

2003 CfAO Summer School for Adaptive Optics
Mathematics for AO Part II. Estimation Theory

Curt Vogel

`vogel@math.montana.edu`

Montana State University
Department of Mathematical Sciences
Bozeman, MT 59717-2400
www.math.montana.edu/~vogel

Goals

Present language and techniques from statistical estimation theory. These have important applications in

- Mathematical Imaging (Peyman Milanfar's talk on Tuesday)
- Wavefront Reconstruction (Lisa Poyneer's talks Mon & Tues)
- Multi-Conjugate Adaptive Optics (Miska Le Louarn's talk on Tues)

Outline

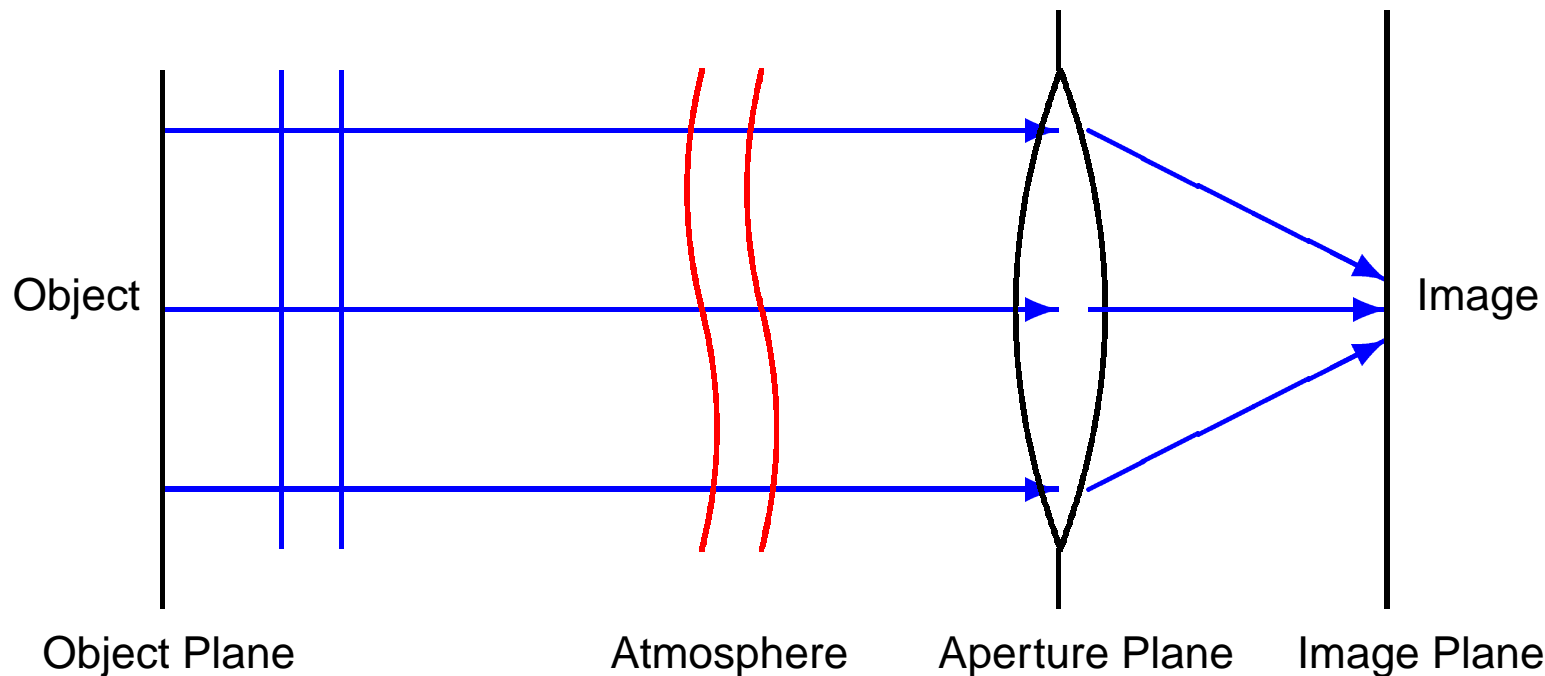
- Illustrative Example—Image Deblurring
 - Stochastic vs deterministic modeling
- Review of Probability (sample space, random variable, probability, expected value, independence, ...)
- Statistical distributions (Gaussian, Poisson)
- Estimation techniques
 - Maximum likelihood estimation.
 - Best linear unbiased estimation.
 - Bayesian statistics and maximum a posteriori (MAP) estimation.
 - Minimum variance estimation.

Incoherent Optical Imaging System

$$\underbrace{I(x, y)}_{\text{Image}} = \underbrace{\int \int}_{\text{Object Plane}} \underbrace{s(x - x', y - y')}_{\text{Point Spread Function (PSF)}} \underbrace{f(x', y')}_{\text{Object}} dx' dy'$$

PSF is intensity distribution due to point source.

$$s(x, y) = \left| \underbrace{\int \int}_{\text{Aperture Plane}} \underbrace{e^{-i2\pi(xx'' + yy'')}}_{\text{Fourier Kernel}} \underbrace{A(x'', y'')}_{\text{Mask}} \underbrace{e^{i\phi(x'', y'')}}_{\text{E-Field}} dx'' dy'' \right|^2$$



Deterministic vs Stochastic Modeling

Image data actually recorded on a CCD array (digital camera) is

$$d_{i,j} = \int \int s(x_i - x', y_j - y') f(x', y') dx' dy' + \eta_{ij}$$

where η_{ij} represents “noise”. For practical computations, convert to discrete linear system

$$\mathbf{d} = A\mathbf{f} + \boldsymbol{\eta}.$$

Typically think of the “system matrix” or “design matrix” A as being fixed, or deterministic. Noise term $\boldsymbol{\eta}$ is “uncertain” or “stochastic” and incorporates

- Measurement error.
- Discretization error (numerical integration).
- Modeling error (uncertainties in optical imaging system, e.g., wavefront aberrations due to atmosphere).

For purposes of estimation, it will be useful to think of \mathbf{f} as stochastic as well.

What is Probability?

Even uncertain events have structure.

- A coin toss can yield either a “Head” or a “Tail”.
- The birth weight of a child will be positive. It is likely to lie in the range of 1 to 10 pounds.

Probability provides a means to quantitatively model uncertain events.

Intuitive “frequentist” view of probability. To illustrate, suppose we want to quantify the outcome of a coin toss. We take the probability of the occurrence of a Head to be following limit, where N denotes the total number of coin tosses:

$$P(H) = \lim_{N \rightarrow \infty} \frac{\text{Number of Heads}}{N}$$

Some conceptual problems.

- Can’t perform an infinite number of coin tosses.
- Even if we could, it is not clear that the limit would exist.

To overcome conceptual problems, scientists developed an **axiomatic framework** for probabilistic modeling.

Probability Space & Random Variable

A **probability space** $(S, \mathcal{B}, \mathcal{P})$ consists of

- A **sample space** S ;
- A **collection of subsets** \mathcal{B} of S ; and
- a **probability function** $\mathcal{P} : S \rightarrow \mathbb{R}$ (mapping from S into real numbers) satisfying

$$\mathcal{P}(\emptyset) = 0; \mathcal{P}(S) = 1; \text{ and } \mathcal{P}(\cup_i A_i) = \sum_i \mathcal{P}(A_i)$$

for any **disjoint**, countable collection of subsets $A_i \in \mathcal{B}$.

A **random variable** is a function $X : S \rightarrow \mathbb{R}$.

Notational Convention: Statisticians use capital letters for random variables. Lower case letters represent values in the range (i.e., “realizations”) of random variables.

Coin Toss Example

To model the outcome of flipping 2 “fair” coins, we simply **define** the following:

- $S = \{HH, HT, TH, TT\}$
- $\mathcal{B} = \{\text{subsets of } S\} = \{\emptyset, \{HH\}, \dots, S\}$.
- $\mathcal{P}(\emptyset) = 0, \mathcal{P}(S) = 1, \mathcal{P}(HH) = \mathcal{P}(HT) = \mathcal{P}(TH) = \mathcal{P}(TT) = 1/4$.

To count number of heads, define random variable $X : S \rightarrow \mathbb{R}$ by

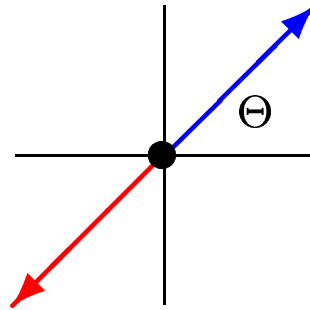
$$X(HH) = 2, X(HT) = X(TH) = 1, X(TT) = 0.$$

Note: X induces a **new sample space** $\mathcal{X} = \{0, 1, 2\}$ with a **new probability function** $P_X(x) = \mathcal{P}\{s \in S \mid X(s) = x\}$. Here

$$P_X(0) = \mathcal{P}(X = 0) = \mathcal{P}(TT) = 1/4,$$

$$P_X(1) = 1/2, P_X(2) = 1/4.$$

Spinning Dial Example



Sample space S consists of all possible orientations of the dial after it has been spun. Random variable Θ gives orientation angle in units of radians. For $0 \leq \alpha < \beta \leq 2\pi$, take

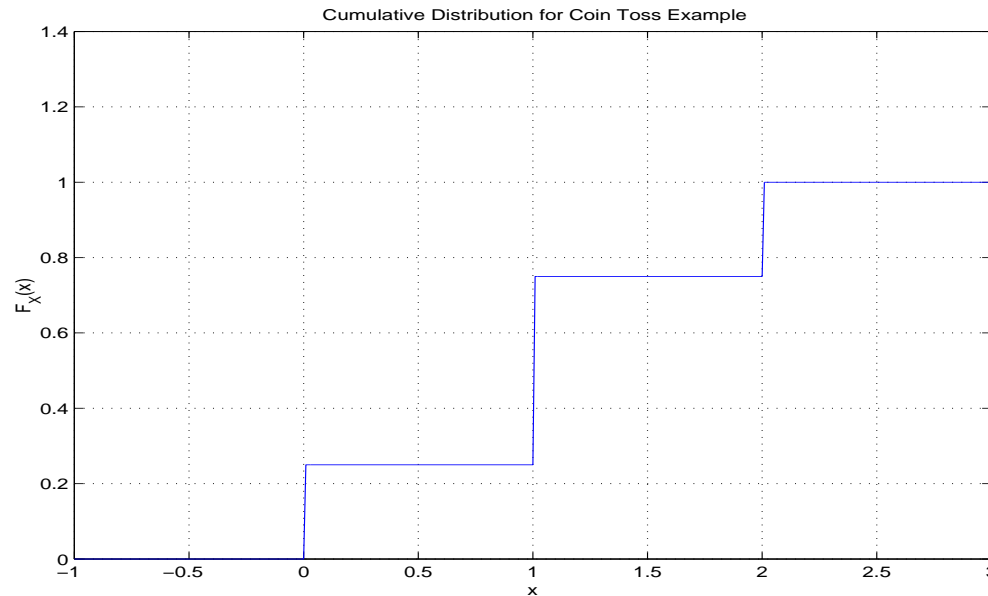
$$\mathcal{P}(\alpha \leq \Theta \leq \beta) = \frac{\beta - \alpha}{2\pi}$$

Cumulative Distribution Function

Notation $\{X \leq x\}$ means $\{s \in S \mid X(s) \leq x\}$. Associated with random variable X is the cumulative distribution function

$$F_X(x) = \mathcal{P}\{X \leq x\}, x \in \mathbb{R}.$$

- X is **continuous** if $F_X(x)$ is continuous in x .
- X is **discrete** if F_X is a step function, e.g., in coin toss example,



Probability Density & Mass Functions

- Continuous case: **Probability density function** (pdf) $\pi_X(x)$ satisfies

$$F_X(x) = \int_{-\infty}^x \pi_X(u) du,$$

so

$$\pi_X = \frac{dF_X}{dx}.$$

- Discrete case: **Probability mass function** (pmf) is

$$\pi_X(x) = \mathcal{P}(X = x), \text{ so } F_X(x) = \sum_{u \leq x} \pi_X(u).$$

Can express pmf as a “generalized” pdf with

$$\pi_X(x) = \sum_i p_i \delta(x - x_i), \quad p_i = \mathcal{P}(X = x_i) > 0.$$

Here $\delta(\cdot)$ denotes the Dirac delta.

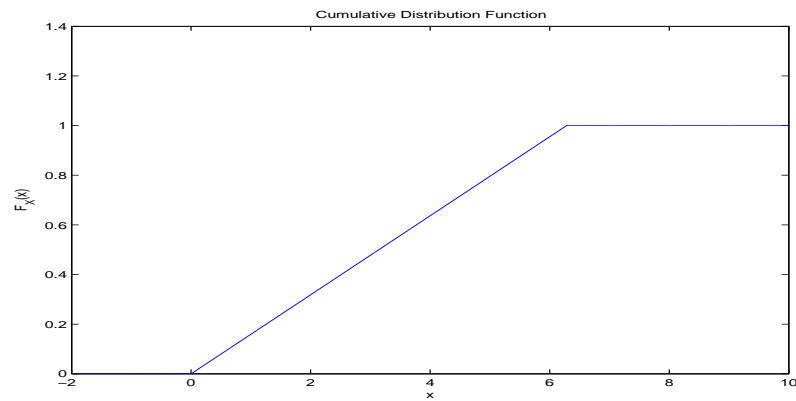
Pop Quiz!

For the spinning dial example,

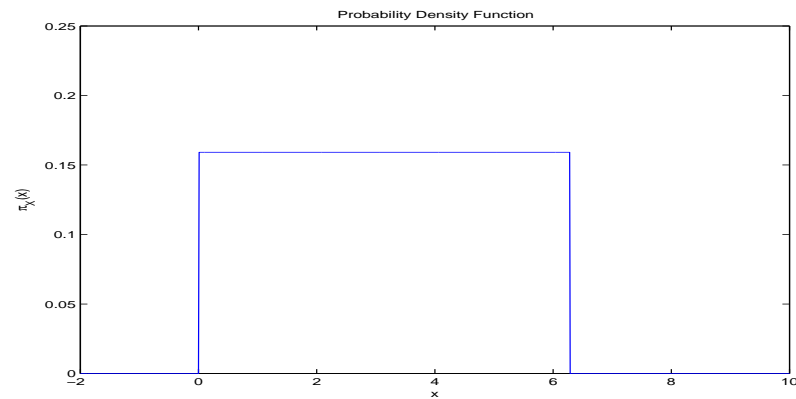
- What is the cumulative distribution function?
- What is the probability density/mass function?

Quiz Solution

Cumulative distribution function for spinning dial.



Probability density function for spinning dial.



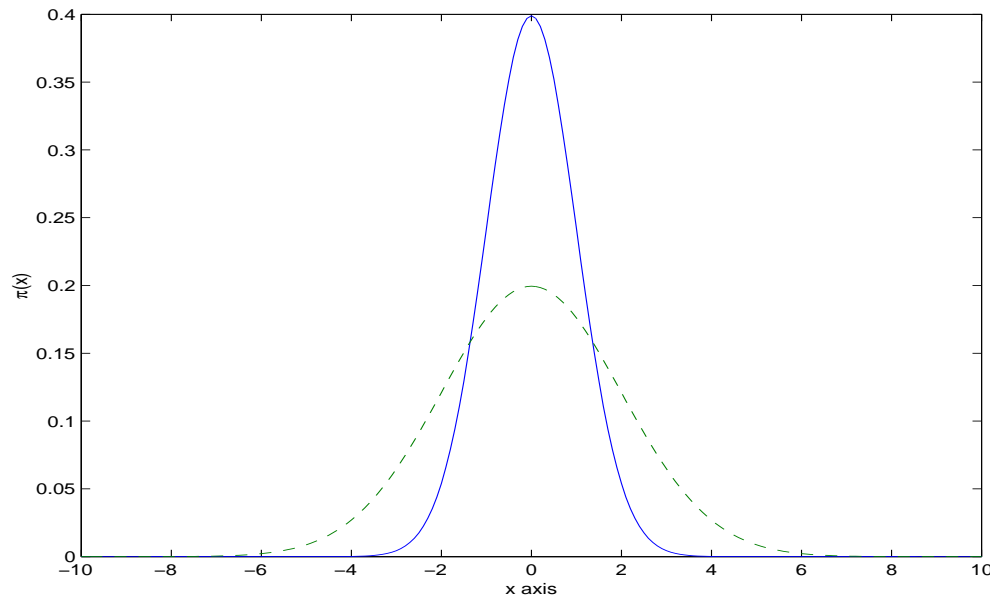
Gaussian Random Variable

This is a **continuous** random variable with pdf

$$\pi_X(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty.$$

Characterized by 2 parameters:

- Mean μ gives location of the “central peak” of the pdf.
- Variance σ^2 quantifies the “spread”.



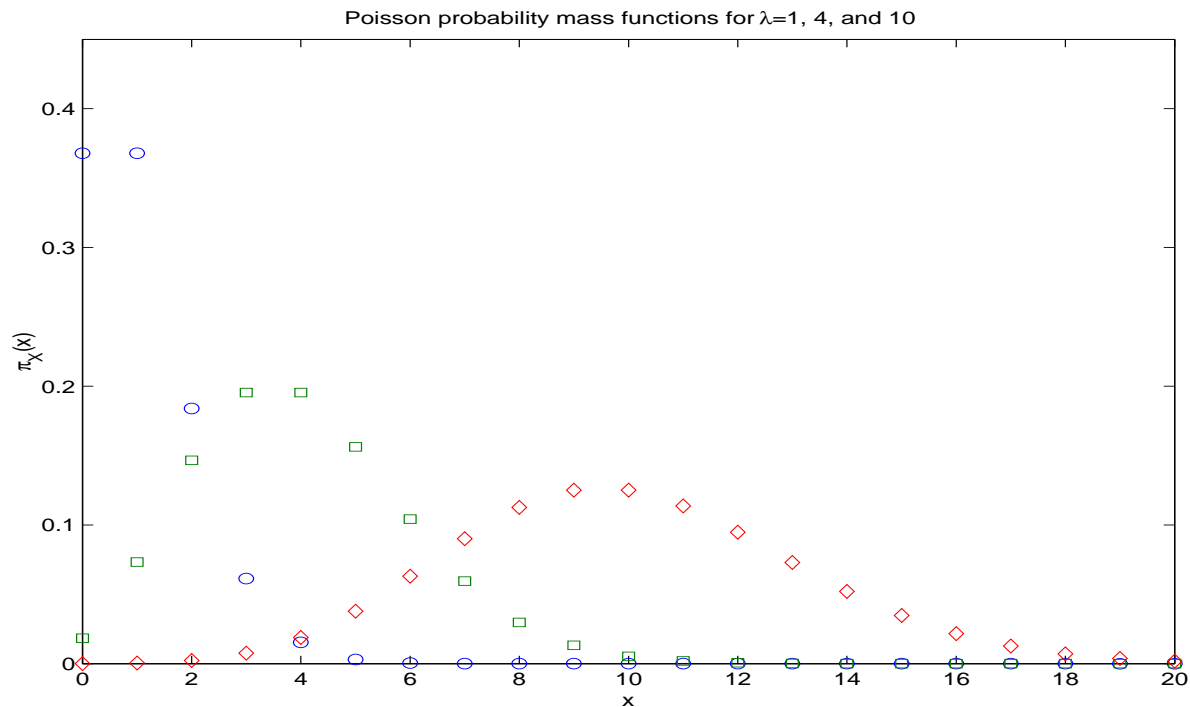
Poisson Random Variable

This is a **discrete** random variable with pmf

$$\pi_X(x; \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \dots$$

Characterized by 1 parameter:

- λ is the mean (as well as the variance).



Transformations & Expectation

If $X : S \rightarrow \mathbb{R}$ is a random variable and $g : \mathbb{R} \rightarrow \mathbb{R}$ is a function, then $Y = g(X) : S \rightarrow \mathbb{R}$ is a new random variable with cumulative distribution

$$\begin{aligned} F_Y(y) &= \mathcal{P}(Y \leq y) = \mathcal{P}\{s \in S \mid g(X(s)) \leq y\} \\ &= \mathcal{P}\{s \in S \mid X(s) \in g^{-1}(-\infty, y]\}. \end{aligned}$$

Expectation (mean, expected value): “Average Value”

$$\langle g(X) \rangle = \begin{cases} \int_{-\infty}^{\infty} g(x) \pi_X(x) dx, & X \text{ continuous} \\ \sum_x g(x) \pi_X(x), & X \text{ discrete} \end{cases}$$

Variance quantifies “spread” or “variability”

$$\text{var}(X) = \langle (X - \langle X \rangle)^2 \rangle.$$

Random Vectors, Independence

Random variables are “jointly distributed” if they are defined on the same probability space $(S, \mathcal{B}, \mathcal{P})$. A **random vector** $\mathbf{X} = (X_1, \dots, X_n) : S \rightarrow \mathbb{R}^n$ has jointly distributed components X_i . The **joint distribution function** is

$$F_{\mathbf{X}}(x_1, \dots, x_n) = \mathcal{P}(X_1 \leq x_1, \dots, X_n \leq x_n).$$

\mathbf{X} is continuous with joint pdf $\pi_{\mathbf{X}}$ if

$$F_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} \pi_{\mathbf{X}}(\mathbf{u}) d\mathbf{u}.$$

In the discrete case, $\pi_{\mathbf{X}}(\mathbf{x}) = \sum_i p_i \delta(\mathbf{x} - \mathbf{x}_i)$, $p_i > 0$.

Components X_i are **independent** if the joint distribution is the product of individual distribution functions, i.e.,

$$F_{\mathbf{X}}(\mathbf{x}) = \prod_i F_{X_i}(x_i).$$

Then $\pi_{\mathbf{X}}(\mathbf{x}) = \prod_i \pi_{X_i}(x_i)$.

Mean & Covariance

Mean, or expected value, of random vector \mathbf{X} is

$$\langle \mathbf{X} \rangle = [\langle X_1 \rangle, \dots, \langle X_n \rangle].$$

Covariance is $n \times n$ matrix with components

$$[\text{cov}(\mathbf{X})]_{ij} = \langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \rangle, \quad 1 \leq i, j \leq n.$$

This quantifies “degree of association” between components of \mathbf{X} .

- If $[\text{cov}(\mathbf{X})]_{ij} > 0$, then $X_i > \langle X_i \rangle$ and $X_j > \langle X_j \rangle$ or $X_i < \langle X_i \rangle$ and $X_j < \langle X_j \rangle$ tend to **happen together**.
- If $[\text{cov}(\mathbf{X})]_{ij} < 0$, then $X_i > \langle X_i \rangle$ and $X_j < \langle X_j \rangle$ or $X_i < \langle X_i \rangle$ and $X_j > \langle X_j \rangle$ tend to **happen together**.

Diagonal entries of covariance matrix are variances σ_i^2 of X_i . If $\langle \mathbf{X} \rangle = \mathbf{0}$, then

$$\text{cov}(\mathbf{X}) = \langle \mathbf{X}\mathbf{X}^T \rangle$$

Gaussian Random Vector

Random vector \mathbf{X} has a (nondegenerate) **Gaussian**, or **normal**, distribution if it has a joint pdf

$$\pi_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\mu}, C) = \frac{1}{\sqrt{(2\pi)^n \det(C)}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T C^{-1}(\mathbf{x} - \boldsymbol{\mu})\right),$$

where $\mathbf{x}, \boldsymbol{\mu} \in \mathbb{R}^n$, C is an $n \times n$ symmetric positive definite matrix, and $\det(\cdot)$ denotes matrix determinant.

Notation:

$$\mathbf{X} \sim \text{Normal}(\boldsymbol{\mu}, C)$$

One can show

$$\langle \mathbf{X} \rangle = \boldsymbol{\mu}, \quad \text{cov}(\mathbf{X}) = C$$

Poisson Random Vector

(Discrete) random vector \mathbf{X} has a Poisson distribution with independent components if it has joint pmf

$$\pi_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\lambda}) = \begin{cases} \prod_{i=1}^n e^{-\lambda_i} \lambda_i^{x_i} / x_i!, & \mathbf{x} \in \mathbf{Z}_+^n, \\ 0, & \text{elsewhere,} \end{cases}$$

where \mathbf{Z}_+^n denotes n-vectors \mathbf{x} whose components x_i are each nonnegative integers. Components λ_i of the Poisson parameter $\boldsymbol{\lambda}$ are each nonnegative real numbers.

Notation:

$$\mathbf{X} \sim \text{Poisson}(\boldsymbol{\lambda})$$

One can show

$$\langle \mathbf{X} \rangle = \boldsymbol{\lambda}, \quad \text{cov}(\mathbf{X}) = \text{diag}(\lambda_i)$$

Application: Two Gaussian Imaging Models

$$\mathbf{Y} = A\mathbf{X} + \mathbf{N}$$

- System, or “design”, matrix A is deterministic and models optical imaging system.
- \mathbf{X} represents the object. Model it as a random vector with a Gaussian distribution, $\mathbf{X} \sim \text{Normal}(\mathbf{0}, C_{\mathbf{X}})$.
- \mathbf{N} represents noise, modeled as a random vector with a Gaussian distribution, $\mathbf{N} \sim \text{Normal}(\mathbf{0}, \sigma_N^2 I)$.
- Assume \mathbf{X}, \mathbf{N} are independent.

Can show (using characteristic functions, or Fourier transforms) that

$$\mathbf{Y} \sim \text{Normal}(\mathbf{0}, C_{\mathbf{Y}}), \quad C_{\mathbf{Y}} = \langle \mathbf{Y}\mathbf{Y}^T \rangle = AC_{\mathbf{X}}A^T + \sigma_N^2 I.$$

If we assume unknown quantity $\mathbf{X} = \mathbf{x}$ is deterministic, then

$$\mathbf{Y} = A\mathbf{x} + \mathbf{N} \sim \text{Normal}(A\mathbf{x}, \sigma_N^2 I).$$

Maximum Likelihood Estimation

Suppose \mathbf{X} has a joint probability density/mass $\pi_{\mathbf{X}}(\mathbf{x}; \alpha)$, where α is an unknown parameter vector, and suppose we observe data $\mathbf{d} = \mathbf{X}(s)$ for some s in the sample space S .

The **maximum likelihood estimator** (MLE) for α given \mathbf{d} is a parameter $\hat{\alpha}$ that maximizes the **likelihood function**

$$L(\alpha) \stackrel{\text{def}}{=} \pi_{\mathbf{X}}(\mathbf{d}; \alpha).$$

Monotone transformations like the logarithm preserve maximizers and can simplify MLE computations. Hence, the MLE is a maximizer of the **log likelihood function**,

$$\ell(\alpha) \stackrel{\text{def}}{=} \log \pi_{\mathbf{X}}(\mathbf{d}; \alpha).$$

MLE Example 1

Suppose $X \sim \text{Normal}(\mu, 1)$, where the mean μ is unknown. Estimate μ from a single observation d of the random variable X .

The likelihood function is

$$L(\mu; d) = \pi_X(d; \mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(d - \mu)^2}{2}\right).$$

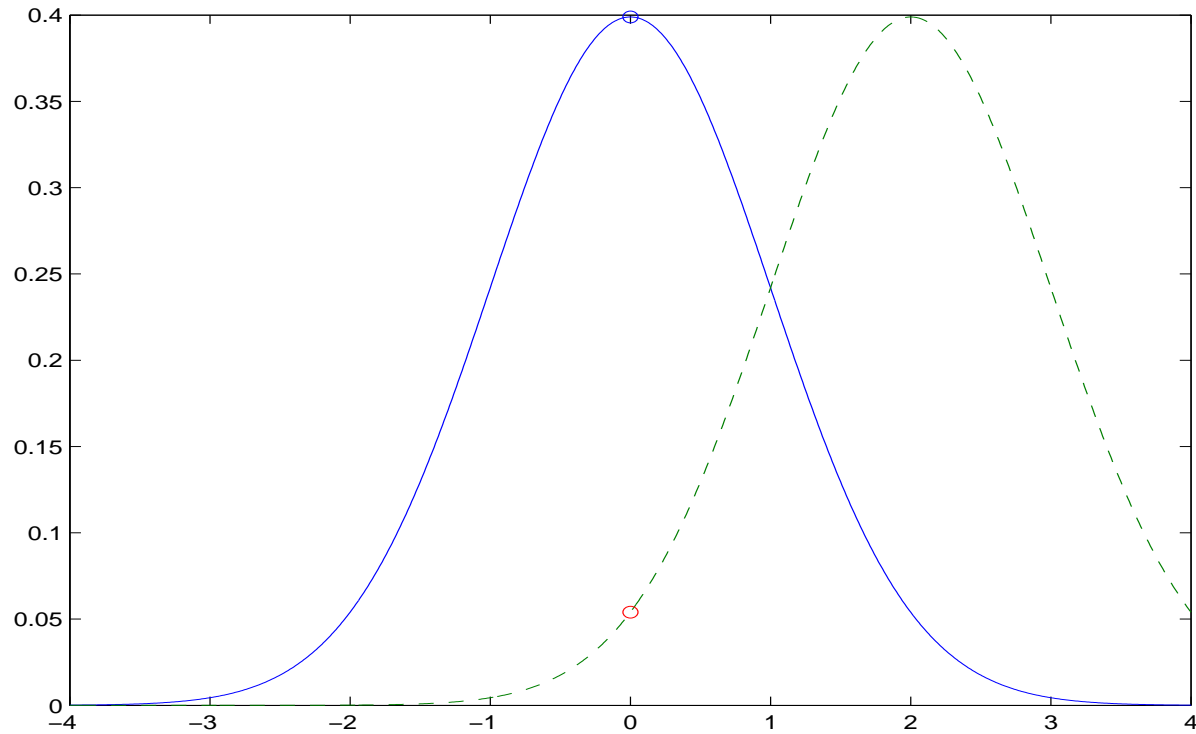
The log likelihood function is

$$\ell(\mu; d) = \log \pi_X(d; \mu) = -\frac{(d - \mu)^2}{2} + \text{const.}$$

The MLE is

$$\hat{\mu} = \operatorname{argmax}_{\mu} \left\{ -\frac{(d - \mu)^2}{2} \right\} = d.$$

Illustration of MLE Example 1



Given observation $d = 0$, red circle denotes value of likelihood function $L(\mu; d)$ for $\mu = 2$; blue circle denotes value for $\mu = 0$.

MLE Example 2

Consider the stochastic model

$$\mathbf{Y} = A\mathbf{x} + \mathbf{N},$$

where $\mathbf{N} \sim \text{Normal}(\mathbf{0}, \sigma_N^2 I_{m \times m})$, A is an $m \times n$ matrix, and $\mathbf{x} \in \mathbb{R}^n$. Assume that A and \mathbf{x} are deterministic, but \mathbf{x} is unknown and is to be estimated from data $\mathbf{y} = \mathbf{Y}(s)$. Since $\mathbf{Y} \sim \text{Normal}(A\mathbf{x}, \sigma_N^2 I)$, the MLE for \mathbf{x} is

$$\begin{aligned}\hat{\mathbf{x}} &= \arg \max_{\mathbf{x}} \log \pi_{\mathbf{Y}}(\mathbf{y}; A\mathbf{x}, \sigma_N^2 I) \\ &= \arg \max_{\mathbf{x}} \log \frac{\exp\left(-\frac{1}{2}(\mathbf{y} - A\mathbf{x})^T (\sigma_N^2 I)^{-1} (\mathbf{y} - A\mathbf{x})\right)}{\sqrt{(2\pi)^m \det(\sigma_N^2 I)}} \\ &= \arg \min_{\mathbf{x}} \underbrace{(\mathbf{y} - A\mathbf{x})^T (2\sigma_N^2 I)^{-1} (\mathbf{y} - A\mathbf{x})}_{\|\mathbf{Ax}-\mathbf{y}\|^2 / 2\sigma_N^2} + \text{const} \\ &= A^\dagger \mathbf{y}.\end{aligned}$$

Best Linear Unbiased Estimator (BLUE)

Motivation:

- MLE may be **hard to compute** for non-Gaussian distributions.
- The **pdf/pmf may be unknown**, but means and covariances are often available or can be accurately approximated.

Assume linear model of MLE Example 2, $\mathbf{Y} = A\mathbf{x} + \mathbf{N}$, where A , \mathbf{x} are deterministic and \mathbf{x} is unknown. Now assume only

$$\langle \mathbf{N} \rangle = \mathbf{0}, \quad C_N = \langle \mathbf{N}\mathbf{N}^T \rangle \text{ is known.}$$

The BLUE is the random vector which **mimimizes the cost functional**

$$J(\mathbf{X}) = \langle \|\mathbf{X} - \mathbf{x}\|^2 \rangle$$

subject to the constraints (linearity, bias free)

$$\begin{aligned} \mathbf{X} &= B\mathbf{Y}, \quad B \text{ an } n \times m \text{ matrix} \\ \langle \mathbf{X} \rangle &= \mathbf{x} \end{aligned}$$

BLUE, Continued

If A has full rank (linearly independent columns) and C_N is nonsingular, then

$$\hat{\mathbf{X}}_{\text{BLUE}} = (A^T C_N^{-1} A)^{-1} A^T C_N^{-1} \mathbf{Y}.$$

Some observations:

- If $C_N = \sigma^2 I$, then $\hat{\mathbf{X}}_{\text{BLUE}} = A^\dagger \mathbf{Y}$.
- If $\mathbf{N} \sim \text{Normal}(\mathbf{0}, C_N)$, then the MLE and the BLUE are the same.
- If A has small singular values, then the MLE and BLUE may be unstable with respect to noise in the data.

Can stabilize by incorporating prior information about solution \mathbf{x} . We will assume $\mathbf{x} = \mathbf{X}$ is a random vector rather than an unknown deterministic vector. This leads to **Maximum A Posteriori (MAP)** estimation.

- Estimate will be **biased**, but will have **smaller variance**.

Frequentist View of Conditional Probability

Suppose events A, B may occur together (e.g., rising carbon dioxide levels and global warming). Let

- $n = \#$ of times A, B occur together in N “experiments”.
- $m = \#$ of times A occurs in N experiments, regardless of B .

In the limit as $N \rightarrow \infty$,

$$P(A, B) = \frac{n}{N} = \underbrace{\frac{n}{m}}_{P(B|A)} \underbrace{\frac{m}{N}}_{P(A)},$$

or equivalently,

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

This is the “conditional” relative frequency of B , given that A occurred.

- If A, B are independent, then $P(B|A) = P(B)$.
- If A, B always occur together, then $P(B|A) = 1$.

Axiomatic Conditional Probability

For simplicity, assume discrete random variables. Replace \sum by \int in the continuous case. Let $\mathbf{X} = (X_1, \dots, X_n)$ and $\mathbf{Y} = (Y_1, \dots, Y_m)$ be jointly distributed random vectors. The joint pmf for (\mathbf{X}, \mathbf{Y}) is

$$\pi_{(\mathbf{X}, \mathbf{Y})}(\mathbf{x}, \mathbf{y}) = \mathcal{P}(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}).$$

The marginal pmf for \mathbf{X} is

$$\pi_{\mathbf{X}}(\mathbf{x}) = \sum_{\mathcal{P}\{\mathbf{Y}=\mathbf{y}\}>0} \pi_{(\mathbf{X}, \mathbf{Y})}(\mathbf{x}, \mathbf{y}).$$

The conditional pmf for \mathbf{Y} given $\mathbf{X} = \mathbf{x}$ is

$$\pi_{(\mathbf{Y}|\mathbf{X})}(\mathbf{y}|\mathbf{x}) = \frac{\pi_{(\mathbf{X}, \mathbf{Y})}(\mathbf{x}, \mathbf{y})}{\pi_{\mathbf{X}}(\mathbf{x})}.$$

Coin Toss Example, Revisited

Toss 2 independent fair coins. Let X_i denote number of heads for coin i , $i=1,2$, and let $Y = X_1 + X_2$ denote total number of heads.

Outcome	X_1	X_2	$Y = X_1 + X_2$	Probability
(T,T)	0	0	0	1/4
(T,H)	0	1	1	1/4
(H,T)	1	0	1	1/4
(H,H)	1	1	2	1/4

The marginal pmf for Y is given by

$$\pi_Y(0) = \mathcal{P}\{Y = 0\} = 1/4, \quad \pi_Y(1) = 1/2, \quad \pi_Y(2) = 1/4.$$

Revisited Coin Toss Example, Continued

By direct computation, or using the fact that X_1, X_2 are independent, we obtain the marginal pmf for X_1 ,

$$\pi_{X_1}(0) = \mathcal{P}(X_1 = 0) = 1/2, \quad \pi_{X_1}(1) = 1/2.$$

Thus the conditional pmf of Y given $X_1 = x_1$, denoted by $\pi_{Y|X_1}(y|x_1)$, is given in the following table.

	$x_1 = 0$	$x_1 = 1$
$y = 0$	1/2	0
$y = 1$	1/2	1/2
$y = 2$	0	1/2

Knowing whether the 1st coin is a H or a T changes the probabilities of the total number of heads being 0, 1, or 2.

Another Pop Quiz

Let random variable Θ represent orientation angle of spinner, in radians. Take 2nd random

$$X = \begin{cases} 1, & \text{if spinner points in 1st quadrant } (0 \leq \theta < \pi/2) \\ 0, & \text{otherwise} \end{cases}$$

What is $\pi_{\Theta|X}(\theta|x)$?

Conditional Expectation

Let $g : \mathbb{R}^n \rightarrow \mathbb{R}^k$ be a (Borel measurable) mapping. The **conditional expectation** of $g(\mathbf{Y})$ given $\mathbf{X} = \mathbf{x}$ is

$$\langle g(\mathbf{Y}) | \mathbf{X} = \mathbf{x} \rangle = \sum_{\mathcal{P}\{\mathbf{Y}=\mathbf{y}\} > 0} g(\mathbf{y}) \pi_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}).$$

In the 2nd coin toss example,

$$\langle Y \rangle = \sum_y y \mathcal{P}(Y = y) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1.$$

On the other hand,

$$\begin{aligned} \langle Y | X_1 = 0 \rangle &= \sum_y y \mathcal{P}(Y = y | X_1 = 0) \\ &= 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} + 2 \cdot 0 = \frac{1}{2}, \\ \langle Y | X_1 = 1 \rangle &= \sum_y y \mathcal{P}(Y = y | X_1 = 1) = \frac{3}{2}. \end{aligned}$$

Bayes Theorem

Basic result from the “calculus of probability”. Relates conditional probability of \mathbf{X} given \mathbf{Y} to the conditional probability of \mathbf{Y} given \mathbf{X} .

$$\pi_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) = \frac{\pi_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) \pi_{\mathbf{X}}(\mathbf{x})}{\pi_{\mathbf{Y}}(\mathbf{y})}.$$

In inverse problems, \mathbf{X} represents the quantity of interest.

- The **prior** $\pi_{\mathbf{X}}(\mathbf{x})$ quantifies a priori info about \mathbf{X} .
- \mathbf{y} represents an **observation** of the data \mathbf{Y} .
- $\pi_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$ is the “**forward model**”.
- $\pi_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})$ gives the **a posteriori** probability of \mathbf{X} given an observation \mathbf{y} of \mathbf{Y} .

Bayesian Solution to Inverse Problems

Given an observation \mathbf{y} of \mathbf{Y} ,

- The “average value” of \mathbf{X} given \mathbf{y} is the **posterior mean**

$$\langle \mathbf{X} | \mathbf{Y} = \mathbf{y} \rangle = \begin{cases} \sum_{\mathbf{x}} \mathbf{x} \pi_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}), & \mathbf{X} \text{ discrete,} \\ \int \mathbf{x} \pi_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) d\mathbf{x}, & \mathbf{X} \text{ cts.} \end{cases}$$

- The “most likely” value of \mathbf{X} given \mathbf{y} is the **maximum a posteriori (MAP) estimator**,

$$\begin{aligned} \mathbf{x}_{\text{MAP}} &= \arg \max_{\mathbf{x}} \pi_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) \\ &= \arg \max_{\mathbf{x}} \{ \pi_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) \pi_{\mathbf{X}}(\mathbf{x}) / \pi_{\mathbf{Y}}(\mathbf{y}) \} \\ &= \arg \max_{\mathbf{x}} \{ \log \pi_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) + \log \pi_{\mathbf{X}}(\mathbf{x}) - \text{const} \} \\ &= \arg \min_{\mathbf{x}} \{ -\log \pi_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) - \log \pi_{\mathbf{X}}(\mathbf{x}) \}. \end{aligned}$$

MAP Example

Let \mathbf{X} , \mathbf{N} be independent, jointly distributed random vectors with $\mathbf{X} \sim \text{Normal}(\mathbf{0}_n, C_{\mathbf{X}})$, $\mathbf{N} \sim \text{Normal}(\mathbf{0}_m, C_{\mathbf{N}})$. Let A be an $m \times n$ matrix, and model data by

$$\mathbf{Y} = A\mathbf{X} + \mathbf{N}.$$

Goal: Derive MAP estimator of \mathbf{x} based on sample $\mathbf{y} = A\mathbf{x} + \mathbf{n}$. One can show

$$\pi_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \pi_{A\mathbf{X}+\mathbf{N}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \pi_{A\mathbf{x}+\mathbf{N}}(\mathbf{y}).$$

But $A\mathbf{x} + \mathbf{N} \sim \text{Normal}(A\mathbf{x}, C_{\mathbf{N}})$, so

$$\pi_{A\mathbf{x}+\mathbf{N}}(\mathbf{y}) = \frac{\exp[-(\mathbf{y} - A\mathbf{x})^T C_{\mathbf{N}}^{-1} (\mathbf{y} - A\mathbf{x})/2]}{\sqrt{(2\pi)^m \det(C_{\mathbf{N}})}}.$$

MAP Example, Continued

The prior is given by

$$\pi_{\mathbf{X}}(\mathbf{x}) = \frac{\exp[-\mathbf{x}^T C_{\mathbf{X}}^{-1} \mathbf{x}/2]}{\sqrt{(2\pi)^n \det(C_{\mathbf{X}})}}.$$

Thus

$$\begin{aligned} \mathbf{x}_{\text{MAP}} &= \arg \min_{\mathbf{x}} \{-\log \pi_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) - \log \pi_{\mathbf{X}}(\mathbf{x})\} \\ &= \arg \min_{\mathbf{x}} \{(\mathbf{y} - A\mathbf{x})^T C_{\mathbf{N}}^{-1} (\mathbf{y} - A\mathbf{x})/2 + \mathbf{x}^T C_{\mathbf{X}}^{-1} \mathbf{x}/2\} \\ &= (A^T C_{\mathbf{N}}^{-1} A + C_{\mathbf{X}}^{-1})^{-1} A^T C_{\mathbf{N}}^{-1} \mathbf{y}. \end{aligned}$$

Now if $C_{\mathbf{N}} = \sigma_{\mathbf{N}}^2 I_m$ and $C_{\mathbf{X}} = \sigma_{\mathbf{X}}^2 I_n$,

$$\mathbf{x}_{\text{MAP}} = (A^T A + \alpha I_n)^{-1} A^T \mathbf{y}, \quad \alpha = \left(\frac{\sigma_{\mathbf{N}}}{\sigma_{\mathbf{X}}} \right)^2.$$

- With a linear model and Gaussian statistics, MAP estimation is equivalent to standard Tikhonov Regularization, or Wiener filtering.

Minimum Variance Estimation

This is the analogue for MAP of BLUE for maximum likelihood estimation.
Again take the linear model

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{N}$$

but assume only

- $\langle \mathbf{N} \rangle = \mathbf{0}$ and $C_{\mathbf{N}} = \langle \mathbf{N}\mathbf{N}^T \rangle$ is known and nonsingular.
- $\langle \mathbf{X} \rangle = \mathbf{0}$ and $C_{\mathbf{X}} = \langle \mathbf{X}\mathbf{X}^T \rangle$ is known and nonsingular.
- \mathbf{X}, \mathbf{N} are independent.

Find $\mathbf{X}_{\text{MV}} = R\mathbf{Y}$ which minimizes (with no other constraints)

$$\begin{aligned} J(\hat{\mathbf{X}}) &= \langle \|\hat{\mathbf{X}} - \mathbf{X}\|^2 \rangle \\ &= \langle \|\mathbf{R}\mathbf{Y} - \mathbf{X}\|^2 \rangle \\ &= \langle \|(R\mathbf{A} - I)\mathbf{X} + R\mathbf{N}\|^2 \rangle \end{aligned}$$

Minimum Variance Estimation, Continued

Can show matrix R (the “reconstructor”, in case of wavefront estimation) is given by

$$\begin{aligned} R &= \langle \mathbf{X}\mathbf{Y}^T \rangle \langle \mathbf{Y}\mathbf{Y}^T \rangle^{-1} \\ &= C_{\mathbf{X}} A^T (A C_{\mathbf{X}} A^T + C_{\mathbf{N}})^{-1} \\ &= (A^T C_{\mathbf{N}}^{-1} A + C_{\mathbf{X}}^{-1})^{-1} A^T C_{\mathbf{N}}^{-1} \\ &= (A^T A + \sigma_{\mathbf{N}}^2 C_{\mathbf{X}}^{-1})^{-1} A^T \quad \text{in case } C_{\mathbf{N}} = \sigma_{\mathbf{N}}^2 I \end{aligned}$$

Same result as MAP with Gaussian statistics!

Realistic Model for Image Data

$$\mathbf{d} = \mathbf{c} + \mathbf{b} + \mathbf{r},$$

where \mathbf{c} is a vector of photon “signal” counts from a CCD detector array, \mathbf{b} is a vector of photon “background” counts, and \mathbf{r} is detector read noise. Assume

- $\mathbf{r} \sim \text{Normal}(\mathbf{0}, \sigma^2 I_n)$, where σ^2 is known.
- $\mathbf{b} \sim \text{Poisson}(\boldsymbol{\lambda}_b)$, where $\boldsymbol{\lambda}_b$ is an n -vector with known constant entries $b > 0$.
- $\mathbf{c} \sim \text{Poisson}(\boldsymbol{\lambda})$, where $\lambda_i = (s \star f_{\text{true}})(\mathbf{x}_i)$, $i = 1, 2, \dots, n$.
- The components of each of \mathbf{c} , \mathbf{b} , and \mathbf{r} are all independent.
- $(s \star f)(\mathbf{x}) = \int \int s(\mathbf{x} - \mathbf{x}') f(\mathbf{x}') d\mathbf{x}'$. Assume the PSF s is known.
- Light source is incoherent, so the object f_{true} and the PSF s are **nonnegative**. This implies the Poisson parameter $\boldsymbol{\lambda}$ for \mathbf{c} has nonnegative entries.

Goal: Apply MAP estimation to estimate f_{true} from observed image data \mathbf{d} .

Preliminary Step: Approximate $(s \star f)(\mathbf{x})$ by $A\mathbf{f}$, e.g., using midpoint quadrature.

Assume A , \mathbf{f} have **nonnegative entries** and the **quadrature approximation error is small** compared to the measurement noise components. Then

$$\mathbf{c} \sim \text{Poisson}(A\mathbf{f}).$$

Posterior Log Likelihood Approximation

Technical Problem: The sum of independent Poisson random variables is a Poisson random variable (the Poisson parameters add). Unfortunately, the sum of a Poisson rv and a Gaussian rv has a really ugly probability density function (it has an infinite series representation, and it has jump discontinuities at the positive integers).

Moment Matching Approximation, due to Snyder: Add σ^2 to each component of both sides of model. Then approximate each $r_i + \sigma^2 \sim \text{Normal}(\sigma^2, \sigma^2)$ by a Poisson rv with Poisson parameter σ^2 (equal to mean and variance of $r_i + \sigma^2$). Combining this approximation with the previous assumptions,

$$\mathbf{d} + \sigma^2 \sim \text{Poisson}(A\mathbf{f} + b + \sigma^2).$$

This gives the log likelihood

$$\ell(A\mathbf{f} + b + \sigma^2; \mathbf{d} + \sigma^2) = - \sum_i ([A\mathbf{f}]_i + b + \sigma^2) + \sum_i (d_i + \sigma^2) \log([A\mathbf{f}]_i + b + \sigma^2)$$

Relabel this quantity as $\ell(A\mathbf{f}; \mathbf{d})$.

Choice of Prior

This is **very much problem dependent**. It also **tends to be very ad hoc**.

A typical “noninformative” prior used in imaging is to take

$$P(\mathbf{f}) = \alpha \|\mathbf{f}\|^2$$

with the additional constraint $f_i \geq 0$, $i = 1, \dots, N$. $\alpha > 0$ gives the “strength” of the prior and is unknown.

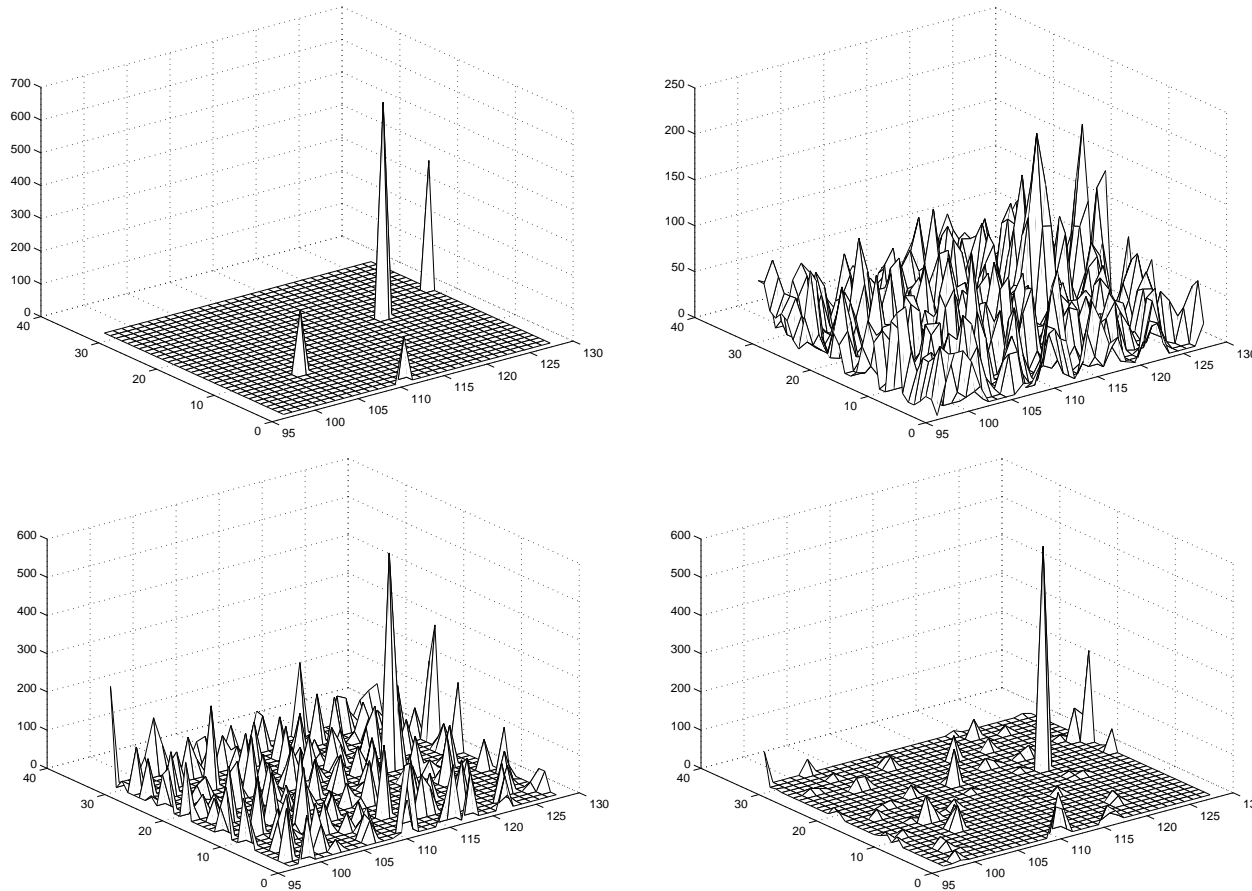
Can also use an entropy prior.

Combine the posterior log likelihood with the prior to get

$$J(\mathbf{f}) = \sum_i ([A\mathbf{f}]_i + b + \sigma^2) - \sum_i (d_i + \sigma^2) \log([A\mathbf{f}]_i + b + \sigma^2) + \alpha \|\mathbf{f}\|^2$$

to be minimized over $f_i > 0$.

Effects of Posterior Likelihood (Fit-to-Data Functional)



Object Reconstructions. Upper left: lower left 32×32 pixels of true image; lower left: reconstruction obtained using regularized, weighted least squares with nonnegativity constraints; upper right: unconstrained least squares reconstruction in which negative pixels have been set to zero; lower right: regularized Poisson likelihood reconstruction with nonnegativity constraints.

Conclusions

- Statistical estimation theory is a powerful tool for applications with uncertainty in the data, the model,
- A variety of estimation methods are available.
- Maximum likelihood estimate (MLE) lacks robustness when there is ill-conditioning (design matrix A has small singular values).
- Best linear unbiased estimate (BLUE) uses less information (mean and covariance) than MLE, but has same robustness problems.
- MLE and BLUE are equivalent for Gaussian linear models with zero-mean noise.
- Maximum a posteriori (MAP) and minimum variance (MV) estimates incorporate prior information about the solution and are more robust.
- MAP and MV are equivalent for Gaussian linear models with zero-mean random vectors.
- Incorporating noise statistics and prior information can dramatically improve accuracy.

References

- M. Bertero and P. Boccacci, *Introduction to Inverse Problems in Imaging*, IOP Publishing, 1998.
- G. Casella and R. L. Berger, *Statistical Inference, 2nd Edition*, Duxbury Press, 2002.
- J. W. Goodman, *Statistical Optics*, John Wiley and Sons, 2000.
- D.L. Snyder, M. Hammoud, and R.L. White, “Image recovery from data acquired with a charge-coupled device camera”, *JOSA-A*, **10** (1993), pp. 1014-1023.
- D.L. Snyder, G.W. Helstrom, A.D. Lanterman, M. Faisal, and R.L. White, “Compensation for readout noise in CCD images”, *JOSA-A*, **12** (1993), pp. 272-283.
- C. R. Vogel, *Computational Methods for Inverse Problems*, SIAM, 2002.