STK 4020: Bayesian Statistics
Autumn 2012

Course Notes and Exercises
by Nils Lid Hjort

– This version: as of 15 September 2012 –

1. Prior to posterior updating with Poisson data

This exercise illustrates the basic prior to posterior updating mechanism for Poisson data.

(a) First make sure that you are reasonably acquainted with the Gamma distribution. We say that $Z \sim \mathrm{Gamma}(a, b)$ if its density is

$$g(z) = \frac{b^a}{\Gamma(a)} z^{a-1} \exp(-bz) \quad \text{on } (0, \infty).$$

Here a and b are positive parameters. Show that

$$\mathrm{E}\, Z = \frac{a}{b} \quad \text{and} \quad \mathrm{Var}\, Z = \frac{a}{b^2} = \frac{\mathrm{E} Z}{b}.$$

In particular, low and high values of b signify high and low variability, respectively.

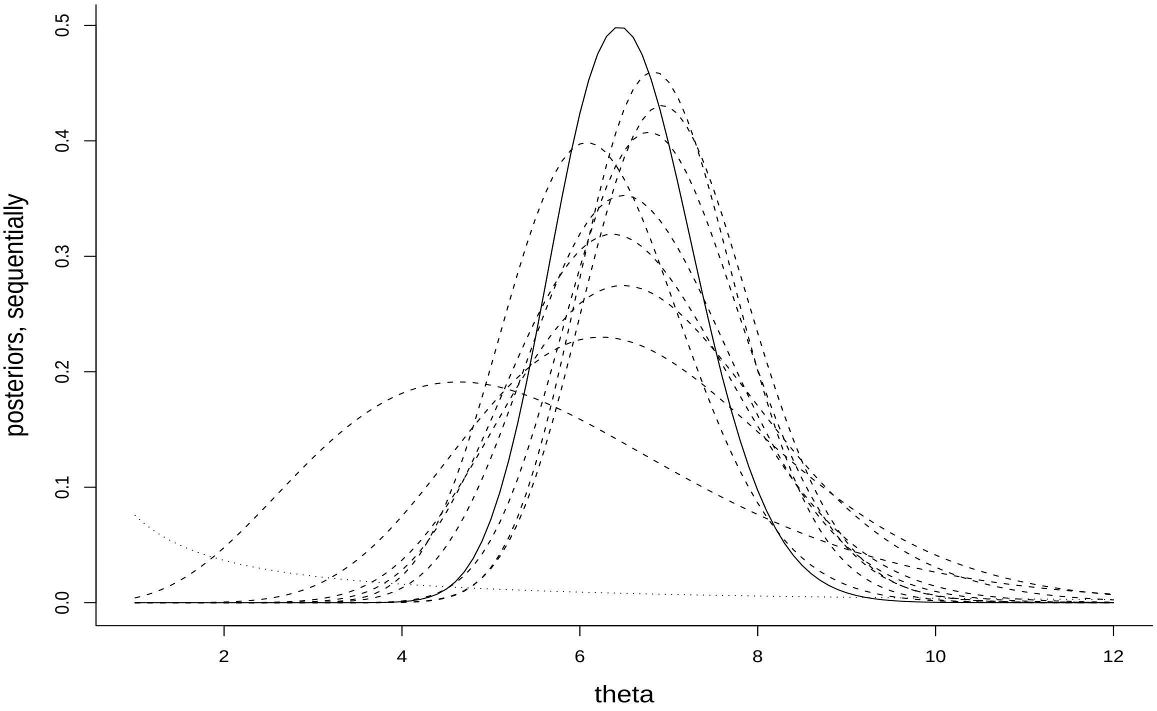

Figure 1: Eleven curves are displayed, corresponding to the $\mathrm{Gamma}(0.1, 0.1)$ *intial prior density for the Poisson parameter* θ *along with the ten updates following each of the observations 6, 8, 7, 6, 7, 4, 11, 8, 6, 3.*

(b) Now suppose $y \mid \theta$ is a Poisson with parameter θ, and that θ has the prior distribution $\mathrm{Gamma}(a, b)$. Show that $\theta \mid y \sim \mathrm{Gamma}(a + y, b + 1)$.

(c) Then suppose there are repeated Poisson observations $y_1, \ldots, y_n$, being i.i.d. $\sim \mathrm{Pois}(\theta)$ for given θ. Use the above result repeatedly, e.g. interpreting $p(\theta \mid y_1)$ as the new prior before observing y_2, etc., to show that

$$\theta \mid y_1, \ldots, y_n \sim \mathrm{Gamma}(a + y_1 + \cdots + y_n, b + n).$$

Also derive this result directly, i.e. without necessarily thinking about the data having emerged sequentially.

(d) Suppose the prior used is a rather flat $\mathrm{Gamma}(0.1, 0.1)$ and that the Poisson data are 6, 8, 7, 6, 7, 4, 11, 8, 6, 3. Reconstruct a version of Figure 1 in your computer, plotting the ten curves $p(\theta \mid \mathrm{data}_j)$, where data_j is $y_1, \ldots, y_j$, along with the prior density. Also compute the ten Bayes estimates $\widehat{\theta}_j = \mathrm{E}(\theta \mid \mathrm{data}_j)$ and the posterior standard deviations, for $j = 0, \ldots, 10$.

(e) The mathematics turned out to be rather uncomplicated in this situation, since the Gamma continuous density matches the Poisson discrete density so nicely. Suppose instead that the initial prior for θ is a uniform over $[0.5, 50]$. Try to compute posterior distributions, Bayes estimates and posterior standard deviations also in this case, and compare with you found above.

2. The Master Recipe for finding the Bayes solution

Consider a general framework with data y, in a suitable sample space $\mathcal{Y}$; having likelihood $p(y \mid \theta)$ for given parameter θ (stemming from an appropriate parametric model), with θ being inside a parameter space Ω; and with loss function $L(\theta, a)$ associate with decision or action a if the true parameter value is θ, with a belonging to a suitable action space $\mathcal{A}$. This could be the real line, if a parameter space is called for; or a two-valued set $\{\mathrm{reject}, \mathrm{accept}\}$ if a hypothesis test is being carried out; or the set of all intervals, if the statistician needs a confidence interval.

A statistical *decision function*, or procedure, is a function $\widehat{a}\colon \mathcal{Y} \to \mathcal{A}$, getting from data y the decision $\widehat{a}(y)$. Its *risk function* is the expected loss, as a function of the parameter:

$$R(\widehat{a}, \theta) = \mathrm{E}_\theta L(\theta, \widehat{a}) = \int L(\theta, \widehat{a}(y)) p(y \mid \theta) \,\mathrm{d}y.$$

(In particular, in this expectation operation the random element is y, having its $p(y \mid \theta)$ distribution for given parameter, and the integration range is that of the sample space $\mathcal{Y}$.)

So far the framework does not include Bayesian components per se, and is indeed a useful one for frequentist statistics, where risk functions for different decision functions (be they estimators, or tests, or confidence intervals, depending on the action space and the loss function) may be compared.

We are now adding one more component to the framework, however, which is that of a *prior distribution* $p(\theta)$ for the parameter. The overall risk, or *Bayes risk*, associated with a decision function $\widehat{a}$, is then the overall expected loss, i.e.

$$\mathrm{BR}(\widehat{a},p)=\mathrm{E}\,R(\widehat{a},\theta)=\int R(\widehat{a},\theta)p(\theta)\,\mathrm{d}\theta.$$

(Here θ is the random quantity, having its prior distribution, making also the risk function $R(\widehat{a},\theta)$ random.) The *minimum Bayes risk* is the smallest possible Bayes risk, i.e.

$$\mathrm{MBR}(p)=\min\{\mathrm{BR}(\widehat{a},p)\colon \text{all decision functions } \widehat{a}\}.$$

The *Bayes solution* for the problem is the strategy or decision function $\widehat{a}_B$ that succeeds in minimising the Bayes risk, with the given prior, i.e.

$$\mathrm{MBR}(p)=\mathrm{BR}(\widehat{a}_B,p).$$

The *Master Theorem* about Bayes procedures is that there is actually a recipe for finding the optimal Bayes solution $\widehat{a}_B(y)$, for the given data y (even without taking into account other values y' that could have been observed).

(a) Show that the *posterior density* of θ, i.e. the distribution of the parameter given the data, takes the form

$$p(\theta\,|\,y)=k(y)^{-1}p(\theta)p(y\,|\,\theta),$$

where $k(y)$ is the required integration constant $\int p(\theta)p(y\,|\,\theta)\,\mathrm{d}\theta$. This is the *Bayes theorem.*

(b) Show also that the *marginal distribution* of y becomes

$$p(y)=\int p(y\,|\,\theta)p(\theta)\,\mathrm{d}\theta.$$

(I follow the GCSR book's convention regarding using the 'p' multipurposedly.)

(c) Show that the overall risk may be expressed as

$$\begin{aligned}\mathrm{BR}(\widehat{a},p)&=\mathrm{E}\,L(\theta,\widehat{a}(Y))\\&=\mathrm{E}\,\mathrm{E}\,\{L(\theta,\widehat{a}(Y))\,|\,Y\}\\&=\int\Bigl\{\int L(\theta,\widehat{a}(y))p(\theta\,|\,y)\,\mathrm{d}\theta\Bigr\}\,p(y)\,\mathrm{d}y.\end{aligned}$$

The inner integral, or 'inner expectation', is $\mathrm{E}\{L(\theta,\widehat{a}(y))\,|\,y\}$, the expected loss given data.

(d) Show then that the optimal Bayes strategy, i.e. minimising the Bayes risk, is achieved by using

$$\widehat{a}_B(y) = \mathrm{argmin} g = \text{the value } a_0 \text{ minimising the function } g,$$

where $g = g(a)$ is the expected posterior loss,

$$g(a) = \mathrm{E}\{L(\theta, a) \,|\, y\}.$$

The g function is evaluated and mininised over all a, for the given data y. This is the Bayes recipe. – For examples and illustrations, with different loss functions, see the Nils 2008 Exercises.

3. Minimax estimators

For a decision function $\widehat{a}$, bringing data y into a decision $\widehat{a}(y)$, its max-risk is

$$R_{\max}(\widehat{a}) = \max_\theta R(\widehat{a}, \theta).$$

We say that a procedure a^* is *minimax* if it minimises the max-risk, i.e.

$$R_{\max}(a^*) \le R_{\max}(\widehat{a}) \quad \text{for all competitors } \widehat{a}.$$

Here I give recipes (that often but not always work) for finding minimax strategies.

(a) For any prior p and strategy $\widehat{a}$, show that

$$\mathrm{MBR}(p) \le R_{\max}(\widehat{a}).$$

(b) Assume a^* is such that there is actually equality in (a), for a suitable prior p. Show that a^* is then minimax.

(c) Assume more generally that a^* is such that $\mathrm{MBR}(p_m) \to R_{\max}(a^*)$, for a suitable sequence of priors p_m. Show that a^* is indeed minimax.

We note that minimax strategies often but not always have constant risk functions, and that they need not be unique – different minimax strategies for the same problem need to have identical max-risks, but the risk functions themselves need not be identical.

4. Minimax estimation of a normal mean [cf. Nils 2008 #3, 6, 9]

A prototype normal mean model is the simple one with a single observation $y \sim \mathrm{N}(\theta, 1)$. We let the loss function be squared error, $L(\theta, a) = (a - \theta)^2$.

(a) Show that the maximum likelihood (ML) solution is simply $\theta^* = y$. Show that its risk function is $R(\theta^*, \theta) = 1$, i.e. constant.

(b) Let θ have the prior $\mathrm{N}(0, \tau^2)$. Show that (θ, y) is binormal, and that $\theta \,|\, y \sim \mathrm{N}(\rho y, \rho)$, with $\rho = \tau^2/(\tau^2 + 1)$. In particular, $\widehat{\theta}_B(y) = \rho y$ is the Bayes estimator.

(c) Find the risk function for the Bayes estimator, and identify where it is smaller than that of the ML solution, and where it is larger. Comment on the situation where τ is small (and hence ρ), as well as on the case of τ being big (and hence ρ close to 1).

(d) Show that $\mathrm{MBR}(\mathrm{N}(0,\tau^2)) = \rho = \tau^2/(\tau^2+1)$. Use the technique surveyed above to show that y is indeed minimax.

(e) This final point is to exhibit a technique for demonstrating, in this particular situation, that y is not only minimax, but the only minimax solution – this was given as Exercise #9(e) in the Nils 2008 collection, but without any hints. Assume that there is a competitor $\widehat\theta$ that is different from y and also a minimax estimator. Then, since risk functions are continuous (show this), there must be a positive ε and a non-empty interval $[c,d]$ with

$$R(\widehat\theta,\theta) \le \begin{cases} 1-\varepsilon & \text{on } [c,d], \\ 1 & \text{everywhere.} \end{cases}$$

Deduce from this that

$$\mathrm{MBR}(\mathrm{N}(0,p_\tau)) \le \mathrm{BR}(\widehat\theta,p_\tau) \le \int_{[c,d]} (1-\varepsilon)p_\tau(\theta)\,\mathrm{d}\theta + \int_{\text{elsewhere}} 1\cdot p_\tau(\theta)\,\mathrm{d}\theta,$$

writing p_τ for the $\mathrm{N}(0,\tau^2)$ prior. This leads to

$$\varepsilon(2\pi)^{-1/2}\frac{1}{\tau}\int_{[c,d]} \exp(-\tfrac12\theta^2/\tau^2)\,\mathrm{d}\theta \le 1-\mathrm{MBR}(p_\tau) = \frac{1}{\tau^2+1}.$$

Show that this leads to a contradiction: hence y is the single minimax estimator in this problem.

(f) Generalise the above to the situation with $y_1,\ldots,y_n \sim \mathrm{N}(\theta,\sigma^2)$.

5. Minimax estimation of a Poisson mean [cf. Nils 2008 #12]

Let $y\,|\,\theta$ be a Poisson with mean parameter θ, which is is to be estimated with weighted squared error loss $L(\theta,t) = (t-\theta)^2/\theta$. This case was treated in Nils 2008 #12, but here I add more, to take care of the more difficult admissibility point #12(g), where the task is to show that y is the only minimax estimator.

(a) Show that the maximum likelihood (ML) estimator is y itself, and that its risk function is the constant 1.

(b) Consider the prior distribution $\mathrm{Gamma}(a,b)$ for θ. Show that $\mathrm{E}\,\theta = a/b$ and that $\mathrm{E}\,\theta^{-1} = b/(a-1)$ if $a>1$, and infinite if $a\le 1$.

(c) Show that $\theta\,|\,y$ is a $\mathrm{Gamma}(a+y,b+1)$, from which follows

$$\mathrm{E}(\theta\,|\,y) = \frac{a+y}{b+1} \quad\text{and}\quad \mathrm{E}(\theta^{-1}\,|\,y) = \frac{b+1}{a-1+y}.$$

The latter formula holds if $a-1+y>0$, which means for all y if $a\ge 1$, but care is needed if $a<1$ and $y=0$. Show that the Bayes solution is

$$\widehat\theta = \frac{a-1+y}{b+1} \quad\text{for all } y\ge 0,$$

provided $a \geq 1$, but that we need the more careful formula

$$\widehat{\theta} = \begin{cases} (a-1+y)/(b+1) & \text{if } y \geq 1, \\ 0 & \text{if } y = 0, \end{cases}$$

in the case of $a < 1$.

(d) Taking care of the simplest case $a > 1$ first, show that

$$\mathrm{MBR}(p_{a,b}) = \frac{1}{b+1},$$

writing $p_{a,b}$ for the Gamma prior (a,b). This is enough to demonstrate that y is indeed minimax, cf. the Nils 2008 #12 Exercise.

(e) Attempt to show that y is the only minimax estimator via the technique of the previous exercise, starting with a competitor $\widetilde{\theta}$ with risk function always bounded by 1 and bounded by say $1-\varepsilon$ on some non-empty parameter interval $[c,d]$. Show that this leads to

$$\varepsilon \int_{[c,d]} p_{a,b}(\theta)\,\mathrm{d}\theta \leq 1 - \mathrm{MBR}(p_{[a,b]}).$$

For the easier case of $a > 1$, this gives a simple right hand side, but, perhaps irritatingly, not a contradiction – one does not yet know, despite certain valid and bold mathematical efforts, whether y is the unique minimax method or not.

(f) Since the previous attempt ended with 'epic fail', we need to try out the more difficult case $a < 1$ too. Show that

$$\mathrm{E}\{L(\theta,\widehat{\theta})\,|\,y\} = \begin{cases} 1/(b+1) & \text{if } y \geq 1, \\ a/(b+1) & \text{if } y = 0. \end{cases}$$

Deduce from this a minimum Bayes risk formula also for the case of $a < 1$:

$$\mathrm{MBR}(p_{a,b}) = \frac{1}{b+1}\Big\{1 - \Big(\frac{b}{b+1}\Big)^a\Big\} + \frac{a}{b+1}\Big(\frac{b}{b+1}\Big)^a.$$

(g) Find a sufficiently clever sequence of Gamma priors (a_m, b_m), with $a_m \to 1$ from the left and $b_m \to 0$ from the right, that succeeds in squeezing a contradiction out of equality in point(e). Conclude that y is not only minimax, but the only minimax strategy.

(h) Generalise these results to the situation where $y_1, \ldots, y_n$ are independent and Poisson with rates $c_1\theta, \ldots, c_n\theta$, and known multipliers $c_1, \ldots, c_n$. Identify a minimax solution and show that it is the only one on board.

6. Computation of marginal distributions

Assume data y stem from a model density $f(y\,|\,\theta)$ and that there is a prior density $\pi(\theta)$ for the model vector parameter. The *marginal distribution* of the data is then

$$f(y) = \int f(y\,|\,\theta)\pi(\theta)\,\mathrm{d}\theta.$$

In many types of Bayesian analysis this marginal density is not really required, as analysis is rather driven by the posterior distribution $\pi(\theta \,|\, y)$; cf. the recipes and illustrations above. Calculation of $f(y)$ is nevertheless of importance in some situations. It is inherently of interest to understand the distribution of data under the assumptions of the model and the prior (leading e.g. to positive correlations even when observations are independent given the parameter); insights provided by such calculations may lead to new types of models; and numerical values of $f(y)$ are often needed when dealing with issues of different candidate models (see the following exercise).

(a) Let $y \,|\, \theta$ be a binomial (n, θ), and assume $\theta \sim \mathrm{Beta}(k\theta_0, k(1-\theta_0))$. Find the marginal distribution of y, and, in particular, its mean and variance. Exhibit the 'extra-binomial variance', i.e. the quantity with which the variance exceeds $n\theta_0(1-\theta_0)$.

(b) Let $y \,|\, \theta$ be a $\mathrm{N}(\theta, \sigma^2)$, and let θ have the $\mathrm{N}(0, \tau^2)$ prior. Find the marginal distribution of y.

(c) Now assume $y_1, \ldots, y_n$ given θ are i.i.d. from the $\mathrm{N}(\theta, \sigma^2)$ distribution, and let as above $\theta \sim \mathrm{N}(0, \tau^2)$. Find the marginal distribution of the data vector. Show also that

$$\mathrm{corr}(y_i, y_j) = \frac{\tau^2}{\sigma^2 + \tau^2},$$

so the data have positive correlations marginally even though they are independent given the mean parameter. This is a typical phenomenon.

(d) Take $y_1, \ldots, y_n$ to be independent and Poisson θ for given mean parameter, and let $\theta \sim \mathrm{Gamma}(a, b)$. Find an expression for the marginal density of a single y_i, for a pair (y_i, y_j), and for the full vector $y_1, \ldots, y_n$. Find also the marginal means, variances and covariances.

(e) We shall now develop a couple of numerical strategies for computing the actual value of $f(y)$; such will be useful in the model comparison settings below. We think of data y as comprising n observations, and write $\ell_n(\theta) = \log L_n(\theta)$ for the log-likelihood function. Letting $\widehat\theta$ be the maximum likelihood estimate, with $\ell_{n,\max} = \ell_n(\widehat\theta)$, verify first that

$$\begin{aligned} f(y) &= L_n(\widehat\theta) \int \exp\{\ell_n(\theta) - \ell_n(\widehat\theta)\}\pi(\theta)\,\mathrm{d}\theta \\ &\doteq \exp(\ell_{n,\max}) \int \exp\{-\tfrac12(\theta - \widehat\theta)^{\mathrm{t}}\widehat J(\theta - \widehat\theta)\}\pi(\theta)\,\mathrm{d}\theta, \end{aligned}$$

with $\widehat J$ the Hessian matrix $-\partial^2\ell_n(\widehat\theta)/\partial\theta\partial\theta^{\mathrm{t}}$, i.e. the observed information matrix. Derive from this that

$$f(y) = L_{n,\max}R_n, \quad \text{or} \quad \log f(y) = \ell_{n,\max} + \log R_n,$$

where

$$R_n \doteq (2\pi)^{p/2}|\widehat J|^{-1/2}\pi(\widehat\theta), \quad \text{or} \quad \log R_n \doteq -\tfrac12 \log|\widehat J| + \tfrac12 p\log(2\pi) + \log\pi(\widehat\theta).$$

(f) Discuss conditions under which the above Laplace type approximation may expect to provide a good approximation, and when it does not. Consider then the case of n independent observations we may typically write $\widehat{J} = nJ_n^*$, say, with $J_n^* = -n^{-1}\partial^2\ell_n(\widehat{\theta})/\partial\theta\partial\theta^{\rm t}$ converging to a suitable matrix as sample size increases. Show that

$$\begin{aligned}\log f(y) &\doteq \ell_{n,\max} - \tfrac{1}{2}p\log n - \tfrac{1}{2}\log|J_n^*| + \tfrac{1}{2}p\log(2\pi) + \log\pi(\widehat{\theta}) \\ &\doteq \ell_{n,\max} - \tfrac{1}{2}p\log n.\end{aligned}$$

The latter is sometimes called 'the BIC approximation'; see below. Note that it is easy to compute and that it does not even involve the prior.

7. Model averaging and model probabilities

Assume that a data set y has been collected and that more than one parametric model is being contemplated. The traditional statistical view may then be that one of these is 'correct' (or 'best') and that the others are 'wrong' (or 'worse'), with various model selection strategies for finding the correct or best model (see e.g. Claeskens and Hjort, *Model Selection and Model Averaging*, Cambridge University Press, 2008). Such problems may also be tackled inside the Bayesian paradigm, if one is able to assign prior probabilities for the models along with prior densities for the required parameter vector inside each model.

Assume that the models under consideration are $M_1,\ldots,M_k$, where model M_j holds that $y \sim f_j(y\,|\,\theta_j)$, with θ_j belonging to parameter region Ω_j; note that y denotes the full data set, e.g. of the type $y_1,\ldots,y_n$, with or without regression covariates $x_1,\ldots,x_n$, so that f_j denotes the full joint probability density of the data given the parameter vector. Let furthermore $\pi_j(\theta_j)$ be the prior for the parameter vector of model M_j, and, finally, assume $p_j = \Pr(M_j)$ is the probability assigned to model M_j before seeing any data.

(a) Show that the marginal distribution of y has density

$$f(y) = p_1 f_1(y) + \cdots + p_k f_k(y),$$

in terms of the marginal distributions inside each model,

$$f_j(y) = \int f_j(y\,|\,\theta_j)\pi_j(\theta_j)\,{\rm d}\theta_j.$$

(b) Show also that the model probabilities $p_1,\ldots,p_k$ are changed to

$$p_j^* = \Pr(M_j\,|\,{\rm data}) = \frac{p_j f_j(y)}{p_1 f_1(y) + \cdots + p_k f_k(y)} = \frac{p_j f_j(y)}{f(y)}$$

when data have been observed.

(c) Use the results above to deduce the following approximations to the posterior model probabilities:

$$\begin{aligned} p_j^* &= \Pr(M_j \,|\, \text{data}) \\ &\doteq p_j \exp\{\ell_{n,j,\max} - \tfrac{1}{2} p_j \log n - \tfrac{1}{2}\log|J^*_{n,j}| + \tfrac{1}{2} p_j \log(2\pi) + \log \pi(\widehat{\theta}_j)\}/f(y) \\ &\doteq p_j \exp\{\ell_{n,j,\max} - \tfrac{1}{2} p_j \log n\}/f(y), \end{aligned}$$

in terms of maximum likelihood estimates $\widehat{\theta}_j$ for the p_j-dimensional model parameter of model M_j, with associated log-likelihood maximum value $\ell_{n,j,\max}$. This is the argument behind the so-called BIC, the *Bayesian Information Criterion*

$$\text{BIC}_j = 2\ell_{n,j,\max} - p_j \log n,$$

where the model with highest BIC value is declared the winner, in that it has the highest posterior probability (to the order of approximation used).

(d) Sometimes the primary interest may be in learning which model is the most appropriate one, in which case the analysis above is pertinent. In other situations the focus lies with a certain parameter, say μ, assumed to have a precise physical interpretation so that it can be relevantly expressed in terms of θ_j of model M_j, for each of the models considered. In that case one needs the posterior distribution of μ. Show that this may be written

$$\pi(\mu \,|\, \text{data}) = p_1^* \pi_1(\mu \,|\, \text{data}) + \cdots + p_k^* \pi_k(\mu \,|\, \text{data}),$$

in terms of the posterior model probabilities already worked with and of the model-conditional posterior densities $\pi_j(\mu \,|\, \text{data})$.

8. Life lengths in Roman era Egypt

Consider the data set consisting of $n = 141$ life lengths from Roman era Egypt, from Claeskens and Hjort (2008), analysed using in Nils Exam stk 4020 2008.

(a) As in the Exam 2008 exercise, provide a Bayesian analysis, using a Weibull (a, b) model, focussing on the median parameter μ – which under Weibull conditions is equal to $\mu = a(\log 2)^{1/b}$. Using the prior on (a, b) which is uniform over $[10, 50] \times [0.1, 3.0]$, compute the posterior density of μ, via sampling say 10^5 values of (a, b) from the posterior distribution. I find a 90% credibility interval of $[22.852, 28.844]$, and posterior median equal to 25.829.

(b) Similarly carry out a Bayesian analysis of the same data set but now employing the Gamma (c, d) model, again focussing on the median, i.e. $\mu = \text{qgamma}(0.50, \text{c}, \text{d})$ in R notation. Use the prior for (c, d) which is uniform on $[0.5, 2.5] \times [0.01, 0.10]$. Here I find a 90% credibility interval of $[21.817, 27.691]$, and posterior median equal to 24.628.

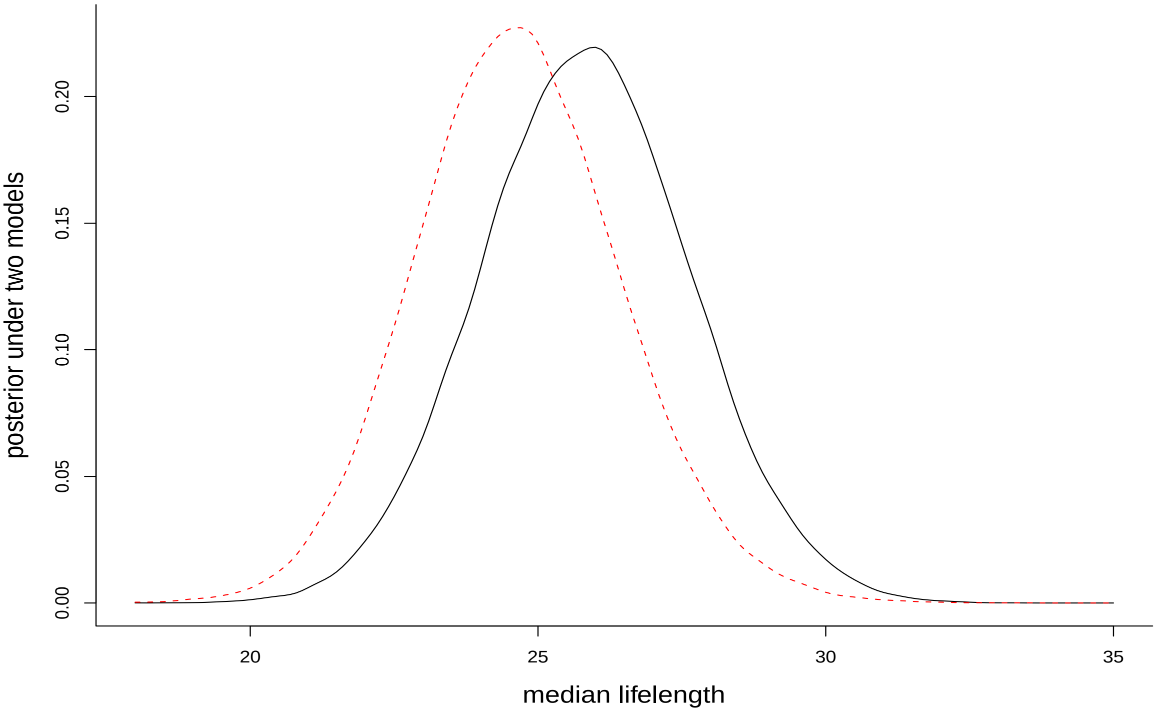

Figure 2: Posterior density for the median life-length in Roman era Egypt, based on respectively the Weibull model (full line) and the Gamma model (dotted line). The posterior model probabilities are respectively 0.825 and 0.175.

(c) Display both posterior distributions (for the same median parameter μ, but computed under respectively the Weibull and the Gamma model) in a diagram, using e.g. histograms or kernel density estimation based on e.g. 10^5 simulations. See Figure 2. These are $\pi_1^*(\mu\,|\,\text{data})$ and $\pi_2^*(\mu\,|\,\text{data})$ in the notation and vocabulary of Exercise 7(d).

(d) Finally compute the posterior model probabilities p_1^* and p_2^*, for the Weibull and the Gamma, using the priors indicated for (a,b) and (c,d). Assume equal probabilities for these two models a priori. Note that these priors do not matter much for the model-based posterior distributions of the median parameter (see Figure 2), but that they do matter quite a bit for the precise computation of p_1^* and p_2^*, via the terms $\log\pi_1(\widehat\theta_w)$ and $\log\pi_2(\widehat\theta_g)$ in the formulae of Exercise 7(c). I find 0.825 and 0.175 for these, with the given priors.

(e) Finally use the methods of Exercise 7(d) to compute and display the overall posterior density of the median life-length, mixing properly over the two parametric models used.

9. The multinormal distribution

'Multivariate statistics' is broadly speaking the area of statistical modelling and analysis where data exhibit dependencies. The most important multivariate distribution is the multinormal one. We say that $X = (X_1, \ldots, X_k)^{\rm t}$ is multinormal with mean vector ξ (a k-vector) and variance matrix Σ (a positive definite $k \times k$ matrix) if its density has the form

$$f(x) = (2\pi)^{-k/2}|\Sigma|^{-1/2}\exp\{-\tfrac{1}{2}(x-\xi)^{\rm t}\Sigma^{-1}(x-\xi)\} \quad \text{for } x \in I\!R^k.$$

We write $X \sim {\rm N}_k(\xi, \Sigma)$ to indicate this. For dimension $k = 1$ this corresponds to the traditional Gaußian ${\rm N}(\xi, \sigma^2)$.

(a) Show that if $X \sim {\rm N}_k(\xi, \Sigma)$ and A is $k \times k$ of full rank, and b a k-vector, then

$$Y = AX + b \sim {\rm N}_k(A\xi + b, A\Sigma A^{\rm t}).$$

Generalise to the situation where A is of dimension $m \times k$ (rather than merely $k \times k$).

(b) Show that if $X \sim {\rm N}_k(\xi, \Sigma)$, then indeed

$${\rm E}\, X = \xi \quad \text{and} \quad {\rm Var}\, X = \Sigma,$$

justifying the semantic terms used above.

(c) Show that X is multinormal if and only if all linear combinations are normal. In particular, if $X \sim {\rm N}_k(\xi, \Sigma)$, then $a^{\rm t}X = a_1X_1 + \cdots + a_kX_k$ is ${\rm N}(a^{\rm t}\xi, a^{\rm t}\Sigma a)$. – We will also allow saying '$X \sim {\rm N}_k(\xi, \Sigma)$' in cases where Σ has less than full rank. in particular, a constant may be seen as a normal distribution with zero variance.

(d) An important property of the multinormal is that a subset of components, conditional on another subset of components, remains multinormal. Show in fact that if

$$X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} \sim {\rm N}_{k_1+k_2}(\begin{pmatrix} \xi^{(1)} \\ \xi^{(2)} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}),$$

then

$$X^{(1)} \,|\, \{X^{(2)} = x^{(2)}\} \sim {\rm N}_{k_1}(\xi^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(x^{(2)} - \xi^{(2)}), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

(e) How tall is Professor Hjort? Assume that the heights of Norwegian men above the age of twenty follows the normal distribution ${\rm N}(\xi, \sigma^2)$, with $\xi = 180$ cm and $\sigma = 9$ cm. Thus, if you have *not yet seen* or bothered to notice this particular aspect of Professor Hjort and his lectures, your point estimate of his height ought to be $\xi = 180$ and a 95% prediction interval for his height would be $\xi \pm 1.96\,\sigma$, or $[162.4, 197.6]$. – Assume now that you learn that his four brothers are actually 195 cm, 207 cm, 196 cm, 200 cm tall, and furthermore that correlations between brothers' heights in the population of Norwegian men is equal to $\rho = 0.80$. Use this information about his four brothers (still assuming that you have not noticed Professor Hjort's height) to revise your initial point estimate of Professor Hjort's height. Is he a five-percent statistical outlier in his family (i.e. outside the 95% prediction interval)?

(f) Assume Professor Hjort has n brothers (rather than merely four) and that you're learning their heights $X_1, \ldots, X_n$. What is the conditional distribution of Professor Hjort's height X_0, given this information? Represent this as a $\mathrm{N}(\xi_n, \sigma_n^2)$ distribution, with proper formulae for its parameters. How small is σ_n for a large number of brothers? (The point here is partly that even if you observe and measure my 99 brothers, there's still a limit to how much you can infer about me.)

10. Simulating from the multinormal distribution

There are special routines that manage to simulate directly from the multinormal distribution, as `mvrnorm` in **R** (preceded by `library(MASS)`, if necessary). These sometimes do not work well for high dimensions. At any rate it is useful to work out different simulation strategies for the multinormal, also for use in Gaußian processes and Gaußian random fields.

(a) Let Σ be a $k \times k$ positive definite symmetric matrix (which is equivalent to saying that it is a covariance matrix, for a suitable k-dimensional probability distribution). Let $\Sigma^{1/2}$ be any matrix square root of Σ, i.e. a symmetric matrix with the property that $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$ (there may in general be several matrices with this property, see the following point). Show that when $U = (U_1, \ldots, U_k)^{\mathrm{t}}$ is a vector of independent standard normals, then

$$X = \Sigma^{1/2} U \sim \mathrm{N}_k(0, \Sigma).$$

This is accordingly a general recipe for simulating from a multinormal vector, via independent standard normals, provided one manages to compute the square root matrix numerically.

(b) By a famous linear algebra theorem, there exist a unitary (or orthonormal) matrix P (with the property that $PP^{\mathrm{t}} = I_k = P^{\mathrm{t}}P$, i.e. its transpose is its inverse) such that

$$P\Sigma P^{\mathrm{t}} = \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_k),$$

where the diagonal Λ matrix has the eigenvalues of Σ along its diagonal (in decreasing order). The P matrix and the $\lambda_1, \ldots, \lambda_k$ values are found numerically in **R** using the `eigen` operation: use

$$\texttt{lambda} = \texttt{eigen(Sigma, symmetric} = \texttt{T)\$values},$$
$$\texttt{P} = \texttt{t(eigen(Sigma, symmetric} = \texttt{T)\$vectors)},$$

and use these to define Λ. (The `symmetric=T` part is not really required, but helps numerical stability for big matrices.) Then indeed the relations above hold, and these imply $\Sigma = P^{\mathrm{t}}\Lambda P$. Show that $\Sigma^{1/2} = P^{\mathrm{t}}\Lambda^{1/2}P$ is symmetric and does the job. Write a few-lined **R** programme, say `squareroot`, which computes `squareroot(Sigma)` for any given `Sigma`.

11. The Ornstein–Uhlenbeck process

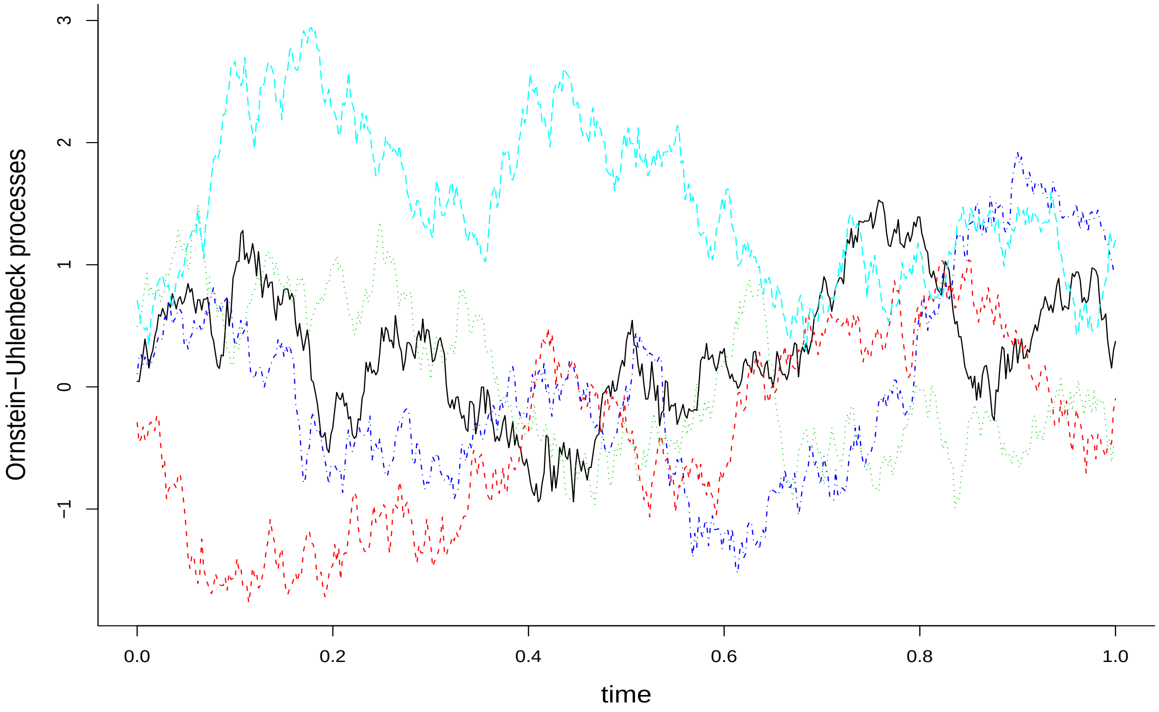

Figure 3: Five simulated Ornstein–Uhlenbeck processes, with dependence parameter $\rho = \exp(-3.00) = 0.0498$. *The grid used for this figure has fineness* $1/m$ *with* $m = 500$.

Consider the so-called Ornstein–Uhlenbeck process $Z = \{Z(t)\colon t \in [0,1]\}$ on the unit interval. This is a zero-mean constant-variance Gaußian process with covariance function

$$\mathrm{cov}\{Z(s), Z(t)\} = \exp\{-a|s-t|\} = \rho^{|s-t|}$$

for a suitable positive parameter a, dictating the degree of autocorrelation.

(a) Your task is now to simulate paths of such a process, say for $a = 3.00$ (which corresponds to correlation $\rho = \exp(-a) = 0.0498$ between pairs distance 1 apart); see Figure 3. Do this by (i) gridding the unit interval to $0/m, 1/m, \ldots, m/m$; (ii) then building the appropriate Σ matrix of size $(m+1)\times(m+1)$ for $Z_{\rm grid} = \{Z(i/m)\colon 0 \le i \le m\}$; (iii) then simulating and plotting such $Z_{\rm grid}$ via the strategy outlined in Exercise 10.

(b) The simulation method used in (a) is 'direct' and 'brute force', involving the square-rooting of a big matrix, and may be slow for a fine grid. Show now that the distribution of $Z(u)$ given $Z(s) = x$ and $Z(t) = y$, where $s < t < u$ are three time-points, in fact does not depend on the $Z(s) = x$ information, only on $Z(t) = y$. This indicates that the Z process is Markovian. Explain how this gives rise to a different simulation strategy, which is in effect much quicker and not hampered by eigen-operations of big matrices.

(c) Suppose one learns that $Z(0) = 0.44, Z(1) = -0.11$. Simulate realisations from the Z process on the unit interval given this information. This may be accomplished via 'brute force' application of the result about conditional multinormal distributions given in Exercise 9(d). It is however instructive and useful to also characterise the distribution of the tied-down Z process. Find in fact formulae for

$$\xi(t) = \mathrm{E}\{Z(t)\,|\,Z(0) = a, Z(1) = b\}, \quad K(s,t) = \mathrm{cov}\{Z(s), Z(t)\,|\,Z(0) = a, Z(1) = b\}.$$

Check in particular $\xi(t)$ and $K(t,t)$ for $t \to 0$ and $t \to 1$.

(d) The Ornstein–Uhlenbeck process may be used as a 'nonparametric prior' for an unknown function. Suppose for illustration that Z is such an unknown function on the unit interval, that the prior used is a process of the above type, with $a = 3.00$, and that Statoil with a few billion Euro has been able to observe that

$$Z(0) = 0.44, Z(0.20) = 0.88, Z(0.70) = -0.55, Z(1) = -0.11.$$

Simulate paths from the posterior distribution of the unknown curve. Use these to compute the probability that the curve has a maximum exceeding 1.50 along with a minimum below -1.50 (up to simulation accuracy). – It is again possible to carry out these simulations 'directly', via the conditioning recipe of Exercise 9(d), but it is more interesting and useful to work out proper characterisations of the conditional Z process given its observed values in a finite number of points. In particular, show that Z splits into independent parts over each of these intervals; it may accordingly be simulated separately over intervals.

12. Alarm or not?

Suppose y is binomial (n, θ), that the action space is $\{\text{alarm}, \text{no alarm}\}$, and that the loss function is as follows:

$$L(\theta, \text{no alarm}) = \begin{cases} 5000 & \text{if } \theta > 0.15, \\ 0 & \text{if } \theta < 0.15, \end{cases},$$

$$L(\theta, \text{alarm}) = \begin{cases} 0 & \text{if } \theta > 0.15, \\ 1000 & \text{if } \theta < 0.15, \end{cases}.$$

Work out when the correct decision is 'alarm', in terms of the posterior distribution, having started with a given prior $p(\theta)$. In particular, for $n = 50$, for which values of y should one decide on 'alarm'? Sort out this for each of the following priors for θ.

(a) θ is uniform on $(0, 1)$.

(b) θ is a Beta $(2, 8)$.

(c) θ is an even mixture of a Beta $(2, 8)$ and a Beta $(8, 2)$.

13. The Dirichlet-multinomial model

The Beta-binomial model, with a Beta distribution for the binomial probability parameter, is on the 'Nice List' where the Bayesian machinery works particularly well: Prior elicitation is easy, as is the updating mechanism. This exercise concerns the generalisation to the Dirichlet-multinomial model, which is certainly also on the Nice List and indeed in broad and frequent use for a number of statistical analyses.

(a) Let $(y_1, \ldots, y_m)$ be the count vector associated with n independent experiments having m different outcomes $A_1, \ldots, A_m$. In other words, y_j is the number of events of type A_j, for $j = 1, \ldots, m$. Show that if the vector of $\Pr(A_j) = p_j$ is constant across the n independent experiments, then the probability distribution governing the count data is

$$f(y_1, \ldots, y_m) = \frac{n!}{y_1! \cdots y_m!} p_1^{y_1} \cdots p_m^{y_m}$$

for $y_1 \geq 0, \ldots, y_m \geq 0, y_1 + \cdots + y_m = n$. This is the multinomial model. Explain how it generalises the binomial model.

(b) Show that

$$\mathrm{E}\, Y_j = np_j, \quad \mathrm{Var}\, Y_j = np_j(1 - p_j), \quad \mathrm{cov}(Y_j, Y_k) = -np_j p_k \text{ for } j \neq k.$$

(c) Now define the Dirichlet distribution over m cells with parameters $(a_1, \ldots, a_m)$ as having probability density

$$\pi(p_1, \ldots, p_{m-1}) = \frac{\Gamma(a_1 + \cdots + a_m)}{\Gamma(a_1) \cdots \Gamma(a_m)} p_1^{a_1 - 1} \cdots p_{m-1}^{a_{m-1} - 1} (1 - p_1 - \cdots - p_{m-1})^{a_m - 1},$$

over the simplex where each $p_j \geq 0$ and $p_1 + \cdots + p_{m-1} \leq 1$. Of course we may choose to write this as

$$\pi(p_1, \ldots, p_{m-1}) \propto p_1^{a_1 - 1} \cdots p_{m-1}^{a_{m-1} - 1} p_m^{a_m},$$

with $p_m = 1 - p_1 - \cdots - p_{m-1}$; the point is however that there are only $m - 1$ unknown parameters in the model as one knows the mth once one learns the values of the other $m - 1$. Show that the marginals are Beta distributed,

$$p_j \sim \mathrm{Beta}(a_j, a - a_j) \quad \text{where } a = a_1 + \cdots + a_m.$$

(d) Infer from this that

$$\mathrm{E}\, p_j = p_{0,j} \quad \text{and} \quad \mathrm{Var}\, p_j = \frac{1}{a + 1} p_{0,j}(1 - p_{0,j}),$$

in terms of $a_j = ap_{0,j}$. Show also that

$$\mathrm{cov}(p_j, p_k) = -\frac{1}{a + 1} p_{0,j} p_{0,k} \quad \text{for } j \neq k.$$

For the 'flat Dirichlet', with parameters $(1, \dots, 1)$ and prior density $(m-1)!$ over the simplex, find the means, variances, covariances.

(e) Now for the basic Bayesian updating result. When $(p_1, \dots, p_m)$ has a $\mathrm{Dir}(a_1, \dots, a_m)$ prior, then, given the multinomial data, show that

$$(p_1, \dots, p_m) \,|\, \mathrm{data} \sim \mathrm{Dir}(a_1 + y_1, \dots, a_m + y_m).$$

Give formulae for the posterior means, variances, and covariances. In particular, explain why

$$\widehat{p}_j = \frac{a_j + y_j}{a + n}$$

is a natural Bayes estimate of the unknown p_j. Also find an expression for the posterior standard deviation of the p_j.

(f) In order to carry out easy and flexible Bayesian inference for $p_1, \dots, p_m$ given observed counts $y_1, \dots, y_m$, one needs a recipe for simulating from the Dirichlet distribution. One such is as follows: Let $X_1, \dots, X_m$ be independent with $X_j \sim \mathrm{Gamma}(a_j, 1)$ for $j = 1, \dots, m$. Then the ratios

$$Z_1 = \frac{X_1}{X_1 + \cdots + X_m}, \dots, Z_m = \frac{X_1}{X_1 + \cdots + X_m}$$

are in fact $\mathrm{Dir}(a_1, \dots, a_m)$. Try to show this from the transformation law for probability distributions: If X has density $f(x)$, and $Z = h(X)$ is a one-to-one transformation with inverse $X = h^{-1}(Z)$, then the density of Z is

$$g(z) = f(h^{-1}(z)) \Big| \frac{\partial h^{-1}(z)}{\partial z} \Big|$$

(featuring the determinant of the Jacobian of the transformation). Use in fact this theorem to find the joint distribution of $(Z_1, \dots, Z_{m-1}, S)$, where $S = Z_1 + \cdots + Z_m$ (one discovers that the Dirichlet vector of Z_j is independent of their sum S).

(g) The Dirichlet distribution has a nice 'collapsibility' property: If say $(p_1, \dots, p_8)$ is $\mathrm{Dir}(a_1, \dots, a_8)$, show that then the collapsed vector $(p_1 + p_2, p_3 + p_4 + p_5, p_6, p_7 + p_8)$ is $\mathrm{Dir}(a_1 + a_2, a_3 + a_4 + a_5, a_6, a_7 + a_8)$.

14. Gott würfelt nicht

but I do so, on demand. I throw a certain moderately strange-looking die 30 times and have counts $(2, 5, 3, 7, 5, 8)$ of outcomes 1, 2, 3, 4, 5, 6. Use either of the priors

. 'flat', $\mathrm{Dir}(1, 1, 1, 1, 1, 1)$,

. 'symmetric but more confident', $\mathrm{Dir}(3, 3, 3, 3, 3, 3)$,

. 'unwilling to guess', $\mathrm{Dir}(0.1, 0.1, 0.1, 0.1, 0.1, 0.1)$

for the probabilities $(p_1, \ldots, p_6)$ to assess the posterior distribution of each of the following quantities:

$$\begin{aligned}
\rho &= p_6/p_1, \\
\alpha &= (1/6)\sum_{j=1}^{6}(p_j - 1/6)^2, \\
\beta &= (1/6)\sum_{j=1}^{6}|p_j - 1/6|, \\
\gamma &= (p_4 p_5 p_6)^{1/3}/(p_1 p_2 p_3)^{1/3}.
\end{aligned}$$