Chapter 31

Probability and random variables

ap,prob

Contents

s,prob,intro

31.1 Introduction (s,prob,intro)

This appendix summarizes some facts from probability and statistics that are used in developing and analyzing image reconstruction algorithms. This is not intended to be a tutorial introduction; for review see [1].

s,prob,cond

31.2 Conditional probability (s,prob,cond)

For two events A and B, the definition of **conditional probability** is:

e,prob,cond

$$\mathsf{P}\{A \,|\, B\} = \frac{\mathsf{P}\{A, B\}}{\mathsf{P}\{B\}}. \tag{31.2.1}$$

Bayes rule is:

e,prob,bayes

$$\mathsf{P}\{A|B\} = \frac{\mathsf{P}\{B|A\}\,\mathsf{P}\{A\}}{\mathsf{P}\{B\}}. \tag{31.2.2}$$

If $\{A_i\}$ are disjoint (mutually exclusive) events and $\mathsf{P}\{\bigcup_i A_i\} = 1$, then the **total probability** relation is:

e,prob,total

$$\mathsf{P}\{B\} = \sum_i \mathsf{P}\{B|A_i\}\,\mathsf{P}\{A_i\}\,. \tag{31.2.3}$$

31.3 Discrete distributions

s,prob,nomial

31.3.1 Binomial and Multinomial (s,prob,nomial)

Random variable X has a **binomial distribution** with parameter p if

$$\mathsf{P}\{X = x\} = \begin{cases} \mathsf{p}, & x = 1 \\ 1 - \mathsf{p}, & x = 0. \end{cases}$$

Random variables $X_1, \ldots, X_M$ have a **multinomial distribution** with parameters $\mathsf{p}_1, \ldots, \mathsf{p}_M$ and N if

e,prob,multinomial

$$\mathsf{P}\{X_1 = x_1, \ldots, X_M = x_M\} = \left(\begin{array}{c} N \\ x_1 \; x_2 \; \ldots \; x_M \end{array} \right) \prod_{m=1}^{M} \mathsf{p}_m \, \mathbb{I}_{\{\sum_{m=1}^M x_m = N\}}, \tag{31.3.1}$$

where $\left(\begin{array}{c} N \\ x_1 \; x_2 \; \ldots \; x_M \end{array} \right) = \frac{N!}{\prod_{m=1}^M x_m!}$.

s,prob,poisson

31.3.2 Poisson (s,prob,poisson)

A **Poisson** [2, p. 206] random variable N with mean μ_0 has the following **probability mass function** (**PMF**):

$$\mathsf{P}\{N = k\} = \frac{1}{k!} \mathrm{e}^{-\mu_0} \mu_0^k, \qquad k = 0, 1, \ldots.$$

We write $N \sim \mathsf{Poisson}\{\mu_0\}$. The above PMF is the usual form, but to be more precise, one should consider the case where $\mu_0 = 0$, and write the PMF as follows:

e,prob,poisson,pmf

$$\mathsf{P}\{N = k\} = \begin{cases} 1, & \mu_0 = 0, \; k = 0 \\ \frac{1}{k!} \mathrm{e}^{-\mu_0} \mu_0^k, & \mu_0 > 0, \; k = 0, 1, \ldots \\ 0, & \text{otherwise}. \end{cases} \tag{31.3.2}$$

The fellowing sections review properties of Poisson random variables are useful for statistical image reconstruction algorithms.

s,prob,poisson,moment

31.3.2.1 Moments

The most important moments of $N \sim \mathsf{Poisson}\{\mu_0\}$ are $\mathsf{E}[N] = \mu_0$, $\mathsf{Var}\{N\} = \mu_0$, $\mathsf{E}\left[N^4\right] = \mu_0 + 7\mu_0^2 + 6\mu_0^3 + \mu_0^4$, $\mathsf{Var}\left\{N^2\right\} = \mu_0 + 6\mu_0^2 + 4\mu_0^3$. [wiki]

s,prob,poisson,thin

31.3.2.2 Bernoulli thinning

Suppose a source transmits N photons of a certain energy along a ray passing through an object towards a specified pixel on the detector. We assume N is a **Poisson** random variable with mean μ_0 as in (31.3.2). Each of the N transmitted photons may either pass unaffected ("survive" passage) or may interact with the object. These are Bernoulli trials since the photons interact independently. From **Beer's law** [3] we know that the probability of surviving passage is given by

$$\mathsf{p} = \mathrm{e}^{-\int \mu(z)\,\mathrm{d}z}.$$

The number of photons M that pass unaffected through the object is a random variable, and this process is called **Bernoulli thinning** or **Bernoulli selection**. From Beer's law [3]:

e,prob,poisson,M|N

$$\mathsf{P}\{M = m \,|\, N = n\} = \binom{n}{m} \mathsf{p}^m (1-\mathsf{p})^{n-m}, \qquad m = 0, \ldots, n. \tag{31.3.3}$$

Using total probability:

$$\begin{aligned}
\mathsf{P}\{M = m\} &= \sum_{n=0}^{\infty} \mathsf{P}\{M = m \,|\, N = n\}\, \mathsf{P}\{N = n\} \\
&= \sum_{n=m}^{\infty} \binom{n}{m} \mathsf{p}^m (1-\mathsf{p})^{n-m} \frac{1}{n!} \mathrm{e}^{-\mu_0} \mu_0^n = \frac{1}{m!} \mathrm{e}^{-\mu_0 \mathsf{p}} (\mu_0 \mathsf{p})^m, \quad m = 0, 1, \ldots.
\end{aligned}$$

Therefore the distribution of photons that survive passage is also Poisson, with mean $\mathsf{E}[M] = \mu_0 \mathsf{p}$.

31.3.2.3 Poisson conditionals

It also is useful to examine the reverse of (31.3.3). By applying Bayes' rule, for $0 \le m \le n$:

$$\begin{aligned}
\mathsf{P}\{N = n \,|\, M = m\} &= \frac{\mathsf{P}\{M = m \,|\, N = n\}\, \mathsf{P}\{N = n\}}{\mathsf{P}\{M = m\}} \\
&= \frac{\binom{n}{m} p^m (1-p)^{n-m} \ \frac{1}{n!} \mathrm{e}^{-\mu_0} \mu_0^n}{\frac{1}{m!} \mathrm{e}^{-\mu_0 p} (\mu_0 p)^m} \\
&= \frac{1}{(n-m)!} (\mu_0 - \mu_0 p)^{n-m} \mathrm{e}^{-(\mu_0 - \mu_0 p)} \\
&= \frac{1}{(n-m)!} (\mathsf{E}[N] - \mathsf{E}[M])^{n-m} \mathrm{e}^{-(\mathsf{E}[N] - \mathsf{E}[M])}.
\end{aligned}$$

Thus, conditioned on M, the random variable $N - M$ has a **Poisson** distribution with mean $\mathsf{E}[N] - \mathsf{E}[M]$. In particular,

$$\mathsf{E}[N - M \,|\, M] = \mathsf{E}[N] - \mathsf{E}[M],$$

which is useful in deriving the transmission EM algorithm [4, 5].

31.3.2.4 Sums of Poisson random variables

Suppose $Y = \sum_{k=1}^K X_k$ where $X_k \sim \mathsf{Poisson}\{\mu_k\}$. Then $Y \sim \mathsf{Poisson}\left\{\sum_{k=1}^K \mu_k\right\}$. Interestingly, there is a converse for this result called **Raikov's theorem** [6]. If two independent nonnegative random variables have a sum that has a Poisson distribution, then the two random variables also have a Poisson distribution.

31.3.2.5 Poisson conditioning and multinomials

Suppose $Y = \sum_{k=1}^K X_k$ where $X_k \sim \mathsf{Poisson}\{\mu_k\}$ with each $\mu_k \ge 0$. Then, conditioned on the event $[Y = y]$, the collection of random variables $(X_1, \ldots, X_K)$ has a multinomial distribution with

e,prob,poisson,multinomial

$$\mathsf{E}[X_j \,|\, Y = y] = \begin{cases} \frac{\mu_j}{\sum_k \mu_k} y, & \sum_k \mu_k > 0,\ y \ge 0 \\ 0, & \sum_k \mu_k = 0,\ y = 0 \\ \text{undefined}, & \sum_k \mu_k = 0,\ y > 0. \end{cases} \tag{31.3.4}$$

Proof:
Using Bayes rule and the Poisson distributions of the independent X_k values:

$$\begin{aligned}
&\mathsf{P}\{X_1 = x_1, \ldots, X_K = x_K \,|\, Y = y\} \\
&\quad = \frac{\mathsf{P}\{Y = y \,|\, X_1 = x_1, \ldots, X_K = x_K\}\, \mathsf{P}\{X_1 = x_1, \ldots, X_K = x_K\}}{\mathsf{P}\{Y = y\}} \\
&\quad = \mathbb{I}_{\{y = x_1 + \cdots + x_K\}} \frac{\prod_{k=1}^K \mu_k^{x_k}\, \mathrm{e}^{-\mu_k} / x_k!}{[\sum_k \mu_k]^y\, \mathrm{e}^{-\sum_k \mu_k} / y!} \\
&\quad = \mathbb{I}_{\{y = x_1 + \cdots + x_K\}} \binom{y}{x_1 \;\cdots\; x_K} \prod_{j=1}^K \left(\frac{\mu_j}{\sum_k \mu_k} \right)^{x_k},
\end{aligned}$$

which is a multinomial distribution with means given by (31.3.4). □

One can also show that

$$\mathsf{E}\left[\prod_j t_j^{X_j} \,|\, Y = y \right] = \left(\frac{\sum_k t_k \mu_k}{\sum_k \mu_k} \right)^y, \tag{31.3.5}$$

e,prob,poisson,multi,ratio

a fact that is somewhat useful in deriving an **α-EM algorithm** for Poisson data; see Problem 18.6 and Problem 31.1.

prob,poisson,sum,multi

31.3.2.6 Poisson sums of multinomials

Let X_l be independently and identically distributed multinomial random variables with parameters $(p_1, \ldots, p_J; 1)$, *i.e.*, $\mathsf{P}\{X_l = j\} = \mathsf{p}_j,\ j = 1, \ldots, J$. Let $N \sim \mathsf{Poisson}\{\mu_0\}$ and define $Y_j = \sum_{l=1}^N \mathbb{I}_{\{X_l = j\}}$. Then the $\{Y_j\}$ values have independent Poisson distributions with means $\mathsf{E}[Y_j] = \mu_0\, \mathsf{p}_j$.
Proof:
By construction, conditioned in $[N = n]$ the Y_j values have a multinomial distribution with parameters $(p_1, \ldots, p_J; n)$. Thus by total probability:

$$\mathsf{P}\{Y_1 = y_1, \ldots, Y_J = y_J\} = \sum_{n=0}^\infty \mathsf{P}\{Y_1 = y_1, \ldots, Y_J = y_J \,|\, N = n\}\, \mathsf{P}\{N = n\}.$$

The only nonzero term in this sum is when $n = \sum_{j=1}^J y_j$. Thus

$$\mathsf{P}\{Y_1 = y_1, \ldots, Y_J = y_J\} = \frac{n!}{\prod_{j=1}^J y_j!} \prod_{j=1}^J \mathsf{p}_j^{y_j}\, \frac{1}{n!}\, \mathrm{e}^{-\mu_0}\, \mu_0^n = \prod_{j=1}^J \frac{(\mu_0\, \mathsf{p}_j)^{y_j}}{y_j!}\, \mathrm{e}^{-\mu_0\, \mathsf{p}_j}. \tag{31.3.6}$$

e,prob,poison,sum,multi

31.3.2.7 Moments of random sums of IID random variables

Let $X_1, X_2, \ldots$ denote IID random variables with mean μ_X and variance σ_X^2. Often in imaging problems we need the moments of a sum of a random number N of these:

$$Y = \sum_{n=1}^N X_n.$$

Often N has a Poisson distribution, but the moments of Y can be found more generally. Using **iterated expectation**, the mean of Y is:

$$\mathsf{E}[Y] = \mathsf{E}[\mathsf{E}[Y \,|\, N]] = \mathsf{E}[N \mu_X] = \mathsf{E}[N]\, \mu_X. \tag{31.3.7}$$

e,prob,sumN,mean

Similarly, the second moment of Y is

$$\begin{aligned}
\mathsf{E}\big[Y^2\big] &= \mathsf{E}\big[\mathsf{E}\big[Y^2 \,|\, N\big]\big] = \mathsf{E}\left[\sum_{n=1}^N \sum_{m=1}^N \mathsf{E}[X_n X_m] \right] = \mathsf{E}\big[N\, \mathsf{E}\big[X_n^2\big] + (N^2 - N)\mu_X^2\big] \\
&= \mathsf{E}\big[N(\sigma_X^2 + \mu_X^2) + (N^2 - N)\mu_X^2\big] = \mathsf{E}\big[N\sigma_X^2 + N^2\mu_X^2\big] = \mathsf{E}[N]\, \sigma_X^2 + (\sigma_N^2 + \mathsf{E}^2[N])\mu_X^2.
\end{aligned}$$

Thus the **variance of a random sum** is

$$\mathsf{Var}\{Y\} = \mathsf{E}\big[Y^2\big] - \mathsf{E}^2[Y] = \mathsf{E}[N]\, \sigma_X^2 + (\sigma_N^2 + \mathsf{E}^2[N])\mu_X^2 - \mathsf{E}^2[N]\, \mu_X^2 = \mathsf{E}[N]\, \sigma_X^2 + \sigma_N^2 \mu_X^2. \tag{31.3.8}$$

e,prob,sumN,var

31.4 Continuous distributions

s,prob,gauss

31.4.1 Normal or gaussian (s,prob,gauss)

For a gaussian random vector $\boldsymbol{X}$, we write $\boldsymbol{X} \sim \mathsf{N}(\boldsymbol{\mu}, \boldsymbol{K})$ as a short hand for the joint normal pdf

e,prob,gauss,jpdf

$$\mathsf{p}(\boldsymbol{x}) = \frac{1}{\sqrt{\det\{2\pi \boldsymbol{K}\}}} \, \mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})' \boldsymbol{K}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \,. \tag{31.4.1}$$

The covariance matrix $\boldsymbol{K}$ is always symmetric positive-semidefinite and is usually positive definite.

We say $\boldsymbol{X} \sim \mathsf{N}(\boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{K}_{\boldsymbol{X}})$ and $\boldsymbol{Y} \sim \mathsf{N}(\boldsymbol{\mu}_{\boldsymbol{Y}}, \boldsymbol{K}_{\boldsymbol{Y}})$ are **jointly distributed** gaussian random vectors with covariance $\mathsf{Cov}\{\boldsymbol{X}, \boldsymbol{Y}\}$ if $\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} \sim \mathsf{N}(\boldsymbol{\mu}, \boldsymbol{K})$ where $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_{\boldsymbol{X}} \\ \boldsymbol{\mu}_{\boldsymbol{Y}} \end{bmatrix}$ and $\boldsymbol{K} = \begin{bmatrix} \boldsymbol{K}_{\boldsymbol{X}} & \mathsf{Cov}\{\boldsymbol{X}, \boldsymbol{Y}\} \\ \mathsf{Cov}\{\boldsymbol{Y}, \boldsymbol{X}\} & \boldsymbol{K}_{\boldsymbol{Y}} \end{bmatrix}$.

If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are jointly distributed gaussian random vectors, then

$$\mathsf{E}[\boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y}] = \mathsf{E}[\boldsymbol{X}] + \mathsf{Cov}\{\boldsymbol{X}, \boldsymbol{Y}\} \left[\mathsf{Cov}\{\boldsymbol{Y}\}\right]^{-1} (\boldsymbol{y} - \mathsf{E}[\boldsymbol{Y}])$$

and

$$\mathsf{Cov}\{\boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y}\} = \mathsf{Cov}\{\boldsymbol{X}\} - \mathsf{Cov}\{\boldsymbol{X}, \boldsymbol{Y}\} \left[\mathsf{Cov}\{\boldsymbol{Y}\}\right]^{-1} \mathsf{Cov}\{\boldsymbol{Y}, \boldsymbol{X}\} \,.$$

If two jointly distributed gaussian random variables are uncorrelated, then they are independent.

Two random variables can each have gaussian distributions yet still not be *jointly* gaussian distributed; see Problem 3.37.

The following property of gaussian random vectors is useful for deriving **Stein's unbiased risk estimate** (**SURE**). If $\boldsymbol{Z} \sim \mathsf{N}\big(\boldsymbol{x}, \sigma^2 \boldsymbol{I}\big)$ and if $\boldsymbol{h}(\boldsymbol{z})$ is a n_{p}-dimensional vector function for which $\mathsf{E}\left[\left|\frac{\partial}{\partial z_j} h_j(\boldsymbol{z})\right|\right] < \infty$ for $j = 1, \ldots, n_{\mathrm{p}}$, then [7]:

e,prob,gauss,sure

$$\mathsf{E}\left[\sum_{j=1}^{n_{\mathrm{p}}} h_j(\boldsymbol{z}) x_j\right] = \mathsf{E}\left[\sum_{j=1}^{n_{\mathrm{p}}} h_j(\boldsymbol{z}) z_j\right] - \sigma^2 \, \mathsf{E}[\mathrm{div}\{\boldsymbol{h}(\boldsymbol{z})\}], \tag{31.4.2}$$

where the divergence is defined as $\mathrm{div}\{\boldsymbol{h}(\boldsymbol{z})\} \triangleq \sum_{j=1}^{n_{\mathrm{p}}} \frac{\partial}{\partial z_j} h_j(\boldsymbol{z})$.

s,prob,xform

31.4.2 Transformations (s,prob,xform)

The following theorem generalizes the usual such formulas found in engineering probability texts for **transformations of random vectors**.

t,transform

Theorem 31.4.1 *(See [8, 9] for proofs.)*
*Let $g : \mathbb{R}^n \to \mathbb{R}^n$ be one-to-one and assume that $h = g^{-1}$ is continuous. Assume that, on an open set $\mathcal{S} \subseteq \mathbb{R}^n$, h is continuously differentiable with **Jacobian determinant**[1] $|\det\{\nabla h(\boldsymbol{x})\}| \triangleq \left|\det\left\{\frac{\partial}{\partial x_j} h_i(\boldsymbol{x})\right\}\right|$. Suppose random vector $\boldsymbol{X}$ has pdf $\mathsf{p}(\boldsymbol{x})$ and*

e,p0

$$\mathsf{P}\{\boldsymbol{X} \in h(\mathcal{S}^c)\} = \int_{\mathcal{S}^c} \mathsf{p}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = 0, \tag{31.4.3}$$

where $\mathcal{S}^c$ denotes the set complement (in $\mathbb{R}^n$) of $\mathcal{S}$, and $h(\mathcal{A}) = \{h(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{A}\}$. Then the pdf of the random vector $\boldsymbol{Y} = g(\boldsymbol{X})$ is given by

e,prob,transform

$$\mathsf{p}(\boldsymbol{y}) = |\det\{\nabla h(\boldsymbol{y})\}| \, \mathsf{p}(h(\boldsymbol{y})), \quad \boldsymbol{y} \in \mathcal{S}, \tag{31.4.4}$$

and is zero otherwise.

s,prob,cov

31.5 Covariances (s,prob,cov)

The covariance of a random vector $\boldsymbol{X}$ is defined by:

$$\mathsf{Cov}\{\boldsymbol{X}\} \triangleq \mathsf{E}[(\boldsymbol{X} - \mathsf{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathsf{E}[\boldsymbol{X}])'] \,.$$

The cross covariance of a random vector $\boldsymbol{X}$ and a random vector $\boldsymbol{Y}$ is defined by:

$$\mathsf{Cov}\{\boldsymbol{X}, \boldsymbol{Y}\} \triangleq \mathsf{E}[(\boldsymbol{X} - \mathsf{E}[\boldsymbol{X}])(\boldsymbol{Y} - \mathsf{E}[\boldsymbol{Y}])'] \,.$$

Properties of covariance

[1] We use $|\det\{\boldsymbol{A}\}|$ to denote the absolute value of the determinant of a matrix $\boldsymbol{A}$.

- Covariances are positive-semidefinite matrices: $\mathsf{Cov}\{\boldsymbol{X}\} \succeq \boldsymbol{0}$
- $\mathsf{Cov}\{\boldsymbol{X},\boldsymbol{Y}\} = \mathsf{E}[\boldsymbol{X}\boldsymbol{Y}'] - \mathsf{E}[\boldsymbol{X}]\,\mathsf{E}[\boldsymbol{Y}']$
- $\mathsf{Cov}\{\boldsymbol{X},\boldsymbol{Y}\} = \mathsf{Cov}\{\boldsymbol{Y},\boldsymbol{X}\}'$
- $\mathsf{Cov}\{\boldsymbol{X},\boldsymbol{X}\} = \mathsf{Cov}\{\boldsymbol{X}\}$
- $\mathsf{Cov}\{a\boldsymbol{X}+b, c\boldsymbol{Y}+d\} = ac^*\,\mathsf{Cov}\{\boldsymbol{X},\boldsymbol{Y}\}$
- $\mathsf{Cov}\{\boldsymbol{A}\boldsymbol{X}, \boldsymbol{B}\boldsymbol{Y}+d\} = \boldsymbol{A}\,\mathsf{Cov}\{\boldsymbol{X},\boldsymbol{Y}\}\,\boldsymbol{B}'$
- In particular, $\mathsf{Var}\{X_i\} = \boldsymbol{e}_i'\,\mathsf{Cov}\{\boldsymbol{X}\}\,\boldsymbol{e}_i$, where $\boldsymbol{e}_i$ denotes the ith unit vector
- If non-random vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ have the same size as $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively, then a version of the **Schwarz Inequality** is:

$$|\boldsymbol{u}'\,\mathsf{Cov}\{\boldsymbol{X},\boldsymbol{Y}\}\,\boldsymbol{v}| \le \sqrt{\mathsf{Var}\{\boldsymbol{u}'\boldsymbol{X}\}\,\mathsf{Var}\{\boldsymbol{v}'\boldsymbol{Y}\}} = \sqrt{\boldsymbol{u}'\,\mathsf{Cov}\{\boldsymbol{X}\}\,\boldsymbol{u}}\sqrt{\boldsymbol{v}'\,\mathsf{Cov}\{\boldsymbol{Y}\}\,\boldsymbol{v}}$$

- In particular, if X and Y are scalar random variables, then the **correlation coefficient** $\rho_{X,Y} \triangleq \frac{\mathsf{Cov}\{X,Y\}}{\sqrt{\mathsf{Var}\{X\}\,\mathsf{Var}\{Y\}}}$ is bounded by unity: $|\rho_{\boldsymbol{X},\boldsymbol{Y}}| \le 1$.
- If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent r.v.s, then $E[\boldsymbol{X}\boldsymbol{Y}] = \mathsf{E}[\boldsymbol{X}]\,\mathsf{E}[\boldsymbol{Y}]]$ so $\mathsf{Cov}\{\boldsymbol{X},\boldsymbol{Y}\} = \boldsymbol{0}$.
- The reverse is *not true* in general (uncorrelated does not ensure independence); an exception is gaussian.
- $\mathsf{Cov}\left\{\sum_i \boldsymbol{X}_i, \sum_j \boldsymbol{Y}_j\right\} = \sum_i \sum_j \mathsf{Cov}\{\boldsymbol{X}_i, \boldsymbol{Y}_j\}$
- If $\mathsf{E}[\boldsymbol{X}] = \boldsymbol{\mu}$, then using (28.1.7):

$$\begin{aligned}\mathsf{E}[\boldsymbol{X}'\boldsymbol{B}\boldsymbol{X}] &= \mathsf{E}[\mathsf{trace}\{\boldsymbol{X}'\boldsymbol{B}\boldsymbol{X}\}] = \mathsf{E}[\mathsf{trace}\{\boldsymbol{B}\boldsymbol{X}\boldsymbol{X}'\}] = \mathsf{trace}\{\boldsymbol{B}\,\mathsf{E}[\boldsymbol{X}\boldsymbol{X}']\} \\ &= \boldsymbol{\mu}'\boldsymbol{B}\boldsymbol{\mu} + \mathsf{trace}\{\boldsymbol{B}\,\mathsf{Cov}\{\boldsymbol{X}\}\}\,. \end{aligned} \tag{31.5.1}$$

e,prob,cov,xBx

s,prob,stderr

31.6 Standard errors (s,prob,stderr)

We often estimate the mean, variance, or standard deviation from a sample of n elements and present the estimates with standard errors or error bars (in plots) as well. A standard error of a statistic (or estimator) is the (estimated) standard deviation of the statistic. An error bar is, in a plot, a line which is centered at the estimate with length that is double the standard error. Standard errors mean the statistical fluctuation of estimators, and they are important particularly when one compares two estimates (for example, whether one quantity is higher than the other in a statistically meaningful way). Here we review the standard errors of frequently used estimators of the mean, variance, and standard deviation [10].

31.6.1 Normal one sample problem

Let $X_1, \ldots, X_n$ denote a random sample from $\mathsf{N}(\mu, \sigma^2)$ where both μ and σ are unknown parameters. Define the following two statistics (sample mean and sample variance):

$$\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \;\text{ and }\; S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2.$$

31.6.2 Mean estimator

The uniformly minimum variance unbiased (UMVU) estimator of μ is $\bar{X}$ [11, p. 92]. Because $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, the standard error of $\bar{X}$ is

$$\sigma_{\bar{X}} = \sqrt{\mathsf{Var}\{\bar{X}\}} = \frac{\sigma}{\sqrt{n}}.$$

Hence the natural estimate of $\sigma_{\bar{X}}$ is $\hat{\sigma}_{\bar{X}} = \hat{\sigma}/\sqrt{n}$ For $\hat{\sigma}$, see §31.6.4.

31.6.3 Variance estimator

From [11, p. 92], S^2 is UMVU for σ^2 and

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}. \tag{31.6.1}$$

e,chi

Because the chi-squared distribution with $n-1$ degrees of freedom $\left(\chi^2_{n-1}\right)$ has variance $2(n-1)$ [11, p. 31], the standard error of S^2 is

$$\sigma_{S^2} = \sqrt{\mathsf{Var}\{S^2\}} = \sigma^2 \sqrt{\frac{2}{n-1}}.$$

Hence the natural estimate is $\hat{\sigma}_{S^2} = S^2\sqrt{2/(n-1)}$. It is useful to note

$$\sigma_{S^2}/\sigma^2 = \hat{\sigma}_{S^2}/S^2 = \sqrt{\frac{2}{n-1}}.$$

Because σ^2 and S^2 have the square of the units of X_i, often it is preferable to report estimates of σ, as described next.

subsec,std

31.6.4 Standard deviation estimator

The UMVU estimator of σ is $K_n S$[11, p. 92] where

$$K_n = \sqrt{\frac{n-1}{2}} \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})} = \sqrt{\frac{n-1}{2}} e^{\ln \Gamma(\frac{n-1}{2}) - \ln \Gamma(\frac{n}{2})},$$

where the second form is more numerically stable for large values of n when using the "ln gamma function." By setting $K_n = 1$, S is a common choice in practice but it is slightly biased. Because

$$\frac{\sqrt{n-1}}{\sigma} S \sim \chi_{n-1}$$

[see (31.6.1)] and the chi distribution with $n-1$ degrees of freedom (χ_{n-1}) has variance [12, p. 49: typo corrected]

$$V_n = 2\left(\frac{n-1}{2} - \frac{\Gamma^2(\frac{n}{2})}{\Gamma^2(\frac{n-1}{2})}\right),$$

the standard error of $K_n S$ is

$$\sigma_{K_n S} = \sqrt{\mathsf{Var}\{K_n S\}} = \sigma K_n \sqrt{\frac{V_n}{n-1}}.$$

To investigate the asymptotic behavior of $\sigma_{K_n S}$, we need the following approximation [13, P. 602]:

e,gamma

$$\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} = \sqrt{\frac{n-1}{2}} \left(1 - \frac{1}{4(n-1)} + O\left(\frac{1}{n^2}\right)\right). \tag{31.6.2}$$

Using (31.6.2), it can be shown that

$$K_n = 1 + O\left(\frac{1}{n}\right)$$

and

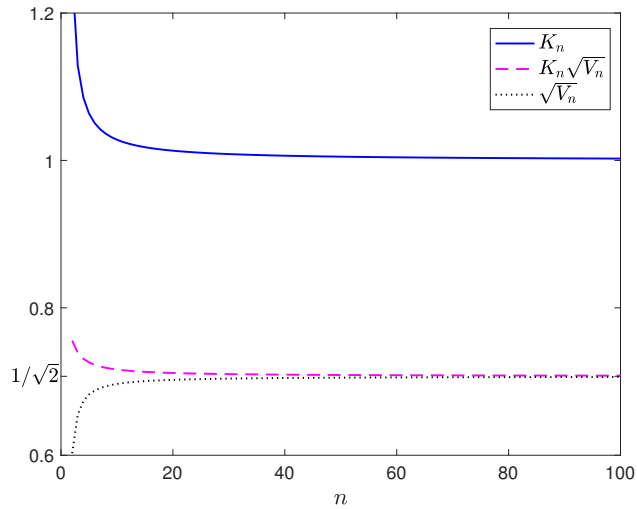
$$\begin{aligned} \sigma_{K_n S} &= \frac{\sigma}{\sqrt{2(n-1)}} \left(1 + O\left(\frac{1}{n}\right)\right) \\ &= \frac{\sigma}{\sqrt{2(n-1)}} + O\left(\frac{1}{n\sqrt{n}}\right). \end{aligned}$$

To summarize,

e,sigsig

$$\begin{aligned} \sigma_{K_n S}/\sigma = \hat{\sigma}_{K_n S}/(K_n S) &= \frac{K_n \sqrt{V_n}}{\sqrt{n-1}} \\ &\approx \frac{1}{\sqrt{2(n-1)}} \quad \text{for large } n. \end{aligned} \tag{31.6.3}$$

Fig. 31.6.1 shows a plot of K_n, $\sqrt{V_n}$, and $K_n\sqrt{V_n}$ versus n. For $n > 10$, it seems reasonable to use $K_n = 1$ and the approximation (31.6.3) for the standard error.



fig_stderr_gamma

Figure 31.6.1: This plot shows that K_n and $\sqrt{V_n}$ approach 1 and $1/\sqrt{2}$, respectively, as n increases.

31.7 Fisher information and Cramér-Rao bound (s,prob,fish)

s,prob,fish

For a statistical model with measurements $\boldsymbol{y}$ distributed according to a distribution $\mathsf{p}(\boldsymbol{y};\boldsymbol{x})$ that depends on a parameter vector $\boldsymbol{x}$, an important quantity is the **Fisher information matrix** defined by

e,prob,fish,1

$$\mathbf{F}(\boldsymbol{x}) \triangleq \mathsf{E}[\nabla \mathsf{L}(\boldsymbol{x}) \nabla \mathsf{L}(\boldsymbol{x})], \tag{31.7.1}$$

where $\mathsf{L}(\boldsymbol{x}) \triangleq \log \mathsf{p}(\boldsymbol{y};\boldsymbol{x})$ denotes the **log likelihood** associated with the measurement statistics. Under suitable regularity conditions, an equivalent expression is

e,prob,fish,2

$$\mathbf{F}(\boldsymbol{x}) = -\mathsf{E}\left[\nabla^2 \mathsf{L}(\boldsymbol{x})\right]. \tag{31.7.2}$$

x,prob,fish,gauss

Example 31.7.1 *For the linear gaussian model* $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\varepsilon}$ *where* $\boldsymbol{\varepsilon} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{\Pi})$, *the log-likelihood is* $\mathsf{L}(\boldsymbol{x}) \equiv -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})'\boldsymbol{\Pi}^{-1}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})$ *and* $\nabla \mathsf{L}(\boldsymbol{x}) = \boldsymbol{A}'\boldsymbol{\Pi}^{-1}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}'\boldsymbol{\Pi}^{-1}\boldsymbol{\varepsilon}$ *so the Fisher information is*

$$\mathbf{F} = \mathsf{E}\left[\boldsymbol{A}'\boldsymbol{\Pi}^{-1}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{\Pi}^{-1}\boldsymbol{A}\right] = \boldsymbol{A}'\boldsymbol{\Pi}^{-1}\,\mathsf{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']\,\boldsymbol{\Pi}^{-1}\boldsymbol{A} = \boldsymbol{A}'\boldsymbol{\Pi}^{-1}\boldsymbol{A}.$$

More generally, for the nonlinear *gaussian model* $\boldsymbol{y} = \boldsymbol{\mu}(\boldsymbol{x}) + \boldsymbol{\varepsilon}$ *where* $\boldsymbol{\varepsilon} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{\Pi})$, *the log-likelihood is* $\mathsf{L}(\boldsymbol{x}) \equiv -\frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{x})\right)'\boldsymbol{\Pi}^{-1}\left(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{x})\right)$ *and* $\nabla \mathsf{L}(\boldsymbol{x}) = \left(\nabla\boldsymbol{\mu}(\boldsymbol{x})\right)\boldsymbol{\Pi}^{-1}\left(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{x})\right) = \left(\nabla\boldsymbol{\mu}(\boldsymbol{x})\right)\boldsymbol{\Pi}^{-1}\boldsymbol{\varepsilon}$, *where* $\nabla\boldsymbol{\mu}(\boldsymbol{x})$ *denotes the* $n_{\mathrm{p}} \times n_{\mathrm{d}}$ *gradient of* $\boldsymbol{\mu} : \mathbb{R}^{n_{\mathrm{p}}} \to \mathbb{R}^{n_{\mathrm{d}}}$. *So the Fisher information is*

e,prob,fish,nonlin,gauss

$$\mathbf{F} = \mathsf{E}\left[\nabla\boldsymbol{\mu}(\boldsymbol{x})\boldsymbol{\Pi}^{-1}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{\Pi}^{-1}\nabla\boldsymbol{\mu}(\boldsymbol{x})\right] = \nabla\boldsymbol{\mu}(\boldsymbol{x})\boldsymbol{\Pi}^{-1}\,\mathsf{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']\,\boldsymbol{\Pi}^{-1}\nabla\boldsymbol{\mu}(\boldsymbol{x}) = \nabla\boldsymbol{\mu}(\boldsymbol{x})\boldsymbol{\Pi}^{-1}\nabla\boldsymbol{\mu}(\boldsymbol{x}). \tag{31.7.3}$$

One reason the Fisher information matrix is important is that under certain regularity conditions, **maximum likelihood estimators** (**MLE**) $\hat{\boldsymbol{x}}$ asymptotically have gaussian distributions with covariance $\mathbf{F}^{-1}$. Furthermore, the **Cramér-Rao lower bound** (**CRB**) [wiki] states that if $\hat{\boldsymbol{x}}$ is an unbiased estimator for $\boldsymbol{x}$, then the covariance of $\hat{\boldsymbol{x}}$ satisfies

e,prob,fish,crb

$$\mathsf{Cov}\{\hat{\boldsymbol{x}}\} \succ \mathbf{F}^{-1}. \tag{31.7.4}$$

See §25.18 for a concise derivation. The only estimator that can achieve the lower bound is the MLE. Even for finite sample sizes, the CRB is often a good approximation to the covariance of a ML estimate:

$$\mathsf{Cov}\{\hat{\boldsymbol{x}}\} \approx \mathbf{F}^{-1}.$$

31.8 Gauss-Markov theorem (s,prob,gauss-markov)

s,prob,gauss-markov

The **Gauss-Markov** theorem is a particularly important result from statistical estimation theory that serves as a guide for designing estimators including image reconstruction methods.

Consider the linear model $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\varepsilon}$ where we assume that the noise is zero mean: $\mathsf{E}[\boldsymbol{\varepsilon}] = \boldsymbol{0}$ and has known covariance: $\mathsf{Cov}\{\boldsymbol{\varepsilon}\} = \boldsymbol{K}$. Note that no other assumptions about the distribution of $\boldsymbol{\varepsilon}$ are made. Now suppose that our goal is to find an **unbiased** linear estimator of $\boldsymbol{x}$ that has the "smallest possible" variability. This is known as the **best linear unbiased estimator** (**BLUE**). Specifically, for a linear estimator of the form

$$\hat{\boldsymbol{x}} = \boldsymbol{B}\boldsymbol{y}$$

for some matrix $\boldsymbol{B}$, we want to minimize $\mathsf{trace}\{\mathsf{Cov}\{\hat{\boldsymbol{x}}\}\}$ subject to the constraint that $\mathsf{E}[\hat{\boldsymbol{x}}] = \boldsymbol{x}$ for any $\boldsymbol{x}$, *i.e.*, that $\boldsymbol{B}\boldsymbol{A} = \boldsymbol{I}$. The solution is [14, 15]

$$\boldsymbol{B} = \left[\boldsymbol{A}'\boldsymbol{K}^{-1}\boldsymbol{A}\right]^{-1}\boldsymbol{A}'\boldsymbol{K}^{-1}.$$

In other words, the best $\hat{\boldsymbol{x}}$ solves the following **weighted least-squares** (**WLS**) problem:

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_{\boldsymbol{W}^{1/2}}^2 = \left[\boldsymbol{A}'\boldsymbol{W}\boldsymbol{A}\right]^{-1}\boldsymbol{A}'\boldsymbol{W}\boldsymbol{y},$$

where the weighting is the inverse of the noise covariance:

$$\boldsymbol{W} = \boldsymbol{K}^{-1}.$$

Therefore, when feasible, we should include weighting based on the inverse of the data covariance for least-squares data consistency terms. Generalizations of this argument to biased estimators are also available [16].

The covariance of this BLUE is

$$\mathsf{Cov}\{\hat{\boldsymbol{x}}\} = \boldsymbol{B}\,\mathsf{Cov}\{\boldsymbol{y}\}\,\boldsymbol{B}' = \left[\boldsymbol{A}'\boldsymbol{K}^{-1}\boldsymbol{A}\right]^{-1}.$$

s,prob,varpro

31.9 Variable-projection methods for nonlinear least-squares (s,prob,varpro)

For problems with gaussian noise, the ML estimate $\hat{\boldsymbol{\theta}}$ of a parameter vector $\boldsymbol{\theta}$ minimizes a WLS cost function:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{2} \|\boldsymbol{y} - \bar{\boldsymbol{y}}(\boldsymbol{\theta})\|^2_{\boldsymbol{W}^{1/2}},$$

where $\bar{\boldsymbol{y}}(\boldsymbol{\theta})$ denotes the model for the measurements. A typical optimization method used for this minimization problem is the **Levenberg-Marquardt** method [17, 18]. There are a variety of nonlinear least-squares estimation problems where the model is linear in some of the unknown parameters and nonlinear in the other parameters. Mathematically, the overall parameter vector has the form $\boldsymbol{\theta} = (\boldsymbol{x}, \boldsymbol{\alpha})$ where

$$\bar{\boldsymbol{y}}(\boldsymbol{\theta}) = \boldsymbol{A}(\boldsymbol{\alpha})\boldsymbol{x}.$$

In these cases, the WLS problem becomes

$$\arg\min_{\boldsymbol{x},\boldsymbol{\alpha}} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}(\boldsymbol{\alpha})\boldsymbol{x}\|^2_{\boldsymbol{W}^{1/2}}.$$

This **joint estimation** problem can be simplified by exploiting the partially linear structure of the model. The resulting approach is called the **variable projection** method [19].

For any given $\boldsymbol{\alpha}$, the minimizer over $\boldsymbol{x}$ is a linear WLS problem with solution

$$\hat{\boldsymbol{x}}(\boldsymbol{\alpha}) = \arg\min_{\boldsymbol{x}} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}(\boldsymbol{\alpha})\boldsymbol{x}\|^2_{\boldsymbol{W}^{1/2}} = \boldsymbol{B}^\dagger(\boldsymbol{\alpha})\boldsymbol{W}^{1/2}\boldsymbol{y},$$

where $\boldsymbol{B}(\boldsymbol{\alpha}) \triangleq \boldsymbol{W}^{1/2}\boldsymbol{A}(\boldsymbol{\alpha})$ and $\boldsymbol{B}^\dagger$ denotes the **pseudo-inverse** of $\boldsymbol{B}$. Usually such problems are **over determined**, *i.e.*, $\boldsymbol{B}$ is a "tall" matrix with full column rank, in which case

$$\boldsymbol{B}^\dagger = [\boldsymbol{B}'\boldsymbol{B}]^{-1}\boldsymbol{B}'.$$

Substituting $\hat{\boldsymbol{x}}(\boldsymbol{\alpha})$ back into the original cost function, the minimization over $\boldsymbol{\alpha}$ becomes

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}(\boldsymbol{\alpha})\,\hat{\boldsymbol{x}}(\boldsymbol{\alpha})\|^2_{\boldsymbol{W}^{1/2}} = \arg\min_{\boldsymbol{\alpha}} \frac{1}{2} \left\|\boldsymbol{y} - \boldsymbol{A}(\boldsymbol{\alpha})\boldsymbol{B}^\dagger(\boldsymbol{\alpha})\boldsymbol{W}^{1/2}\boldsymbol{y}\right\|^2_{\boldsymbol{W}^{1/2}}.$$

Now note that (after some simplification):

$$\begin{aligned}\left\|\boldsymbol{y} - \boldsymbol{A}(\boldsymbol{\alpha})\boldsymbol{B}^\dagger(\boldsymbol{\alpha})\boldsymbol{W}^{1/2}\boldsymbol{y}\right\|^2_{\boldsymbol{W}^{1/2}} &= \boldsymbol{y}'\left(\boldsymbol{I} - \boldsymbol{A}(\boldsymbol{\alpha})\boldsymbol{B}^\dagger(\boldsymbol{\alpha})\boldsymbol{W}^{1/2}\right)'\boldsymbol{W}\left(\boldsymbol{I} - \boldsymbol{A}(\boldsymbol{\alpha})\boldsymbol{B}^\dagger(\boldsymbol{\alpha})\boldsymbol{W}^{1/2}\right)\boldsymbol{W}^{1/2}\boldsymbol{y} \\ &= \boldsymbol{y}'\boldsymbol{W}\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{W}^{1/2}\boldsymbol{B}(\boldsymbol{\alpha})\boldsymbol{B}^\dagger(\boldsymbol{\alpha})\boldsymbol{W}^{1/2}\boldsymbol{y}.\end{aligned}$$

The first term is independent of $\boldsymbol{\alpha}$, so the nonlinear optimization problem simplifies to

$$\begin{aligned}\hat{\boldsymbol{\alpha}} &= \arg\max_{\boldsymbol{\alpha}} \boldsymbol{y}'\boldsymbol{W}^{1/2}\boldsymbol{B}(\boldsymbol{\alpha})\boldsymbol{B}^\dagger(\boldsymbol{\alpha})\boldsymbol{W}^{1/2}\boldsymbol{y} \\ &= \arg\max_{\boldsymbol{\alpha}} \boldsymbol{y}'\boldsymbol{W}\boldsymbol{A}(\boldsymbol{\alpha})\left[\boldsymbol{A}'(\boldsymbol{\alpha})\boldsymbol{W}\boldsymbol{A}(\boldsymbol{\alpha})\right]^{-1}\boldsymbol{A}'(\boldsymbol{\alpha})\boldsymbol{W}\boldsymbol{y}.\end{aligned}$$

Because $\dim(\boldsymbol{\alpha}) < \dim(\boldsymbol{\theta})$, less computation is needed to find $\hat{\boldsymbol{\alpha}}$ using this approach. After one finds $\hat{\boldsymbol{\alpha}}$, the final estimate for $\boldsymbol{x}$ is $\hat{\boldsymbol{x}}(\hat{\boldsymbol{\alpha}})$. In the special case where the matrix $\boldsymbol{B}(\boldsymbol{\alpha})$ has orthonormal columns *for every possible parameter* $\boldsymbol{\alpha}$, then $\boldsymbol{B}^\dagger = \boldsymbol{B}'$ and the computation further simplifies to

e,proj,varpro,orth

$$\arg\max_{\boldsymbol{\alpha}} \|\boldsymbol{A}'(\boldsymbol{\alpha})\boldsymbol{W}\boldsymbol{y}\|^2. \tag{31.9.1}$$

The variable projection method has been generalized beyond least squares problems [20]. To avoid expensive exhaustive search over $\boldsymbol{\alpha}$, one can use **cover trees** methods [21].

x,prob,varpro,cos

Example 31.9.1 *Consider the problem of finding the phase of a sinusoid from equally spaced samples:* $\bar{y}_m(\boldsymbol{\theta}) = \sqrt{\frac{1}{M}}\, x_1 + \sqrt{\frac{2}{M}}\, x_2 \cos(2\pi m/M + \alpha), \quad m = 0, \ldots, M-1$, *where* $\boldsymbol{\theta} = (x_1, x_2, \alpha)$. *In this case, if* $\boldsymbol{W} = \boldsymbol{I}$, *then* $\boldsymbol{B} = [\boldsymbol{u} \;\; \boldsymbol{c}(\alpha)]$ *where* $\boldsymbol{u} = \sqrt{\frac{1}{M}}\mathbf{1}$ *and* $c_m(\alpha) = \sqrt{\frac{2}{M}} \cos(2\pi m/M + \alpha)$. *One can verify that this matrix* $\boldsymbol{B}(\alpha)$ *is orthonormal for any value of* α. *Thus, from (31.9.1) the LS estimate of* α *is given by by the maximizer of* $\|\boldsymbol{B}'(\alpha)\boldsymbol{y}\|^2 = |\boldsymbol{u}'\,\boldsymbol{y}|^2 + |\boldsymbol{c}'(\alpha)\boldsymbol{y}|^2$. *The first term is independent of* α, *so* $\hat{\alpha}$ *is the maximizer of* $|\boldsymbol{c}'(\alpha)\boldsymbol{y}| = \left|\sum_{m=0}^{M-1} c_m^*(\alpha) y_m\right| = \left|\sum_{m=0}^{M-1} \cos(2\pi m/M + \alpha)\, y_m\right| = \left|\mathrm{real}\left\{\sum_{m=0}^{M-1} \mathrm{e}^{-\imath 2\pi m/M}\, \mathrm{e}^{-\imath\alpha}\, y_m\right\}\right| = |\mathrm{real}\{\mathrm{e}^{-\imath\alpha}\, Y_1\}| = ||Y_1| \cos(\angle Y_1 - \alpha)|$. *where the first DFT coefficient is* $Y_1 = \sum_{m=0}^{M-1} \mathrm{e}^{-\imath 2\pi m/M}\, y_m$ *and we have assumed that* $\boldsymbol{y}$ *is real. Thus the LS estimate is* $\hat{\alpha} = \angle Y_1 + k\pi$ *where* k *is any integer.*

s,prob,entropy

31.10 Entropy (s,prob,entropy)

The **entropy** of a random variable, as defined by Shannon [22, 23], is often used (and abused) in imaging problems. In particular, it is useful for multi-modality image registration problems. It has also been studied extensively as a regularization method for image recovery problems.

31.10.1 Shannon entropy

For a **discrete random variable** X having **probability mass function** (**PMF**) $\{\mathsf{p}_k,\ k = 1, \ldots, K\}$, the **entropy** of X is defined by[2]

e,prob,entropy,H

$$\mathsf{H}\{X\} = -\sum_{k=1}^{K} \mathsf{p}_k \log \mathsf{p}_k \,. \tag{31.10.1}$$

Note that like expectation $\mathsf{E}[X]$ and variance $\mathsf{Var}\{X\}$, entropy is defined in terms of the *distribution* of X and is thus a constant, not a random variable. This definition of entropy has several desirable properties. For example, one can show that

e,prob,entropy,bound

$$0 \le \mathsf{H} \le \log K. \tag{31.10.2}$$

The upper bound is achieved by the (discrete) **uniform distribution** where $\mathsf{p}_k = 1/K$. This is the "most random" distribution for a discrete random variable taking K distinct values. The lower bound is achieved when X takes only one value (with probability one), *e.g.*, when $\mathsf{p}_1 = 1,\ \mathsf{p}_2 = \cdots = \mathsf{p}_K = 0$. Such a random variable is not random at all.

In some image recovery problems, notably those based on **maximum entropy** methods, the 2D array of image values $f[m,n]$ is treated as if it represents a 2D probability mass function (after suitable normalization), and the "entropy" of such an image is defined as

e,prob,entropy,H,fmn=pmf

$$\mathsf{H} = -\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} \frac{|f[m,n]|}{s} \log\left(\frac{|f[m,n]|}{s}\right), \tag{31.10.3}$$

where $s \triangleq \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} |f[m,n]|$. Often the normalization factor s is ignored. Given an imaging model $g = \boldsymbol{A}f$, typically under-determined, one can try to find $f[m,n]$ to maximize H subject to $\|g = \boldsymbol{A}f\| \le \varepsilon$. Because the upper bound in (31.10.2) is achieved for uniform, rather than impulsive, distributions, one can expect that the maximum entropy image will be among the smoother images that are sufficiently consistent with the data.

Consider a $N \times M$ digital image $f[m,n]$ with a finite number gray levels $\{f_k\}$, *e.g.*, $\{0, 1, \ldots, 255\}$. Whether this image should be viewed as a deterministic function or as a random process is a question of modeling and philosophy. However, one can always define a random variable X by selecting one pixel at random and letting X be the value of the selected pixel. This will be a discrete-valued random variable having **probability mass function**

e,prob,entropy,pk,pmf

$$\mathsf{p}_k = \mathsf{P}\{X = f_k\} = \frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} \mathbb{I}_{\{f[m,n]=f_k\}}. \tag{31.10.4}$$

MIRT *See* `hist_bin_int.m`.

This PMF is simply the **histogram** of the values of the image $f[m,n]$, normalized by MN. Another definition of the **entropy** of the image $f[m,n]$ uses this PMF:

$$\mathsf{H}_f = \mathsf{H}\{X\} = -\sum_{k=1}^{K} \mathsf{p}_k \log \mathsf{p}_k \tag{31.10.5}$$

e,prob,entropy,H,fmn,hist

$$= -\sum_{k=1}^{K}\left(\frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} \mathbb{I}_{\{f[m,n]=f_k\}}\right) \log\left(\frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} \mathbb{I}_{\{f[m,n]=f_k\}}\right). \tag{31.10.6}$$

Note that this definition differs considerably from (31.10.3), as illustrated in the following example.

x,prob,entropy,uniform

Example 31.10.1 *Consider an $N \times M$ image $f[m,n]$ that is completely uniform: $f[m,n] = c$. In this case, (31.10.3) evaluates to its maximum $\mathsf{H} = \log(MN)$, whereas (31.10.6) evaluates to its minimum $\mathsf{H} = 0$.*

[2] In digital communications, often the logarithm base 2 is used, but this is just a scale factor that is unimportant for imaging problems.

Now suppose that the image $f[m,n]$ is continuous valued, *e.g.*, arbitrary real or complex numbers. (This situation arises in many image registration problems because even if the initial images are discrete valued, interpolation operations can lead to arbitrary gray scale values.) In this situation, the standard "histogram" definition of entropy given in (31.10.6) may not be useful. For example, if every image pixel has a different value, then $K = MN$ and (31.10.6) reduces to $\mathsf{H} = \log(MN)$.

For continuous-valued images, one might be inclined to seek a continuous analogue of (31.10.1). For a **continuous random variable** with **probability density function** $\mathsf{p}(x)$, its **differential entropy** is defined by

e,prob,entropy,h

$$\mathsf{h} = -\int \mathsf{p}(x) \log \mathsf{p}(x) \, \mathrm{d}x \,. \tag{31.10.7}$$

x,prob,entropy,gauss

Example 31.10.2 *If $X \sim \mathsf{N}(\mu, \sigma^2)$ then, using (31.4.1), its differential entropy is $\mathsf{h} = \frac{1}{2}\left(1 + \log(2\pi\sigma^2)\right)$, which increases monotonically with σ^2. More generally, if $\boldsymbol{X} \sim \mathsf{N}(\boldsymbol{\mu}, \boldsymbol{K}) \in \mathbb{R}^d$ then $\mathsf{h} = \frac{1}{2}\left(d + \log(\det\{2\pi\boldsymbol{K}\})\right)$.*

The definition (31.10.7) lacks some of the desirable properties of (31.10.1). For example it can take negative values, as illustrated in Example 31.10.2 when $\sigma^2 < 1/(2\pi e)$. Furthermore, the distribution $\mathsf{p}(x)$ is rarely known in practice for images so this expression seems to be of limited use for inverse problems. One could *estimate* $\mathsf{p}(x)$ from an image $f[m,n]$ using a **kernel density estimate**, also known as the **Parzen window method**:

$$\hat{\mathsf{p}}(x) = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} q(f[m,n] - x),$$

for some nonnegative function $q(\cdot)$ (called the **kernel** or **window**) that integrates to unity. Choosing the width of the kernel requires care [24–26]. However, one would still need to sample x to perform numerical calculations, so it seems more direct to work with discrete x values directly as described next.

A simple way to apply (31.10.1) to continuous-valued images would be to **quantize** those values into K bins. For example, if the image values lie in the range $[0, 100)$ then one could use K bins of the form $\mathcal{B}_k = [k\Delta, (k+1)\Delta)$ where in this case $\Delta = 100/K$, covering that interval. By analogy with (31.10.4), define the PMF of the quantized values as

e,prob,entropy,pk,Bk

$$\mathsf{p}_k = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \mathbb{I}_{\{f[m,n] \in \mathcal{B}_k\}} \tag{31.10.8}$$

and then use (31.10.1) to define entropy. (Even for discrete-valued images one can use (31.10.8) by neglecting the least-significant bits [27].

MIRT *See* `jf_histn.m`.

A limitation of this definition is that $\mathsf{H} \to \log(MN)$ as $\Delta \to 0$, if the image values are distinct. Therefore, choosing the bin size Δ is delicate. Note that if Δ is small, then $\mathsf{p}_k \approx \Delta\, \mathsf{p}(x_k)$ where $x_k = k\Delta$, so

$$\mathsf{H} = -\sum_k \mathsf{p}_k \log \mathsf{p}_k \approx -\sum_k \Delta\, \mathsf{p}(x_k) \log(\Delta\, \mathsf{p}(x_k)) \approx \mathsf{h} - \log \Delta.$$

Unfortunately, the simple quantization method (31.10.8) is not a continuous function of the image, which prevents the use of gradient-based methods for optimization.

Inspired by the interpolation method of [28], an alternative approach is to use a differentiable kernel $\psi(t)$ to "interpolate" the histogram as follows:

e,prob,entropy,pk,kernel

$$\mathsf{p}_k = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \psi(f[m,n]/\Delta - k), \quad k = k_{\min}, \ldots, k_{\max}, \tag{31.10.9}$$

where we assume $\psi(\cdot)$ is nonnegative and satisfies the **interpolation property**

$$\sum_{k=-\infty}^{\infty} \psi(t-k) = 1, \; \forall t \in \mathbb{R},$$

so that $\sum_{k=k_{\min}}^{k_{\max}} \mathsf{p}_k = 1$ (provided we choose $k_{\min}$ and $k_{\max}$ properly as discussed below). Choosing Δ and ψ is still important because typically $\mathsf{H} \to 0$ as $\Delta \to \infty$ and $\mathsf{H} \to \log(MN)$ as $\Delta \to 0$. It might appear that (31.10.9) requires $O(MNK)$ computation. But if we use a kernel ψ with finite support $(-W/2, W/2)$, then for each k we need only

evaluate $\psi(t-k)$ for values of t for which $t-k$ lies in the support interval, *i.e.*, $k - W/2 < t < k + W/2$. For a given data value, the only pertinent values of k are those integers for which $t - W/2 < k < t + W/2$, *i.e.*,

e,prob,entropy,kde,range

$$\lfloor t - W/2 \rfloor + 1 \le k \le \lceil t + W/2 \rceil - 1. \tag{31.10.10}$$

The number of such k values is $\lceil t + W/2 \rceil - \lfloor t - W/2 \rfloor - 1 \le \lceil W \rceil$ using (29.12.2). So this method requires $O(MN \lceil W \rceil)$ computation if implemented efficiently.

Based on (31.10.10), the smallest and largest relevant values of k are respectively $k_{\min} = \lfloor f_{\min}/\Delta - W/2 \rfloor + 1$ and $k_{\max} = \lfloor f_{\min}/\Delta + W/2 \rfloor - 1$.

If $\psi(\cdot)$ is differentiable, then p_k is a differentiable function of the image, facilitating gradient-based optimization. Substituting (31.10.9) into (31.10.1) and differentiating the resulting entropy definition yields

$$\begin{aligned}
\frac{\partial}{\partial f[m,n]} \mathsf{H} &= -\sum_k \left[\left(\frac{\partial}{\partial f} \mathsf{p}_k \right) \log \mathsf{p}_k + \frac{\mathsf{p}_k}{\mathsf{p}_k} \left(\frac{\partial}{\partial f} \mathsf{p}_k \right) \right] = -\sum_k \left(\frac{\partial}{\partial f} \mathsf{p}_k \right) \log \mathsf{p}_k - \frac{\partial}{\partial f} \sum_k \mathsf{p}_k \\
&= -\sum_k \left(\frac{\partial}{\partial f} \mathsf{p}_k \right) \log \mathsf{p}_k = \frac{1}{MN\Delta} \left(\sum_k \dot{\psi}(f[m,n]/\Delta - k) \log \mathsf{p}_k \right).
\end{aligned}$$

Again, by using an interpolator ψ with finite support, gradient computation is also $O(MN \lceil W \rceil)$. In particular, by using a quadratic B-spline for ψ, the derivative operation is akin to linear interpolation.

MIRT *See* `kde_pmf1.m` *and* `kde_pmf2.m`

The literature describes many methods for choosing the parameter Δ. One simple rule of thumb that has been advocated for a gaussian window $q(\cdot)$ is to choose

$$\Delta = 0.9 \min(\sigma, \mathrm{IQR}/1.34)/(MN)^{1/5},$$

where σ is the **sample standard deviation** of the image values and IQR is the **inter-quartile range** [29, p. 48]. For another kernel ψ, such as a quadratic B-spline, one can scale Δ according to the relative FWHM of ψ and a gaussian kernel.

MIRT *See* `kde_pmf_width`.

31.10.2 Joint entropy

For a pair of discrete random variables X, Y having joint probability mass function $\mathsf{p}_{kl} = \mathsf{P}\{X = x_k,\ Y = y_l\}$, for $k = 1, \ldots, K,\ l = 1, \ldots, L$, the **joint entropy** of X and Y is defined by

e,prob,entropy,joint

$$\mathsf{H}\{X,Y\} = -\sum_{k=1}^{K} \sum_{l=1}^{L} \mathsf{p}_{kl} \log \mathsf{p}_{kl}\,. \tag{31.10.11}$$

An important property of this definition is

e,prob,entropy,subadd

$$\max(\mathsf{H}\{X\}, \mathsf{H}\{Y\}) \le \mathsf{H}\{X,Y\} \le \mathsf{H}\{X\} + \mathsf{H}\{Y\}\,. \tag{31.10.12}$$

The latter inequality is called **subadditivity**. The upper bound is reached when X and Y are stastistically independent. Joint entropy has been studied for multi-modality image registration problems [27, 30–32] and for certain multi-modality regularization methods [33–35].

Given a pair of images $f[m,n]$ and $g[m,n]$ (possibly continuous valued), the joint histogram analog of (31.10.9) is

$$\mathsf{p}_{kl} = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \psi(f[m,n]/\Delta_f - k)\, \psi(g[m,n]/\Delta_g - l), \quad k = 1, \ldots, K, \quad l = 1, \ldots, L,$$

where now it may be appropriate to allow Δ_f and Δ_g to differ.

31.10.3 Mutual information

A related quantity for a pair of jointly distributed random variables is their **mutual information**:

e,prob,entropy,mi

$$\mathsf{MI}(X,Y) = \mathsf{H}\{X\} + \mathsf{H}\{Y\} - \mathsf{H}\{X,Y\}\,. \tag{31.10.13}$$

This quantity has been used for multi-modality image registration problems [28, 36, 37]. It also has been explored for multi-modality regularization methods [33, 38, 39]. Numerous variations have been proposed, *e.g.*, [40, 41], including **normalized mutual information**

e,prob,entropy,nmi

$$\mathsf{NMI}(X,Y) = \frac{\mathsf{H}\{X\} + \mathsf{H}\{Y\}}{\mathsf{H}\{X,Y\}} \tag{31.10.14}$$

and **entropy correlation coefficient**

e,prob,entropy,ecc

$$\mathsf{ECC}(X,Y) = \sqrt{2 - 2\frac{\mathsf{H}\{X,Y\}}{\mathsf{H}\{X\} + \mathsf{H}\{Y\}}}. \tag{31.10.15}$$

An important practical consideration in all such **similarity measures** for image registration is whether they exhibit **overlap invariance** [42, 43].

31.10.4 Cross entropy

Yet another related quantity that has been used in imaging problems is the **cross entropy** of two random variables [44, 45], defined by (*cf.* §18.4.3):

$$\mathsf{H}\{X\} + \mathsf{D}(\boldsymbol{p} \,\|\, \boldsymbol{q}) = \left(-\sum_k \mathsf{p}_k \log \mathsf{p}_k \right) + \left(\sum_k \mathsf{p}_k \log \frac{\mathsf{p}_k}{\mathsf{q}_k} - \mathsf{p}_k + \mathsf{q}_k \right) = -\sum_k \mathsf{p}_k \log \mathsf{q}_k,$$

where p_k denotes the PMF of X and q_k denotes the PMF of Y. This quantity has been used for multi-modality regularization methods [46–48] and for image reconstruction [46, 47, 49–53].

31.11 Problems (s,prob,prob)

s,prob,prob

ob,poisson,multi,ratio

Problem 31.1 *Prove (31.3.5).* *(Solve?)*

p,prob,mix,mmse

Problem 31.2 *Random variable $\boldsymbol{x}$ has a mixture distribution: $\mathsf{p}(\boldsymbol{x}) = \sum_k \pi_k \, \mathsf{p}_k(\boldsymbol{x})$ where $\pi_k \geq 0$ and $\sum_k \pi_k = 1$ and $\mathsf{p}_k(\boldsymbol{x})$ is the distribution of $\boldsymbol{x}$ for the kth class. Assume that the conditional distribution $\mathsf{p}(\boldsymbol{y} \,|\, \boldsymbol{x})$ of the measurements $\boldsymbol{y}$ is independent of the class of $\boldsymbol{x}$. Show that the **MMSE** estimate of $\boldsymbol{x}$ from $\boldsymbol{y}$ is*

$$\hat{\boldsymbol{x}} = \sum_k \pi_k \frac{\mathsf{p}(\boldsymbol{y} \,|\, \mathcal{A}_k)}{\mathsf{p}(y)} \, \mathsf{E}[\boldsymbol{x} \,|\, \boldsymbol{y}, \mathcal{A}_k],$$

where $\mathcal{A}_k$ denotes the event that $\boldsymbol{x}$ is drawn from the kth class, i.e., $\mathsf{p}_k(\boldsymbol{x}) = \mathsf{p}(\boldsymbol{x} \,|\, \mathcal{A}_k)$. *This estimator expression simplifies further when $\boldsymbol{x}$ is a **mixture of gaussians** and $\boldsymbol{y} \,|\, \boldsymbol{x} \sim \mathsf{N}(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{\Pi})$.*

31.12 Bibliography

leongarcia:94

[1] A. Leon-Garcia. *Probability and random processes for electrical engineering*. 2nd ed. New York: Addison-Wesley, 1994 (cit. on p. 31.1).

poisson:1837

[2] S. D. Poisson. *Recherches sur la Probabilité des Jugements en Matière Criminelle et en Matière Civile*. Paris: Bachelier, 1837. URL: http://books.google.com/books?id=uovoFE3gt2EC (cit. on p. 31.2).

beer:1852:bda

[3] A. Beer. "Bestimmung der absorption des rothen lichts in farbigen flüssigkeiten." In: *Annalen der Physik* 162.5 (1852), 78–88. DOI: 10.1002/andp.18521620505 (cit. on p. 31.3).

lange:84:era

[4] K. Lange and R. Carson. "EM reconstruction algorithms for emission and transmission tomography." In: *J. Comp. Assisted Tomo.* 8.2 (Apr. 1984), 306–16 (cit. on p. 31.3).

fessler:00:sir

[5] J. A. Fessler. "Statistical image reconstruction methods for transmission tomography." In: *Handbook of Medical Imaging, Volume 2. Medical Image Processing and Analysis*. Ed. by M. Sonka and J. Michael Fitzpatrick. Bellingham: SPIE, 2000, pp. 1–70. DOI: 10.1117/3.831079.ch1 (cit. on p. 31.3).

galambos:06:rt

[6] J. Galambos. *Raikov's theorem*. In Encyclopedia of Statistical Sciences, Wiley. 2006. DOI: 10.1002/0471667196.ess2160.pub2 (cit. on p. 31.3).

blu:07:tsl
[7] T. Blu and F. Luisier. "The SURE-LET approach to image denoising." In: *IEEE Trans. Im. Proc.* 16.11 (Nov. 2007), 2778–86. DOI: 10.1109/TIP.2007.906002 (cit. on p. 31.5).

bickel:77
[8] P. J. Bickel and K. A. Doksum. *Mathematical statistics*. Oakland, CA: Holden-Day, 1977 (cit. on p. 31.5).

fessler:98:oto
[9] J. A. Fessler. *On transformations of random vectors*. Tech. rep. 314. Available from http://web.eecs.umich.edu/~fessler. Univ. of Michigan, Ann Arbor, MI, 48109-2122: Comm. and Sign. Proc. Lab., Dept. of EECS, Aug. 1998. URL: http://web.eecs.umich.edu/~fessler/papers/files/tr/98,314,oto.pdf (cit. on p. 31.5).

ahn:03:seo
[10] S. Ahn and J. A. Fessler. *Standard errors of mean, variance, and standard deviation estimators*. Tech. rep. 413. Univ. of Michigan, Ann Arbor, MI, 48109-2122: Comm. and Sign. Proc. Lab., Dept. of EECS, July 2003. URL: http://web.eecs.umich.edu/~fessler/papers/files/tr/stderr.pdf (cit. on p. 31.6).

lehmann:98
[11] E. L. Lehmann and G. Casella. *Theory of point estimation*. New York: Springer-Verlag, 1998 (cit. on pp. 31.6, 31.7).

evans:93
[12] M. Evans, N. Hastings, and B. Peacock. *Statistical distributions*. New York: Wiley, 1993 (cit. on p. 31.7).

graham:94
[13] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete mathematics: a foundation for computer science*. Reading: Addison-Wesley, 1994 (cit. on p. 31.7).

aitken:1935:ols
[14] A. C. Aitken. "On least squares and linear combinations of observations." In: *Proceedings of the Royal Society of Edinburgh* 55 (1935), 42–8 (cit. on p. 31.9).

humpherys:12:afl
[15] J. Humpherys, P. Redd, and J. West. "A fresh look at the Kalman filter." In: *SIAM Review* 54.4 (2012), 801–23. DOI: 10.1137/100799666 (cit. on p. 31.9).

eldar:04:mvi
[16] Y. C. Eldar. "Minimum variance in biased estimation: bounds and asymptotically optimal estimators." In: *IEEE Trans. Sig. Proc.* 52.7 (July 2004), 1915–30. DOI: 10.1109/TSP.2004.828929 (cit. on p. 31.9).

levenberg:1944:amf
[17] K. Levenberg. "A method for the solution of certain non-linear problems in least squares." In: *Quart. Appl. Math.* 2.2 (July 1944), 164–8. URL: http://www.jstor.org/stable/43633451 (cit. on p. 31.10).

marquardt:63:aaf
[18] D. W. Marquardt. "An algorithm for least-squares estimation of nonlinear parameters." In: *J. Soc. Indust. Appl. Math.* 11.2 (June 1963), 431–41. URL: http://www.jstor.org/stable/2098941 (cit. on p. 31.10).

golub:03:snl
[19] G. Golub and V. Pereyra. "Separable nonlinear least squares: the variable projection method and its applications." In: *Inverse Prob.* 19.2 (Apr. 2003), R1–26. DOI: 10.1088/0266-5611/19/2/201 (cit. on p. 31.10).

shearer:13:ago
[20] P. Shearer and A. C. Gilbert. "A generalization of variable elimination for separable inverse problems beyond least squares." In: *Inverse Prob.* 29.4 (Apr. 2013), p. 045003. DOI: 10.1088/0266-5611/29/4/045003 (cit. on p. 31.10).

beygelzimer:06:ctf
[21] A. Beygelzimer, S. Kakade, and J. Langford. "Cover trees for nearest neighbor." In: *Proc. Intl. Conf. Mach. Learn.* 2006, 97–104. DOI: 10.1145/1143844.1143857 (cit. on p. 31.10).

shannon:1948:amt-1
[22] C. E. Shannon. "A mathematical theory of communication." In: *Bell Syst. Tech. J.* 27.3 (July 1948), 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x (cit. on p. 31.12).

shannon:53:cte
[23] C. Shannon. "Communication theory–Exposition of fundamentals." In: *IEEE Trans. Info. Theory* 1.1 (Feb. 1953), 44–7. DOI: 10.1109/TIT.1953.1188568 (cit. on p. 31.12).

sheather:91:ard
[24] S. J. Sheather and M. C. Jones. "A reliable data-based bandwidth selection method for kernel density estimation." In: *J. Royal Stat. Soc. Ser. B* 53.3 (1991), 683–90. URL: http://www.jstor.org/stable/2345597 (cit. on p. 31.13).

jones:96:abs
[25] M. C. Jones, J. S. Marron, and S. J. Sheather. "A brief survey of bandwidth selection for density estimation." In: *J. Am. Stat. Assoc.* 91.433 (Mar. 1996), 401–7. URL: http://www.jstor.org/stable/2291420 (cit. on p. 31.13).

shimazaki:09:kbo
[26] H. Shimazaki and S. Shinomoto. "Kernel bandwidth optimization in spike rate estimation." In: *J. Computational Neuroscience* (2009). DOI: 10.1007/s10827-009-0180-4 (cit. on p. 31.13).

collignon:95:3mm
[27] A. Collignon et al. "3D multi-modality medical image registration using feature space clustering." In: *Computer Vision, Virtual Reality, and Robotics in Medicine*. Vol. LNCS-905. 1995, 195–204. DOI: 10.1007/BFb0034948 (cit. on pp. 31.13, 31.14).

maes:97:mir
[28] F. Maes et al. "Multimodality image registration by maximization of mutual information." In: *IEEE Trans. Med. Imag.* 16.2 (Apr. 1997), 187–98. DOI: 10.1109/42.563664 (cit. on pp. 31.13, 31.15).

silverman:86
[29] B. W. Silverman. *Density estimation for statistics and data analysis*. New York: Chapman and Hall, 1986 (cit. on p. 31.14).

studholme:95:mvs
[30] C. Studholme, D. L. G. Hill, and D. J. Hawkes. "Multiresolution voxel similarity measures for MR-PET registration." In: *Information Processing in Medical Im.* 1995, 287–98 (cit. on p. 31.14).

studholme:97:atd
[31] C. Studholme, D. L. G. Hill, and D. J. Hawkes. "Automated three-dimensional registration of magnetic resonance and positron emission tomography brain images by multiresolution optimization of voxel similarity measures." In: *Med. Phys.* 24.1 (Jan. 1997), 25–35. DOI: 10.1118/1.598130 (cit. on p. 31.14).

pluim:03:mib
[32] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever. "Mutual-information-based registration of medical images: a survey." In: *IEEE Trans. Med. Imag.* 22.8 (Aug. 2003), 986–1004. DOI: 10.1109/TMI.2003.815867 (cit. on p. 31.14).

nuyts:07:tuo
[33] J. Nuyts. "The use of mutual information and joint entropy for anatomical priors in emission tomography." In: *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.* Vol. 6. 2007, 4194–54. DOI: 10.1109/NSSMIC.2007.4437034 (cit. on pp. 31.14, 31.15).

tang:08:bpi
[34] J. Tang, B. M. W. Tsui, and A. Rahmim. "Bayesian PET image reconstruction incorporating anato-functional joint entropy." In: *Proc. IEEE Intl. Symp. Biomed. Imag.* 2008, 1043–6. DOI: 10.1109/ISBI.2008.4541178 (cit. on p. 31.14).

tang:09:bpi
[35] J. Tang and A. Rahmim. "Bayesian PET image reconstruction incorporating anato-functional joint entropy." In: *Phys. Med. Biol.* 54.23 (Dec. 2009), 7063–76. DOI: 10.1088/0031-9155/54/23/002 (cit. on p. 31.14).

kim:96:mif
[36] B. Kim, J. L. Boes, and C. R. Meyer. "Mutual information for automated multimodal image warping." In: *NeuroImage* 3.1 (June 1996). Second Intl. Conf. on Functional Mapping of the Human Brain, S158. DOI: 10.1016/S1053-8119(96)80160-1 (cit. on p. 31.15).

wells:96:mmv
[37] W. M. Wells et al. "Multi-modal volume registration by maximization of mutual information." In: *Med. Im. Anal.* 1.1 (Mar. 1996), 35–51. DOI: 10.1016/S1361-8415(01)80004-9 (cit. on p. 31.15).

somayajula:05:pir
[38] S. Somayajula, E. Asma, and R. M. Leahy. "PET image reconstruction using anatomical information through mutual information based priors." In: *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.* 2005, 2722–26. DOI: 10.1109/NSSMIC.2005.1596899 (cit. on p. 31.15).

somayajula:07:pir
[39] S. Somayajula, A. Rangarajan, and R. M. Leahy. "PET image reconstruction using anatomical information through mutual information based priors: A scale space approach." In: *Proc. IEEE Intl. Symp. Biomed. Imag.* 2007, 165–8. DOI: 10.1109/ISBI.2007.356814 (cit. on p. 31.15).

studholme:99:aoi
[40] C. Studholme, D. L. G. Hill, and D. J. Hawkes. "An overlap invariant entropy measure of 3D medical image alignment." In: *Pattern Recognition* 32.1 (Jan. 1999), 71–86. DOI: 10.1016/S0031-3203(98)00091-0 (cit. on p. 31.15).

cahill:09:oio
[41] N. D. Cahill et al. "Overlap invariance of cumulative residual entropy measures for multimodal image alignment." In: *Proc. SPIE 7259 Medical Imaging: Image Proc.* 2009, p. 72590I. DOI: 10.1117/12.811585 (cit. on p. 31.15).

cahill:08:roi
[42] N. D. Cahill et al. "Revisiting overlap invariance in medical image alignment." In: *mmbia*. 2008, 1–8. DOI: 10.1109/CVPRW.2008.4562989 (cit. on p. 31.15).

cahill:10:afc
[43] N. Cahill, A. Noble, and D. Hawkes. "Accounting for changing overlap in variational image registration." In: *Proc. IEEE Intl. Symp. Biomed. Imag.* 2010, 0384–7. DOI: 10.1109/ISBI.2010.5490328 (cit. on p. 31.15).

shore:80:ado
[44] J. E. Shore and R. W. Johnson. "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy." In: *IEEE Trans. Info. Theory* 26.1 (Jan. 1980), 26–37. DOI: 10.1109/TIT.1980.1056144 (cit. on p. 31.15).

burr:89:ici
[45] R. L. Burr. "Iterative convex I-projection algorithms for maximum entropy and minimum cross-entropy computations." In: *IEEE Trans. Info. Theory* 35.3 (May 1989), 695–8. DOI: 10.1109/18.30998 (cit. on p. 31.15).

ardekani:96:mce
[46] B. A. Ardekani et al. "Minimum cross-entropy reconstruction of PET images using prior anatomical information." In: *Phys. Med. Biol.* 41.11 (Nov. 1996), 2497–517. DOI: 10.1088/0031-9155/41/11/018 (cit. on p. 31.15).

som:98:pom

[47] S. Som, B. F. Hutton, and M. Braun. "Properties of minimum cross-entropy reconstruction of emission tomography with anatomically based prior." In: *IEEE Trans. Nuc. Sci.* 45.6 (Dec. 1998), 3014–21. DOI: 10.1109/23.737658 (cit. on p. 31.15).

zhu:02:vir

[48] Y-M. Zhu. "Volume image registration by cross-entropy optimization." In: *IEEE Trans. Med. Imag.* 21.2 (Feb. 2002), 174–80. DOI: 10.1109/42.993135 (cit. on p. 31.15).

byrne:93:iir

[49] C. L. Byrne. "Iterative image reconstruction algorithms based on cross-entropy minimization." In: *IEEE Trans. Im. Proc.* 2.1 (Jan. 1993). Erratum and addendum: 4(2):226-7, Feb. 1995., 96–103. DOI: 10.1109/83.210869 (cit. on p. 31.15).

hess:99:mce

[50] C. P. Hess, Z-P. Liang, and P. C. Lauterbur. "Maximum cross-entropy generalized series reconstruction." In: *Intl. J. Imaging Sys. and Tech.* 10.3 (1999), 258–65. DOI: '10.1002/(SICI)1098-1098(1999)10:3<258::AID-IMA6>3.0.CO;2-7' (cit. on p. 31.15).

wang:01:mce

[51] Y. Wang. "Multicriterion cross-entropy minimization approach to positron emission tomographic imaging." In: *J. Opt. Soc. Am. A* 18.5 (May 2001), 1027–32. DOI: 10.1364/JOSAA.18.001027 (cit. on p. 31.15).

zhang:01:aat

[52] S. Zhang and Y. M. Wang. "An approach to positron emission tomography based on penalized cross-entropy minimization." In: *Signal Processing* 81.5 (May 2001), 1069–74. DOI: 10.1016/S0165-1684(00)00245-0 (cit. on p. 31.15).

zhu:05:aep

[53] H. Zhu et al. "A edge-preserving minimum cross-entropy algorithm for PET image reconstruction using multiphase level set method." In: *Proc. IEEE Conf. Acoust. Speech Sig. Proc.* Vol. 2. 2005, 469–72. DOI: 10.1109/ICASSP.2005.1415443 (cit. on p. 31.15).