Probability Distributions

Prob. distributions form the building blocks for more complex models.

Density Estimation: model $p(x)$ given a finite set $\{x_1, \dots, x_N\}$ of observations. This is an ill-posed problem, since any $p(X)$ which is non-zero at each data point is a potential candidate.

Parametric: the distr. is governed by a small number of adaptive parameters. Ex.: Gaussian depends on mean μ, and covariance Σ. One limitation is that assumes a specific functional form.

Nonparametric: the form of the distr. depends on the size of the data set. The parameters control the model complexity rather than the form of the prob. distr.

Binary Variables

$x \in \{0,1\}$. $\quad p(x=1|\mu) = \mu \qquad 0 \leq \mu \leq 1$

$\qquad\qquad p(x=0|\mu) = 1-\mu$

$$\boxed{\text{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x}}$$

Bernoulli Distr.

$$\mathbb{E}[x] = p(x=1|\mu).1 + p(x=0|\mu).0$$
$$= \mu$$

$$\text{Var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \mu - \mu^2 = \mu(1-\mu)$$

Suppose we have data $D = \{x_1, \dots, x_N\}$. The likelihood function is

$$P(D|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

In the frequentist approach we maximize the likelihood, or equivalently, its logarithm:

$$\log P(D|\mu) = \sum_{n=1}^{N} x_n \log\mu + (1-x_n)\log(1-\mu)$$

$$\frac{\partial \log P}{\partial \mu} = 0 = \sum_{n=1}^{N} \left\{ \frac{x_n}{\mu} + \frac{(1-x_n)}{1-\mu}(-1) \right\}$$

$$= \sum_{n=1}^{N} \frac{x_n - \cancel{\mu x_n} - \mu + \cancel{x_n \mu}}{\mu(1-\mu)}$$

$$\boxed{\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n} \text{ or } \boxed{\mu_{ML} = \frac{m}{N}}$$ if m is the number of $x=1$ occurrences.

What is the distribution for the number of $x=1$ observations, denoted m? It is proportional to $\mathrm{Bern}(x|\mu)$, and of N possibilities, we can get m times $x=1$ in $\binom{N}{m}$ distinct ways. Thus the answer is

$$\boxed{\mathrm{Bin}(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}}$$ Binomial Distr.

$$m = x_1 + x_2 + \ldots + x_N$$

$\mathbb{E}[m] = N\,\mathbb{E}[x] = N\mu$ since they are Bernoulli iid.

Another way:

$$\mathbb{E}[m] = \sum_{m=0}^{N} m \frac{N!}{(N-m)!\,m!}\mu^m(1-\mu)^{N-m}$$

$$= \sum_{m=1}^{N} \frac{N!}{(N-m)!\,(m-1)!}\mu^m(1-\mu)^{N-m}$$

$$m \to m+1 \quad = \sum_{m=0}^{N-1} \frac{N!}{(N-1-m)!\,m!}\mu\,\mu^m(1-\mu)^{N-1-m}$$

$$N \to N+1 \quad = \mu N \underbrace{\sum_{m=0}^{N} \frac{N!}{(N-m)!\,m!}\mu^m(1-\mu)^{N-m}}_{=1} = \mu N$$

$$\mathrm{Var}[m] = \mathbb{E}[(m - \mathbb{E}m)^2]$$
$$= \mathbb{E}\Big[\Big(\sum_{i=1}^{N} x_i - N\mu\Big)^2\Big] = \mathbb{E}\Big[\Big(\sum_{i=1}^{N}(x_i-\mu)\Big)^2\Big]$$
$$= \mathbb{E}\Big[\sum_{i,j=1}^{N}(x_i-\mu)(x_j-\mu)\Big]$$
$$= \mathbb{E}\Big[\sum_{i=1}^{N}(x_i-\mu)^2 + \sum_{\substack{i,j=1\\ i\neq j}}^{N}(x_i-\mu)(x_j-\mu)\Big] \rightarrow \text{iid}$$
$$= \sum_{i=1}^{N}\mathbb{E}[(x_i-\mu)^2] + \sum_{\substack{i,j=1\\ i\neq j}}\mathbb{E}(x_i-\mu)\,\mathbb{E}(x_j-\mu)$$
$$= N\,\mathrm{Var}[x] + 0$$
$$= N\mu(1-\mu)$$

Another way:

$$\mathbb{E}[m^2] = \sum_{m=0}^{N} m^2 \frac{N!}{(N-m)!\,m!}\mu^m(1-\mu)^{N-m}$$
$$= \sum_{m=1}^{N}\frac{N!\,m}{(N-m)!\,(m-1)!}\mu^m(1-\mu)^{N-m}$$
$$m\to m+1 \quad = \sum_{m=0}^{N-1}\frac{N!\,(m+1)}{(N-1-m)!\,m!}\mu\mu^m(1-\mu)^{N-1-m}$$
$$N-1\to N' \quad = N\mu\sum_{m=0}^{N-1}\frac{(N-1)!\,(m+1)}{(N-1-m)!\,m!}\mu^m(1-\mu)^{N-1-m}$$
$$= N\mu\Big\{\sum_{m=0}^{N'}\frac{N'!\,(m+1)}{(N'-m)!\,m!}\mu^m(1-\mu)^{N'-m}\Big\}$$
$$= N\mu\Big\{\underbrace{\sum_{m=0}^{N'}\frac{N'!}{(N'-m)!\,m!}\mu^m(1-\mu)^{N'-m}}_{=1} + \underbrace{\sum_{m=0}^{N'} m\frac{N'!}{(N'-m)!\,m!}\mu^m(1-\mu)^{N'-m}}_{\mu N' = \mu(N-1)}\Big\}$$

$$\mathbb{E}[m^2] = N\mu(1 + \mu(N-1))$$

$$\text{Var}[m] = \mathbb{E}[m^2] - (\mathbb{E}m)^2 = N\mu + N\mu^2(N-1) - N^2\mu^2$$

$$= N\mu + \cancel{N^2\mu^2} - N\mu^2 - \cancel{N^2\mu^2}$$

$$= N\mu(1-\mu)$$

The Beta Distribution

We saw that ML overfits for small N. In a Bayesian treatment we need a prior $p(\mu)$ such that

$$p(\mu|D) = \frac{p(D|\mu)\, p(\mu)}{p(D)}$$

In the Bernolli case, if $p(\mu) \sim \mu^a (1-\mu)^b$ the prior and the posterior will have the same form. This is known as conjugacy. So we introduce

$$\boxed{\text{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1}}$$

One can check $\int_0^1 \text{Beta}(\mu|a,b)\, d\mu = 1$.

Now

$$\mathbb{E}[\mu] = \left(\int_0^1 d\mu\, \mu^a (1-\mu)^{b-1}\right) \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

$$= \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+1+b)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \qquad \Gamma(n+1) = n\Gamma(n)$$

$$= \frac{a\, \cancel{\Gamma(a)}\, \cancel{\Gamma(b)}}{(a+b)\cancel{\Gamma(a+b)}} \frac{\cancel{\Gamma(a+b)}}{\cancel{\Gamma(a)}\cancel{\Gamma(b)}}$$

$$= \frac{a}{a+b}$$

$$\mathbb{E}[\mu^2] = \int_0^1 d\mu \ \mu^{a+1}(1-\mu)^{b-1} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

$$= \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

$$= \frac{(a+1)\,a\,\cancel{\Gamma(a)}\cancel{\Gamma(b)}}{(a+b+1)(a+b)\cancel{\Gamma(a+b)}} \frac{\cancel{\Gamma(a+b)}}{\cancel{\Gamma(a)}\cancel{\Gamma(b)}}$$

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

$$\therefore \mathrm{Var}[\mu] = \frac{a(a+1)}{(a+b)(a+b+1)} - \frac{a^2}{(a+b)^2}$$

$$= \frac{a(a+1)(a+b) - a^2(a+b+1)}{(a+b)^2(a+b+1)}$$

$$= \frac{(a^2+a)(a+b) - a^3 - a^2b - a^2}{\cdots}$$

$$= \frac{\cancel{a^3} + \cancel{a^2b} + \cancel{a^2} + ab - \cancel{a^3} - \cancel{a^2b} - \cancel{a^2}}{\cdots}$$

$$= \frac{ab}{(a+b)^2(a+b+1)}$$

(a, b) are called Hyperparameters because the control the distr. of the parameter μ.

The likelihood function for Bernoulli is

$$P(D|\mu) = \prod_{i=1}^{N} \mu^{x_i}(1-\mu)^{1-x_i} = \mu^{\sum_{i=1}^{N} x_i}(1-\mu)^{N - \sum_{i=1}^{N} x_i}$$

if $m = \sum_{i=1}^{N} x_i$ is the # times we get $x=1$, and $l = N - m$ the # times we get $x=0$, the

$$P(D|\mu) = \mu^m (1-\mu)^l = p(m, l|\mu)$$

The posterior is the

$$P(\mu|m,l,a,b) \propto p(m,l|\mu)\, p(\mu|a,b)$$
$$= \mu^{m+a-1}(1-\mu)^{l+b-1}$$

which is simply another Beta distr. Introducing the normalization:

$$p(\mu|m,l,a,b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\,\Gamma(l+b)} \mu^{m+a-1}(1-\mu)^{l+b-1}$$

Note that this exactly as the original Beta with $a \to a+m$, $b \to b+l$. So the prior corresponds to an effective number of observations of $x=1$, and $x=0$, and when more data is accessible we get the posterior, which is just an update of our current information. This is a <u>sequential</u> <u>approach to "learning"</u>.

Only assumption: iid data.

This is usefull where we have a real-time stream of Data, for instance.

If we want to predict the outcome of the next trial we have from the sum rule:

$$P(x=1|D) = \int d\mu\, P(x=1|\mu)\, P(\mu|D)$$
$$= \int d\mu\, \mu\, P(\mu|D) = \mathbb{E}[\mu|D]$$
$$= \frac{m+a}{m+a+l+b}$$

When $m, l \to \infty$, this reproduces μ_{ML} from maximum likelihood. It's a general property that ML and Bayesian agree in the limit of infinitely large data.

When $a, b \to \infty$, $\mathrm{Var}[\mu] \to 0$. Is it a general property of Bayesian learning that, as we observe more data, the uncertainty contained in the posterior will decrease? Let us consider the following:

$$\boxed{\mathbb{E}_D[\mathbb{E}_\theta[\theta|D]]} = \mathbb{E}_D\left\{\int d\theta\, p(\theta|D)\theta\right\}$$
$$= \int dD\, p(D) \int d\theta\, p(\theta|D)\theta$$
$$= \int d\theta\, \theta \left\{\int dD\, p(\theta|D) p(D)\right\}$$
$$= \int d\theta\, \theta\, p(\theta) = \boxed{\mathbb{E}_\theta[\theta]}$$

The posterior mean, averaged over the data, is equal to the prior. In other words, the mean of the posterior is the same as the prior, considering the distr. generating the data.

Now lets consider the variance. First the mean, with respect to the data, of the posterior variance:

$$\mathbb{E}_D[\mathrm{Var}_\theta[\theta|D]] = \mathbb{E}_D\left\{\mathbb{E}_\theta[\theta^2|D] - (\mathbb{E}_\theta[\theta|D])^2\right\}$$
$$= \mathbb{E}_D[\mathbb{E}_\theta[\theta^2|D]] - \mathbb{E}_D[\mathbb{E}_\theta[\theta|D]^2]$$
$$= \mathbb{E}_\theta[\theta^2] - \mathbb{E}_D[\mathbb{E}_\theta[\theta|D]^2]$$

Now consider the variance of the posterior mean:

$$\mathrm{Var}_D[\mathbb{E}_\theta[\theta|D]] = \mathbb{E}_D[\mathbb{E}_\theta[\theta|D]^2] - \mathbb{E}_D[\mathbb{E}_\theta[\theta|D]]^2$$
$$= \mathbb{E}_D[\mathbb{E}_\theta[\theta|D]^2] - \mathbb{E}_\theta[\theta]^2$$

Thus summing these results we get

$$\mathbb{E}_D[\mathrm{Var}_\theta[\theta|D]] + \mathrm{Var}_D[\mathbb{E}_\theta[\theta|D]] = \mathrm{Var}_\theta[\theta]$$

Since $\mathrm{Var}_\theta[\theta] \geq 0$, it means that, on average, the posterior variance decreases.

Multinomial Variables

Consider a variable that can take K distinct states. One way to represent a state is through the vector $x \in \mathbb{R}^K$ such that $x_k = 1$ for only one entry, and $x_j = 0$ for $j \neq k$.

Ex.: $x = (0, 0, 1, 0, 0, \ldots)^T$

Note that $\sum_{k=1}^{K} x_k = 1$. Let $p(x_k = 1) = \mu_k$, then the distribution of x is

$$\boxed{p(x|\mu) = \prod_{k=1}^{K} \mu_k^{x_k}} \qquad \mu = (\mu_1, \ldots, \mu_K)^T, \quad \mu_k \geq 0, \quad \sum_k \mu_k = 1.$$

Analogous to Bernoulli, but for K possibilities.

$\sum_x p(x|\mu) = \sum_k \mu_k = 1$, since for each vector x we have $p(x|\mu) = \mu_k$ for some k. Summing over all possibilities gives the result. By analogous reasoning

$$\mathbb{E}[x|\mu] = \sum_x p(x|\mu)x = (\mu_1, \ldots, \mu_K)^T = \mu$$

Now consider data $D = \{x_1, \ldots, x_N\}$. The likelihood function is

$$p(D|\mu) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{\left(\sum_{n=1}^{N} x_{nk}\right)} = \prod_{k=1}^{K} \mu_k^{m_k}$$

where $m_k = \sum_{n=1}^{N} x_{nk}$ is the number of observations where $x_k = 1$. Maximizing this we have

$$\ell = \sum_{k=1}^{K} m_k \log \mu_k + \lambda \left(\sum_{k=1}^{K} \mu_k - 1 \right)$$

$$\frac{\partial \ell}{\partial \mu_j} = 0 = \frac{m_j}{\mu_j} + \lambda \quad \therefore \quad \mu_j = -\frac{m_j}{\lambda}$$

$$\sum_{k=1}^{K} \mu_k = 1 = \frac{1}{\lambda} \sum_{k=1}^{K} - m_k = -\frac{N}{\lambda} \quad \therefore \quad \lambda = -\frac{1}{N} \quad \therefore \quad \boxed{\mu_k^{ML} = \frac{m_k}{N}}$$

which is the fraction where $x_k = 1$ occurs.

For the joint distribution of m_k's we have:

$$\mathrm{Mult}(m_1, \dots, m_K \mid \mu, N) = \binom{N}{m_1, m_2, \dots, m_K} \prod_{k=1}^{K} \mu_k^{m_k}$$

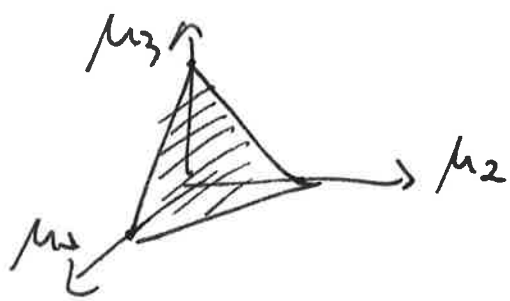
which is the Multinomial distr.

The Dirichlet Distribution

Prior for $\{\mu_k\}$. The conjugate prior is

$$p(\mu \mid \alpha) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

$\alpha = (\alpha_1, \dots, \alpha_K)^T$ are the hyperparameters. Because of $\sum_{k=1}^{K} \mu_k = 1$, the distr. is confined to a simplex (bounded linear manifold) of dimension $K-1$.



The normalized form is

$$\mathrm{Dir}(\mu|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\dots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$ Dirichlet Distr.

$$\alpha_0 = \sum_{k=1}^{K} \alpha_k$$

Now the posterior has the form

$$p(\mu|D,\alpha) \propto p(D|\mu)\,p(\mu|\alpha)$$
$$\propto \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1}$$

Normalizing:

$$p(\mu|D,\alpha) = \mathrm{Dir}(\mu|\alpha + m)$$
$$= \frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + m_1)\dots\Gamma(\alpha_K + m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1}$$

The Gaussian Distribution

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$x \in \mathbb{R}$. For D-dimensional case

$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

Its the continuous distr. that maximizes the entropy. The sum of random variables, under mild assumptions, also follows a Gaussian distribution (Central Limit Theorem).
The binomial distr. when $N \to \infty$ tends to a Gaussian.

$\Delta^2 = (x-\mu)^T \Sigma^{-1} (x-\mu)$ Mahalanobis distance

$N = cte$ on surfaces where $\Delta^2 = cte$.

Σ can be assumed symmetric, without loss of generality. In effect, we have

$$\Sigma = \Sigma_S + \Sigma_A = \frac{\Sigma + \Sigma^T}{2} + \frac{\Sigma - \Sigma^T}{2}$$

Now $z^T \Sigma z = z_i \Sigma_{ij} z_j = z_i \Sigma_{S\,ij} z_j + z_i \Sigma_{A\,ij} z_j$.

$$z_i \Sigma_{A\,ij} z_j = z_i \Sigma_{A\,ji} z_j = - z_i \Sigma_{A\,ij} z_j \therefore z_i \Sigma_{A\,ij} z_j = 0.$$

Consider the eigenvalue eq:

$$\Sigma u_i = \lambda_i u_i$$

Since $\Sigma = \Sigma^T \Rightarrow \lambda_i$ are real, and its eigenvectors form an orthonormal set, $u_i^T u_j = \delta_{ij}$.

Let's show these things:

$$\left.\begin{array}{l} \Sigma u_i = \lambda_i u_i \\ u_j^T \Sigma^T = \lambda_j^* u_j^T \end{array}\right\} \Rightarrow \begin{array}{l} u_j^T \Sigma^T \Sigma u_i = \lambda_i \lambda_j^* u_j^T u_i \\ u_j^T \Sigma^2 u_i = \lambda_i \lambda_j^* u_j^T u_i \\ \lambda_i^2 u_j^T u_i = \lambda_i \lambda_j^* u_j^T u_i \\ (\lambda_i^2 - \lambda_i \lambda_j^*) u_j^T u_i = 0 \end{array}$$

if $i = j \Rightarrow \lambda_i^2 = |\lambda_i|^2 \therefore \underline{\lambda_i \in \mathbb{R}}$

if $i \neq j \Rightarrow u_j^T u_i = 0$ <u>orthogonal</u>.

So by appropriate normalization, $\boxed{u_i^T u_j = \delta_{ij}}$

Now multiply this by u_i: $u_i u_i^T u_j = \delta_{ij} u_i$

or $\left(\sum_i u_i u_i^T\right) u_j = \left(\sum_i \delta_{ij} u_i\right) = u_j \therefore \boxed{I = \sum_i u_i u_i^T}$

Now we have

$$\Sigma = \Sigma I = \Sigma \sum_i u_i u_i^T = \boxed{\sum_i \lambda_i u_i u_i^T}$$

Thus

$$\Delta^2 = (x-\mu)^T \left(\sum_i \frac{1}{\lambda_i} u_i u_i^T \right) (x-\mu)$$

$$= \sum_i \frac{1}{\lambda_i} (x-\mu)^T u_i \, u_i^T (x-\mu)$$

Define $y_i = u_i^T (x-\mu)$ $\Rightarrow$ $\Delta^2 = \sum_i \frac{1}{\lambda_i} y_i^T y_i$

↖ Number

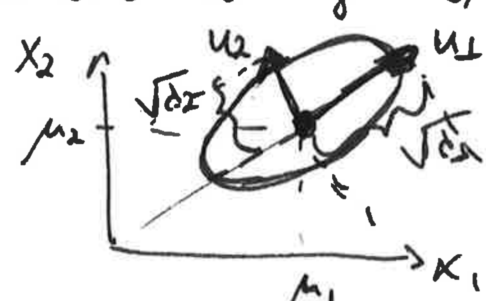
$$\boxed{\Delta^2 = \sum_i \frac{y_i^2}{\lambda_i}}$$

Let $y = (y_1, \dots, y_D)^T \in \mathbb{R}^D$, and U the matrix $U = \begin{pmatrix} u_1^T \\ \vdots \\ u_D^T \end{pmatrix}$. Then $y = U(x-\mu)$. It follows that ~~$U^T U = (u_1 \dots u_D) \begin{pmatrix} u_1^T \\ \vdots \\ u_D^T \end{pmatrix} = U^T U$~~

$$U^T U = \begin{pmatrix} | & | & & | \\ u_1 & u_2 & \dots & u_D \\ | & | & & | \end{pmatrix} \begin{pmatrix} - u_1^T - \\ - u_2^T - \\ \vdots \\ - u_D^T - \end{pmatrix} = I = U U^T$$

so U is <u>orthogonal</u>.

Δ will be constant when y_i's are constant. If all $\lambda_i > 0$, these surfaces are ellipsoids with center at μ, and axes oriented along u_i, with scaling factor $\sqrt{\lambda_i}$.



$d_i > 0$ otherwise it's not possible to normilize. Thus Σ must be positive definite.
If some $d_i = 0$ (singular distr.) the Σ is positive semi-definite.

Change of basis: $y = U(x-\mu)$. $U^T y = x - \mu$

$$x = U^T y + \mu. \qquad x_i = U^T_{ij}(y_j) + \mu_i$$
$$= U_{ji}\, y_j + \mu_i$$

$$J = \frac{\partial x}{\partial y}, \quad J_{ij} = \frac{\partial x_i}{\partial y_j} = U_{ji} \quad \therefore \boxed{J = U^T}$$

$$|J|^2 = |J|\,|J^T| = |J J^T| = |I| = 1.$$

$$|J| = 1.$$

Also, $|\Sigma|^{1/2} = \prod_{j=1}^{D} d_j^{1/2}$. Therefore

$$P(y) = p(x)\,|J| = \prod_{j=1}^{D} \frac{1}{(2\pi d_j)^{1/2}} e^{-\frac{y_j^2}{2d_j}}$$

which is a product of independent gaussians.
So the eigenvectors of the covariance Σ define a system of coordinates such that the joint $p(y)$ is factorized.

Now we compute the moments.

$$\mathbb{E}[x] = \int dx\, x \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$
$$= \int dz\, (z+\mu) \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} e^{-\frac{1}{2} z^T \Sigma^{-1} z}$$
$$= \mu + \underbrace{\int dz\, z \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} e^{-1/2\, z^T \Sigma^{-1} z}}_{\text{odd } = 0}$$

$\boxed{\mathbb{E}[x] = \mu}$

There are D^2 terms $\mathbb{E}[x_i x_j]$ which can be grouped in the matrix $\mathbb{E}[x x^T]$, thus

$$\mathbb{E}[x x^T] = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \int dz\, (z+\mu)(z+\mu)^T e^{-\frac{1}{2} z^T \Sigma^{-1} z}$$

The terms $z\mu^T$ and μz^T vanish by symmetry. The term $\mu\mu^T$ is constant. It remains zz^T.

$$z = x - \mu = \left(\sum_{i=1}^{D} u_i u_i^T\right)(x-\mu) = \sum_{i=1}^{D} u_i y_i$$ ← scalar

$$z z^T = \sum_{i,j} y_i y_j \, u_i u_j^T$$

$$z^T \Sigma^{-1} z = \sum_{k=1}^{D} \frac{y_k^2}{d_k}$$

$z = U^T y$ and $|J| = 1$, thus

$$\mathbb{E}[x x^T] = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \int dy \sum_{i,j} y_i y_j u_i u_j^T e^{-\frac{1}{2}\sum_k \frac{y_k^2}{d_k}} + \mu\mu^T$$

↑ the integral vanishes unless $i=j$

$$= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \sum_{i=1}^{D} \int dy \, u_i u_i^T y_i^2 e^{-\frac{1}{2}\sum_k \frac{y_k^2}{d_k}} + \mu\mu^T$$

The integral over the components that are not y_i will just give the normalization factor $(2\pi)^{D-1} \prod_{j\neq i} d_j^{1/2}$. And the integral $\int dy_i \, y_i^2 e^{-\frac{1}{2}\frac{y_i^2}{d_i}} = \sqrt{2\pi d_i}\, d_i$

Thus

$$\mathbb{E}[xx^T] = \mu\mu^T + \cancel{\prod_{j=1}^{D} \frac{1}{d_j^{1/2}} \sum_{i=1}^{D} \Big(\prod_{k\neq i} d_j^{1/2}\Big) u_i u_i^T d_i^{1/2}}\, d_i$$

$$= \mu\mu^T + \sum_{i=1}^{D} d_i u_i u_i^T$$

$$= \mu\mu^T + \Sigma$$

$$\boxed{\mathrm{cov}[x]} = \mathbb{E}[(x-\mu)(x-\mu)^T]$$

$$= \mathbb{E}[xx^T - x\mu^T - \mu x^T + \mu\mu^T]$$

$$= \mathbb{E}[xx^T] - \mu\mu^T$$

$$\boxed{= \Sigma}$$

Number of parameters:

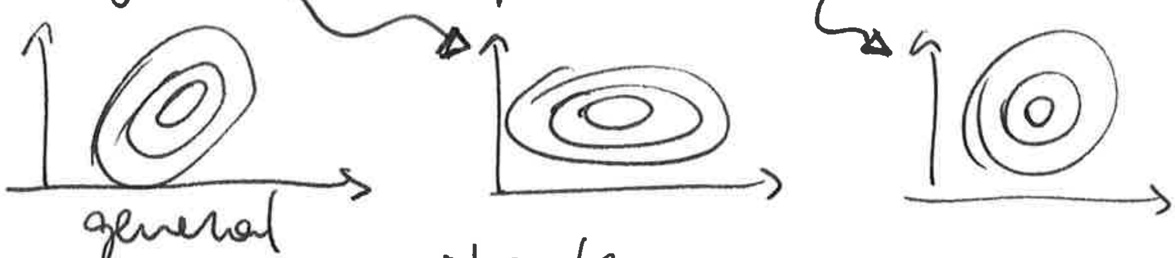
$\Sigma \to \frac{D(D+1)}{2}$ +

$\mu \to D$

$\boxed{\frac{D(D+3)}{2}}$ For large D, grows quadratically with D.

Inverting large matrices are expensive!

One way to deal with this is to assume some restriction on the form of Σ. For instance, $\Sigma = \mathrm{diag}(\sigma_i^2) \Rightarrow 2D$ params, $\Sigma = \sigma^2 I$



general

$N = cte.$

conditional Gaussian

If $p(x,y) \sim \mathcal{N}(\mu, \Sigma) \Rightarrow \begin{cases} \text{conditional dist. is } \mathcal{N} \\ \text{marginal dist. is } \mathcal{N} \end{cases}$

$x \in \mathbb{R}^D \sim \mathcal{N}(x|\mu,\Sigma)$ and we partition x into $\{x_a, x_b\}$ $\quad x_a \cap x_b = \emptyset$

$$x_a = (x_1, \dots, x_M)^T \qquad x = \begin{pmatrix} x_a \\ x_b \end{pmatrix} \; ; \; \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}$$

$$x_b = (x_{M+1}, \dots, x_D)^T$$

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \qquad \Sigma_{ba} = \Sigma_{ab}^T$$

$\Lambda = \Sigma^{-1}$ is the precision matrix

$$\Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix} \qquad \Lambda_{ba} = \Lambda_{ab}^T$$

We then have

$$-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) = -\frac{1}{2}\left((x_a-\mu_a)^T \quad (x_b-\mu_b)^T \right) \cdot \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix} \begin{pmatrix} x_a-\mu_a \\ x_b-\mu_b \end{pmatrix}$$

$$= -\frac{1}{2}\left((x_a-\mu_a)^T \quad (x_b-\mu_b)^T \right) \cdot \begin{pmatrix} \Lambda_{aa}(x_a-\mu_a) + \Lambda_{ab}(x_b-\mu_b) \\ \Lambda_{ba}(x_a-\mu_a) + \Lambda_{bb}(x_b-\mu_b) \end{pmatrix}$$

$$= -\frac{1}{2}(x_a-\mu_a)^T \Lambda_{aa}(x_a-\mu_a)$$
$$-\frac{1}{2}(x_a-\mu_a)^T \Lambda_{ab}(x_b-\mu_b)$$
$$-\frac{1}{2}(x_b-\mu_b)^T \Lambda_{ba}(x_a-\mu_a)$$
$$-\frac{1}{2}(x_b-\mu_b)^T \Lambda_{bb}(x_b-\mu_b)$$

For fixed x_b, this is a quadratic form on x_a. Thus $p(x_a|x_b) \propto p(x_a, \underset{\text{fixed}}{x_b})$ is a Gaussian.

Notice that for a general Gaussian

$$-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) = -\frac{1}{2}x^T\Sigma^{-1}x + \frac{1}{2}x^T\Sigma^{-1}\mu + \frac{1}{2}\mu^T\Sigma^{-1}x - \frac{1}{2}\mu^T\Sigma^{-1}\mu$$

$$= -\frac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu \underbrace{-\frac{1}{2}\mu^T\Sigma^{-1}\mu}_{\text{cte}}$$

We want the mean and covariance of $p(x_a|x_b)$: $\underline{\mu_{a|b},\ \Sigma_{a|b}}$.

Since x_b is fixed, the only term quadratic in x_a is

$$-\frac{1}{2}x_a^T\Lambda_{aa}x_a$$

Thus $\boxed{\Sigma_{a|b} = \Lambda_{aa}^{-1}}$. Now consider the linear terms on x_a:

$$+\frac{1}{2}x_a^T\Lambda_{aa}\mu_a + \frac{1}{2}\mu_a^T\Lambda_{aa}x_a$$

$$-\frac{1}{2}x_a^T\Lambda_{ab}x_b - \frac{1}{2}x_a^T\Lambda_{ab}\mu_b$$

$$-\frac{1}{2}x_b^T\Lambda_{ba}x_a - \frac{1}{2}\mu_b^T\Lambda_{ba}x_a$$

$$= x_a^T\Lambda_{aa}\mu_a - \frac{1}{2}x_a^T\Lambda_{ab}(x_b-\mu_b) - \frac{1}{2}(x_b+\mu_b)^T\underbrace{\Lambda_{ba}}_{\Lambda_{ab}^T}x_a$$

$$= x_a^T\{\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b-\mu_b)\}$$

Thus $\Sigma_{a|b}^{-1}\mu_{a|b} = \Lambda_{aa}\mu_a - \Lambda_{ab}(x_b-\mu_b)$

$$\mu_{a|b} = \Sigma_{a|b}(\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b-\mu_b))$$

we already concluded that $\Sigma_{a|b} = \Lambda_{aa}^{-1}$, thus

$$\boxed{\mu_{a|b} = \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b-\mu_b)}$$

To express these results in terms of Σ_{ab} we can use (inverse of a partitioned matrix)

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1}+D^{-1}CMBD^{-1} \end{pmatrix}$$

$M = (A-BD^{-1}C)^{-1}$, M^{-1} is the Schur complement.

Proof. Just multiply both sides.

$$I = \begin{pmatrix} A & B \\ C & D \end{pmatrix}\begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1}+D^{-1}CMBD^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} AM - BD^{-1}CM & -AMBD^{-1}+BD^{-1}+BD^{-1}CMBD^{-1} \\ \cancel{CM} - \cancel{CM} & -\cancel{CMBD^{-1}} + \cancel{CMBD^{-1}} + I \end{pmatrix}$$

$$-AMBD^{-1}+BD^{-1}+BD^{-1}CMBD^{-1} =$$
$$= (-A+BD^{-1}C)MBD^{-1}+BD^{-1}$$
$$= -M^{-1}MBD^{-1}+BD^{-1} = 0$$

$$(A-BD^{-1}C)M = M^{-1}M = I$$

Thus $I = I$ ◻

Using this for

$$\begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$$

$$\Lambda_{aa} = M = (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}$$

$$\Lambda_{ab} = - M \Sigma_{ab}\Sigma_{bb}^{-1}$$

$$\cancel{= -\Sigma_{aa}\Sigma_{ab}\Sigma_{bb}^{-1} + \Sigma_{ab}\Sigma_{bb}\Sigma_{ba}\Sigma_{ab}\Sigma_{bb}^{-1}}$$

$$= -(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}\Sigma_{ab}\Sigma_{bb}^{-1}$$

Thus $\boxed{\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}}$

$$\mu_{a|b} = \cancel{\Sigma_{a|b}((\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}}$$

$$= \mu_a + (\Sigma_{aa} - \cancel{\Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}}) \cdot$$

$$\cdot (\Sigma_{aa} - \cancel{\Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}})^{-1}\Sigma_{ab}\Sigma_{bb}^{-1} \cdot$$

$$\cdot (x_b - \mu_b)$$

$$\boxed{\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b)}$$

Marginal Gaussian

$$p(x_a) = \int p(x_a, x_b)\, dx_b$$

From the partitioned quadratic form, pick the terms involving x_b:

$$-\frac{1}{2}(x_a - \mu_a)^T \Lambda_{ab} x_b - \frac{1}{2} x_b^T \Lambda_{ba}(x_a - \mu_a)$$

$$-\frac{1}{2} x_b^T \Lambda_{bb} x_b + \frac{1}{2}\mu_b^T \Lambda_{bb} x_b + \frac{1}{2} x_b^T \Lambda_{bb}\mu_b$$

$$-\frac{1}{2}x_b^T\Lambda_{bb}x_b - x_b^T\Lambda_{ba}(x_a-\mu_a) + x_b^T\Lambda_{bb}\mu_b$$

Thus we have

$$-\frac{1}{2}x_b^T\Lambda_{bb}x_b + x_b^T m$$

where $m = \Lambda_{bb}\mu_b - \Lambda_{ba}(x_a-\mu_a)$. This is also

$$-\frac{1}{2}(x_b-\Lambda_{bb}^{-1}m)^T\Lambda_{bb}(x_b-\Lambda_{bb}^{-1}m) + \frac{1}{2}m^T\Lambda_{bb}^{-1}m$$

So the integral over x_b is

$$\int e^{-\frac{1}{2}(x_b-\Lambda_{bb}^{-1}m)^T\Lambda_{bb}(x_b-\Lambda_{bb}^{-1}m)}dx_b$$
$$= (2\pi)^{\frac{D-M}{2}}|\Lambda_{bb}^{-1}|$$

Now picking the terms that do not depend on x_b we have

$$= \frac{1}{2}m^T\Lambda_{bb}^{-1}m - \frac{1}{2}x_a^T\Lambda_{aa}x_a + \frac{1}{2}x_a^T\Lambda_{ab}\mu_b$$
$$+\frac{1}{2}x_a^T\Lambda_{aa}\mu_a + \frac{1}{2}\mu_a^T\Lambda_{aa}x_a + \frac{1}{2}\mu_b^T\Lambda_{ba}x_a + cte$$

$$= \frac{1}{2}m^T\Lambda_{bb}^{-1}m - \frac{1}{2}x_a^T\Lambda_{aa}x_a + x_a^T\Lambda_{ab}\mu_b + x_a^T\Lambda_{aa}\mu_a + cte$$

$$= \frac{1}{2}(\Lambda_{bb}\mu_b - \Lambda_{ba}(x_a-\mu_a))^T\Lambda_{bb}^{-1}(\Lambda_{bb}\mu_b - \Lambda_{ba}(x_a-\mu_a))$$
$$-\frac{1}{2}x_a^T\Lambda_{aa}x_a + x_a^T(\Lambda_{aa}\mu_a + \Lambda_{ab}\mu_b) + cte$$

$$= -\frac{1}{2}(\Lambda_{bb}\mu_b)^T\Lambda_{bb}^{-1}\Lambda_{ba}x_a - \frac{1}{2}(\Lambda_{ba}x_a)^T\Lambda_{bb}^{-1}\Lambda_{bb}\mu_b$$
$$+\frac{1}{2}(\Lambda_{ba}(x_a-\mu_a))^T\Lambda_{bb}^{-1}\Lambda_{ba}(x_a-\mu_a) + \ldots$$

$$= -\frac{1}{2}\mu_b^T \cancel{\Lambda_{bb}\Lambda_{bb}^{-1}}\Lambda_{ba}x_a - \frac{1}{2}x_a^T\Lambda_{ab}\cancel{\Lambda_{bb}^{-1}\Lambda_{bb}}\mu_b$$

$$+\frac{1}{2}x_a^T\Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba}x_a - \frac{1}{2}x_a^T\Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba}\mu_a$$

$$-\frac{1}{2}\mu_a^T\Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba}x_a - \frac{1}{2}x_a^T\Lambda_{aa}x_a$$

$$+ x_a^T(\Lambda_{aa}\mu_a + \Lambda_{ab}\mu_b) + cte$$

$$= -\frac{1}{2}x_a^T(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})x_a$$

$$+ x_a^T(\Lambda_{aa}\mu_a + \cancel{\Lambda_{ab}\mu_b} - \cancel{\Lambda_{ab}\mu_b} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba}\mu_a) + cte$$

$$= -\frac{1}{2}x_a^T(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})x_a$$

$$+ x_a^T(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})\mu_a + cte.$$

Therefore

$$\boxed{\Sigma_a^{-1} = \Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba}}$$

$$\Sigma_a^{-1}\mu_a = (\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})\mu_a \quad \text{(just an identity)}$$

$$\begin{pmatrix}\Lambda_{aa} & \Lambda_{ab}\\ \Lambda_{ba} & \Lambda_{bb}\end{pmatrix}^{-1} = \begin{pmatrix}\Sigma_{aa} & \Sigma_{ab}\\ \Sigma_{ba} & \Sigma_{bb}\end{pmatrix}$$

$$\boxed{\Sigma_{aa} = (\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})^{-1} = \Sigma_a}$$

Thus $\mathbb{E}[x_a] = \mu_a$

$\mathrm{cov}[x_a] = \Sigma_{aa}$ //

Summing up: give $N(x|\mu,\Sigma)$, $\Lambda=\Sigma^{-1}$
with the following partition

$$x=\begin{pmatrix}x_a\\x_b\end{pmatrix}\quad \mu=\begin{pmatrix}\mu_a\\\mu_b\end{pmatrix}\quad \Sigma=\begin{pmatrix}\Sigma_{aa}&\Sigma_{ab}\\\Sigma_{ba}&\Sigma_{bb}\end{pmatrix}\quad \Lambda=\begin{pmatrix}\Lambda_{aa}&\Lambda_{ab}\\\Lambda_{ba}&\Lambda_{bb}\end{pmatrix}$$

The conditional distribution is

$$p(x_a|x_b)=N(x_a|\mu_{a|b},\Lambda_{aa}^{-1})$$

$$\mu_{a|b}=\mu_a-\Lambda_{aa}^{-1}\Lambda_{ab}(x_b-\mu_b)$$

linear in x_b

The marginal distribution is

$$p(x_a)=N(x_a|\mu_a,\Sigma_{aa})$$

Bayes' Theorem for Gaussians

Suppose we are given a Gaussian marginal $p(x)$, and a Gaussian conditional $p(y|x)$ – the mean is linear in x (linear Gaussian Model). We want to find $p(y)$ and ~~$p(x,y)$~~. $p(x|y)$. Thus

$$p(x)=N(x|\mu,\Lambda^{-1})$$

$$p(y|x)=N(y|Ax+b,L^{-1})$$

$$\overset{*}{p}(y|x)=\frac{p(x,y)}{p(x)}$$

if $x\in\mathbb{R}^M$, $y\in\mathbb{R}^D$, then $A\in\mathbb{R}^{D\times M}$. First we find the joint distr.* in terms of $z=\begin{pmatrix}x\\y\end{pmatrix}$.

$$\log p(z)=\log p(x)+\log p(y|x)$$
$$=-\frac{1}{2}(x-\mu)^T\Lambda(x-\mu)-\frac{1}{2}(y-Ax-b)^TL(y-Ax-b)+\text{cte}.$$

This is quadratic in the components of z so it is again a Gaussian.

$$\log p(z) = -\frac{1}{2}x^T\Lambda x - \frac{1}{2}y^T L y - \frac{1}{2}x^T A^T L A x + \frac{1}{2}x^T\Lambda\mu + \frac{1}{2}\mu^T\Lambda x + \frac{1}{2}y^T L A x + \frac{1}{2}y^T L b + \frac{1}{2}x^T A^T L y - \frac{1}{2}x^T A^T L b + \frac{1}{2}b^T L y - \frac{1}{2}b^T L A x + cte$$

quadratic

$$= -\frac{1}{2}x^T(\Lambda + A^T L A)x - \frac{1}{2}y^T L y + \frac{1}{2}y^T L A x + \frac{1}{2}x^T A^T L y$$

$$+ x^T\Lambda\mu + y^T L b - x^T A^T L b \rightsquigarrow \text{linear}$$

$$+ cte$$

The quadratic terms can be written as

$$-\frac{1}{2}\begin{pmatrix} x \\ y \end{pmatrix}^T \underbrace{\begin{pmatrix} \Lambda + A^T L A & -A^T L \\ -LA & L \end{pmatrix}}_{R} \begin{pmatrix} x \\ y \end{pmatrix} = -\frac{1}{2}z^T R z$$

So the joint $p(z)$ has precision R, and

$$\mathrm{cov}[z] = R^{-1} = \begin{pmatrix} \Lambda^{-1} & +\Lambda^{-1}A^T \\ A\Lambda^{-1} & L^{-1} + A\Lambda^{-1}A^T \end{pmatrix}$$

use the inversion formula

$$\cancel{M = (\Lambda + A^T L A L^{-1}(-LA))^{-1}}$$
$$\cancel{= (\Lambda + A^T L A A)^{-1}}$$

$$M = (\Lambda + A^T L A - A^T \cancel{L L^{-1}} L A)^{-1}$$
$$= (\Lambda + \cancel{A^T L A} - \cancel{A^T L A})^{-1}$$
$$= \Lambda^{-1}$$

$$+ \Lambda^{-1}A^T \cancel{L L^{-1}}$$

$$-(L)^{-1}(-LA)\Lambda^{-1}$$

$$L^{-1} + \cancel{L^{-1}}(-\cancel{L}A)\Lambda^{-1}(-A^T\cancel{L})\cancel{L^{-1}}$$

$$L^{-1} + A\Lambda^{-1}A^T$$

Doing the same for the linear terms:

$$x^T(\Lambda\mu - A^T L b) + y^T L b = \begin{pmatrix} x \\ y \end{pmatrix}^T \begin{pmatrix} \Lambda\mu - A^T L b \\ L b \end{pmatrix}$$

$$\Sigma^{-1}\mu =$$

So
$$\mathbb{E}[z] = R^{-1}\begin{pmatrix} \Lambda\mu - A^T L b \\ L b \end{pmatrix}$$

$$= \begin{pmatrix} \Lambda^{-1} & +\Lambda^{-1}A^T \\ A\Lambda^{-1} & L^{-1} + A\Lambda^{-1}A^T \end{pmatrix}\begin{pmatrix} \Lambda\mu - A^T L b \\ L b \end{pmatrix}$$

$$= \begin{pmatrix} \mu - \cancel{\Lambda^{-1}A^T L b} + \cancel{\Lambda^{-1}A^T L b} \\ A\mu - \cancel{A\Lambda^{-1}A^T L b} + b + \cancel{A\Lambda^{-1}A^T L b} \end{pmatrix}$$

$$\mathbb{E}[z] = \begin{pmatrix} \mu \\ A\mu + b \end{pmatrix}$$

To obtain the marginal $p(y)$ we just integrate this last result over x, and according to previous results gives a Gaussian with

$$\mathbb{E}[y] = A\mu + b$$
$$\mathrm{cov}[y] = R_{yy} = L^{-1} + A\Lambda^{-1}A^T$$

Particular case: $A = I \Rightarrow \mathbb{E}[y] = \mu + b$ Convolution

$$\mathrm{cov}[y] = L^{-1} + \Lambda^{-1}$$

For the conditional $p(x|y)$ we have:

$$p(x|y) = \mathcal{N}(x \mid \mu_{x|y}, \Lambda_{xx}^{-1})$$

$$\mu_{x|y} = \mu_x - \Lambda_{xx}^{-1}\Lambda_{xy}(y-\mu_y)$$
$$= \mu - (\Lambda + A^T L A)^{-1}(-A^T L)(y - A\mu - b)$$
$$= \mu + (\Lambda + A^T L A)^{-1}(A^T L(y-b) - A^T L A\mu)$$

+ moving this inside

$$(\Lambda + A^T L A)\mu - A^T L A\mu = \Lambda\mu$$

$$\begin{cases} \mathbb{E}[x|y] = (\Lambda + A^T L A)^{-1}\{A^T L(y-b) + \Lambda\mu\} \\ \mathrm{cov}[x|y] = (\Lambda + A^T L A)^{-1} \end{cases}$$

Summing up:

$$p(x) = \mathcal{N}(x|\mu, \Lambda^{-1})$$
$$p(y|x) = \mathcal{N}(y|Ax+b, L^{-1})$$

Then

$$p(y) = \mathcal{N}(y|A\mu+b, L^{-1} + A\Lambda^{-1}A^T)$$
$$p(x|y) = \mathcal{N}(x|\Sigma(A^T L(y-b) + \Lambda\mu), \Sigma)$$
$$\Sigma = (\Lambda + A^T L A)^{-1}$$

Maximum Likelihood for Gaussian

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

given iid data drawn from this distribution $X = (x_1, \ldots, x_N)^T$ we form the log likelihood function

$$\log p(X|\mu, \Sigma) = -\frac{ND}{2}\log 2\pi - \frac{N}{2}\log|\Sigma| - \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^T\Sigma^{-1}(x_n-\mu)$$

using that $\frac{\partial (x^T A x)}{\partial x} = Ax + A^T x$
$= 2Ax$ (symmetric)

we have $\frac{\partial}{\partial \mu} \log p = 0 = \sum_{n=1}^{N} \Sigma^{-1}(x_n - \mu)$

$$\boxed{\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n}$$

Now we use the following identities

$$\frac{\partial \log |X|}{\partial X} = (X^{-1})^T$$

$$\frac{\partial\, a^T X^{-1} b}{\partial X} = -(X^{-1})^T a b^T (X^{-1})^T = -(X^{-1} b a^T X^{-1})^T$$

Thus

$$\frac{\partial \log p}{\partial \Sigma} = 0 = -\frac{N}{2}\Sigma^{-1} + \frac{1}{2}\sum_{n=1}^{N} \Sigma^{-1}(x_n - \mu)(x_n - \mu)^T \Sigma^{-1}$$

$$\boxed{\Sigma_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})(x_n - \mu_{ML})^T}$$

Now taking the expectation over the true distribution we have

$$\boxed{\mathbb{E}[\mu_{ML}] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[x_n] = \frac{1}{N}N\mu = \mu}$$

To compute $\mathbb{E}[\Sigma_{ML}]$ use

$$\Sigma_{ML} = \frac{1}{N}\sum_n x_n x_n^T - \underbrace{\left(\frac{1}{N}\sum_n x_n\right)}_{\mu_{ML}}\mu_{ML}^T - \mu_{ML}\underbrace{\left(\frac{1}{N}\sum_n x_n^T\right)}_{\mu_{ML}}$$

$$+ \mu_{ML}\mu_{ML}^T$$

$$\Sigma_{ML} = \frac{1}{N}\sum_n x_n x_n^T - \mu_{ML}\mu_{ML}^T$$

$$= \frac{1}{N}\sum_n x_n x_n^T - \frac{1}{N^2}\sum_n\sum_m x_n x_m^T$$

We computed before that $\mathbb{E}[xx^T] = \mu\mu^T + \Sigma$.
Now if $n \neq m$, since the data is iid, we have $\mathbb{E}[x_n x_m^T] = \mathbb{E}[x_n]\,\mathbb{E}[x_m^T] = \mu\mu^T$. Therefore,

$$\mathbb{E}[x_n x_m^T] = \mu\mu^T + \delta_{nm}\Sigma$$

From this we have

$$\boxed{\mathbb{E}[\Sigma_{ML}]} = \frac{1}{N}\sum_n(\mu\mu^T + \Sigma) - \frac{1}{N^2}\sum_{n,m}(\mu\mu^T + \delta_{nm}\Sigma)$$

$$= \cancel{\mu\mu^T} + \Sigma - \cancel{\mu\mu^T} - \frac{1}{N}\Sigma$$

$$= \boxed{\frac{N-1}{N}\Sigma}$$ biased

We can correct this by the unbiased estimator

$$\tilde{\Sigma} = \frac{N}{N-1}\Sigma_{ML} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{ML})(x_n - \mu_{ML})^T$$

thus $\mathbb{E}[\tilde{\Sigma}] = \Sigma$.

Sequential Estimation

Allow data points to be processed one at a time. This is important for on-line applications, and for very large data sets.

This can be done through Robbins-Monro algorithm. consider the joint $p(z, \theta)$.

$$f(\theta) = \mathbb{E}[z|\theta] = \int z\, p(z|\theta)\, dz$$

this is a deterministic function of θ. This is called a regression function. The goal is to find the roots

$$f(\theta^*) = 0$$

under the assumption that we receive one z at a time, and

$$\mathbb{E}[(z-f)^2|\theta] < \infty$$

without loss of generality we assume $f(\theta) > 0$ for $\theta > \theta^*$ and $f(\theta) < 0$ for $\theta < \theta^*$. Then Robbins-Monro procedure is

$$\boxed{\theta^{(N)} = \theta^{(N-1)} + a_{N-1}\, z(\theta^{(N-1)})}$$

where

$$\lim_{N\to\infty} a_N = 0$$

$$\sum_{N=1}^{\infty} a_N = \infty$$

$$\sum_{N=1}^{\infty} a_N^2 < \infty$$

They can show that the sequence $\{\theta^{(N)}\}$ converge to θ^* with probability one.

In a ML estimation we have

$$\frac{\partial}{\partial\theta}\left\{\frac{1}{N}\sum_{n=1}^{N} \log p(x_n|\theta)\right\} = 0$$

$$\frac{1}{N}\sum_{n=1}^{N} \frac{\partial}{\partial\theta} \log p(x_n|\theta) = \mathbb{E}_x\left[\frac{\partial}{\partial\theta} \log p(x|\theta)\right]$$

Now $\theta^{(N)} = \theta^{(N-1)} + a_{N-1} \frac{\partial}{\partial \theta^{(N-1)}} \log p(x_N | \theta^{(N-1)})$

Consider a 1D gaussian $p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$

Then $\boxed{z = \frac{\partial}{\partial \mu} \log p(x|\mu_{ML},\sigma^2) = \frac{(x-\mu_{ML})}{\sigma^2}}$ so z is normally distributed with mean $\mu - \mu_{ML}$.

$$\mu^{(N)} = \mu^{(N-1)} + \frac{a_{N-1}(x_N - \mu^{(N-1)})}{\sigma^2}$$

choosing $a_{N-1} = \frac{\sigma^2}{N}$, $\mu^{(N)} = \mu^{(N-1)} + \frac{1}{N}(x_N - \mu^{(N-1)})$.

It's not hard to see that the solution to this recursive equation is

$$\mu^{(N)} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

which is the ML estimate.

Bayesian Inference for the Gaussian

We need to introduce priors to the parameters.
Suppose we have $N(x|\mu,\sigma^2)$ (1D), and σ is known, and we want to estimate μ.

The data likelihood func. is

$$p(X|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2}$$

Choose the conjugate prior for μ as

$$p(\mu) = N(\mu|\mu_0, \sigma_0^2)$$

The posterior is the

$$p(\mu|X) \propto p(X|\mu)\,p(\mu)$$

both functions are quadratic in μ^2 so the posterior is a Gaussian. We can compute its parameters by considering

$$-\frac{1}{2\sigma^2}\sum_{n=1}^{N}\{x_n^2 - 2x_n\mu + \mu^2\} - \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2 =$$

$$= -\frac{1}{2\sigma^2}\sum_{n=1}^{N} x_n^2 + \frac{\mu}{\sigma^2}N\mu_{ML} - \frac{1}{2\sigma^2}N\mu^2 - \frac{1}{2\sigma_0^2}\mu^2 + \frac{\mu\mu_0}{\sigma_0^2} - \frac{1}{2\sigma_0^2}\mu_0^2$$

$$= \mu^2\left(-\frac{N}{2\sigma^2} - \frac{1}{2\sigma_0^2}\right) + \mu\left(\frac{N\mu_{ML}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right) + cte$$

General form $-\frac{1}{2\sigma_N^2}(\mu-\mu_N)^2 = -\frac{1}{2\sigma_N^2}\mu^2 + \frac{1}{\sigma_N^2}\mu\mu_N + cte$

$$-\frac{1}{2\sigma_N^2} = -\frac{1}{2}\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}\right) \quad \therefore \quad \boxed{\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}}$$

$$= \frac{\sigma^2 + N\sigma_0^2}{\sigma_0^2 \sigma^2}$$

$$\frac{\mu_N}{\sigma_N^2} = \frac{N\mu_{ML}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}$$

$$\mu_N = \frac{\sigma_0^2\sigma^2}{\sigma^2 + N\sigma_0^2}\left(\frac{N\mu_{ML}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right) = \frac{N\sigma_0^2}{\sigma^2 + N\sigma_0^2}\mu_{ML} + \frac{\sigma^2}{\sigma^2 + N\sigma_0^2}\mu_0$$

$$\boxed{\mu_N = \frac{\sigma^2}{\sigma^2 + N\sigma_0^2}\mu_0 + \frac{N\sigma_0^2}{\sigma^2 + N\sigma_0^2}\mu_{ML}}$$

so $\quad p(\mu|X) = \mathcal{N}(\mu|\mu_N, \sigma_N^2)$

Notice that $\begin{cases} N=0 \Rightarrow \mu_N=\mu_0 \\ N\to\infty \Rightarrow \mu_N\to\mu_{ML} \end{cases}$, thus $\boxed{\mu_N \in [\mu_0, \mu_{ML})}$

$$\sigma_N^2 = \frac{\sigma_0^2\sigma^2}{\sigma^2+N\sigma_0^2}. \quad \begin{array}{l} N=0 \Rightarrow \\ N\to\infty \Rightarrow \end{array} \begin{cases} \sigma_N=\sigma_0 \\ \sigma_N\to 0 \end{cases}$$

so the posterior becomes infinitely peaked around the ML solution

So ML is recovered when $N\to\infty$.

Notice also, if N is fixed, and $\sigma_0\to\infty$ the $\mu_N\to\mu_{ML}$, $\sigma_N^2\to\frac{\sigma^2}{N}$.

Notice that we can view   <u>sequential view</u>

$$p(\mu|X) \propto \Big[\underbrace{p(\mu)\prod_{n=1}^{N-1} p(x_n|\mu)}_{p(X^{(N-1)}|\mu)}\Big]\, \underbrace{p(x_N|\mu)}_{\text{likelihood}}$$

$\propto$ prior with $N-1$ points.

Now assume the mean is known and we estimate the variance.

$$p(X|\lambda) \propto \lambda^{N/2} e^{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2}$$

precision $\lambda=\frac{1}{\sigma^2}$

The conjugate prior must be $e^{a\lambda}\cdot\lambda^b$ so we can use

$$\mathrm{Gam}(\lambda|a,b)=\frac{1}{\Gamma(a)}b^a\lambda^{a-1}e^{-b\lambda}$$

$\int d\lambda \, \mathrm{Gam}(\lambda|a,b) = 1$ (6-1)

$$\boxed{\mathbb{E}[\lambda]} = \int d\lambda \, \lambda \, \mathrm{Gam}(\lambda|a,b)$$

$$= \int d\lambda \, \frac{b^a \lambda^a}{\Gamma(a)} e^{-b\lambda} = \boxed{\frac{a}{b}}$$

$$\int d\lambda \, \frac{b^{a+1}}{\Gamma(a+1)} \lambda^a e^{-b\lambda} = 1 = \frac{b}{a} \int d\lambda \, \frac{b^a \lambda^a e^{-b\lambda}}{\Gamma(a)}$$

$$\mathbb{E}[\lambda^2] = \int d\lambda \, \frac{b^a \lambda^{a+1} e^{-b\lambda}}{\Gamma(a)} = \frac{a+1}{b} \cdot \frac{a}{b}$$

$$\frac{a}{b} = \int d\lambda \, \frac{b^a \lambda^a e^{-b\lambda}}{\Gamma(a)} \rightarrow \frac{a+1}{b} \int d\lambda \, \frac{b \, b^a \lambda^{a+1} e^{-b\lambda}}{a \Gamma(a)}$$

Thus $\boxed{\mathrm{Var}[\lambda] = \frac{a+1}{b} \cdot \frac{a}{b} - \frac{a^2}{b^2} = \frac{a}{b^2}}$

The posterior is

$$p(\lambda|X) \propto p(X|\lambda) \, p(\lambda)$$

$$= \lambda^{N/2} e^{-\frac{\lambda}{2} \sum_n (x_n - \mu)^2} \lambda^{a_0 - 1} e^{-b_0 \lambda}$$

$$= \lambda^{N/2 + a_0 - 1} e^{-b_0 \lambda - \frac{\lambda}{2} \sum_n (x_n - \mu)^2}$$

which is also a $\mathrm{Gam}(\lambda|a_N, b_N)$ distribution with

$$a_N - 1 = \frac{N}{2} + a_0 - 1, \quad \boxed{a_N = a_0 + \frac{N}{2}}$$

$$-b_N = -b_0 - \frac{1}{2} \sum_n (x_n - \mu)^2, \quad \boxed{b_N = b_0 + \frac{N}{2} \sigma^2_{ML}}$$

Now suppose that both, the precision and the mean are unknown.

$$p(X|\mu,\lambda) = \prod_{n=1}^{N} \left(\frac{\lambda}{2\pi}\right)^{1/2} e^{-\frac{\lambda}{2}(x_n-\mu)^2}$$

$$= \prod_{n=1}^{N} \frac{\lambda^{1/2}}{(2\pi)^{1/2}} e^{-\frac{\lambda}{2}(x_n^2 - 2x_n\mu + \mu^2)}$$

$$= \left(\frac{\lambda^{1/2}}{(2\pi)^{1/2}} e^{-\frac{\lambda\mu^2}{2}}\right)^N e^{\lambda\mu\sum_n x_n - \frac{\lambda}{2}\sum_n x_n^2}$$

So the conjugate prior should be of the form

$$p(\mu,\lambda) \propto \left(\lambda^{1/2} e^{-\frac{\lambda\mu^2}{2}}\right)^\beta e^{c\lambda\mu - d\lambda}$$

$$= \lambda^{\beta/2} e^{-\frac{\lambda\beta}{2}(\mu-\frac{c}{\beta})^2} e^{+\frac{\lambda c^2}{2\beta} - d\lambda}$$

$$= e^{-\frac{\lambda\beta}{2}(\mu-\frac{c}{\beta})^2} \lambda^{\beta/2} e^{-\lambda(d - \frac{c^2}{2\beta})}$$

$$p(\mu,\lambda) = p(\mu|\lambda)\, p(\lambda)$$

gaussian gamma

$$p(\mu,\lambda) = N(\mu|\mu_0, (\beta\lambda)^{-1})\, \mathrm{Gam}(\lambda|a,b)$$

Normal-gamma

$$\mu_0 = \frac{c}{\beta}$$

$$a = 1 + \frac{\beta}{2}$$

$$b = d - \frac{c^2}{2\beta}$$

For
$$N(x|\mu,\Lambda) = \frac{1}{(2\pi)^{D/2}} |\Lambda|^{1/2} e^{-\frac{1}{2}(x-\mu)^T\Lambda(x-\mu)}$$

if $\Sigma^{-1} = \Lambda$ is known and μ is unknown, the conjugate prior $p(\mu)$ is also a Gaussian. If μ is known, and Λ is unknown then

$$p(X|\Lambda) = \frac{1}{(2\pi)^{ND/2}} |\Lambda|^{N/2} e^{-\frac{1}{2}\sum_n (x_n-\mu)^T\Lambda(x_n-\mu)}$$

Notice that
$$\sum_n (x_n-\mu)^T\Lambda(x_n-\mu) = \sum_n Tr(z_n^T \Lambda z_n)$$

where $z_n \equiv x_n - \mu$.
$$= \sum_n Tr(z_n z_n^T \Lambda)$$

Since $Tr(A+B) = Tr(A) + Tr(B)$
$$= Tr\left(\left(\sum_n z_n z_n^T\right)\Lambda\right)$$
$$= Tr(Z\Lambda)$$

Thus the conjugate prior is the Wishart distr.

$$W(\Lambda|W,\nu) = B\,|\Lambda|^{(\nu-D-1)/2} e^{-\frac{1}{2}Tr(W^{-1}\Lambda)}$$

ν is the number of degrees of freedom of the distr. $W \in \mathbb{R}^{D\times D}$? the normalization is given by

$$B(W,\nu) = |W|^{-\nu/2}\left(2^{\nu D/2}\pi^{D(D-1)/4}\prod_{i=1}^{D}\Gamma\left(\frac{\nu+1-i}{2}\right)\right)^{-1}$$

If both the mean and precision are unknown we have the conjugate prior

$$p(\mu,\Lambda|\mu_0,\beta,W,\nu) = N(\mu|\mu_0,(\beta\Lambda)^{-1})\,W(\Lambda|W,\nu)$$

which is the normal-Wishart distr.

Student's t-Distribution

If μ is known, but $\sigma^2 = \frac{1}{\lambda}$ is not, the conjugate prior to the Gaussian is the gamma distr. Here we consider:

$$p(x|\mu,a,b) = \int_0^\infty d\lambda\, N(x|\mu,\lambda^{-1})\, \mathrm{Gam}(\lambda|a,b)$$

$$= \int_0^\infty d\lambda \left(\frac{\lambda}{2\pi}\right)^{1/2} e^{-\frac{\lambda}{2}(x-\mu)^2} \frac{b^a \lambda^{a-1} e^{-b\lambda}}{\Gamma(a)}$$

$$= \frac{b^a}{(2\pi)^{1/2}} \frac{1}{\Gamma(a)} \int_0^\infty d\lambda\, \lambda^{a-1/2} e^{-\lambda\left(b + \frac{(x-\mu)^2}{2}\right)}$$

solving the integral.

$$\int_0^\infty d\lambda\, \lambda^{a-1/2} e^{-\lambda c} = c^{-a-1/2} \int_0^\infty dt\, t^{a-1/2} e^{-t} = c^{-a-1/2}\,\Gamma(a+1/2)$$

$$\lambda c = t,\quad d\lambda = \frac{1}{c}dt \quad\Big|\quad \frac{1}{c}dt \cdot c^{-a+1/2} = dt\, c^{-a-1/2} \quad\Big|\quad \Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$$

Thus

$$p(x|\mu,a,b) = \frac{b^a}{(2\pi)^{1/2}} \frac{1}{\Gamma(a)} \left(b + \frac{(x-\mu)^2}{2}\right)^{-a-1/2} \Gamma(a+1/2)$$

Now introduce parameters $\nu = 2a$, $\lambda = a/b$:

$$p(x|\mu,\nu,\lambda) = \mathrm{St}(x|\mu,\lambda,\nu)$$

$$= \frac{1}{(2\pi)^{1/2}} \frac{\Gamma(a+1/2)}{\Gamma(a)} \cancel{b^a}\, b^{\cancel{-a}-1/2} \left(1 + \frac{(x-\mu)^2}{2b}\right)^{-a-1/2}$$

$$b^{-1/2} = \left(\frac{a}{\lambda}\right)^{-1/2} = \left(\frac{\nu}{2\lambda}\right)^{-1/2} = 2^{1/2}\left(\frac{\lambda}{\nu}\right)^{1/2}$$

Thus

$$St(x|\mu,\lambda,\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}\left(\frac{\lambda}{\pi\nu}\right)^{1/2}\left(1+\frac{\lambda(x-\mu)^2}{\nu}\right)^{-\frac{\nu}{2}-\frac{1}{2}}$$

λ is called the precision
ν is called the ~~[illegible]~~ degrees of freedom

$\nu = 1 \Rightarrow St \to$ Cauchy distr.
$\nu \to \infty \Rightarrow St \to N(x|\mu,\lambda^{-1})$. Let us show this:

$$= \frac{\Gamma(\frac{\nu}{2}+\frac{1}{2})}{\Gamma(\frac{\nu}{2})\,\nu^{1/2}}\left(\frac{\lambda}{\pi}\right)^{1/2}\left(1+\frac{\lambda(x-\mu)^2}{\nu}\right)^{-\frac{\nu}{2}-\frac{1}{2}}$$

$$\overset{\nu\to 2z}{=} \frac{\Gamma(z+1/2)}{\Gamma(z)\,z^{1/2}}\left(\frac{\lambda}{2\pi}\right)^{1/2}\left(1+\frac{\lambda(x-\mu)^2}{2z}\right)^{-z-1/2}$$

Now using $\lim_{z\to\infty}\frac{\Gamma(z+\alpha)}{\Gamma(z)\,z^{\alpha}} = 1$ and $\lim_{z\to\infty}\left(1+\frac{x}{z}\right)^{-z} =$

$$= \lim_{z\to\infty}\left(\left(1+\frac{x}{z}\right)^{z}\right)^{-1}$$

$$= (e^x)^{-1} = e^{-x}$$

we have

$$\lim_{\nu\to\infty} St(x|\mu,\lambda,\nu) = \left(\frac{\lambda}{2\pi}\right)^{1/2} e^{-\frac{\lambda(x-\mu)^2}{2}} = N(x|\mu,\lambda^{-1})$$

The t-distribution is obtained by adding an infinite number of Gaussians with the same mean but different precisions. This is an infinite mixture. This gives a distribution which has <u>longer tails</u> than a Gaussian. $\Rightarrow$ <u>Robustness</u>, less sensitive to outliers.

In higher dim. this generalizes to

$$\mathrm{St}(x|\mu,\Lambda,\nu) = \int_0^\infty \mathcal{N}(x|\mu,(\eta\Lambda)^{-1})\,\mathrm{Gam}(\eta|\nu/2,\nu/2)\,d\eta$$

Computing this integral gives

$$\mathrm{St}(x|\mu,\Lambda,\nu) = \frac{\Gamma(D/2+\nu/2)}{\Gamma(\nu/2)} \frac{|\Lambda|^{1/2}}{(\pi\nu)^{D/2}} \left(1+\frac{\Delta^2}{\nu}\right)^{-D/2-\nu/2}$$

$$\Delta^2 = (x-\mu)^T\Lambda(x-\mu)$$

$$\mathbb{E}[x] = \mu$$

$$\mathrm{cov}[x] = \frac{\nu}{\nu-2}\Lambda^{-1}$$

$$\mathrm{mode}[x] = \mu$$

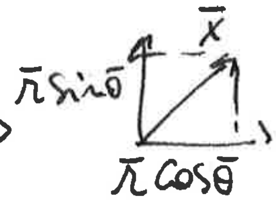
Periodic Variables

$D = \{\theta_1, \dots, \theta_N\}$ periodic. Mean?

$\frac{1}{N}\sum \theta_i$ will be strongly coordinate dependent.

These points are on the unit circle in 2D so we can average vectors $\{x_i\}$ and then invert to find $\bar{\theta}$.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^N x_i \longrightarrow$$



$$\bar{\theta} = \arctan\frac{\bar{x}_2}{\bar{x}_1}$$

or $$\bar{\theta} = \arctan\left(\frac{\sum_i \sin\theta_i}{\sum_i \cos\theta_i}\right)$$

Periodic generalization of a Gaussian is the von Misses distr.

conditions

$$\begin{cases} p(\theta) \geq 0 \\ \int_0^{2\pi} p(\theta)\, d\theta = 1 \\ p(\theta + 2\pi) = p(\theta) \end{cases}$$

gaussian 2D: $p(x_1, x_2) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x_1-\mu_1)^2 + (x_2-\mu_2)^2}{2\sigma^2}}$

$\Sigma = \sigma^2 I$

$$\begin{cases} x_1 = r\cos\theta \\ x_2 = r\sin\theta \\ \mu_1 = r_0\cos\theta_0 \\ \mu_2 = r_0\sin\theta_0 \end{cases} \Rightarrow \frac{r_0\cos(\theta-\theta_0)}{\sigma^2} + cte$$

indep. of θ

$$\boxed{p(\theta) = \frac{1}{2\pi I_0(m)} e^{m\cos(\theta-\theta_0)}}$$

$\equiv m$

von Misses circular Normal

$$I_0(m) = \frac{1}{2\pi}\int_0^{2\pi} e^{m\cos\theta}\, d\theta$$

For large m this becomes approximately Gaussian since $\cos\theta = 1 - \frac{\theta^2}{2} + \ldots$

MLE: $\log p(\Theta \mid \theta_0, m) = -N(\log 2\pi + \log I_0(m)) + m\sum_{n=1}^{N}\cos(\theta_n - \theta_0)$

$$\frac{\partial}{\partial\theta_0}(\ldots) = m\sum_{n=1}^{N} + \sin(\theta_n - \theta_0) = 0$$

$$\sum_{n=1}^{N}\sin(\theta_n - \theta_0) = 0$$

$$\sum_{n=1}^{N}(\sin\theta_n\cos\theta_0 - \sin\theta_0\cos\theta_n) = 0$$

$$\boxed{\frac{\sin\theta_0}{\cos\theta_0} = \frac{\sum_n \sin\theta_n}{\sum_n \cos\theta_n}}$$

$\bar{\theta} = \theta_{0ML}$

$$\frac{\partial}{\partial m}(\cdots) = 0 = -\frac{N}{I_0(m)} \underbrace{\frac{\partial I_0(m)}{\partial m}}_{\equiv I_1(m)} + \sum_{n=1}^{N} \cos(\theta_n - \theta_0)$$

$$\frac{I_1(m)}{I_0(m)} = \frac{1}{N} \sum_{n=1}^{N} \cos(\theta_n - \theta_0^{ML}) \equiv A(m)$$

$$A(m) = \cos\theta_0^{ML}\left(\frac{1}{N}\sum \cos\theta_n\right) + \sin\theta_0^{ML}\left(\frac{1}{N}\sum \sin\theta_n\right)$$

$$A^{-1}(m) = m$$

can be inverted numerically.

We can form mixtures of von Mises distributions to account for multimodality.

Mixture of Gaussians

superposition of 3 gaussians!

Mixture

$$p(x) = \sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \Sigma_k)$$

mixing coefficients (π_k)

component k of the mixture ($N(x \mid \mu_k, \Sigma_k)$)

$$\int p(x)\,dx = 1 \Rightarrow \sum_{k=1}^{K} \pi_k = 1$$

since both sides are positive $\pi_k \geq 0$,

thus $0 \leq \pi_k \leq 1$. probabilities

$$p(x) = \sum_{k=1}^{K} \underbrace{p(k)}_{\pi \text{ prior}} \underbrace{p(x|k)}_{N \text{ cond. likelihood}}$$

$p(k|X)$ posterior

$$\gamma_k(x) = p(k|X) = \frac{p(x|k)\,p(k)}{\sum_k p(k)p(x|k)}$$

responsibilities

$$\gamma_k(x) = \frac{\pi_k N(x|\mu_k, \Sigma_k)}{\sum_{l=1}^{K} \pi_l N(x|\mu_l, \Sigma_l)}$$

MLE: $$p(D|\pi, \mu, \Sigma) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k)$$

$\pi = \{\pi_1, \dots, \pi_K\}$

$\mu = \{\mu_1, \dots, \mu_K\}$

$\Sigma = \{\Sigma_1, \dots, \Sigma_K\}$

$D = \{x_1, \dots, x_N\}$

$$\log P = \sum_{n=1}^{N} \log \left\{ \sum_{k=1}^{K} \pi_k N_k(x_n) \right\}$$

More complicated
Does not have a closed form solution.

Exponential Family

Except for gaussian, all other previous prob. distr. are members of the exponential family.

$$\boxed{p(x|\eta) = h(x)\, g(\eta)\, e^{\eta^T u(x)}}$$

η is called natural parameters.

$g(\eta)$ ensures normalization.

Ex.: Bernoulli

$$p(x) = \mu^x (1-\mu)^{1-x}$$
$$= e^{x \log \mu + (1-x)\log(1-\mu)}$$
$$= (1-\mu)\, e^{x \log \mu - x \log(1-\mu)}$$
$$= \underbrace{(1-\mu)}_{g(\mu)}\, e^{x \underbrace{\log\left(\frac{\mu}{1-\mu}\right)}_{\eta = \log\frac{\mu}{1-\mu}}}$$

$$e^{\eta} = \frac{\mu}{1-\mu} \quad \therefore \quad \frac{1}{\mu} - 1 = e^{-\eta}$$
$$\frac{1}{\mu} = e^{-\eta} + 1$$

logistic sigmoid func. $\boxed{\mu = \frac{1}{1+e^{-\eta}} = \sigma(\eta)}$

$$g(\mu) = 1-\mu$$
$$= 1 - \frac{1}{1+e^{-\eta}}$$
$$= \frac{\cancel{1}+e^{-\eta}\cancel{-1}}{1+e^{-\eta}}$$
$$= \frac{1}{1+e^{\eta}} = \sigma(-\eta)$$

$$\therefore \boxed{\mathrm{Bern}(x|\mu) = \sigma(-\eta)\, e^{\eta x}}$$ where $\mu = \sigma(\eta)$

Ex.: Multinomial

$$p(x|\mu) = \prod_{k=1}^{M} \mu_k^{x_k} = \prod_{k=1}^{M} e^{x_k \log \mu_k}$$

$\sum_k \mu_k = 1$

$$= e^{\sum_k x_k \log \mu_k} = e^{\eta^T x}$$

$$\eta = (\log \mu_1, \dots, \log \mu_M)^T$$
$$x = (x_1, \dots, x_M)^T$$

(72)

Incorporating $\sum_{k=1}^{M} \mu_k = 1$ and $\sum_{k=1}^{M} x_k = 1$,

since $x = (x_1, \dots, x_M)^T$ is a binary vector, we have

$$\sum_{k=1}^{M} x_k \log \mu_k = \sum_{k=1}^{M-1} x_k \log \mu_k + x_M \log \mu_M$$

$$= \sum_{k=1}^{M-1} x_k \log \mu_k + \left(1 - \sum_{k=1}^{M-1} x_k\right) \log\left(1 - \sum_{k=1}^{M-1} \mu_k\right)$$

$$= \sum_{k=1}^{M-1} x_k \log\left(\frac{\mu_k}{1 - \sum_{j=1}^{M-1} \mu_j}\right) + \log\left(1 - \sum_{k=1}^{M-1} \mu_k\right)$$

Let $\eta_k = \log \dfrac{\mu_k}{1 - \sum_{j=1}^{M-1} \mu_j}$ $\therefore$ $e^{\eta_k} = \dfrac{\mu_k}{1 - \sum_{j=1}^{M-1} \mu_j}$

Ansatz $\boxed{c e^{\eta_k} = \mu_k}$ $\rightarrow$ $\cancel{e^{\eta_k}} = \dfrac{c e^{\eta_k}}{1 - c \sum_j e^{\eta_j}}$, $c = \dfrac{1}{1 + \sum_{j=1}^{M-1} e^{\eta_j}}$

Thus $\mu_k = \dfrac{e^{\eta_k}}{1 + \sum_{j=1}^{M-1} e^{\eta_j}}$ softmax function or normalized exponential

So $p(x) = \left(1 - \sum_{k=1}^{M-1} \mu_k\right) e^{\sum_{k=1}^{M-1} x_k \eta_k}$

$$\hookrightarrow = \frac{\mu_k}{e^{\eta_k}} = \frac{1}{1 + \sum_{j=1}^{M-1} e^{\eta_j}}$$

$$p(x|\eta) = \underbrace{\left(1 + \sum_{j=1}^{M-1} e^{\eta_j}\right)^{-1}}_{g(\eta)} e^{\eta^T x}$$

$\eta = (\eta_1, \dots, \eta_M)^T$

Simple Gaussian:

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

$$= \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} e^{-\frac{1}{2}\frac{x^2-2x\mu+\mu^2}{\sigma^2}}$$

$$-\frac{1}{2}\frac{x^2}{\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{1}{2}\frac{\mu^2}{\sigma^2} = \eta^T u(x)$$

$$\eta = \begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix}, \quad u(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$$

$$e^{-\frac{1}{2}\frac{\mu^2}{\sigma^2}} = e^{\frac{\eta_1^2}{4\eta_2}}$$

$$\frac{1}{\sigma} = (-2\eta_2)^{1/2}$$

Thus $$\mathcal{N}(x|\eta) = \underbrace{\frac{1}{\sqrt{2\pi}}}_{h(x)} \underbrace{(-2\eta_2)^{1/2} e^{\frac{\eta_1^2}{4\eta_2}}}_{g(\eta)} e^{\eta^T u(x)}$$

Maximum Likelihood

$$p(x|\eta) = h(x) g(\eta) e^{\eta^T u(x)}$$

$$\int p(x|\eta)\,dx = 1 = g(\eta) \int h(x) e^{\eta^T u(x)}\,dx$$

$$0 = \nabla g(\eta) \int h(x) e^{\eta^T u(x)}\,dx + g(\eta) \int h(x) e^{\eta^T u(x)} u(x)\,dx$$

$$0 = \frac{\nabla g(\eta)}{g(\eta)} + g(\eta) \int h(x) e^{\eta^T u(x)} u(x)\,dx$$

$$\boxed{-\nabla \log g(\eta) = \mathbb{E}[u(x)]}$$

The covariance comes from 2nd derivatives.
As long as we can normalize an exponential distr.
we can compute its moments in this way.

Likelihood function

$$p(D|\eta) = \prod_{n=1}^{N} h(x_n)\; g(\eta)^N\; e^{\eta^T \sum_{n=1}^{N} u(x_n)}$$

$$\log p(D|\eta) = \sum_{n=1}^{N} \log h(x_n) + N \log g(\eta) + \eta^T \sum_{n=1}^{N} u(x_n)$$

$$\nabla_\eta \log p(D|\eta) = 0 = N \nabla \log g(\eta) + \sum_{n=1}^{N} u(x_n)$$

$$\therefore \boxed{-\nabla \log g(\eta)_{ML} = \frac{1}{N} \sum_{n=1}^{N} u(x_n)}$$

Sufficient statistic

Conjugate Priors

Given $p(X|\eta)$ we want a prior $p(\eta)$ such that the posterior $p(\eta|X)$ has the same form as the prior. For the exponential family, $\exists$ prior in the form

$$p(\eta|\chi, \nu) = f(\chi, \nu)\, g(\eta)^\nu e^{\nu \eta^T \chi}$$

Thus

$$p(X|\eta)\, p(\eta|\chi,\nu) \Rightarrow p(\eta|X,\chi,\nu) \propto g(\eta)^{\nu+N} \cdot e^{\nu \eta^T X + \eta^T \sum_{n=1}^{N} u(x_n)}$$

$$p(\eta|X,\chi,\nu) \propto g(\eta)^{\nu+N} e^{\eta^T \left\{ \sum_{n=1}^{N} u(x_n) + \nu \chi \right\}}$$

which has the same form as the prior.

When we don't have much prior information, we don't want to influence the posterior, so we can use uniform distr. as a prior. However, in the continuous case this can lead to an improper prior, which cannot be normalized. This is ok, as long as the posterior is proper.
A second problem may appear in a change of variables: suppose $h(d) = \text{const.}$ $d = \eta^2 \Rightarrow$
$h(\eta^2) = \text{const}$, but $p_\eta(\eta) = p_d(d)\left|\frac{dd}{d\eta}\right| = p_d(d)\, 2\eta$

$\uparrow$ not const. $\qquad$ $\uparrow$ const

Consider the noninformative prior

$$p(x|\mu) = f(x-\mu)$$

$\hat{x} = x + c \Rightarrow p(\hat{x}|\hat{\mu}) = f(\hat{x}-\hat{\mu})$ indep. choice of origin.

prior must assign equal prob. to the interval $A \leq \mu \leq B$ as $A-c \leq \mu \leq B-c$:

$$\int_A^B p(\mu)d\mu = \int_{A-c}^{B-c} p(\mu)d\mu = \int_A^B p(\mu-c)d\mu$$

$$\therefore \underline{p(\mu-c) = p(\mu)} \Rightarrow p(\mu) = \text{cte.}$$

Ex.: mean μ of a gaussian

Consider $\quad p(x|\sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right)$

$\hat{x} = cx \Rightarrow p(\hat{x}|\hat{\sigma}) = \frac{1}{\hat{\sigma}} f\left(\frac{\hat{x}}{\hat{\sigma}}\right)$

$$\int_A^B p(\sigma)\,d\sigma = \int_{A/c}^{B/c} p(\sigma)\,d\sigma = \int_A^B p\left(\frac{\sigma}{c}\right)\frac{1}{c}\,d\sigma$$

scale invariant.

$$\therefore p(\sigma) = \frac{1}{c}\,p\left(\frac{\sigma}{c}\right). \quad p(\sigma) \propto \frac{1}{\sigma}$$

Nonparametric Methods

Parametric Approach: prob. distr. has a specific form depending on few parameters whose values are determined from the data.

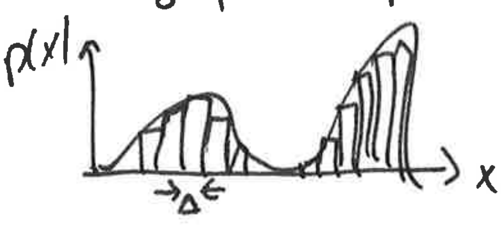
Nonparametric Approach: few assumptions about the form of the distribution.

Histogram method: partition X into bins Δ_i and count the number of observations of X, n_i, falling into Δ_i:

$$p_i = \frac{n_i}{N\Delta_i} = \text{prob}(x \in \Delta_i)$$

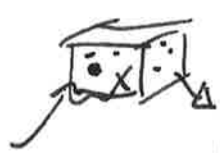
$$\sum_i p_i \Delta_i = 1 = \int p(x)\,dx.$$

Density $p(x)$ is constant inside each Δ_i.



problems:
- discontinuities due to the edges of the bins
- For high dimensions $\#(\Delta_i) = M^D$ need lots of data. curse of dimensionality.

Locality



Δ distance measure around X.

Δ cannot be too small nor too large.
* choice of model complexity.

Kernel Density Estimators

$x \in \mathbb{R}^D$. Prob mass around x is

$$P = \int_R p(x)\,dx$$

$x_1, \dots, x_N \sim p(x)$. $P(x_i \in R) = P$. The total number K of points that lie in R has prob distribution $\mathrm{Bin}(K|N,P) = \binom{N}{K} P^K (1-P)^{1-K}$

$$\mathbb{E}[K/N] = P \quad \xrightarrow{N\to\infty} \quad K \approx NP$$

$$\mathrm{var}[K/N] = \frac{P(1-P)}{N}$$

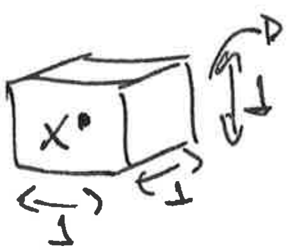
If R is small, $p(x) \approx$ const. inside R, thus

$$P \approx p(x)V \qquad V = \mathrm{vol}(R)$$

Thus $$p(x) = \frac{P}{V} = \frac{K}{NV}$$

1. Fix K and determine V from data.
   K-nearest-neighbour
2. Fix V and determine K from data.
   Kernel Approach

Both converge to $p(x)$ when $N \to \infty$ provided V shrinks with N, and K grows with N.



$$K(u) = \begin{cases} 1, & |u_i| \leq 1/2 \quad i = 1, \dots, D \\ 0, & \text{otherwise} \end{cases}$$

Kernel function
Parzen window

$$K\left(\frac{x-x_n}{h}\right) = \begin{cases} 1 & \text{if } x_n \in \text{cube size } h \text{ around } x \\ 0 & \text{otherwise} \end{cases}$$

Total # of points inside this h-cube $= k = \sum_{n=1}^{N} K\left(\frac{x-x_n}{h}\right)$

Thus $$p(x) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{h^D} K\left(\frac{x-x_n}{h}\right)$$
$\hookrightarrow V$

We can also see this as a sum of N cubes centered at each data point x_n.

We can "smooth" this model to remove discontinuities in

$$p(x) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{(2\pi h^2)^{D/2}} e^{-\frac{\|x-x_n\|^2}{2h^2}}$$

$\Sigma = h^2 I$ we place a gaussian at each point and add the whole contribution over the data.

h plays the whole of smoothing parameter.

choice of h = model complexity.

any $K(u)$ such that $\begin{cases} \bullet\ K(u) \geq 0 \\ \bullet\ \int K(u)du = 1 \end{cases}$ can be used.

the computation cost for the density grows linearly with N.

Nearest-neighbour Methods

Kernel, h is the same. Bad if data is too concentrated in a region (too smoothing). If h is too small then we get a lot of Noise in regions where data is scarce.

Recall $p(x) = \frac{K}{NV}$.

$x^{\cdot}$ sphere around x

Fix K

We allow the radius to grow until it captures K points

- K controls the smoothing now.
- Not a true density because the integral over all space diverges.

Example (classification): N_K points in class C_K. $\sum_K N_K = N$. Given a new point x, classify it!

Sphere around x containing K points independent of their classes. Suppose K_K points of C_K in this sphere of vol. V. The

$$p(x|C_K) = \frac{K_K}{N_K V}$$

$$p(x) = \frac{K}{NV}$$

$$p(C_K) = \frac{N_K}{N}$$

Bayes $\Rightarrow$ $p(C_K|x) = \frac{p(x|C_K)\,p(C_K)}{p(x)} = \frac{K_K}{K}$

Assign x to the largest posterior!