The central, Gaussian or normal distribution

← Back to Chapters

Exercise 7.1

TO DO

Exercise 7.2

Consider the family of distributions $p_{\mu,\sigma}(v)$.

We want to express fact that convolution of $p_{\mu,\sigma}(v)$ with $q(v)$ still belongs to the same family. I will prefer to work with parameter $\nu = \sigma^2$ instead. To avoid confusion the variable v will be replaced by x. So we work with the family $p_{\mu,\nu}(x)$.

Let $\mu_q = \langle\epsilon\rangle_q$ be the mean and $\nu_q = \langle\epsilon^2\rangle_q - \langle\epsilon\rangle_q^2$ the variance of q. Then the new distribution must be $p_{\mu+\mu_q,\nu+\nu_q}(x)$. At the same time the expansion 7.20 becomes

$$p_{\mu+\mu_q,\nu+\nu_q}(x) = p_{\mu,\nu}(x) - \mu_q \frac{\partial}{\partial x} p_{\mu,\nu}(x) + \frac{1}{2}(\nu_q + \mu_q^2)\frac{\partial^2}{\partial^2 x} p_{\mu,\nu}(x) + \dots$$

Taylor expanding around μ, ν

$$p_{\mu+\mu_q,\nu+\nu_q}(x) = p_{\mu,\nu}(x) + \mu_q \frac{\partial}{\partial \mu} p_{\mu,\nu}(x) + \nu_q \frac{\partial}{\partial \nu} p_{\mu,\nu}(x) +$$

$$\frac{1}{2}\mu_q^2 \frac{\partial^2}{\partial^2 \mu} p_{\mu,\nu}(x) + \frac{1}{2}\nu_q^2 \frac{\partial^2}{\partial^2 \nu} p_{\mu,\nu}(x) + \mu_q \nu_q \frac{\partial^2}{\partial \mu \partial \nu} p_{\mu,\nu}(x)$$

Now **if** we wanted this to be true for arbitrary (small) μ_q, ν_q we would have equality of Taylor coefficients:

$$-\frac{\partial}{\partial x} p_{\mu,\nu}(x) = \frac{\partial}{\partial \mu} p_{\mu,\nu}(x)$$

$$\frac{1}{2}\frac{\partial^2}{\partial^2 x} p_{\mu,\nu}(x) = \frac{\partial}{\partial \nu} p_{\mu,\nu}(x) = \frac{1}{2}\frac{\partial^2}{\partial^2 \mu} p_{\mu,\nu}(x)$$

where the first equation says that $p_{\mu,\nu}(x)$ is a function of $x-\mu$ and not of μ and x separately, $p_{\mu,\nu}(x) = f_\nu(x-\mu)$. From this $\frac{\partial^2}{\partial^2 x}p_{\mu,\nu}(x) = \frac{\partial^2}{\partial^2 \mu}p_{\mu,\nu}(x)$ follows, and we simply recover the more general Gaussina family $p_{\mu,\nu}(x) = \frac{1}{\sqrt{2\pi\nu}}\exp\{-\frac{(x-\mu)^2}{2\nu}\}$ as in 7.23.

However, **if** we instead think of μ and ν as $\mu(t)$ and $\nu(t)$ so that the family $p_t(x) = p_{\mu(t),\nu(t)}(x)$ is a single-parameter family, then the expansions become expansions in terms of t: with $\mu_q(t) = \mu_q'(0)t + o(t^2)$, $\nu_q(t) = \nu_q'(0)t + o(t^2)$

$$p_{\mu+\mu_q(t),\nu+\nu_q(t)}(x) = p_{\mu,\nu}(x) + [-\mu_q'(0)\frac{\partial}{\partial x}p_{\mu,\nu}(x) + \frac{1}{2}\nu_q'(0)\frac{\partial^2}{\partial^2 x}p_{\mu,\nu}(x)]t +$$

$$p_{\mu+\mu_q(t),\nu+\nu_q(t)}(x) = p_{\mu,\nu}(x) + \frac{\partial}{\partial t}p_{\mu,\nu}(x)t + o(t^2)$$

Equating Taylor coefficients:

$$\frac{\partial}{\partial t}p_t(x) = -\mu_q'\frac{\partial}{\partial x}p_t(x) + \frac{1}{2}\nu_q'\frac{\partial^2}{\partial^2 x}p_t(x)$$

This is a Fokker-Plank equation, albeit a very special one, with $\mu(x,t) = \mu'(0)$, $\sigma^2(x,t) = \nu'(0)$, corresponding to the stochastic process where the drift μ and diffusion coefficient $\nu/2$ are both constant. Denote $\mu'(0) = m$ and $\nu'(0) = v$.

Changing coordinates to $y(x,t) = x - mt$ aka $x(y,t) = y + mt$, we have

$$p_t(x(y,t)) = p_t(y+mt) =: q_t(y)$$

and compute by chin rule

$$\frac{\partial}{\partial t}q_t(y) = \frac{\partial}{\partial t}p_t(x(y,t)) = \frac{\partial}{\partial t}p_t(y+mt) + m\frac{\partial}{\partial x}p_t(y+mt)$$

while

$$\frac{1}{2}v\frac{\partial^2}{\partial^2 y}q_t(y) = \frac{1}{2}v\frac{\partial^2}{\partial^2 y}p_t(x(y,t)) = \frac{1}{2}v\frac{\partial^2}{\partial^2 x}p_t(y+mt)$$

So the substitution we made reduces the Fokker-Plank equation we have (with drift) to the diffusion equation (without drift) i.e. 7.22 (with $\sigma^2 = t$), which by 7.23 has solution $q_t(y) = \frac{1}{\sqrt{2\pi t}}\exp\{-\frac{y^2}{2t}\}$, or, after substitution

$$p_t(x) = \frac{1}{\sqrt{2\pi t}}\exp\{-\frac{(x-mt)^2}{2t}\}$$

This has variance $\sigma^2 = t$ so we can rewrite it as $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-m\sigma^2)^2}{2\sigma^2}\}$, as in the formulation of the exercise.

Exercise 7.3 preliminary remarks.

I interpret this as follows. We are now considering evolution of beliefs about noise-generating process, and want to update this based on data and see that as the amount of data grows our posterior beliefs about the noise-generating process will imply beliefs about frequencies of noise that will converge to those produced by true frequencies $f(e)$.

In order to talk about the above meaningfully in mathematical sense, one needs to define some space of noise-genera processes in such a way that formulating a prior probability over it is possible, and also define in what sense the posterior beliefs about frequencies "converge".

Since we only observe real-valued noise data (and no other type of information about the noise-generating processes) the space of noise-generating processes can be taken to be the set of real-valued stochastic processes. Now this is still very large. Now one could either 1) take only noise-generating processes that generate noise in i.i.d. way 2) focus only on the beliefs about "frequencies" i.e. how often various values (close to) various es are observed - or both.

The intuition is that those processes that produce incorrect frequency of es will be suppressed in he update (due to a mechanism like that in the Borel's law of large numbers), leaving only the process that do have the correct frequencies in the "posterior distribution over processes", whatever that means. To illustrate some of the difficulties, consider the simple case of discrete time and processes that just i.i.d. sample from some probability distribution. Consider a distribution which moreover has a density p. Then a particular dataset $e_1, \ldots, e_n$ has likelihood $\prod_{i=1}^N p(e_i)$. At each finite N the closer our distribution p is to the empirical distribution(of $\frac{1}{N}\sum_i \delta e_i$) the higher the likelihood; this seems problematic. Perhaps one has to discretize the set of possible es and then take limit, or simply use some more sophisticated analysis.

Footnote 12

Suppose the sampling distribution is Cauchy $p(y|\mu) = \frac{1}{\pi}\frac{1}{1+(y-\mu)^2}$ and the sample is $y_1, \ldots, y_n$. Then likelihood is $\prod_{i=1}^n \frac{1}{\pi}\frac{1}{1+(y_i-\mu)^2}$, and log likeliehood is up to a constant $L_{\vec{y}}(\mu) = \sum_i -\ln(1 + (y_i - \mu)^2)$. The extremality condition is $\frac{d}{d\mu}L_{\vec{y}}(\mu) = 0$ i.e. $\sum_i^n \frac{y_i-\mu}{1+(y_i-\mu)^2} = 0$. This is in general equivalent to a degree $2n-1$ polynomial equation in μ - there are many local optima for the likelihood.

We only treat the case $n = 2$ fully.

In that case, since the whole problem is equivariant to shifts, without loss of generality we can assume $y_1 = -y_2 = y$ (then the general solution is obtained by substituting $y = \frac{y_1-y_2}{2}$ and shifting by $\frac{y_1+y_2}{2}$).

The optimality equation becomes $\frac{y-\mu}{1+(y-\mu)^2} - \frac{y+\mu}{1+(y+\mu)^2} = 0$ so $f(x) = \frac{x}{1+x^2} = \frac{1}{x+\frac{1}{x}}$

has $f(y-\mu) = f(y+\mu)$. Either $y - \mu = y + \mu$, i.e. $\mu = 0$ or $(y-\mu)(y+\mu) = 1$, $\mu = \pm\sqrt{y^2 - 1}$. This last pair of solution is real only if $|y| > 1$.

Supose that in fact $|y| > 1$. Then one over the likeliehood is a positive fourth degree polynomial which we now know has three local extrema - $-\sqrt{y^2-1}, 0, \sqrt{y^2-1}$. Therefore these extrema must be non-degenerate and be min, max, min. Correspondingly, the (log)likeliehood extrema must be max, min, max.

The MLE estimate is thus indifferent between $\mu = \frac{y_1+y_2}{2} + \sqrt{(\frac{y_1-y_2}{2})^2 - 1}$ and $\mu = \frac{y_1+y_2}{2} - \sqrt{(\frac{y_1-y_2}{2})^2 - 1}$. Let's see how this can be written in the form $\mu = y_1 w_1(y_1 - y_2) + y_2 w_2(y_1 - y_2)$.

Supposing

$$\frac{y_1 + y_2}{2} + \sqrt{(\frac{y_1 - y_2}{2})^2 - 1} =$$

$$= y_1 w_1(y_1 - y_2) + y_2 w_2(y_1 - y_2)$$

Plug in $y_2 = 0$ to get

$$\frac{y_1}{2} + \sqrt{(\frac{y_1}{2})^2 - 1} = y_1 w_1(y_1)$$

$$w_1(y_1) = \frac{1}{2} + sgn(y_1)\sqrt{\frac{1}{4} - \frac{1}{y_1^2}}$$

Plug in $y_1 = 0$ to get

$$w_2(-y_2) = \frac{1}{2} + sgn(y_2)\sqrt{\frac{1}{4} - \frac{1}{y_2^2}}$$

or

$$w_2(y) = \frac{1}{2} - sgn(y)\sqrt{\frac{1}{4} - \frac{1}{y^2}}$$

Now plugging back we should have if $y_1 > y_2$

$$\frac{y_1 + y_2}{2} + \sqrt{(\frac{y_1 - y_2}{2})^2 - 1} =$$

$$y_1 w_1(y_1-y_2)+y_2 w_2(y_1-y_2)=$$

$$y_1\left(\frac{1}{2}+\sqrt{\frac{1}{4}-\frac{1}{(y_1-y_2)^2}}\right)+$$

$$+y_2\left(\frac{1}{2}-\sqrt{\frac{1}{4}-\frac{1}{(y_1-y_2)^2}}\right)$$

$$\sqrt{(\tfrac{y_1-y_2}{2})^2-1}=(y_1-y_2)\left(\sqrt{\tfrac{1}{4}-\tfrac{1}{(y_1-y_2)^2}}\right)$$

$$y_1-y_2=y_1-y_2$$

which indeed holds. Other cases are similar. Observe, moreover, that $w_1+w_2=1$.

For general n, the inverse of likelihood is a polynomial in μ of degree $2n$ with coefficients some symmetric polynomials in y_i, and its derivative is a polynomial in μ of degree $2n-1$ with coefficients some symmetric polynomials in y_i, which we can write as $Q(\mu, y_1, \ldots, y_n)$. The hypersurface $Q(\mu, y_1, \ldots, y_n) = Q(\mu, \vec{y}) = 0$ in $\mathbb{R}^{n+1}$ projects to $\vec{y} = y_1, \ldots, y_n$. In the $\vec{y}$ space and if we exclude those $\vec{y}$ tuples for which the discriminant of Q is zero we have over $\mathbb{R}^n \setminus D$ a covering map, with varying number of sheets k, being the number of soultions of $Q(\mu, \vec{y}) = 0$ in μ for given $\vec{k}$. Locally in $\mathbb{R} \setminus D$, there are k "inverse of projection" maps $M : C \to \mathbb{R}$ that pick out a particular solution out of the k in a continuous (and in fact smooth) way (a solution of $Q(\mu, \vec{y}) = 0$ is a local extremum of the likelihood function). Our task is to show that at least those M that correspond to maxima of likeliehood can be written as $M(\vec{y}) = \sum_i y_i w_i(\vec{d})$ where $\vec{d}$ is the vector of pairwise differences $y_k - y_l$ and $\sum w_i(\vec{d}) = 1$. I have failed to prove this or find it in the literature.

Exercise 7.4

We are minimizing $\vec{w}^T C \vec{w}$. Let's minimize over the hyperplane $\sum w_i = 1$. Since C is positive definite on $\mathbb{R}^n$, at infinity the values are large and positive. So the minimum is achieved at finite distance and must satisfy the Lagrange multiplier equation is $C\vec{w} = \lambda\vec{1}$, so $\vec{w} = \lambda C^{-1}\vec{1}$ and $\sum w_i = 1$ gives, with $C^{-1} = K$ the answer $w_i = \sum_j K_{ij} / \sum_{i,j} K_{ij}$, as wanted (note that the denominator is $\vec{1}^T K \vec{1} > 0$).

Corresponding value is

$$\vec{w}^T C \vec{w} = \lambda^2 (C^{-1}\vec{1})^T C (C^{-1}\vec{1}) = \lambda = (\sum_{ij} K_{ij})^{-1}$$

HOWEVER this answer satisfies does not always satisfy the constraints $w_i \geq 0$: consider $C = \begin{pmatrix} 1 & 4 \\ 4 & 17 \end{pmatrix}$ so that $K = \begin{pmatrix} 17 & -4 \\ -4 & 1 \end{pmatrix}$; then $w = (1.3, -0.3)^T$.

Exercise 7.5

See https://www.cs.toronto.edu/~yuvalf/CLT.pdf

7.84

$$\exp\{xa - \frac{a^2}{2}\} = \exp\{\frac{x^2}{2}\}\exp\{-\frac{(x-a)^2}{2}\}$$

so

$$\frac{d^n}{da^n}\exp\{xa - \frac{a^2}{2}\} = \exp\{\frac{x^2}{2}\}\frac{d^n}{da^n}\left(\exp\{-\frac{(x-a)^2}{2}\}\right)$$

7.85

$$\frac{\phi(x-a)\phi(x-b)}{\phi(x)} = \phi(x)\left(\sum_n R_n(x)\frac{a^n}{n!}\right)\left(\sum_m R_m(x)\frac{b^m}{m!}\right)$$

LHS: Observe

$$\phi(x-a)\phi(x-b) = \phi(x)\phi(x-(a+b))\exp\{ab\}$$

Then

$$\int \frac{\phi(x-a)\phi(x-b)}{\phi(x)}dx = \int \phi(x-(a+b))\exp\{ab\}dx = \exp\{ab\}$$

As a power series in ab it is $\sum_i \frac{a^i b^i}{i!}$.

RHS:

$\sum_{n,m}(\int \phi(x)R_n(x)R_m(x)dx)\frac{a^n b^m}{n!m!}$

Equating coefficients we get 7.85.

7.86

From 7.83 $\phi(x-y) = \phi(x)\sum_m R_m(x)\frac{y^m}{m!}$ so

$$\phi(x-y)R_n(x) = R_n(x)\phi(x)\sum_m R_m(x)\frac{y^m}{m!}$$

and integrating and using 7.85 we get

$$\int \phi(x-y)R_n(x)dx = y^n.$$

7.89

$$\exp\{xa\}\exp\{-a^2/2\} = \sum_k \frac{x^k a^k}{k!}\sum_m \frac{a^{2m}}{(-2)^m m!}$$

Isolating the term in front of $n = k + 2m$ gives 7.89