1 Density Estimation

1.1 Characterizing multivariate data

A multivariate random variable $\mathbf{y}$ can consist of p dimensions. Assuming we have n such observation vectors,

$$\mathbf{y}_i^\mathsf{T} = (y_{i1}, y_{i2}, \ldots, y_{ip}) \tag{1.1}$$

The sample mean $\bar{\mathbf{y}}$ is

$$\bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i = \begin{pmatrix}\bar{y}_1\\ \bar{y}_2\\ \vdots\\ \bar{y}_p\end{pmatrix} \tag{1.2}$$

The matrix $\mathbf{Y}$ may be used to represent the entire data set as

$$\mathbf{Y} = \begin{pmatrix}\mathbf{y}_1^\mathsf{T}\\ \vdots\\ \mathbf{y}_n^\mathsf{T}\end{pmatrix} = \begin{pmatrix}y_{11} & y_{12} & \cdots & y_{1p}\\ \vdots & \vdots & & \vdots\\ y_{n1} & y_{n2} & \cdots & y_{np}\end{pmatrix} \tag{1.3}$$

Note that another way to define the mean vector now is to use $\mathbf{j}$ which is a vector of ones:

$$\bar{\mathbf{y}} = \frac{1}{n}\mathbf{Y}^\mathsf{T}\mathbf{j} \tag{1.4}$$

The expectation of $\mathbf{y}$ is

$$\mathcal{E}[\mathbf{y}] = \mathcal{E}\begin{bmatrix}y_1\\ \vdots\\ y_p\end{bmatrix} = \begin{pmatrix}\mathcal{E}[y_1]\\ \vdots\\ \mathcal{E}[y_p]\end{pmatrix} = \begin{pmatrix}\mu_1\\ \vdots\\ \mu_p\end{pmatrix} = \boldsymbol{\mu} \tag{1.5}$$

Similarly,

$$\mathcal{E}[\bar{\mathbf{y}}] = \mathcal{E}\begin{bmatrix}\bar{y}_1\\ \vdots\\ \bar{y}_p\end{bmatrix} = \begin{pmatrix}\mathcal{E}[\bar{y_1}]\\ \vdots\\ \mathcal{E}[\bar{y_p}]\end{pmatrix} = \begin{pmatrix}\mu_1\\ \vdots\\ \mu_p\end{pmatrix} = \boldsymbol{\mu} \tag{1.6}$$

The MLE of $\boldsymbol{\mu}$ is $\bar{\mathbf{y}}$.

The sample covariance matrix is

$$\mathbf{S} = (s_{jk}) = \begin{pmatrix}s_{11} & \cdots & s_{1p}\\ \vdots & & \vdots\\ s_{p1} & \cdots & s_{pp}\end{pmatrix} \tag{1.7}$$

Note that the covariance matrix is denoted $\mathbf{S}$ and not $\mathbf{S}^2$.

where

$$s_{jj} = \frac{1}{n-1}\sum_{i=1}^{n}(y_{ij}-\bar{y}_j)^2 = \frac{1}{n-1}\left(\sum_i y_{ij}^2 - n\bar{y}_j^2\right) \tag{1.8}$$

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(y_{ij}-\bar{y}_j)(y_{ik}-\bar{y}_k) = \frac{1}{n-1}\left(\sum_i y_{ij}y_{ik} - n\bar{y}_j\bar{y}_k\right) \tag{1.9}$$

$\mathbf{S}$ may also be written as

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^\mathsf{T} = \frac{1}{n-1}\left(\sum_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i^\mathsf{T} - n\bar{\mathbf{y}}\bar{\mathbf{y}}^\mathsf{T}\right) \tag{1.10}$$

$$= \frac{1}{n-1}\left[\mathbf{Y}^\mathsf{T}\mathbf{Y} - \mathbf{Y}^\mathsf{T}\left(\frac{1}{n}\mathbf{J}\right)\mathbf{Y}\right] = \frac{1}{n-1}\mathbf{Y}^\mathsf{T}\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y} \tag{1.11}$$

The population covariance matrix is

$$\mathbf{\Sigma} = \mathrm{cov}\left[\mathbf{y}\right] = \begin{pmatrix}\sigma_{11} & \cdots & \sigma_{1p}\\ \vdots & & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp}\end{pmatrix} \tag{1.12}$$

$$= \mathcal{E}\left[(\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})^\mathsf{T}\right] = \mathcal{E}\left[\mathbf{y}\mathbf{y}^\mathsf{T}\right] - \boldsymbol{\mu}\boldsymbol{\mu}^\mathsf{T} \tag{1.13}$$

$\mathbf{S}$ is an unbiased estimator of $\mathbf{\Sigma}$:

$$\mathcal{E}\left[\mathbf{S}\right] = \mathbf{\Sigma} \tag{1.14}$$

The sample correlation coefficient between the jth and kth variables is

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}} = \frac{s_{jk}}{s_j s_k} \tag{1.15}$$

which suggests a sample correlation matrix analogous to the covariance matrix:

$$\mathbf{R} = (r_{jk}) = \begin{pmatrix}1 & r_{12} & \cdots & r_{1p}\\ r_{21} & 1 & \cdots & r_{2p}\\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & 1\end{pmatrix} \tag{1.16}$$

The MLE of $\mathbf{P}_\rho$ is $\mathbf{R}$

The population correlation matrix is

$$\mathbf{P}_\rho = (\rho_{jk}) = \begin{pmatrix}1 & \rho_{12} & \cdots & \rho_{1p}\\ \rho_{21} & 1 & \cdots & \rho_{2p}\\ \vdots & \vdots & & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1\end{pmatrix} \tag{1.17}$$

It can be useful to have scalar measures of overall multivariate scatter. The *generalized sample variance* is defined as $|\mathbf{S}|$. Note that the p-dimensional ellipsoid $(\mathbf{y}-\bar{\mathbf{y}})^\mathsf{T}\mathbf{S}^{-1}(\mathbf{y}-\bar{\mathbf{y}}) = a^2$ has an ellipsoid volume proportional to $|\mathbf{S}|^{1/2}$. The *total* sample variance is defined as the trace of $\mathbf{S}$ and is the sum of the sample variances. This would be more useful when the covariances are unimportant, for example, when interpreting variation along principal components.

Note that $|\mathbf{S}| = \lambda_1\lambda_2\ldots\lambda_p$, the product of its eigenvalues.

1.2 Multivariate normal distribution

For independent x_i $(i = 1, ..., p)$ (i.e. orthogonal coordinates)

$$p(\mathbf{x}) = \prod_{i=1}^{p} p(x_i) = \prod_{i=1}^{p}\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left[-\frac{1}{2}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right]$$

$$= \frac{1}{(2\pi)^{p/2}\prod_{i=1}^{p}\sigma_i}\exp\left[-\frac{1}{2}\sum_{i=1}^{p}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right] \tag{1.18}$$

The covariance matrix is diagonal:

$\mathbf{\Sigma}$ is symmetric.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & . & 0 \\ 0 & \sigma_2^2 & . & 0 \\ . & . & . & . \\ 0 & 0 & . & \sigma_p^2 \end{bmatrix} \quad \Rightarrow \quad \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} 1/\sigma_1^2 & 0 & . & 0 \\ 0 & 1/\sigma_2^2 & . & 0 \\ . & . & . & . \\ 0 & 0 & . & 1/\sigma_p^2 \end{bmatrix} \tag{1.19}$$

Therefore

$$\sum_{i=1}^{p} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 = (\mathbf{x} - \boldsymbol{\mu})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

Also, $\left(\prod_{i=1}^{p} \sigma_i^2\right)^{1/2} = |\boldsymbol{\Sigma}|^{1/2}$. Therefore

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right] \tag{1.20}$$

In general $\boldsymbol{\Sigma}$ need not be a diagonal matrix.

The distance from $\mathbf{x}$ to $\boldsymbol{\mu}$ is the square of the Mahalanobis distance.

$$r^2 = (\mathbf{x} - \boldsymbol{\mu})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \tag{1.21}$$

In transforming coordinates/calculating distances, the $\boldsymbol{\Sigma}^{-1}$ middle term accounts for covariances between variables in x and generates a set of orthonormal vectors (eigenvectors). The product of the eigenvalues $= |\boldsymbol{\Sigma}|$. We have therefore achieved a rescaling and orthonormalization of our coordinate system.

Properties of multivariate normally distributed variables

Some properties of multivariate normal random variables are listed below: We assume that the $p \times 1$ vector $\mathbf{y}$ arises from $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. If $\mathbf{a}$ is a vector of constants, then

The subscript to $\mathcal{N}$ indicates the size of dimensions of the multivariate.

$$\mathbf{a}^\mathsf{T} \mathbf{y} \sim \mathcal{N}_1(\mathbf{a}^\mathsf{T} \boldsymbol{\mu}, \mathbf{a}^\mathsf{T} \boldsymbol{\Sigma} \mathbf{a}) \tag{1.22}$$

If $\mathbf{A}$ is a $q \times p$ matrix of constants, of rank q, where $q \leq p$, then the q independent linear combinations in $\mathbf{A}$ will follow

$$\mathbf{A}\mathbf{y} \sim \mathcal{N}_q(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\mathsf{T}) \tag{1.23}$$

The standardized vector $\mathbf{z}$ can be obtained from $\mathbf{y}$ in two ways:

$\boldsymbol{\Sigma} = \mathbf{T}^\mathsf{T}\mathbf{T}$ using Cholesky factorization. $\mathbf{T}$ is a nonsingular upper triangular matrix.

$$\mathbf{z} = (\mathbf{T}^\mathsf{T})^{-1} (\mathbf{y} - \boldsymbol{\mu}) \tag{1.24}$$

or

$$\mathbf{z} = (\boldsymbol{\Sigma}^{1/2})^{-1} (\mathbf{y} - \boldsymbol{\mu}) \tag{1.25}$$

$\boldsymbol{\Sigma}^{1/2}$ is the symmetric square root of $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2} \boldsymbol{\Sigma}^{1/2}$$

The resultant vector $\mathbf{z} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I})$. Further,

$$\sum_{j=1}^{p} z_j^2 = \mathbf{z}^\mathsf{T} \mathbf{z} = (\mathbf{y} - \boldsymbol{\mu})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \sim \chi_p^2 \tag{1.26}$$

If $\mathbf{y}$ is multivariate normal, and if $\mathbf{y}_1$ is a subset of $\mathbf{y}$, then $\mathbf{y}_1$ is also multivariate normal. Specifically, if $\mathbf{y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $y_j \sim \mathcal{N}_1(\mu_j, \sigma_{jj})$ for $j = 1, \ldots, p$. The converse is not true in general: if each y_j in $\mathbf{y}$ is normal, $\mathbf{y}$ need not be multivariate normal.

If the observations are partitioned into two subvectors $\mathbf{y}$ $(p \times 1)$and $\mathbf{x}$ $(q \times 1)$, then

$$\mathcal{E} \begin{bmatrix} \mathbf{y} \\ \mathbf{x} \end{bmatrix} = \begin{pmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_x \end{pmatrix}, \qquad \text{cov} \begin{bmatrix} \mathbf{y} \\ \mathbf{x} \end{bmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{pmatrix}$$

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix} \sim \mathcal{N}_{p+q} \left[\begin{pmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_x \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{pmatrix} \right] \tag{1.27}$$

If $\mathbf{y}$ and $\mathbf{x}$ are independent, $\boldsymbol{\Sigma}_{yx} = \mathbf{0}$. If $\mathbf{y}$ and $\mathbf{x}$ are not independent, then

$$\mathcal{E}[\mathbf{y}|\mathbf{x}] = \boldsymbol{\mu}_y + \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{x} - \boldsymbol{\mu}_x) \tag{1.28}$$

$$\text{cov}[\mathbf{y}|\mathbf{x}] = \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy} \tag{1.29}$$

$\bar{\mathbf{y}}$ and $\mathbf{S}$

For n random samples $\mathbf{y}_i \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ($i = 1$ to n), if $\bar{\mathbf{y}} = \sum_i \mathbf{y}_i/n$,

$$\bar{\mathbf{y}} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n) \tag{1.30}$$

We previously had

$$(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$$

There are $p(p+1)/2$ distinct entries (variances and covariances) in $\mathbf{S}$. The joint distribution of these $p(p+1)/2$ variables in $\mathbf{W} = (n-1)\mathbf{S} = \sum_i (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^\mathsf{T}$ is the multivariate extension of the χ^2-distribution and is called the Wishart distribution with $n-1$ degrees of freedom, $W_p(n-1, \boldsymbol{\Sigma})$.

$$\sum_i (\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})^\mathsf{T} \sim W_p(n, \boldsymbol{\Sigma}) \tag{1.31}$$

$$\sum_i (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^\mathsf{T} \sim W_p(n-1, \boldsymbol{\Sigma}) \tag{1.32}$$

1.3 Multivariate hypothesis testing

Testing $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ with Σ known

We wish to test $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ vs. $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. H_0 implies that $\mu_j = \mu_{0j}$ for all p components of $\boldsymbol{\mu}$. H_1 requires at least one $\mu_j \neq \mu_{0j}$ to hold. Given the n measurement vectors $\mathbf{y}_i$ which follows $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we find $\bar{\mathbf{y}}$ and then the following test statistic:

$$Z^2 = n(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu}_0) \tag{1.33}$$

which is expected to follow the χ^2_p distribution.

Why is this better than testing the p components separately?

Example 1.1. Consider $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ where $n = 20$,

$$\boldsymbol{\mu}_0 = \begin{bmatrix} 70 \\ 170 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 20 & 100 \\ 100 & 1000 \end{bmatrix} \qquad \bar{\mathbf{y}} = \begin{bmatrix} 71.45 \\ 164.7 \end{bmatrix}$$

Solution:

$$Z^2 = (20)(1.45, -5.3)\begin{pmatrix} 0.1 & -0.01 \\ -0.01 & 0.002 \end{pmatrix}\begin{pmatrix} 1.45 & -5.3 \end{pmatrix} = 8.4026$$

using $\alpha = 0.05$, $\chi^2_{2,0.05} = 5.99$ which implies we reject H_0. If we had tested each component separately,

$$z_1 = \frac{\bar{y}_1 - \mu_{01}}{\sigma_1/\sqrt{n}} = 1.450 < 1.96 \qquad z_2 = \frac{\bar{y}_2 - \mu_{02}}{\sigma_2/\sqrt{n}} = -0.7495 > -1.96$$

which implies that both individual tests accept H_0! The difference between these two approaches is the utilization of the positive correlation that exists between y_1 and y_2 in the multivariate test.

When the multivariate test disagrees with tests of individual variables, go with the multivariate test result.

Testing $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ with $\boldsymbol{\Sigma}$ unknown

We wish to test $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ vs. $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. Given the n measurement vectors $\mathbf{y}_i$ which follows $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we find $\bar{\mathbf{y}}$ (as an estimate of $\boldsymbol{\mu}$) and $\mathbf{S}$ (as an estimate of $\boldsymbol{\Sigma}$). The test statistic for the multivariate situation is the squared version of the univariate t statistic.

$$T^2 = n(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)^\mathsf{T} \mathbf{S}^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu}_0) \tag{1.34}$$

This T^2 statistic follows a generalization of the t-distribution called Hotelling's T^2-distribution. The specific distribution (of a family) to be used in this test clearly depends on the number of dimensions (p), and the degrees of freedom ($n-1$). H_0 is rejected if $T^2 > T^2_{\alpha,p,n-1}$. The T^2 test has the following properties:

- $n-1$ must be greater than p. (Else, $\mathbf{S}$ is singular and T^2 cannot be computed).
- The degrees of freedom remains $n-1$ as it was for the univariate test. It would be n_1+n_2-2 for the multivariate two sample test of $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.
- H_1 is written with a two-sided test in mind, since all p components need not follow, for example, $\mu_i > \mu_{i0}$. However, the critical region is one-tailed for convenience.

Remember the T^2 distribution depends on p and n and it would be a pain to keep track of two thresholds per distribution.

- For the univariate case, $t^2_{n-1} = F_{1,n-1}$. The T^2 statistic can be mapped on to an F-statistic:

$$\frac{\nu - p + 1}{\nu p} T^2_{p,\nu} = F_{p,\nu-p+1} \tag{1.35}$$

  For a one-sample test, the degrees of freedom $\nu = n - 1$.

Comparing two mean vectors

The multivariate two-sample T^2 test involves testing $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ vs. $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$. We have n_1 points in one random sample $(\mathbf{y}_{11}, \ldots, \mathbf{y}_{1n_1})$ from $\mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$. Similarly $(\mathbf{y}_{21}, \ldots, \mathbf{y}_{2n_2})$ are the n_2 points in sample 2, all following $\mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. In the T^2 test we assume that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, with $\boldsymbol{\Sigma}$ unknown. The sample mean vectors are $\bar{\mathbf{y}}_1 = \sum_{i=1}^{n_1} y_{1i}/n_1$ and $\bar{\mathbf{y}}_2 = \sum_{i=1}^{n_2} y_{2i}/n_2$. The sums of square matrices are

The test for $H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ will be described later.

$$\mathbf{W}_1 = \sum_{i=1}^{n_1} (\mathbf{y}_{1i} - \bar{\mathbf{y}}_1)(\mathbf{y}_{1i} - \bar{\mathbf{y}}_1)^\mathsf{T} = (n_1 - 1)\mathbf{S}_1 \tag{1.36}$$

$$\mathbf{W}_2 = \sum_{i=1}^{n_2} (\mathbf{y}_{2i} - \bar{\mathbf{y}}_2)(\mathbf{y}_{2i} - \bar{\mathbf{y}}_2)^\mathsf{T} = (n_2 - 1)\mathbf{S}_2 \tag{1.37}$$

A pooled, unbiased estimator of the population covariance matrix would be

$\mathcal{E}[\mathbf{S}_p] = \boldsymbol{\Sigma}$

$$\mathbf{S}_p = \frac{1}{n_1 + n_2 - 2}(\mathbf{W}_1 + \mathbf{W}_2) \tag{1.38}$$

For a univariate two sample test

$$t = \frac{\bar{y}_1 - \bar{y}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad \Rightarrow \qquad t^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{y}_1 - \bar{y}_2)(S_p^2)^{-1}(\bar{y}_1 - \bar{y}_2) \tag{1.39}$$

this can be generalized to p variables as

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)^\mathsf{T} \mathbf{S}_p^{-1} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2) \tag{1.40}$$

which is distributed as T^2_{p,n_1+n_2-2} when H_0 is true. Therefore, for a chosen α, H_0 is rejected if $T^2 > T^2_{\alpha,p,n_1+n_2-1}$.

Paired sample test

This is the multivariate extension of the paired t test. We have n multivariate measurements of treatment 1 ($\mathbf{y}_i$) and treatment 2 ($\mathbf{x}_i$) which generates the multivariate difference variable $\mathbf{d}_i = \mathbf{y}_i - \mathbf{x}_i$. We then wish to test $H_0 : \boldsymbol{\mu}_d = \mathbf{0}$ using

$H_0 : \boldsymbol{\mu}_y = \boldsymbol{\mu}_x$

$$\bar{\mathbf{d}} = \frac{\sum \mathbf{d}_i}{n} \qquad \mathbf{S}_d = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{d}_i - \bar{\mathbf{d}})(\mathbf{d}_i - \bar{\mathbf{d}})^\mathsf{T} \tag{1.41}$$

The test statistic T^2 is compared against $T^2_{p,n-1}$.

$$T^2 = \bar{\mathbf{d}}^\mathsf{T}\left(\frac{\mathbf{S}_d}{n}\right)^{-1}\bar{\mathbf{d}} = n\bar{\mathbf{d}}^\mathsf{T}\mathbf{S}_d^{-1}\bar{\mathbf{d}} \tag{1.42}$$

1.4 Conditional and Marginal Gaussians

Conditional Gaussian distributions

Consider the d-dimensional vector $\mathbf{x}$ which follows $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. $\mathbf{x}$ can be partitioned into $\mathbf{x}_a$ and $\mathbf{x}_b$:

$$\mathbf{x} = \begin{bmatrix}\mathbf{x}_a \\ \mathbf{x}_b\end{bmatrix} \qquad \Rightarrow \qquad \boldsymbol{\mu} = \begin{bmatrix}\boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b\end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb}\end{bmatrix} \tag{1.43}$$

where $\boldsymbol{\Sigma}_{ba} = \boldsymbol{\Sigma}_{ab}^\mathsf{T}$ because $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^\mathsf{T}$. The precision, $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$, is usually more convenient to use than the covariance. We can partition $\boldsymbol{\Lambda}$ as

We wish to find $p(\mathbf{x}_a|\mathbf{x}_b)$.

$$\boldsymbol{\Lambda} = \begin{bmatrix}\boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb}\end{bmatrix} \tag{1.44}$$

Since the inverse of a symmetric matrix is symmetric, $\boldsymbol{\Lambda}_{aa}$ and $\boldsymbol{\Lambda}_{bb}$ are symmetric. Also, $\boldsymbol{\Lambda}_{ab}^\mathsf{T} = \boldsymbol{\Lambda}_{ba}$. The exponent in the joint distribution $p(\mathbf{x})$ can be rearranged as follows:

Important: $\boldsymbol{\Lambda}_{aa} \neq \boldsymbol{\Sigma}_{aa}^{-1}$ etc.

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^\mathsf{T}\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^\mathsf{T}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b)$$

$$-\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^\mathsf{T}\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^\mathsf{T}\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b) \tag{1.45}$$

This expression, assuming that $\mathbf{x}_b$ is given, is a quadratic in $\mathbf{x}_a$, and hence $p(\mathbf{x}_a|\mathbf{x}_b)$ is a Gaussian.

$$p(\mathbf{x}_a|\mathbf{x}_b) \sim \mathcal{N}(\mu_{a|b}, \boldsymbol{\Sigma}_{a|b}) \tag{1.46}$$

There is a quick way to identify the parameters of a Gaussian as summarized below, by completing the square.

Finding the parameters of a multivariate Gaussian.

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}\mathbf{x}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \text{const.} \tag{1.47}$$

The first term on the right gives us $\boldsymbol{\Sigma}^{-1}$ and the second term yields $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. Collecting all the second order terms in $\mathbf{x}_a$ gives

The constant term is a function of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

$$-\frac{1}{2}\mathbf{x}_a^\mathsf{T}\boldsymbol{\Lambda}\mathbf{x}_a \tag{1.48}$$

which gives

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1} \tag{1.49}$$

Grouping together terms linear in $\mathbf{x}_a$ gives

Using $\boldsymbol{\Lambda}_{ab}^\mathsf{T} = \boldsymbol{\Lambda}_{ba}$.

$$\mathbf{x}_a^\mathsf{T}\left[\mathbf{\Lambda}_{aa}\boldsymbol{\mu}_a - \mathbf{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\right] \tag{1.50}$$

Since the coefficient of $\mathbf{x}_a$ is $\mathbf{\Sigma}_{a|b}^{-1}\boldsymbol{\mu}_{a|b} = \mathbf{\Lambda}_{aa}\boldsymbol{\mu}_{a|b}$, we get

$$\boldsymbol{\mu}_{a|b} = \mathbf{\Sigma}_{a|b}(\mathbf{\Lambda}_{aa}\boldsymbol{\mu}_a - \mathbf{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)) = \boldsymbol{\mu}_a - \mathbf{\Lambda}_{aa}^{-1}\mathbf{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \tag{1.51}$$

M^{-1} is the Schur complement of the LHS w.r.t. D.

A convenient relation exists to determine inverses of a matrix in terms of submatrices:

$$\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{bmatrix}^{-1} = \begin{bmatrix}\mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1}\\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1}\end{bmatrix} \tag{1.52}$$

where

$$\mathbf{M} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} \tag{1.53}$$

Then since

$$\begin{bmatrix}\mathbf{\Sigma}_{aa} & \mathbf{\Sigma}_{ab}\\ \mathbf{\Sigma}_{ba} & \mathbf{\Sigma}_{bb}\end{bmatrix}^{-1} = \begin{bmatrix}\mathbf{\Lambda}_{aa} & \mathbf{\Lambda}_{ab}\\ \mathbf{\Lambda}_{ba} & \mathbf{\Lambda}_{bb}\end{bmatrix} \tag{1.54}$$

we get

$$\begin{aligned}\mathbf{\Lambda}_{aa} &= (\mathbf{\Sigma}_{aa} - \mathbf{\Sigma}_{ab}\mathbf{\Sigma}_{bb}^{-1}\mathbf{\Sigma}_{ba})^{-1}\\ \mathbf{\Lambda}_{ab} &= -(\mathbf{\Sigma}_{aa} - \mathbf{\Sigma}_{ab}\mathbf{\Sigma}_{bb}^{-1}\mathbf{\Sigma}_{ba})^{-1}\mathbf{\Sigma}_{ab}\mathbf{\Sigma}_{bb}^{-1}\end{aligned} \tag{1.55}$$

These results indicate why it is easier to use the precision.

This then yields

$$\begin{aligned}\boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a + \mathbf{\Sigma}_{ab}\mathbf{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b)\\ \mathbf{\Sigma}_{a|b} &= \mathbf{\Sigma}_{aa} - \mathbf{\Sigma}_{ab}\mathbf{\Sigma}_{bb}^{-1}\mathbf{\Sigma}_{ba} = \mathbf{\Lambda}_{aa}^{-1}\end{aligned} \tag{1.56}$$

Notice that $\boldsymbol{\mu}_{a|b}$ is a linear function of $\mathbf{x}_b$ and $\mathbf{\Sigma}_{a|b}$ is independent of $\mathbf{x}_b$. This result can be seen to be equivalent to Eqns. 1.28 and 1.29 which were equations simply stated before.

This defines a linear Gaussian model.

Marginal Gaussian distributions

When the joint distribution $p(\mathbf{x}_a, \mathbf{x}_b)$ is Gaussian, we have seen above that $p(\mathbf{x}_a|\mathbf{x}_b)$ is also Gaussian. The question that next arises is whether $p(\mathbf{x}_a) = \int p(\mathbf{x}_a|\mathbf{x}_b)d\mathbf{x}_b$ is also Gaussian. Starting with Eq. 1.45, we pick out terms with $\mathbf{x}_b$

Preparing for integration...

$$-\frac{1}{2}\mathbf{x}_b^\mathsf{T}\mathbf{\Lambda}_{bb}\mathbf{x}_b + \mathbf{x}_b^\mathsf{T}\mathbf{m} = -\frac{1}{2}(\mathbf{x}_b - \mathbf{\Lambda}_{bb}^{-1}\mathbf{m})^\mathsf{T}\mathbf{\Lambda}_{bb}(\mathbf{x}_b - \mathbf{\Lambda}_{bb}^{-1}\mathbf{m}) + \frac{1}{2}\mathbf{m}^\mathsf{T}\mathbf{\Lambda}_{bb}^{-1}\mathbf{m} \tag{1.57}$$

where

$$\mathbf{m} = \mathbf{\Lambda}_{bb}\boldsymbol{\mu}_b - \mathbf{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a) \tag{1.58}$$

The first term on the RHS of Eq. 1.57 is in the standard quadratic form; the second term is a function of $\mathbf{x}_a$ but not of $\mathbf{x}_b$. This second term will not matter when we take the exponential of Eq. 1.57 and integrate, as it ultimately gets absorbed into the normalization constant. Therefore we can expect a term of the form

$$\int \exp\left[-\frac{1}{2}(\mathbf{x}_b - \mathbf{\Lambda}_{bb}^{-1}\mathbf{m})^\mathsf{T}\mathbf{\Lambda}_{bb}(\mathbf{x}_b - \mathbf{\Lambda}_{bb}^{-1}\mathbf{m})\right]d\mathbf{x}_b \tag{1.59}$$

The integral of an unnormalized Gaussian should give us the inverse of that normalization constant. That constant is not a function of the mean, but of the determinant of the covariance. Hence, completing the square,

$$\begin{aligned}&\frac{1}{2}\left[\mathbf{\Lambda}_{bb}\boldsymbol{\mu}_b - \mathbf{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a)\right]^\mathsf{T}\mathbf{\Lambda}_{bb}^{-1}\left[\mathbf{\Lambda}_{bb}\boldsymbol{\mu}_b - \mathbf{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a)\right]\\ &\qquad -\frac{1}{2}\mathbf{x}_a^\mathsf{T}\mathbf{\Lambda}_{aa}\mathbf{x}_a + \mathbf{x}_a^\mathsf{T}(\mathbf{\Lambda}_{aa}\boldsymbol{\mu}_a + \mathbf{\Lambda}_{ab}\boldsymbol{\mu}_b) + \text{const.}\\ = -\frac{1}{2}\mathbf{x}_a^\mathsf{T}&(\mathbf{\Lambda}_{aa} - \mathbf{\Lambda}_{ab}\mathbf{\Lambda}_{bb}^{-1}\mathbf{\Lambda}_{ba})\mathbf{x}_a + \mathbf{x}_a^\mathsf{T}(\mathbf{\Lambda}_{aa} - \mathbf{\Lambda}_{ab}\mathbf{\Lambda}_{bb}^{-1}\mathbf{\Lambda}_{ba})^{-1}\boldsymbol{\mu}_a + \text{const.}\end{aligned} \tag{1.60}$$

This expression is not a function of $\mathbf{x}_a$. On comparing this with Eq. 1.47 we get

$$\boldsymbol{\Sigma}_a = (\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba})^{-1} \tag{1.61}$$

$$\text{mean} = \boldsymbol{\Sigma}_a(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba})^{-1})\boldsymbol{\mu}_a = \boldsymbol{\mu}_a \tag{1.62}$$

Equation 1.61 can be rewritten in terms of precisions using

$$\begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix} \tag{1.63}$$

A neat simple result.

Using Eq. 1.52 we get

$$(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba})^{-1} = \boldsymbol{\Sigma}_{aa} \tag{1.64}$$

Hence

$$p(\mathbf{x}_a) = \mathcal{N}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a) = \mathcal{N}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}) \tag{1.65}$$

where $\mathcal{E}[\mathbf{x}_a] = \boldsymbol{\mu}_a$ and $\text{cov}[\mathbf{x}_a] = \boldsymbol{\Sigma}_{aa}$.

Summary of conditional and marginal Gaussian distributions

Given a joint Gaussian $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} \qquad \mathbf{x} = \begin{bmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix} \tag{1.66}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix} \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{bmatrix} \tag{1.67}$$

Conditional distributions

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{aa}^{-1}) \tag{1.68}$$

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b) \tag{1.69}$$

Marginal distributions

$$p(\mathbf{x}_a) = \mathcal{N}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}) \tag{1.70}$$

Bayes theorem for multivariate Gaussians

Given a Gaussian marginal $p(\mathbf{x})$ and a Gaussian conditional $p(\mathbf{y}|\mathbf{x})$ where $p(\mathbf{y}|\mathbf{x})$ has a mean linear in $\mathbf{x}$ and a covariance which is not a function of $\mathbf{x}$, we would wish to find $p(\mathbf{y})$ and $p(\mathbf{x}|\mathbf{y})$. Let

i.e. a linear Gaussian model.

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \tag{1.71}$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \tag{1.72}$$

where $\boldsymbol{\Lambda}$ and $\mathbf{L}$ are precision matrices.

If the dimensions of $\mathbf{x}$ and $\mathbf{y}$ are m and d, then $\mathbf{A}$ is a matrix of size $d \times m$.

What is $p(\mathbf{x}, \mathbf{y})$?
Let

First get the joint distribution.

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \tag{1.73}$$

Then

$$\ln p(\mathbf{z}) = \ln p(\mathbf{x}) + \ln p(\mathbf{y}|\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu}) - \frac{1}{2}(\mathbf{y}-\mathbf{A}\mathbf{x}-\mathbf{b})^\mathsf{T}\mathbf{L}(\mathbf{y}-\mathbf{A}\mathbf{x}-\mathbf{b}) + \text{const.} \tag{1.74}$$

$$\begin{aligned} &-\frac{1}{2}\mathbf{x}^\mathsf{T}(\boldsymbol{\Lambda}+\mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A})\mathbf{x} - \frac{1}{2}\mathbf{y}^\mathsf{T}\mathbf{L}\mathbf{y} + \frac{1}{2}\mathbf{y}^\mathsf{T}\mathbf{L}\mathbf{A}\mathbf{x} + \frac{1}{2}\mathbf{x}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{y} \\ &\quad = \frac{1}{2}\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}^\mathsf{T} \begin{bmatrix} \boldsymbol{\Lambda}+\mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A} & -\mathbf{A}^\mathsf{T}\mathbf{L} \\ -\mathbf{L}\mathbf{A} & \mathbf{L} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = -\frac{1}{2}\mathbf{z}^\mathsf{T}\mathbf{R}\mathbf{z} \end{aligned} \tag{1.75}$$

where

$$\mathbf{R} = \begin{bmatrix} \boldsymbol{\Lambda}+\mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A} & -\mathbf{A}^\mathsf{T}\mathbf{L} \\ -\mathbf{L}\mathbf{A} & \mathbf{L} \end{bmatrix} \tag{1.76}$$

Then,

$$\text{cov}\left[\mathbf{z}\right] = \mathbf{R}^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^\mathsf{T} \\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\mathsf{T} \end{bmatrix} \tag{1.77}$$

The mean can be found by arranging the linear terms in Eq. 1.74.

$$\mathbf{x}^\mathsf{T}\boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{x}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{b} + \mathbf{y}^\mathsf{T}\mathbf{L}\mathbf{b} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{b} \\ \mathbf{L}\mathbf{b} \end{bmatrix} \tag{1.78}$$

which gives

$$\mathcal{E}\left[\mathbf{z}\right] = \mathbf{R}^{-1}\begin{bmatrix} \boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{b} \\ \mathbf{L}\mathbf{b} \end{bmatrix} \tag{1.79}$$

which finally gives the simple result

$$\mathcal{E}\left[\mathbf{z}\right] = \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu}+\mathbf{b} \end{bmatrix} \tag{1.80}$$

Marginal:
We can now find the marginal from the joint distribution

$$\mathcal{E}\left[\mathbf{y}\right] = \mathbf{A}\boldsymbol{\mu}+\mathbf{b} \tag{1.81}$$

$$\text{cov}\left[\mathbf{y}\right] = \mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\mathsf{T} \tag{1.82}$$

which are the parameters of $p(\mathbf{y})$. Note that if $\mathbf{A} = \mathbf{I}$, then this is a simple convolution of two Gaussians, where the mean of the convolution is the sum of the means, and the covariance of the convolution is the sum of covariances.

Conditional:
Finally, the conditional can be easily worked out:

$$\mathcal{E}\left[\mathbf{x}|\mathbf{y}\right] = (\boldsymbol{\Lambda}+\mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A})^{-1}\left[\mathbf{A}^\mathsf{T}\mathbf{L}(\mathbf{y}-\mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right] \tag{1.83}$$

$$\text{cov}\left[\mathbf{x}|\mathbf{y}\right] = (\boldsymbol{\Lambda}+\mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A})^{-1} \tag{1.84}$$

Summary of Bayes forms

Given

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \tag{1.85}$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \tag{1.86}$$

then

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{B}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\top}) \tag{1.87}$$

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}(\mathbf{A}^{\top}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}), \boldsymbol{\Sigma}) \tag{1.88}$$

where $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^{\top}\mathbf{L}\mathbf{A})^{-1}$.

1.5 Bayesian vs. Frequentist estimation

Probabilistic models are systems where a class of objects is simulated, with each object being assigned a probability. These methods are *supervised* methods: training sets of data are used. There are two approaches to defining probability, Bayesian and Frequentist, and they result in different approaches towards using the same available experimental data. The Frequentist insists that parameters have a unique true value which is observable in the long run. A Bayesian talks of a range of probability values (i.e. a distribution) and talks in term of a degree of belief. These approaches have been historically confrontational; however the Bayesian approach is more relevant to us than the Frequentist one.

The concept of frequency is useless when we talk of a belief/probability that global warming would result in ice caps melting.

Bayesian theory:
The Bayesian definition of probability in the context of classification refers to a degree of belief in the truth of some hypothesis: 'There is an 80% chance of rain today'. This belief is a function of whatever data/information is available at the current time. Thus if D refers to information that is currently available (i.e. data), there are two rules that Bayesian inference is based on.

$$p(A|D) + p(\bar{A}|D) = 1 \tag{1.89}$$

$$p(A, B|D) = p(A|B, D)p(B|D) = p(B|A, D)p(A|D) \tag{1.90}$$

Eq. 1.90 gives us the more conventional statement of Bayes' rule, where we drop the condition D

$$\begin{aligned} P(A, B) &= p(A|B)p(B) \\ &= p(B|A)p(A) \end{aligned}$$

In the regression context, we look for those model parameter estimates which best explain the data that we have. In the classification context, the probability of several hypotheses H_i given data D are denoted $p(H_i|D)$. We may have prior information about the various hypotheses in the form of $p(H_i)$. Using Bayes' rule we have

$$p(H_i|D) = \frac{p(D|H_i)p(H_i)}{p(D)} \tag{1.91}$$

$p(D|H_i)$ is the likelihood of observing the data assuming that hypothesis H_i is true. $p(D)$ is the marginal distribution of the data D and does not depend on H_i. Since the hypotheses $\{H_i\}$ must be mutually exclusive and exhaustive,

$$p(D) = \sum_i p(D|H_i)p(H_i) \tag{1.92}$$

Obviously $p(H_i|D)$ (the posterior distribution) is what we want. New examples are observed and prior estimates of hypothesis (accuracy) probabilities are modified. This works better than throwing out a hypothesis that is inconsistent because of one single example.

Bayesian parameter estimation:

Usually, the hypothesis H is about the value of a continuous parameter θ. θ could be μ, σ^2, the proportion p of a binomial etc. In general $\boldsymbol{\theta}$ refers to the simultaneous estimation of several of these parameters. Then the posterior distribution may be written specifically in terms of θ, given data D as

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int_\theta p(D|\theta)p(\theta)d\theta} \tag{1.93}$$

We use an integral rather than a summation because θ is continuous. The integral runs over the allowed range of values of θ. If $\theta = p$, then the range would be 0 to 1.

$p(\theta)$ is the prior distribution and may be assumed to be flat over the range of θ (i.e. the prior is uninformative). Obviously, if any (even speculative) information is available to suggest that θ may be more likely in a given range, then that prior would be 'informative'. The following must obviously hold:

$$\int_\theta p(\theta)d\theta = 1 \tag{1.94}$$

$$\int_D p(D|\theta)dD = 1 \tag{1.95}$$

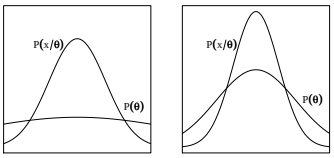

Fig. 1.1: Uninformative prior (left) and informative prior (right).

The second equation indicates that the likelihood of seeing the data given a particular model θ is 1! The posterior probability $p(\theta|D)$ is what we want from an inference standpoint. A 95% confidence interval is created by choosing θ_1 and θ_2 such that

$$\int_{\theta_1}^{\theta_2} p(\theta|D)d\theta = 0.95 \tag{1.96}$$

The probability that $\theta > \theta_0$ given data D is $p(\theta > \theta_0|D) = \int_{\theta_0}^{\theta_{max}} p(\theta|D)d\theta$.

Frequentist theory.

Probability is defined as the relative occurence of an event given an infinite number of repeated measurements. In the context of classification where a hypothesis (or a set of hypotheses) is being evaluated, after infinite sampling there would be no uncertainty in the parameter being evaluated. Therefore, a hypothesis is true or false, without a degree of belief being attributable to it. However, the observed data does vary (due to experimental error) even according to a specific hypothesis. Hence, *data can be assigned probabilities given unique parameter values*! This therefore is the approach that we followed in the previous chapter on estimation. To summarize:

Remember that S is a function of $D = S(D)$.

1. Given the data D, and the parameter θ, identify a relevant statistic S to use in a hypothesis test. For example, S could be Z, t, or X^2. Define a null hypothesis H_0 which fixes the value of the parameter to some particular value (e.g. $H_0 : \mu = \mu_0$).
2. The data D obtained is random, and hence $S(D)$ is also random. Identify the probability distribution that S should follow, using simulated data if need be (e.g. the statistic t is expected to follow the t_{n-1} distribution).
3. Compare S based on actual data against the distribution $p(S(D_{sim})|H_0)$.
   If $\int_{S(D_{actual})}^{\infty} p(S(D_{sim}|H_0)dS(D_{sim})$ is very small (< 0.05), then H_0 may be rejected.

Upper one-sided test.

At this confidence level, if the process is repeated many times, the outcome would be correct 95% of the time.

Bayesian vs. Frequentist approaches:

Frequentists prefer the term 'confidence interval'.

The two approaches to defining probability now result in an interesting difference in how confidence intervals are defined: the Frequentist confidence interval defines a range of values for the data average (e.g. $\bar{x}$) that would arise 95% of the time given a specific value of θ (e.g. $\mu = \mu_0$). To a Frequentist, there can only be one true value of μ given H_0. The Bayesian

approach permits a multitude of θ values to exist, but then requires a prior distribution of θ (μ). The Bayesian would then have prior estimates of the mean (m) and standard deviation (s) and on seeing the Normally distributed data, would then update to the posterior estimates m' and s'.

$$\text{Frequentist} : p\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu_0 < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95 \tag{1.97}$$

$$\text{Bayesian} : p\left(m' - 1.96\frac{s'}{\sqrt{n}} < \mu_0 < m' + 1.96\frac{s'}{\sqrt{n}}\right) = 0.95 \tag{1.98}$$

More on the posterior values of parameters later when we discuss MAP and Bayesian learning.

Therefore, the Bayesian simply observes that there is a 95% probability (as a degree of belief) that the stated interval (based on the posterior parameters) contains the true mean.

Bayesians prefer the term 'credible interval' to 'confidence interval'.

The Frequentist however must resort to a more convoluted argument: since there is only one possible value to the true mean (attainable on infinite sampling), he cannot talk of a 95% probability of the true mean being in this particular interval: it is either in or out. Therefore the Frequentist resorts to "Given that sampled data can vary from one set to another (today's sample, tomorrow's etc.) each set would end up with an interval whose limits depend upon the data average for that day. Then there is a 95% probability (i.e. a long-run fraction) that these intervals would contain the true mean". We do not of course repeat the entire experiment so often, and therefore this reliance on D_{sim} (which is data not really seen) makes the Frequentist approach problematic.

Humans seem to think using the Bayesian approach: if an experiment gives us an absurd parameter value, we are likely to reject it and redo the experiment. In that sense, we utilize a parameter value range based on a prior distribution where data may be trusted (i.e. we have an informative prior). We are not so much interested in a specific value of some parameter θ_0 but would prefer to know of a range of values for that parameter (i.e. specify a pdf for that parameter).

The Frequentist approach would use that data to reject the hypothesis outright.

The stopping rule problem:

Stopping rule problem.

An example of the differences between the two approaches is the stopping rule problem. To decide whether a coin is fair, consider two different experiments:

1. You toss a coin N times and observe the number of heads. For example, $N = 100$, $n_h = 55$.
2. You keep tossing a coin until some predefined number of heads (n_h) is obtained. For example, $n_h = 55$ takes 100 throws.

A Frequentist would come up with different inferences for the two experiments, since the likelihoods used are based on different distributions: a binomial for the first, and a negative binomial for the second. To a Bayesian, there is no difference between the two experiments given that the data obtained is the same: probability of heads for a fair coin = 0.55. For both experiments, the posterior (after normalization) would be of the form below; clearly the stopping rule is irrelevant.

$$P(p|k,r) = \frac{p^r\,(1-p)^k}{\int_0^1 p^r\,(1-p)^k\,dp} \tag{1.99}$$

Negative binomial: If p is the probability of success in one Bernoulli trial, then the probability of k failures and r successes in $k + r$ trials, with success on the last trial is

$${}^{k+r-1}C_{r-1}p^r(1-p)^k$$

Mean $= r(1-p)/p$ and variance $= r(1-p)/p^2$. Clearly, when $r = 1$ we have the geometric distribution.

The Bayesian approach needs a prior:

This can be a good thing as it forces us to look at any available information if we are to choose informative priors. The posterior distribution may be thought of as a compromise between the prior and the likelihood (Fig. 1.2). Given a small data set, the prior would (should) dominate. Given a large data sample, the prior will have little influence, in which case, the posterior resembles the likelihood. A bad choice of prior would typically result in poor estimation/classification results with high confidence. There are now sampling algorithms

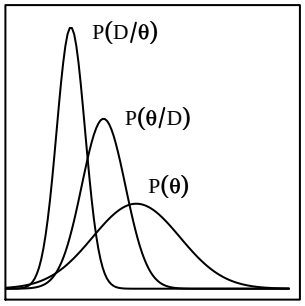

Fig. 1.2: The posterior as a compromise.

like the Markov Chain Monte Carlo (MCMC) which given improved computing capabilities provide the ability ot come up with reasonable choice for priors. Frequentist methods perform poorly, on the other hand, given few data points.

We now look at two frequentist approaches to parameter estimation: maximum likelihood and Maximum a posteriori estimation. We follow this up by demonstrating Bayesian estimation for uni and multivariate distributions.

1.6 Maximum likelihood estimation (MLE)

The maximum likelihood (ML) model/hypothesis is that model (i.e. the value of i) which maximizes $p(\mathbf{D}|M_i)$

$$M_{ML} = \underset{M \in \mathbf{M}}{\operatorname{argmax}}\, p(\mathbf{D}|M) \tag{1.100}$$

Maximum Likelihood Estimation (MLE) of model parameters involves finding those parameter values which result in the model best fitting the training set. These parameter values are unknown, but are assumed to be 'fixed' (i.e. not a distribution of values). Usually MLE of parameters results in intuitive values/conclusions. A model is represented by a distribution with parameters $\boldsymbol{\theta}$, and therefore desiring a maximum likelihood model is equivalent to obtaining a parameter set which maximizes the likelihood of obtaining the data:

$$\boldsymbol{\theta}_{ML} = \underset{M \in \mathbf{M}}{\operatorname{argmax}}\, p(\mathbf{D}|\boldsymbol{\theta}) \tag{1.101}$$

Usually the data set $\mathbf{D}$ consists of a set of training samples $\{x_1, x_2, ..., x_k, ...x_n\}$. Assuming that these samples are statistically independent,

$$p(\mathbf{D}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(x_k|\boldsymbol{\theta}) \tag{1.102}$$

If $l(\boldsymbol{\theta}) = \ln p(\mathbf{D}|\boldsymbol{\theta})$, then $\boldsymbol{\theta}_{ML} = \underset{M \in \mathbf{M}}{\operatorname{argmax}}\, l(\boldsymbol{\theta})$ Therefore a necessary requirement for MLE is

A sum of logs is more convenient to handle than the product of several terms.

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0 \tag{1.103}$$

MLE of the Binomial parameter p:

Let X = random variable describing a coin flip, $X = 1$ if heads, $X = 0$ if tails. For one coin flip,

As before, the Bernoulli and the Binomial.

$$f(x) = p^x (1-p)^{1-x} \tag{1.104}$$

For n coin flips,

$$f(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} \tag{1.105}$$

The log-likelihood may be written as

$$l(p|x_i, ..., x_n) = \sum_{i=1}^{n} [x_i \ln p + (1 - x_i) \ln(1-p)] \tag{1.106}$$

The MLE of $p = \hat{p}$ and is obtained from

$$\frac{d\, l(p)}{dp} = 0 \qquad \Rightarrow \qquad \frac{\sum x_i}{\hat{p}} - \frac{\sum (1 - x_i)}{1 - \hat{p}} = 0$$

The answer was obvious. MLE does this quite often: it gives us intuitive answers. However as the next example shows, MLE could be biased.

which can be rearranged to give

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1.107}$$

MLE of $\mathcal{N}(\mu, \sigma^2)$.

Let $x_1, ..., x_n$ be i.i.d. samples from $\mathcal{N}(\mu, \sigma^2)$. Then

Notation: i.i.d. variables = independent, identically distributed variables. Example of i.i.d.: successive tosses of a coin with a Bernoulli variable describing the result of each coin.

$$L\left((\mu, \sigma^2)'|x_1, ..., x_n\right) = \prod_{i=1}^{n} \frac{e^{-(x_i-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}} = \frac{1}{\sigma^n(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2\right] \tag{1.108}$$

$$l = \ln L = -n\ln\sigma - \frac{n}{2}\ln 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \tag{1.109}$$

To get ML estimates of μ and σ^2, set

$$\frac{\partial l}{\partial \mu} = 0 \qquad\qquad \frac{\partial l}{\partial \sigma^2} = 0 \tag{1.110}$$

$$\begin{aligned}
\frac{\partial l}{\partial \mu} = 0 &= \frac{\partial\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)}{\partial\mu} = -\frac{1}{2\sigma^2}\frac{\partial\left((x_1-\mu)^2+(x_2-\mu)^2+...+(x_n-\mu)^2\right)}{\partial\mu} \\
&= -\frac{1}{2\sigma^2}(-1)\left[2(x_1-\mu)+2(x_2-\mu)+...+2(x_n-\mu)\right] \\
0 &= \frac{1}{\sigma^2}[x_1+x_2+...+x_n-n\mu] = \frac{n}{\sigma^2}[\bar{x}-\mu]
\end{aligned}$$

$$\mu_{\mathrm{ML}} = \hat{\mu} = \bar{x} \tag{1.111}$$

$$\frac{\partial l}{\partial \sigma^2} = 0 = -\frac{n}{2\sigma^2} + \frac{\partial\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)}{\partial\sigma^2}$$

$$\frac{\partial\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)}{\partial\sigma^2} = \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\begin{aligned}
\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 &= \frac{1}{2\sigma^4}\left[\sum_{i=1}^{n}\left[(x_i-\bar{x})+(\bar{x}-\mu)\right]^2\right] \\
&= \frac{1}{2\sigma^4}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2 + 2((\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\bar{x})\right] \\
&= \frac{1}{2\sigma^4}\left[nS^2 + n(\bar{x}-\mu)^2\right] \qquad \left[\text{where } S^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}\right] \\
\frac{\partial l}{\partial\sigma^2} = 0 = -\frac{n}{2\sigma^2} + \frac{\partial\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)}{\partial\sigma^2} &= -\frac{n}{2\sigma^2} + \frac{n}{2\sigma^4}\left[S^2+(\bar{x}-\mu)^2\right]
\end{aligned}$$

which simplifies down to and suggests a possible estimate for σ^2:

$$\widehat{\sigma^2} = S^2 + (\bar{x}-\mu)^2$$

But $\bar{x} = \mu$ from before, and hence a point estimate of σ^2 in the maximum likelihood sense is

MLE of σ^2 is biased.

$$\sigma^2_{\mathrm{ML}} = S^2 \tag{1.112}$$

MLE of $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

We are given the n independent observations $\{\mathbf{x}_n\}$, each in p dimensions. The log likelihood function then is

$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^\mathsf{T}$.

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{np}{2}\ln(2\pi) - \frac{n}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) \tag{1.113}$$

The RHS may be seen to depend on the two sufficient statistics $\sum_{i=1}^{b}\mathbf{x}_i$ and $\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}$. The derivative of this log likelihood function w.r.t. $\boldsymbol{\mu}$ is

$$\frac{\partial}{\partial\boldsymbol{\mu}}\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) \tag{1.114}$$

and if this is set to zero, we get the sample arithmetic mean.

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \tag{1.115}$$

Maximization w.r.t $\boldsymbol{\Sigma}$ would give

$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_i - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathsf{T}} \tag{1.116}$$

which is biased:

$$\mathcal{E}\left[\boldsymbol{\mu}_{\mathrm{ML}}\right] = \boldsymbol{\mu} \qquad \mathcal{E}\left[\boldsymbol{\Sigma}_{\mathrm{ML}}\right] = \frac{n-1}{n}\boldsymbol{\Sigma} \tag{1.117}$$

An unbiased estimator would therefore be

$$\tilde{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_i - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathsf{T}} \tag{1.118}$$

MLE: Summary.
Given a dataset D made up of n values of x_i,

$$p(D|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta}) \tag{1.119}$$

MLE of $\boldsymbol{\theta} = \hat{\theta}$ = value that maximizes $P(D|\theta)$.

$$\theta = \underset{\theta}{\operatorname{argmax}}\; l(\theta) \tag{1.120}$$

As more samples become available, the likelihood would be very narrow, and in the limit becomes a Dirac delta. To determine MLE, set

Careful: the solution may be a maximum, minimum, inflection point, or the boundary of the parameter space.

$$\nabla_{\boldsymbol{\theta}} l = \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \log p(x_i|\boldsymbol{\theta}) = 0 \tag{1.121}$$

If the assumed model is wrong, then any attempted classification will be off. Example: $\mathcal{N}(\mu, 1)$ will classify differently compared to $\mathcal{N}(\mu, 10)$.

1.7 Maximum a posteriori estimation (MAP)

If we do not know any better, then every model is equiprobable and $p(M_i) = p(M_j), \quad \forall\, i, j \in \mathbf{M}$. The prior is uniform and uninformative. In such cases

$$\begin{aligned} M_{MAP} &= \underset{M\in\mathbf{M}}{\operatorname{argmax}}\; p(\mathbf{D}|M) \\ &= M_{ML} \end{aligned}$$

The most probable hypothesis is the *maximum a posteriori* model.

$$M_{MAP} = \underset{M\in\mathbf{M}}{\operatorname{argmax}}\; p(M|\mathbf{D}) = \underset{M\in\mathbf{M}}{\operatorname{argmax}}\; \frac{p(\mathbf{D}|M)p(M)}{p(\mathbf{D})} = \underset{M\in\mathbf{M}}{\operatorname{argmax}}\; p(\mathbf{D}|M)p(M) \tag{1.122}$$

since $p(\mathbf{D})$ is constant and independent of M. As before, if a model is characterized by parameters $\boldsymbol{\theta}$, then

$$\boldsymbol{\theta}_{MAP} = \underset{M\in\mathbf{M}}{\operatorname{argmax}}\; p(\boldsymbol{\theta}|D) = \underset{M\in\mathbf{M}}{\operatorname{argmax}}\; p(D|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{1.123}$$

Further, taking the logarithm, and writing the data set D as a product of training samples, a necessary requirement for MAP is

MAP finds the peak (mode) of the posterior density.

$$\frac{\partial}{\partial\boldsymbol{\theta}} \ln\left(\prod_{k=1}^{n} p(x_k|\boldsymbol{\theta})p(\boldsymbol{\theta})\right) = 0 \tag{1.124}$$

MAP of $\mathcal{N}(\mu, \sigma^2)$:

Consider a univariate distribution $p(x|\mu) \sim \mathcal{N}(\mu, \sigma^2)$. Prior knowledge of μ may be expressed as $p(\mu) \sim \mathcal{N}(\mu_0, \sigma_0^2)$ where μ_0 and σ_0 are known. Let D = data set of n samples: $D = \{x_1, ..., x_n\}$

This may be the best prior guess.

$$p(\mu|D) = \frac{p(D|\mu)p(\mu)}{\int p(D|\mu)p(\mu)d\mu} = \alpha \prod_{k=1}^{n} p(x_k|\mu)p(\mu) \tag{1.125}$$

where α = normalization factor and is a $f(D)$ but not $f(\mu)$. As stated before, decomposing D into a product assumes that the training samples are statistically independent. But $p(x_k|\mu) \sim \mathcal{N}(\mu, \sigma^2)$ and $p(\mu) \sim \mathcal{N}(\mu_0, \sigma_0^2)$.

$$\begin{aligned} p(\mu|D) &= \alpha \prod_{k=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_k - \mu}{\sigma}\right)^2\right] \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right] \\ &= \alpha' \exp\left[-\frac{1}{2}\left(\sum_{k=1}^{n}\left(\frac{\mu - x_k}{\sigma}\right)^2 + \left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right)\right] \\ &= \alpha'' \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{i=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right] \end{aligned} \tag{1.126}$$

But this is itself in the form of a normal distribution in terms of posterior parameters (μ_n, σ_n^2), i.e.

$$p(\mu|D) \sim \mathcal{N}(\mu_n, \sigma_n^2) = \frac{1}{\sigma_n\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_n}{\sigma_n}\right)^2\right] \tag{1.127}$$

The coefficients of μ^2 and μ inside the exponent of Eq. 1.127 above are $1/\sigma_n^2$ and $-2\mu_n/\sigma_n^2$ respectively. On comparison to the coefficients in Eq. 1.126, we get

The inverse of a variance is called 'precision'. So it is convenient to remember that the precision of a posterior equals the sum of precisions of a prior and the observed data. The posterior mean is the precision weighted sum of the means of the prior and data set.

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \tag{1.128}$$

$$\frac{\mu_n}{\sigma_n^2} = \frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_0}{\sigma_0^2} \tag{1.129}$$

$$\hat{\mu}_n = \bar{x}_n = \frac{1}{n}\sum_{k=1}^{n} x_k \tag{1.130}$$

The updated values now are

Notice the convenient trick employed, based on coefficients of μ^2 and μ which we have used to derive an updated (posterior) distribution. Notice also that it is not necessary to precisely derive (and keep track) of the coefficients α, α' and α''.

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right)\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 \tag{1.131}$$

$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2} \tag{1.132}$$

Implementing the MAP approach now involves taking the logarithm of the posterior distribution $\mathcal{N}(\mu_n, \sigma_n^2)$, and setting its derivative w.r.t. μ to 0. Clearly, this would result in

The MAP estimator of the mean is conveniently = μ_n.

$$\mu_{MAP} = \mu_n \tag{1.133}$$

The following may be observed from the posterior normal distribution.

- As n increases, $\mathcal{N}(\mu_n, \sigma_n^2)$ tends to a Dirac delta.
- Note that $\mu_n = c_1\hat{\mu}_n + c_2\mu_0$ where $c_1 + c_2 = 1$ and c_1,$c_2 > 0$. Hence μ_n lies between μ_0 and $\hat{\mu}_n$.
- If $\sigma_0 \neq 0$, then $\mu_n \longrightarrow \hat{\mu}_n$ as $n \longrightarrow \infty$.
- If $\sigma_0 = 0$, then the prior info ($\mu = \mu_0$) is so strong that no further observations result in a change in belief.
- If $\sigma_0 \gg \sigma$, then $\mu_n = \hat{\mu}_n$ and hence prior info is very uncertain, and is not of much use.
- If σ^2/σ_0^2 is not infinite, then after enough sampling, μ_n converges.

MAP of $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:
The prior is

$$p(\boldsymbol{\mu}) \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \tag{1.134}$$

and the likelihood of a new observation is

$$p(\mathbf{x}|\boldsymbol{\mu}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{1.135}$$

Given the n new observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, we get

$$p(\boldsymbol{\mu}|D) = \alpha \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\mu}) p(\boldsymbol{\mu}) = \alpha^{'} \exp\left[-\frac{1}{2}\left(\boldsymbol{\mu}^{\mathsf{T}}(n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1})\boldsymbol{\mu} - 2\boldsymbol{\mu}^{\mathsf{T}}\left(\boldsymbol{\Sigma}^{-1}\sum_{k=1}^{n}\mathbf{x}_k + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)\right)\right] \tag{1.136}$$

which is of the form

$$p(\boldsymbol{\mu}|D) = \alpha^{''} \exp\left[-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_n)^{\mathsf{T}}\boldsymbol{\Sigma}_n^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_n)\right] \tag{1.137}$$

Then, using $\hat{\boldsymbol{\mu}}_n = \sum_{k=1}^{n} \mathbf{x}_k / n$ gives

$$\boldsymbol{\Sigma}_n^{-1} = n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1} \tag{1.138}$$

$$\boldsymbol{\Sigma}_n^{-1}\boldsymbol{\mu}_n = n\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\mu}}_n + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 \tag{1.139}$$

These expressions can be simplified using the matrix identity (valid for a pair of nonsingular matrices of similar square dimensions)

$$(\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1} = \mathbf{A}(\mathbf{A} + \mathbf{B})^{-1}\mathbf{B} = \mathbf{B}(\mathbf{A} + \mathbf{B})^{-1}\mathbf{A} \tag{1.140}$$

to get

Compare these expressions to Eqs. 1.131 and 1.132 for the univariate case.

$$\boldsymbol{\mu}_n = \boldsymbol{\Sigma}_0\left(\boldsymbol{\Sigma}_0 + \frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\hat{\boldsymbol{\mu}}_n + \frac{1}{n}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}_0 + \frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\mu}_0 \tag{1.141}$$

$$\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}_0\left(\boldsymbol{\Sigma}_0 + \frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\frac{1}{n}\boldsymbol{\Sigma} \tag{1.142}$$

Example 1.2. MLE vs. MAP example 1. Let M_1 = patient has HIV, and M_2 = patient does not have HIV. A diagnostic kit for HIV detection is available and on use, identifies a result as positive (+) or negative (−). Prior information is available: only 0.8% of the population is infected with HIV. The kit is imperfect: It correctly identifies HIV$^+$ individuals 98% of the time, and correctly classifies healthy individuals 97% of the time. A new patient tests positive. Should we diagnose that patient as HIV$^+$? The data provided is summarized as

¬ is the negation symbol. So ¬HIV$^+$ reads 'not HIV$^+$'.

$$\begin{array}{ll} p(\text{HIV}^+) = 0.008 & p(\neg\text{HIV}^+) = 0.992 \\ p(+|\text{HIV}^+) = 0.98 & p(-|\text{HIV}^+) = 0.02 \\ p(-|\neg\text{HIV}^+) = 0.97 & p(+|\neg\text{HIV}^+) = 0.03 \end{array}$$

ML involves comparing $p(+|\text{HIV}^+) = 0.98$ versus $p(+|\neg\text{HIV}^+) = 0.03$. Hence the patient would be considered HIV$^+$. MAP involves comparing $p(\text{HIV}^+|+)$ versus $p(\neg\text{HIV}^+|+)$. Using Bayes theorem and ignoring $p(+)$ which would be a common term in the denominator

$M_{ML} = \text{HIV}^+$.

$$p(\text{HIV}^+|+) = \frac{p(+|\text{HIV}^+)p(\text{HIV}^+)}{p(+)} \quad \Longrightarrow \quad p(+|\text{HIV}^+)p(\text{HIV}^+) = 0.98 \times 0.008 = 0.0078$$

$$p(\neg\text{HIV}^+|+) = \frac{p(+|\neg\text{HIV}^+)p(\neg\text{HIV}^+)}{p(+)} \quad \Longrightarrow \quad p(+|\neg\text{HIV}^+)p(\neg\text{HIV}^+) = 0.03 \times 0.992 = 0.0298$$

Hence the patient should be considered HIV$^-$. Note that

$M_{MAP} = \neg\text{HIV}^+ = \text{HIV}^-$.

$$p(+) = p(+|\text{HIV}^+)p(\text{HIV}^+) + p(+|\neg\text{HIV}^+)p(\neg\text{HIV}^+) = 0.0078 + 0.0298 = 0.0376$$

Flipping a coin would have been more accurate!

Priors have a very strong influence on the results.

Therefore

$$p(\text{HIV}^+|+) = \frac{0.0078}{0.0376} = 0.21$$

Note that $p(\text{model M= HIV}^+|\text{data D = kit says +}) = 0.21 \quad \gg \quad p(\text{HIV}) = 0.008$. Yet not being HIV^+ is the preferred model according to MAP.

1.8 Bayesian estimation for a Binomial

$\theta = \pi$.

The binomial is of the form

$$p(x|\pi) = {}^nC_x \pi^x (1-\pi)^{n-x} \tag{1.143}$$

This is normally evaluated for different x $(x = 1, .., n)$ but also may be evaluated at constant n, x for different π. Bayes' theorem then implies that

Try to not get confused by the 'p's and the $\pi's$. A π inside a bracket or another term refers to the proportion.

$$p(\pi|x) = \frac{p(x|\pi)p(\pi)}{\int_0^1 p(x|\pi)p(\pi)d\pi} \tag{1.144}$$

Clearly the choice of $p(\pi)$ decides whether there exists a closed form to the integral or whether numerical integration will have to be performed. Two simple choices are discussed below.

Uniform prior:

If we wish that no personal belief influence estimation, then giving equal weightage to all probability values implies that $p(\pi) = 1$ for $0 \leq \pi \leq 1$. Obviously the posterior is identical to the likelihood.

Beta prior:

The Beta distribution has as its pdf

Unlike the binomial, a and b need not be integers.

$$\text{Beta}(y|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1} \qquad 0 \leq y \leq 1 \tag{1.145}$$

When $a = b = 1$, Beta(y;1,1) = Uniform distribution.

$\Gamma(a+1) = a\Gamma(a)$ in general and when a is an integer, using this relationship recursively gives us $\Gamma(a+1) = a!$. It is important to appreciate that by varying a and b, the pdf could attain various shapes. This is particularly useful when we look for a pdf of an appropriate shape (based on any existing insight) to represent our beliefs in different models (i.e. to represent the prior $p(\theta)$).

$$\mathcal{E}[x] = \int_0^1 x f(x) dx = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 x^a (1-x)^{b-1} dx \tag{1.146}$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \int_0^1 \frac{\Gamma(a+b+1)}{\Gamma(a+1)\Gamma(b)} x^a (1-x)^{b-1} dx$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} = \frac{a}{a+b} \tag{1.147}$$

Similarly

$$\mathcal{E}[x^2] = \frac{a(a+1)}{(a+b+1)(a+b)}$$

$$\text{Var}[x] = \frac{ab}{(a+b)^2(a+b+1)} \tag{1.148}$$

Suppose that $p(\pi)$ is according to a Beta prior

$$p(\pi; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \pi^{a-1}(1-\pi)^{b-1} \qquad 0 \leq \pi \leq 1 \tag{1.149}$$

Therefore the posterior may be written as

We add successes together to get $a + x$, and failures together to get $b+n-x$. Easy!

$$p(\pi|x) \propto p(x|\pi)p(\pi) = \pi^{a+x-1}(1-\pi)^{b+n-x-1} \qquad 0 \le \pi \le 1 \tag{1.150}$$

This itself is a Beta$(a+x, b+n-x)$ distribution and therefore the constants can quickly be identified without resorting to integration:

$$p(\pi|x) = \frac{\Gamma(n+a+b)}{\Gamma(x+a)\Gamma(n-x+b)}\pi^{x+a-1}(1-\pi)^{n-x+b-1} \tag{1.151}$$

Any of several similarly shaped priors may be chosen: they will all give similar posteriors. In the event that the prior mean (π_0) and the prior standard deviation (σ_0) are known, the identities developed earlier may be used:

$$\pi_0 = \frac{a}{a+b}, \qquad \sigma_0 = \sqrt{\frac{ab}{(a+b)^2(a+b+1)}} = \sqrt{\frac{\pi_0(1-\pi_0)}{a+b+1}} \tag{1.152}$$

Obviously the values of a and b (and hence Beta(a, b) may now be obtained in terms of π_0 and σ_0. Notice also that $a+b+1 = n_{eq}$, the equivalent sample size. The MAP estimate of the proportion of successes therefore is

Check to see if that n_{eq} does not exceed a common-sense value.

$$\mathcal{E}[\pi] = \frac{a+x}{(a+x)+(b+n-x)} = \frac{a+x}{a+b+n} \tag{1.153}$$

Interval for x. For the sampling variable $\bar{x}$, replace s' by $s'/\sqrt{n}$.

The mode of a Beta(a, b) distribution may be obtained on setting the derivative of $p(\pi|x)$ w.r.t. p to zero. The mode is $(a-1)/(a+b-2)$. This would be a MAP estimate. Remember also the credible interval for π that Bayesian hypothesis testing would utilize: the $(1-\alpha)\times 100\%$ credible region is $m' \pm z_{1-\alpha/2}s'$.

1.9 Extending the Binomial distribution

Multinomial distributions

$\mathbf{x}$ is of the form $[0, 0, 1, 0, 0, ...]^\mathsf{T}$. Obviously only one x_k value is nonzero at a time and is equal to 1. Clearly $\sum x_k = 1$.

The Bernoulli distribution can be generalized as follows:

$$P(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} \tag{1.154}$$

where μ_k is the probability of outcome k such that $\mu_k \ge 0$ and $\sum \mu_k = 1$. The distribution is a PMF: $\sum P(\mathbf{x}|\boldsymbol{\mu}) = \sum \mu_k = 1$.

$$\mathcal{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_x P(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = [\mu_1, \mu_2, \ldots, \mu_m]^\mathsf{T} = \boldsymbol{\mu} \tag{1.155}$$

Given data $\mathbf{D}$ consisting of the n independent observations $x_1, \ldots, x_n$,

$$P(\mathbf{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{\sum_n x_{nk}} = \prod_{k=1}^{K} \mu_k^{m_k} \tag{1.156}$$

where $m_k = \sum x_{nk}$ is the number of observations of $x_k = 1$.

The maximum likelihood estimate of $\boldsymbol{\mu}$ can be found by maximizing $\ln P(\mathbf{D}|\boldsymbol{\mu})$ w.r.t. μ_k. Differentiating the following expression w.r.t. μ_k and setting to zero

We use a Lagrange multiplier to facilitate the constrained maximization.

$$\max\left[\sum_{k=1}^{K} m_k \ln \mu_k + \lambda\left(\sum_{k=1}^{K}\mu_k - 1\right)\right] \qquad \Rightarrow \qquad \mu_k = -\frac{m_k}{\lambda} \tag{1.157}$$

But $\sum \mu_k = 1$ which gives $\lambda = -\sum_k m_k = -N$ and hence

$$\mu_k^{\mathrm{ML}} = \frac{m_k}{N} \tag{1.158}$$

The multinomial distribution is

$$\binom{N}{m_1 m_2 \cdots m_k} = \frac{N!}{m_1! m_2! \cdots m_k!} \qquad \text{mult}(m_1, m_2, \ldots, m_k | \boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \cdots m_k} \prod_{k=1}^{K} \mu_k^{m_k} \tag{1.159}$$

Dirichlet distribution

The Dirichlet distribution is the multivariate generalization of the Beta distribution (and the continuous counterpart of the multinomial), and is as follows:

$\alpha_0 = \sum \alpha_k$.

$$\text{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_k)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \tag{1.160}$$

Recall that we had identified how a binomial likelihood and a Beta prior combine to generate a Beta posterior distribution. Similarly, given a Dirichlet prior, and a multinomial likelihood, the posterior is

$\sum m_k = N$

$$P(\boldsymbol{\mu}|\mathbf{D}, \boldsymbol{\alpha}) \propto P(\mathbf{D}|\boldsymbol{\mu}) P(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1} \tag{1.161}$$

.

which is also a Dirichlet distribution. This

$$P(\boldsymbol{\mu}|\mathbf{D}) = \text{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha} + \mathbf{m}) = \frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + m_1)\cdots\Gamma(\alpha_k + m_k)} \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1} \tag{1.162}$$

The MAP estimates of the parameters are

$$\mu_k^{\text{MAP}} = \frac{\alpha_k + m_k}{\sum_{k=1}^{K}(\alpha_k + m_k)} \tag{1.163}$$

1.10 Bayesian estimation

Given prior probabilities of classes $w_i = p(w_i)$, and their class-conditional densities, $p(\mathbf{x}|w_i)$, we wish to compute posterior probabilities $p(w_i|\mathbf{x})$. Let D = dataset of samples. Then

$p(w_i|D)$ may be written as $p(w_i)$.

$$p(w_i|\mathbf{x}, D) = \frac{p(\mathbf{x}|w_i, D)p(w_i|D)}{\sum_{j=1}^{c} p(\mathbf{x}|w_j, D)p(w_j|D)} = \frac{p(\mathbf{x}|w_i, D)p(w_i)}{\sum_{j=1}^{c} p(\mathbf{x}|w_j)p(w_j|D_j)} \tag{1.164}$$

where D_j is the data set associated with class w_j. Usually $p(\mathbf{x})$ is unknown. When fitting a model to it with parameters $\boldsymbol{\theta}$, we have $p(\mathbf{x}|\boldsymbol{\theta})$. Some info on $\boldsymbol{\theta}$ may be available *prior* to observing samples, and hence a prior density is $p(\boldsymbol{\theta})$. Given new samples, we get $p(\boldsymbol{\theta}|D)$ which hopefully is more sharply peaked about the true value of $\boldsymbol{\theta}$. Then

$p(\mathbf{x}|D)$ is a predictive distribution. In process data analysis we would want to know

$p(\mathbf{x}_{n+1}|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$

$$p(\mathbf{x}|D) = \int p(\mathbf{x}, \boldsymbol{\theta}|D) d\boldsymbol{\theta} \tag{1.165}$$

Using Bayes' rule,

$$p(\mathbf{x}, \boldsymbol{\theta}|D) = p(\mathbf{x}|\boldsymbol{\theta}, D)p(\boldsymbol{\theta}|D) = p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|D) \tag{1.166}$$

$$p(\mathbf{x}|D) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|D) d\boldsymbol{\theta} \tag{1.167}$$

The two terms $p(\mathbf{x}|\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|D)$ need to be determined on assuming a particular distribution type (pdf).

As stated before, MLE and MAP compute a specific estimate of $\boldsymbol{\theta}$.

MAP and Bayesian estimation differ specifically in how $p(\boldsymbol{\theta}|D)$ is utilized. MAP looks for unique $\boldsymbol{\theta}$ values which maximize $p(\boldsymbol{\theta}|D)$. Bayesian estimation on the other hand focusses on obtaining $p(\mathbf{x}|D)$ (which if you think about it is what we really want from a predictive perspective) by integrating over $\boldsymbol{\theta}$.

Bayesian estimation for a univariate Gaussian

We need to find $p(\boldsymbol{\theta}|D)$ and $p(\mathbf{x}|D)$ when $p(\mathbf{x}|\boldsymbol{\theta}) \sim \mathcal{N}(\mu, \Sigma)$.

Evaluating $p(\boldsymbol{\theta}|D)$:

Consider a univariate distribution $p(x|\mu) \sim \mathcal{N}(\mu, \sigma^2)$. Prior knowledge of μ may be expressed as $p(\mu) \sim \mathcal{N}(\mu_0, \sigma_0^2)$ where μ_0 and σ_0 are known. Let D = data set of n samples: $D = \{x_1, ..., x_n\}$

This is exactly what we did with MAP estimation of a Normal. However we do not bother setting the derivative of the posterior to zero. So instead of a unique parameter estimate, we continue with a distribution of parameter values.

$$p(\mu|D) = \frac{p(D|\mu)p(\mu)}{\int p(D|\mu)p(\mu)d\mu} = \alpha \prod_{k=1}^{n} p(x_k|\mu)p(\mu) \tag{1.168}$$

where α = normalization factor and is a $f(D)$ but not $f(\mu)$. As stated before, decomposing D into a product assumes that the training samples are statistically independent. But $p(x_k|\mu) \sim \mathcal{N}(\mu, \sigma^2)$ and $p(\mu) \sim \mathcal{N}(\mu_0, \sigma_0^2)$.

$$\begin{aligned} p(\mu|D) &= \alpha \prod_{k=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_k - \mu}{\sigma}\right)^2\right] \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right] \\ &= \alpha' \exp\left[-\frac{1}{2}\left(\sum_{k=1}^{n}\left(\frac{\mu - x_k}{\sigma}\right)^2 + \left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right)\right] \\ &= \alpha'' \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{i=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right] \end{aligned}$$

But this is itself a normal distribution. i.e.

$$p(\mu, D) \sim \mathcal{N}(\mu_n, \sigma_n^2) = \frac{1}{\sigma_n\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_n}{\sigma_n}\right)^2\right] \tag{1.169}$$

Hence

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \tag{1.170}$$

$$\frac{\mu_n}{\sigma_n^2} = \frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_0}{\sigma_0^2} \tag{1.171}$$

$$\hat{\mu}_n = \bar{x}_n = \frac{1}{n}\sum_{k=1}^{n} x_k \tag{1.172}$$

The updated values now are

The observations made about the posterior distribution in the MAP section obviously hold true here.

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right)\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 \tag{1.173}$$

$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2} \tag{1.174}$$

Evaluating $p(\mathbf{x}|D)$:

$$\begin{aligned} p(\mathbf{x}|D) &= \int p(x|\mu)p(\mu|D)d\mu \\ &= \int \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right] \frac{1}{\sigma_n\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_n}{\sigma_n}\right)^2\right] d\mu \\ &= \frac{1}{2\pi\sigma\sigma_n} \exp\left[-\frac{1}{2}\frac{(x - \mu_n)^2}{\sigma^2 + \sigma_n^2}\right] f(\sigma, \sigma_n) \end{aligned} \tag{1.175}$$

$$f(\sigma, \sigma_n) = \int \exp\left[-\frac{1}{2}\frac{\sigma^2 + \sigma_n^2}{\sigma^2\sigma_n^2}\left(\mu - \frac{\sigma_n^2 x + \sigma^2\mu_n}{\sigma^2 + \sigma_n^2}\right)^2\right] d\mu \tag{1.176}$$

Hence

$$p(\mathbf{x}|D) \propto \exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right] \sim \mathcal{N}(\mu_n, \sigma^2+\sigma_n^2) \tag{1.177}$$

Simple!

To get $p(x|D)$ from $p(x|\mu)$ which is $\sim \mathcal{N}(\mu, \sigma^2)$, replace μ by μ_n and σ^2 by $\sigma^2+\sigma_n^2$. Usually the prior mean is known or guessed. It is more difficult to intuit a prior variance (σ_0^2). In that case, evaluate n_{eq} from $\sigma_0^2 = \sigma^2/n_{eq}$. If n_{eq} is large it implies that a very strong prior belief exists about μ and a lot of 'contrary' data is needed to shake that belief.

Bayesian estimation for a multivariate Gaussian

For a multivariate Gaussian, we need to perform the integration

$$p(\mathbf{x}|D) = \int p(\mathbf{x}|\boldsymbol{\mu})p(\boldsymbol{\mu}|D)d\boldsymbol{\mu} \tag{1.178}$$

which would finally give (using Eqns. 1.141 and 1.142)

$$p(\mathbf{x}|D) \sim \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}+\boldsymbol{\Sigma}_n) \tag{1.179}$$

This may be interpreted as follows: $\mathbf{x}$ is the sum of two mutually independent random variables, one of which is $\boldsymbol{\mu}$ with $p(\boldsymbol{\mu}|D) \sim \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$, and an independent random vector $\mathbf{y}$ with $p(\mathbf{y}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Then since these are independent and normally distributed, their means and covariances must add up.

Duda et al. Chapter 3, example 1.

Example 1.3. An example to demonstrate recursive Bayesian learning. We make measurements of x where we know only that x is bounded by θ ($0 \le x \le \theta$) where θ itself is unknown. We assume that $0 \le \theta \le 10$. We also assume that the prior distribution for θ is flat.

$$p(\theta) \sim U(0, 10) \qquad 0 \le \theta \le 10$$

$$p(x|\theta) \sim U(0, \theta) = \frac{1}{\theta} \qquad 0 \le x \le \theta$$

4 successive measurements are made = {4, 7, 2, 8}. When we see the first point (4), we learn that $\theta \ge 4$. Hence

$$p(\theta|D^1 = 4) \propto p(x|\theta)p(\theta|D^0) \quad \propto \quad \frac{1}{\theta} \qquad 4 \le \theta \le 10$$

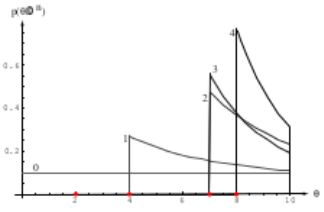

Fig. 1.3: The posterior as a compromise.

When the next point (7) is seen, the interval changes again

$$p(\theta|D^2 = 7) \propto p(x|\theta)p(\theta|D^1) \quad \propto \quad \frac{1}{\theta^2} \qquad 7 \le \theta \le 10$$

This definition of θ as an upper bound is unusual. We are more familiar with it as the expectation of x, variance of x, or as a proportion p.

The general form of the solution is

$$p(\theta|D^n) \propto \frac{1}{\theta^n} \qquad \max_x [D^n] \le \theta \le 10$$

1.11 Estimation for the Exponential family of distributions

The exponential family of distributions is of the form

$$p(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x})g(\boldsymbol{\theta}) \exp\left[\boldsymbol{\theta}^\mathsf{T}\mathbf{u}(\mathbf{x})\right] \tag{1.180}$$

$g(\boldsymbol{\theta})$ ensures normalization:

$$g(\boldsymbol{\theta}) \int h(\mathbf{x}) \exp\left[\boldsymbol{\theta}^\mathsf{T}\mathbf{u}(\mathbf{x})\right] d\mathbf{x} = 1 \tag{1.181}$$

The Bernoulli is an exponential:

$$p(x|\mu) = \mu^x(1-\mu)^{1-x} = \exp\left[x\ln\mu + (1-x)\ln(1-\mu)\right]$$
$$= (1-\mu)\exp\left[\ln\left(\frac{\mu}{1-\mu}\right)x\right] \tag{1.182}$$

Then

$$\theta = \ln\left(\frac{\mu}{1-\mu}\right) \tag{1.183}$$

which can be arranged to get the logistic sigmoid function:

$$\mu = \sigma(\theta) = \frac{1}{1+\exp(-\theta)} \tag{1.184}$$

using $\sigma(-\theta) = 1 - \sigma(\theta)$

$$p(x|\theta) = \sigma(-\theta)\exp(\theta x) \tag{1.185}$$

$$u(x) = x \qquad h(x) = 1 \qquad g(\theta) = \sigma(-\theta) \tag{1.186}$$

The multinomial is an exponential

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{m}\mu_k^{x_k} = \exp\left(\sum_{k=1}^{m}x_k\ln\mu_k\right) \tag{1.187}$$

$$\mathbf{u}(\mathbf{x}) = \mathbf{x} \qquad h(\mathbf{x}) = 1 \qquad g(\boldsymbol{\theta}) = 1 \tag{1.188}$$

The parameters θ_k are not independent because μ_k are subject to $\sum\mu_k = 1$. Hence, we eliminate μ_m and operate with $m-1$ parameters.

$$0 \le \mu_k \le 1 \qquad \sum_{k=1}^{m-1}\mu_k \le 1 \tag{1.189}$$

$$\exp\left(\sum_{k=1}^{m}x_k\ln\mu_k\right) = \exp\left[\textstyle\sum_{k=1}^{m-1}x_k\ln\mu_k + \left(1-\sum_{k=1}^{m-1}x_k\right)\ln\left(1-\sum_{k=1}^{m-1}\mu_k\right)\right]$$
$$= \exp\left[\textstyle\sum_{k=1}^{m-1}x_k\ln\left(\frac{\mu_k}{1-\sum_{j=1}^{m-1}\mu_j}\right) + \ln\left(1-\sum_{k=1}^{m-1}\mu_k\right)\right] \tag{1.190}$$

Therefore, by inspection

$$\ln\left(\frac{\mu_k}{1-\sum_j\mu_j}\right) = \theta_k \tag{1.191}$$

Summing this over k and rearranging gives

The softmax function; also known as the normalized exponential.

$$\mu_k = \frac{\exp(\theta_k)}{1+\sum_j\exp(\theta_j)} \tag{1.192}$$

$$p(\mathbf{x}|\boldsymbol{\theta}) = \left(1+\sum_{k=1}^{m-1}\exp(\theta_k)\right)^{-1}\exp(\boldsymbol{\theta}^{\mathsf{T}}\mathbf{x}) \tag{1.193}$$

$$\mathbf{u}(\mathbf{x}) = \mathbf{x} \qquad h(\mathbf{x}) = 1 \qquad g(\boldsymbol{\theta}) = \left(1+\sum_{k=1}^{m-1}\exp(\theta_k)\right)^{-1} \tag{1.194}$$

The univariate Gaussian as an exponential

$$\boldsymbol{\theta} = \begin{bmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{bmatrix} \tag{1.195}$$

$$\mathbf{u}(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix} \qquad h(x) = (2\pi)^{-1/2} \qquad g(\boldsymbol{\theta}) = (-2\theta_2)^{1/2} \exp\left(\frac{\theta_1^2}{4\theta_2}\right) \tag{1.196}$$

ML and the exponential

To determine the MLE of $\boldsymbol{\theta}$, take the gradient of the exponential w..r.t. $\boldsymbol{\theta}$

$$\nabla g(\boldsymbol{\theta}) \int h(\mathbf{x}) \exp\left(\boldsymbol{\theta}^\mathsf{T} \nu(\mathbf{x})\right) d\mathbf{x} + g(\boldsymbol{\theta}) \int h(\mathbf{x}) \exp\left(\boldsymbol{\theta}^\mathsf{T} \mathbf{u}(\mathbf{x})\right) \mathbf{u}(\mathbf{x}) d\mathbf{x} = 0 \tag{1.197}$$

Rearranging,

$$-\frac{1}{g(\boldsymbol{\theta})} \nabla g(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) \int h(\mathbf{x}) \exp\left(\boldsymbol{\theta}^\mathsf{T} \mathbf{u}(\mathbf{x})\right) \mathbf{u}(\mathbf{x}) d\mathbf{x} = \mathcal{E}\left[\mathbf{u}(\mathbf{x})\right] \tag{1.198}$$

which gives

$$-\nabla \ln g(\boldsymbol{\theta}) = \mathcal{E}\left[\mathbf{u}(\mathbf{x})\right] \tag{1.199}$$

The moments of an exponential can be found by differentiation.

Similarly, the covariance and other higher moments can be expressed in terms of second and higher derivatives. When a data set D consists of the n points,

$$p(D|\boldsymbol{\theta}) = \left(\prod_{k=1}^{n} h(x_k)\right) g(\boldsymbol{\theta})^n \exp\left(\boldsymbol{\theta}^\mathsf{T} \sum_{k=1}^{n} \mathbf{u}(x_k)\right) \tag{1.200}$$

which gives

$$-\nabla \ln g(\boldsymbol{\theta}_{\mathrm{ML}}) = \frac{1}{n} \sum_{k=1}^{n} \mathbf{u}(x_k) \tag{1.201}$$

which can usually be solved for $\boldsymbol{\theta}_{\mathrm{ML}}$. The solution depends only on $\sum \mathbf{u}(x_k)$ which is therefore called the sufficient statistic. For a Bernoulli, $\mathbf{u}(x)$ is x and we need to keep the sum of x_k. For a Gaussian, $\mathbf{u}(x) = [x, x^2]^\mathsf{T}$ and we need to keep track of the sums of x_k and x_k^2.

1.12 Other estimators

Other approaches to estimating model parameters include MMSE (minimum mean square estimation, also called minimum variance estimation) and maximum entropy estimation.

Minimum mean square estimation

The connection to a variance is obvious.

The minimum mean square estimator is of the form

$$\boldsymbol{\theta}_{MMSE} = \underset{\hat{\boldsymbol{\theta}}}{\operatorname{argmin}} \left(\int_{\boldsymbol{\theta}} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\mathsf{T} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) p(\boldsymbol{\theta}|D) d\boldsymbol{\theta} \right) \tag{1.202}$$

where $\boldsymbol{\theta}$ is an m dimensional parameter. Differentiating this integral w.r.t. $\hat{\boldsymbol{\theta}}$ and equating to zero gives the solution to a system of m linear equations as

$$\boldsymbol{\theta}_{MMSE} = \int_{\boldsymbol{\theta}} \boldsymbol{\theta} p(\boldsymbol{\theta}|D) d\boldsymbol{\theta} = \mathcal{E}\left[\boldsymbol{\theta}|z\right] \tag{1.203}$$

The conditional risk of this solution is

$$R(\boldsymbol{\theta}_{MMSE}) = \int_{\boldsymbol{\theta}} (\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^{\mathsf{T}}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})p(\boldsymbol{\theta}|D)d\boldsymbol{\theta} = \int_{\boldsymbol{\theta}} (\mathcal{E}[\boldsymbol{\theta}|D]-\boldsymbol{\theta})^{\mathsf{T}}(\mathcal{E}[\boldsymbol{\theta}|D]-\boldsymbol{\theta})p(\boldsymbol{\theta}|D)d\boldsymbol{\theta} = \sum_{i=0}^{m-1} \mathrm{Var}[\theta_i|D] \tag{1.204}$$

When the data is normally distributed, MLE = MMSE.

Maximum entropy estimation

The concept of entropy comes from information theory where it is required to describe the uncertainty of observed information (and therefore reflects randomness). If $p(x)$ is a pdf, the entropy associated with it is

We will discuss entropy later when trying to quantify unnatural patterns/trends in data. Here we focus on a neat result using the entropy concept.

$$H = -\int_x p(x)\ln p(x)dx \tag{1.205}$$

The maximum entropy estimate of $p(x)$ is that $p(x)$ which maximizes H subject to any constraints that are specified about $p(x)$ such as a mean value, variance etc. This estimate identifies a distribution that exhibits the maximum randomness subject to available constraints.

As a example, consider a situation where all that is known (and constrained) of $p(x)$ is that it is nonzero between x_1 and x_2 and zero elsewhere. Then for $p(x)$ to be a pdf, the following constraint must apply:

$$\int_{x_1}^{x_2} p(x)dx = 1 \tag{1.206}$$

Maximizing H subject to a constraint requires use of the Lagrange multiplier approach: we maximize an equivalent function

$$H_L = -\int_{x_1}^{x_2} p(x)\ln p(x)dx + \lambda\left(\int_{x_1}^{x_2} p(x) - 1\right) = -\int_{x_1}^{x_2} p(x)(\ln p(x) - \lambda)dx + \lambda \tag{1.207}$$

A derivative w.r.t. a distribution!

Taking the derivative w.r.t. $p(x)$

$$\frac{\partial H_L}{\partial p(x)} = -\int_{x_1}^{x_2} [(\ln p(x) - \lambda) + 1]\, dx \tag{1.208}$$

Equating to zero gives

$$\hat{p}(x) = \exp(\lambda - 1) \tag{1.209}$$

Applying the pdf constraint implies that $\exp(\lambda - 1) = 1/(x_2 - x_1)$ which implies that the maximum entropy distribution is the uniform distibution. Since no other constraint about the mean or the variance was required, it turns out that the maximum entropy distribution which maximizes randomness is one where all points are equally probable.

Exponential distribution!

- The maximum entropy estimate of a pdf which is constrained by a mean ($\mu = \int_{-\infty}^{\infty} xp(x)dx$) is the following distribution:

$$p(x) = \frac{1}{\mu}e^{-x/\mu} \qquad \text{for } x > 0 \tag{1.210}$$

One more reason to use Gaussians!

- The maximum entropy estimate of a pdf which is constrained by a mean ($\mu = \int_{-\infty}^{\infty} xp(x)dx$) and a variance ($\sigma^2 = \int_{-\infty}^{\infty}(x-\mu)^2 p(x)dx$) is the Gaussian $\mathcal{N}(\mu, \sigma^2)$ distribution:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] \tag{1.211}$$

The maximum value of H can be shown to be $(1/2)[1 + \ln(2\pi\sigma^2)]$. As σ^2 increases, H increases.Interestingly H can be negative:

$$H(x) < 0 \quad \text{for} \quad \sigma^2 < \frac{1}{2\pi e}$$

Discrete variables and entropy:
When the random variable is discrete,

$$H = -\sum_i p(x_i) \ln p(x_i) \tag{1.212}$$

As above, we would need to maximize this using the appropriate constraints and Lagrange multipliers.

$$\max H = -\sum_i p(x_i) \ln p(x_i) + \lambda \left(\sum_i p(x_i) - 1 \right) \tag{1.213}$$

and this happens when all the $p(x_i)$ are equal.

$$\frac{\partial^2 H}{\partial p(x_i) \partial p(x_j)} = -I_{ij} \frac{1}{p_i} < 0 \tag{1.214}$$

and hence this is a maximum. Unlike for continuous variables, discrete entropy cannot be negative.