\mathbf{g}|\theta} \, . \qquad (13.280)$$

The *ensemble mean-square error* or *EMSE* is found by taking an additional average over the parameter:

$$\mathrm{EMSE} = \left\langle \left\langle |\hat{\theta} - \theta|^2 \right\rangle_{\mathbf{g}|\theta} \right\rangle_{\theta} \, . \qquad (13.281)$$

## Vector parameters – bias

A $P$-dimensional parameter vector $\boldsymbol{\theta}$ has an estimate $\hat{\boldsymbol{\theta}}$ with mean $\langle\hat{\boldsymbol{\theta}}\rangle$ given by

$$\overline{\hat{\boldsymbol{\theta}}}(\mathrm{g}) = \int d^M g \; \mathsf{pr}(\mathrm{g}|\boldsymbol{\theta})\, \hat{\boldsymbol{\theta}}(\mathrm{g}) = \int d^P \hat{\theta} \; \mathsf{pr}(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta})\, \hat{\boldsymbol{\theta}}\,. \tag{13.282}$$

The bias $\mathrm{b}(\boldsymbol{\theta})$ is now a vector quantity:

$$\mathrm{b}(\boldsymbol{\theta}) \equiv \overline{\hat{\boldsymbol{\theta}}} - \boldsymbol{\theta} \equiv \int_\infty d^M g \; \Big[\hat{\boldsymbol{\theta}}(\mathrm{g}) - \boldsymbol{\theta}\Big]\, \mathsf{pr}(\mathrm{g}|\boldsymbol{\theta}) = \int_\infty d^P \theta \; \Big[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\Big]\, \mathsf{pr}(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta})\,. \tag{13.283}$$

The average bias is now written $\overline{\mathrm{b}} = \langle \mathrm{b}(\boldsymbol{\theta})\rangle_{\boldsymbol{\theta}}$.

## Vector parameters – variance and covariance

If we denote the mean of the $p^{th}$ element of the random vector $\hat{\boldsymbol{\theta}}$ by $\langle\hat{\theta}\rangle_p = \overline{\hat{\theta}}_p$, the variance of the $p^{th}$ element is given by

$$\mathsf{Var}(\hat{\theta}_p) \equiv \left\langle \left[\hat{\theta}_p - \langle\hat{\theta}_p\rangle\right] \left[\hat{\theta}_p - \langle\hat{\theta}_p\rangle\right]^* \right\rangle_{\mathbf{g}|\boldsymbol{\theta}}$$

$$= \int_\infty d^M g \; |\hat{\theta}_p(\mathbf{g}) - \langle\hat{\theta}_p(\mathbf{g})\rangle|^2 \, \mathsf{pr}(\mathbf{g}|\boldsymbol{\theta}) = \int_\infty d^P\theta \; |\hat{\theta}_p - \langle\hat{\theta}_p\rangle|^2 \, \mathsf{pr}(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}) \,, \tag{13.284}$$

and the full covariance matrix is given by

$$\mathbf{K}_{\hat{\boldsymbol{\theta}}} = \left\langle \left(\hat{\boldsymbol{\theta}} - \overline{\hat{\boldsymbol{\theta}}}\right) \left(\hat{\boldsymbol{\theta}} - \overline{\hat{\boldsymbol{\theta}}}\right)^\dagger \right\rangle = \langle \Delta\hat{\boldsymbol{\theta}} \Delta\hat{\boldsymbol{\theta}}^\dagger \rangle \,. \tag{13.285}$$

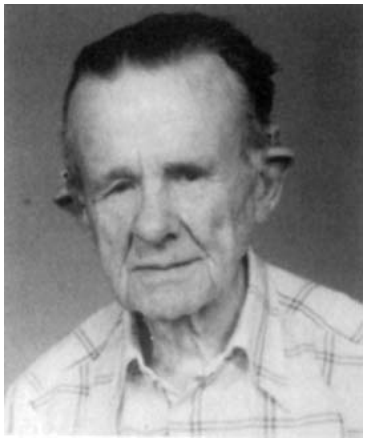
The MSE in the vector case is:

$$\mathsf{MSE} = \left\langle ||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2 \right\rangle_{\mathbf{g}|\boldsymbol{\theta}} = \int_\infty d^M g \; ||\hat{\boldsymbol{\theta}}(\mathbf{g}) - \boldsymbol{\theta}||^2 \, \mathsf{pr}(\mathbf{g}|\boldsymbol{\theta}) = \mathsf{tr}\left[\mathbf{K}_{\hat{\boldsymbol{\theta}}}\right] + \mathsf{tr}\left[\mathbf{b}\mathbf{b}^\dagger\right] \,. \tag{13.287}$$

Key point:

There are fundamental limits to the variance and covariance of any estimator.

# Carl Harald Cramér



Born: 25 Sept 1893 in Stockholm, Sweden

Died: 5 Oct 1985 in Stockholm, Sweden

Seminal book: Mathematical Methods of Statistics (Uppsala, 1945)

## Cramér-Rao bound for a scalar random variable

The variance of any unbiased estimate of a scalar must satisfy

$$\mathrm{Var}\{\hat{\theta}\} \geq \frac{1}{\left\langle \left[\frac{\partial}{\partial\theta} \ln \mathrm{pr}(\mathbf{g}|\theta)\right]^2 \right\rangle_{\mathbf{g}|\theta}}, \tag{13.372}$$

where the denominator is called the *Fisher information*.

Similarly, for a biased estimator,

$$\mathrm{Var}\{\hat{\theta}\} \geq \frac{\left(\frac{db(\theta)}{d\theta} + 1\right)^2}{\left\langle \left[\frac{\partial}{\partial\theta} \ln \mathrm{pr}(\mathbf{g}|\theta)\right]^2 \right\rangle_{\mathbf{g}|\theta}}. \tag{13.377}$$

This result tells us that a stopped clock can have zero variance (infinite precision)!

## Fisher information in the vector case

For estimation of a scalar parameter, the Fisher information $F$ is defined by:

$$F = \left\langle \left[ \frac{\partial}{\partial \theta} \ln \mathrm{pr}(\mathrm{g}|\theta) \right]^2 \right\rangle_{\mathrm{g}|\theta} .$$

For a vector parameter with $P$ components, the Fisher information $\mathbf{F}$ is a $P \times P$ Hermitian matrix with components:

$$F_{jk} = \left\langle \left[ \frac{\partial}{\partial \theta_j} \ln \mathrm{pr}(\mathrm{g}|\boldsymbol{\theta}) \right] \left[ \frac{\partial}{\partial \theta_k} \ln \mathrm{pr}(\mathrm{g}|\boldsymbol{\theta}) \right] \right\rangle_{\mathrm{g}|\boldsymbol{\theta}}$$

$$= \int_\infty d^M g \; \mathrm{pr}(\mathrm{g}|\boldsymbol{\theta}) \left[ \frac{1}{\mathrm{pr}(\mathrm{g}|\boldsymbol{\theta})} \frac{\partial}{\partial \theta_j} \mathrm{pr}(\mathrm{g}|\boldsymbol{\theta}) \right] \left[ \frac{1}{\mathrm{pr}(\mathrm{g}|\boldsymbol{\theta})} \frac{\partial}{\partial \theta_k} \mathrm{pr}(\mathrm{g}|\boldsymbol{\theta}) \right] . \quad (13.361)$$

## CR bound in the vector case

The $nn$ component of the covariance matrix of any random vector is the variance of the $n^{th}$ component. For any unbiased estimate,

$$[\mathbf{K}_{\hat{\boldsymbol{\theta}}}]_{nn} = \mathsf{Var}\{\hat{\theta}_n\} \geq \left[\mathbf{F}^{-1}\right]_{nn} . \tag{13.371}$$

Note that inversion of the Fisher information is required to find the lower bound on the variance of a component of the estimate.

Other forms of the CR bound set limits on the covariance matrix of the estimate.

Terminology: An unbiased estimator that meets the CR bound is *efficient*.

## Maximum-likelihood estimation

So far we have not talked about ways of actually finding an estimate. One general method is *maximum-likelihood estimation*:

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} \equiv \underset{\boldsymbol{\theta}}{\mathrm{argmax}} \; \mathrm{pr}(\mathbf{g}|\boldsymbol{\theta}) \,. \tag{13.348}$$

This procedure can be written equivalently as

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \underset{\boldsymbol{\theta}}{\mathrm{argmax}} \; \ln[\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta})] \,. \tag{13.349}$$

Note that we are *not* maximizing the probability of $\boldsymbol{\theta}$; we are choosing the value of $\boldsymbol{\theta}$ that maximizes the probability of occurrence of the $\mathbf{g}$ that we actually observed.

## What's the score?

If there are no constraints on the estimate, $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ occurs at a point where the gradient (with respect to $\boldsymbol{\theta}$) of the log-likelihood is zero:

$$\nabla_{\boldsymbol{\theta}} \ln[\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta})] = 0 \text{ at } \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\mathrm{ML}}$$

The gradient of the log-likelihood is a random vector called the *score*:

$$\mathbf{s}(\mathbf{g}) = \nabla_{\boldsymbol{\theta}} \ln[\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta})] \,. \tag{13.358}$$

Basically, the score tells us how sensitive the log-likelihood is to changes in the parameters.

The Fisher information matrix is the covariance matrix of the score.

## Why ML?

An ML estimate is:

- Efficient if an efficient estimate exists
- Asypmptotically efficient (as you get more or better data)
- Asypmptotically unbiased
- Asypmptotically consistent
- Usually easy to compute
- A way of rigorously enforcing agreement with the data
- A way of doing estimation with no prior information

## Why not ML?

ML estimation is:

- A way of rigorously enforcing agreement with the data
- A way of doing estimation with no prior information

Data are noisy. Rigorous agreement with noisy data will give noisy estimates – *even though that is the best you can do without bias*!

You always have some prior information – and you should use it, even though it might introduce bias.

One way to use a prior $\mathrm{pr}(\boldsymbol{\theta})$ is with the *weighted likelihood*:

$$\hat{\boldsymbol{\theta}}_{\mathrm{WL}} \equiv \underset{\boldsymbol{\theta}}{\mathrm{argmax}} \ \mathrm{pr}(\mathrm{g}|\boldsymbol{\theta})\ \mathrm{pr}(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\mathrm{argmax}} \ \mathrm{pr}(\boldsymbol{\theta}|\mathrm{g})\,.$$

This estimate is also called the *maximum a posteriori* or MAP estimate, but that is a subject for another lecture.

# TEXTBOOK EXAMPLES OF ML ESTIMATION

- Probability of heads in coin flipping
- Rate of a Poisson random process
- Mean and variance of a normal PDF

## Example 1: Flipping a coin

How do you know a coin is "fair" (probability of heads $= \frac{1}{2}$)? Flip it many times and see how many heads you get.

Suppose you flip a coin 10 times and get 7 heads. What is the ML estimate of the probability of heads?

Need the Bernoulli or binomial probability law. If the probability of heads in a single flip is $p$, the probability of $n$ heads in $N$ flips is

$$\Pr(n|p) = \binom{N}{n} p^n (1-p)^{N-n}, \tag{C.161}$$

where the binomial coefficient is given by

$$\binom{N}{n} \equiv \frac{N!}{n!\,(N-n)!}. \tag{C.162}$$

The mean and variance of a binomial random variable are given by

$$\langle n \rangle = Np \qquad \text{and} \qquad \sigma^2 = Np(1-p). \tag{C.163}$$

## Example 1 – Computing the ML estimate

Given an observed number of heads $n$, maximize $\mathsf{Pr}(n|p)$ (or its logarithm) with respect to $p$ to find the ML estimate of $p$.

$$\mathsf{Pr}(n|p) = \binom{N}{n} p^n (1-p)^{N-n}\,, \tag{C.161}$$

$$\ln \mathsf{Pr}(n|p) = constant + n \ln p + (N-n) \ln(1-p)\,,$$

$$\frac{d}{dp} \ln \mathsf{Pr}(n|p) = \frac{n}{p} - \frac{N-n}{1-p} = 0 \text{ at } p = \hat{p}_{ML}\,,$$

Thus

$$\hat{p}_{ML} = \frac{n}{N} = \frac{\text{Number of heads observed}}{\text{Number of flips}}$$

## Example 1 – Bias, variance and efficiency

$$\hat{p}_{ML} = \frac{n}{N} = \frac{\text{Number of heads observed}}{\text{Number of flips}}$$

Recall:

$$\langle n \rangle = Np \qquad \text{and} \qquad \sigma^2 = Np\,(1-p)\,. \tag{C.163}$$

Thus the bias and variance of the ML estimate are:

$$\langle \hat{p}_{ML} \rangle = \frac{\langle n \rangle}{N} = p\,, \qquad \text{Var}\,(\hat{p}_{ML}) = \frac{\text{Var}\,n}{N^2} = \frac{p\,(1-p)}{N}$$

Is the estimate efficient?

$$F = \left\langle \left[ \frac{n}{p} - \frac{N-n}{1-p} \right]^2 \right\rangle = \frac{1}{p^2\,(1-p)^2} \left\langle [n - Np\,]^2 \right\rangle = \frac{\text{Var}\,n}{p^2\,(1-p)^2} = \frac{N}{p\,(1-p)}$$

Thus

$$\text{Var}\,(\hat{p}_{ML}) = F^{-1}$$

and the estimate is efficient.

## Example 1 – Repeated trials

Now suppose you repeat the experiment $J$ times, observing $n_j$ heads out of $N$ flips in the $j^{th}$ repetition. Denote the set of observations by the vector $\mathbf{n}$ with component $\{n_j, j = 1, ..., J\}$. Since the repetitions are independent, the likelihood function is

$$\Pr(\mathbf{n}|p) = \prod_{j=1}^{J} \Pr(n_j|p) = \prod_{j=1}^{J} \binom{N}{n_j} p^{n_j} (1-p)^{N-n_j} ,$$

and the log-likelihood is

$$\ln \Pr(\mathbf{n}|p) = \sum_{j=1}^{J} \left[ n_j \ln p + (N - n_j) \ln(1-p) \right] + constant ,$$

There is still just one parameter $(p)$ to estimate.

$$\frac{d}{dp} \sum_{j=1}^{J} \left[ n_j \ln p + (N - n_j) \ln(1-p) \right] = 0 \text{ at } p = \hat{p}_{ML} ,$$

from which

$$\hat{p}_{ML} = \frac{\sum_j n_j}{NJ} = \frac{\text{Total number of heads observed}}{\text{Total number of flips}}$$

Thus one long run of $NJ$ flips gives the same ML estimate (and the same Fisher information) as $J$ short runs of $N$ flips each. The total number of heads, $\sum_j n_j$, is a *sufficient statistic* for estimation of $p$.

## Example 2: Estimating the rate of a Poisson random process

Consider photons that arrive at an ideal counting detector at an average rate of $a$ photons/sec. If the photons are independent, the random number of counts observed in time $T$ has a probability law given by

$$\mathsf{Pr}(n|a) = \mathsf{exp}(-aT)\,\frac{(aT)^n}{n!}\,.$$

Suppose we observe a particular value of $n$ in one trial of duration $T$ and want to use this single observation to estimate the rate $a$. The log-likelihood is

$$\mathsf{ln}\,\mathsf{Pr}(n|a) = -aT + n\,\mathsf{ln}\,aT - \mathsf{ln}\,n!\,.$$

Thus

$$\frac{d}{da}\,\mathsf{ln}\,\mathsf{Pr}(n|a) = -T + \frac{n}{a} = 0 \text{ at } a = \hat{a}_{ML} = \frac{n}{T}\,.$$

## Example 2 – bias, variance and efficiency

We know:

$$\hat{a}_{ML} = \frac{n}{T}\,, \qquad \overline{n} = \mathrm{Var}(n) = aT$$

Thus

$$\langle \hat{a}_{ML} \rangle = \frac{\overline{n}}{T} = a\,,$$

so the ML estimate is unbiased. Is if efficient? The variance is

$$\mathrm{Var}\,[\hat{a}_{ML}] = \frac{\mathrm{Var}(n)}{T^2} = \frac{a}{T}\,,$$

and the Fisher information is

$$F = \frac{1}{a^2}\left\langle [n - aT]^2 \right\rangle = \frac{\mathrm{Var}(n)}{a^2} = \frac{T}{a}\,.$$

Thus $\mathrm{Var}\,[\hat{a}_{ML}] = F^{-1}$ and the estimate is efficient.

## Example 3 – Independent samples of a normal RV

Let $x$ be a univariate normal random variable, so that

$$\text{pr}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\overline{x})^2\right]$$

and let $\{x_j, j = 1, ..., J\}$ be samples drawn independently from this distribution. The samples are said to be i.i.d. (independent and identically distributed).

The PDF for the whole set of samples, represented by the $J \times 1$ vector $\mathbf{x}$, is

$$\text{pr}(\mathbf{x}) = (2\pi\sigma^2)^{-J/2} \prod_{j=1}^{J} \exp\left[-\frac{1}{2\sigma^2}(x_j-\overline{x})^2\right] .$$

The log of this PDF is

$$\ln \text{pr}(\mathbf{x}) = -\frac{J}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{J}(x_j-\overline{x})^2 .$$

## Example 3 – estimation of the mean

Suppose we want to estimate the mean $\overline{x}$, assuming for the moment that the variance is known. Then the PDF on the last slide gets reinterpreted as the likelihood $\text{pr}(\mathbf{x}|\overline{x})$ for estimation of the scalar $\overline{x}$. Differentiating the log-likelihood yields:

$$\frac{d}{d\overline{x}} \ln \text{pr}(\mathbf{x}|\overline{x}) = \frac{1}{\sigma^2} \sum_{j=1}^{J} (x_j - \overline{x}) = 0 \,,$$

or

$$\hat{\overline{x}}_{ML} = \frac{1}{J} \sum_{j=1}^{J} x_j = \text{arithmetic average of the data} \,,$$

Efficient, unbiased, all that good stuff.

## Example 3 – estimation of the mean and variance

Suppose now we want to estimate both $\overline{x}$ and $\sigma^2$. The same PDF as before is now the likelihood $\mathrm{pr}(\mathbf{x}|\overline{x},\sigma^2)$ for the 2D estimation problem. Two derivatives must vanish simultaneously:

$$\frac{\partial}{\partial \overline{x}} \ln \mathrm{pr}(\mathbf{x}|\overline{x},\sigma^2) = 0\,, \qquad \frac{\partial}{\partial \sigma^2} \ln \mathrm{pr}(\mathbf{x}|\overline{x},\sigma^2) = 0\,.$$

Two equations in two unknowns. Algebra is messy, but the result is

$$\hat{\overline{x}}_{ML} = \frac{1}{J}\sum_{j=1}^{J} x_j\,, \qquad \widehat{\sigma^2}_{ML} = \frac{1}{J}\sum_{j=1}^{J} \left[x_j - \hat{\overline{x}}_{ML}\right]^2\,.$$

Now the ML estimator is *biased*. It is shown in elementary statistics texts that an unbiased estimate of the variance is given by

$$\widehat{\sigma^2}_{unbiased} = \frac{1}{J-1}\sum_{j=1}^{J} \left[x_j - \hat{\overline{x}}_{ML}\right]^2\,.$$

# Barrett's Theorem

The maximum-likelihood procedure in any problem is what you are most likely to do if you don't know any statistics.

## A multivariate tool kit

Before proceeding to real-world examples of ML estimation, we shall derive some general expressions for log-likelihoods, scores and Fisher information matrices. Problems considered will be:

- Multivariate normal data, mean an arbitrary function of $\boldsymbol{\theta}$
- Multivariate normal data, mean a linear function of $\boldsymbol{\theta}$
- Multivariate normal data, diagonal covariance
- Poisson data, mean an arbitrary function of $\boldsymbol{\theta}$
- Poisson data, mean a linear function of $\boldsymbol{\theta}$

## The general multivariate normal problem

From the last lecture, a general multivariate normal PDF has the form:

$$\text{pr}(\mathbf{g}) = \left[(2\pi)^M \det(\mathbf{K})\right]^{-1/2} \exp\left[-\tfrac{1}{2}(\mathbf{g}-\overline{\mathbf{g}})^t\,\mathbf{K}^{-1}(\mathbf{g}-\overline{\mathbf{g}})\right], \quad (8.185)$$

where $\overline{\mathbf{g}}$ is the mean vector and $\mathbf{K}$ is the covariance matrix of $\mathbf{g}$. The most general likelihood function is obtained by letting the mean and covariance both be functions of $\boldsymbol{\theta}$:

$$\text{pr}(\mathbf{g}|\boldsymbol{\theta}) = \left[(2\pi)^M \det[\mathbf{K}(\boldsymbol{\theta})]\right]^{-1/2} \exp\left\{-\tfrac{1}{2}[\mathbf{g}-\overline{\mathbf{g}}(\boldsymbol{\theta})]^t\,[\mathbf{K}(\boldsymbol{\theta})]^{-1}[\mathbf{g}-\overline{\mathbf{g}}(\boldsymbol{\theta})]\right\}.$$

In many problems, the noise is independent of $\boldsymbol{\theta}$, and the log-likelihood in that case is:

$$\ln \text{pr}(\mathbf{g}|\boldsymbol{\theta}) = constant - \tfrac{1}{2}[\mathbf{g}-\overline{\mathbf{g}}(\boldsymbol{\theta})]^t\,\mathbf{K}^{-1}[\mathbf{g}-\overline{\mathbf{g}}(\boldsymbol{\theta})],$$

which is the most general form of a weighted least-squares functional. Maximum likelihood is the same as minimum least-squares residual.

Basic rule of ML estimation: Gaussian noise $\Rightarrow$ least-squares fitting

## Multivariate normal data, mean a linear function of $\boldsymbol{\theta}$

In many imaging problems with Gaussian noise, we can assume that the covariance is independent of $\boldsymbol{\theta}$ and that

$$\overline{\mathbf{g}}(\boldsymbol{\theta}) = \mathbf{H}\,\boldsymbol{\theta}$$

Then the log-likelihood is given by

$$\ln \mathrm{pr}(\mathbf{g}|\boldsymbol{\theta}) = constant - \tfrac{1}{2}[\mathbf{g} - \mathbf{H}\,\boldsymbol{\theta}]^t\,\mathbf{K}^{-1}[\mathbf{g} - \mathbf{H}\,\boldsymbol{\theta}]$$

$$= constant - \tfrac{1}{2}\sum_{m=1}^{M}\sum_{m'=1}^{M}[\mathbf{g} - \mathbf{H}\,\boldsymbol{\theta}]_m\,[\mathbf{K}^{-1}]_{mm'}[\mathbf{g} - \mathbf{H}\,\boldsymbol{\theta}]_{m'}\,,$$

and the Fisher information matrix is

$$\mathbf{F} = \mathbf{H}^t\mathbf{K}^{-1}\mathbf{H}\,.$$

## Multivariate normal data, i.i.d. noise

In many imaging problems (e.g., CCD arrays), there are no noise correlations between detector pixels, and the pixels can be modeled as identical. Then the covariance matrix is

$$\mathbf{K} = \sigma^2 \mathbf{I} \,.$$

With this noise assumption and $\overline{\mathrm{g}}(\boldsymbol{\theta}) = \mathbf{H}\,\boldsymbol{\theta}$, the log-likelihood becomes

$$\ln \mathrm{pr}(\mathrm{g}|\boldsymbol{\theta}) = constant - \tfrac{1}{2\sigma^2} \,||\mathrm{g} - \mathbf{H}\,\boldsymbol{\theta}\,||^2 \,,$$

and the Fisher information matrix is an old friend:

$$\mathbf{F} = \frac{1}{\sigma^2}\,\mathbf{H}^t\mathbf{H} \,.$$

With these assumptions, ML estimation is identical to ordinary linear least-squares regression. Maximizing $\ln \mathrm{pr}(\mathrm{g}|\boldsymbol{\theta})$ is the same as *minimizing* the residual norm, $||\mathrm{g} - \mathbf{H}\,\boldsymbol{\theta}\,||^2$

## Poisson data, mean an arbitrary function of $\boldsymbol{\theta}$

From the last lecture, a general multivariate Poisson data set has a probability (not PDF) given by

$$\mathsf{Pr}(\mathbf{g}) = \prod_{\mathbf{m=1}}^{\mathbf{M}} \exp(-\overline{g}_{\mathbf{m}}) \, \frac{[\overline{g}_{\mathbf{m}}]^{\mathbf{g_m}}}{\mathbf{g_m!}} \, .$$

Since this probability has only one parameter (the mean) for each $m$, the general Poisson likelihood is

$$\mathsf{Pr}(\mathbf{g}|\boldsymbol{\theta}) = \prod_{\mathbf{m=1}}^{\mathbf{M}} \exp[-\overline{g}_{\mathbf{m}}(\boldsymbol{\theta})] \, \frac{[\overline{g}_{\mathbf{m}}(\boldsymbol{\theta})]^{\mathbf{g_m}}}{\mathbf{g_m!}} \, ,$$

and the log-likelihood is

$$\ln \mathsf{Pr}(\mathbf{g}|\boldsymbol{\theta}) = \sum_{m=1}^{M} \{-\overline{g}_m(\boldsymbol{\theta}) \, + g_m \ln[\overline{g}_m(\boldsymbol{\theta})] - \ln g_m!\} \, .$$

## Poisson data, arbitrary mean – cont.

When the mean is an arbitrary function of $\boldsymbol{\theta}$, it turns out (derivation on request) that the Fisher information has components:

$$F_{jk} = \sum_{m=1}^{M} \frac{1}{\overline{g}_m(\boldsymbol{\theta})} \frac{\partial \overline{g}_m(\boldsymbol{\theta})}{\partial \theta_j} \frac{\partial \overline{g}_m(\boldsymbol{\theta})}{\partial \theta_k}$$

All you ever need to know to compute the likelihood or the Fisher information with Poisson data is $\overline{g}_m(\boldsymbol{\theta})$.

## Poisson data, mean a linear function of $\boldsymbol{\theta}$

This is the usual case in image reconstruction from photon-limited data:

$$\overline{g}_m(\boldsymbol{\theta}) = \sum_{n=1}^{N} H_{mn}\,\theta_n\,.$$

The log-likelihood is now

$$\ln \Pr(\mathrm{g}|\boldsymbol{\theta}) = \sum_{m=1}^{M} \left\{ -\sum_{n=1}^{N} H_{mn}\,\theta_n \; + g_m \ln \left[ \sum_{n=1}^{N} H_{mn}\,\theta_n \right] - \ln g_m! \right\}\,.$$

The $k^{th}$ component of the score is

$$\frac{\partial}{\partial \theta_k} \ln \Pr(\mathrm{g}|\boldsymbol{\theta}) = \sum_{m=1}^{M} \left\{ -H_{mk} + \frac{g_m H_{mk}}{\left[ \sum_{n=1}^{N} H_{mn}\,\theta_n \right]} \right\}\,.$$

Now all we need is an algorithm to find nonnegative values of each $\theta_k$ so that all of these partial derivatives vanish.

## The expectation-maximization (EM) algorithm

From last slide,

$$\frac{\partial}{\partial \theta_k} \ln \Pr(\mathrm{g}|\boldsymbol{\theta}) = \sum_{m=1}^{M} \left\{ -H_{mk} + \frac{g_m H_{mk}}{\left[\sum_{n=1}^{N} H_{mn}\,\theta_n\right]} \right\} = 0\,,\ \forall\, k\,,\ \text{at } \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\mathsf{ML}}\,.$$

A tad of algebra shows that

$$\sum_{m=1}^{M} H_{mk} \frac{g_m}{[\mathbf{H}\,\boldsymbol{\theta}\,]_m} = \sum_{m=1}^{M} H_{mk}$$

Now multiply both sides by $\theta_k$:

$$\theta_k \sum_{m=1}^{M} H_{mk} \frac{g_m}{[\mathbf{H}\,\boldsymbol{\theta}\,]_m} = \theta_k \sum_{m=1}^{M} H_{mk}$$

and rewrite as

$$\theta_k = \theta_k \frac{1}{S_k} \sum_{m=1}^{M} H_{mk} \frac{g_m}{[\mathbf{H}\,\boldsymbol{\theta}\,]_m}\,, \qquad \text{where } S_k = \sum_{m=1}^{M} H_{mk}\,.$$

## The EM algorithm – punchline

From last slide, the ML estimate is obtained if

$$\theta_k = \theta_k \frac{1}{S_k} \sum_{m=1}^{M} H_{mk} \frac{g_m}{[\mathbf{H}\,\boldsymbol{\theta}\,]_m}, \qquad \text{where } S_k = \sum_{m=1}^{M} H_{mk}\,.$$

Now treat this as an iteration rule. Let $\hat{\theta}_k^{(n)}$ be the estimate of $\theta_k$ at the $n^{th}$ iteration and write

$$\theta_k^{(n+1)} = \theta_k^{(n)} \frac{1}{S_k} \sum_{m=1}^{M} H_{mk} \frac{g_m}{[\mathbf{H}\,\hat{\boldsymbol{\theta}}^{(n)}]_m}\,.$$

This iteration rule is called the EM (expectation-maximization) algorithm or, in the optics literature, the Richardson-Lucy algorithm.

Discuss convergence

## Still to come – applications in astronomy

- ML estimation with Shack-Hartman wavefront sensors
- The optimal generic wavefront sensor
- Image reconstruction in SCIDAR
- Estimation of fringe visibility in interferometry
- Gauss-Markov estimation for stellar photometry