STK 4020: Bayesian Statistics
Autumn 2012

Course Notes and Exercises
by Nils Lid Hjort

– This version: as of 10/11/12 –

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 $Z \sim \text{Gamma}(a, b)$ 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

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

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

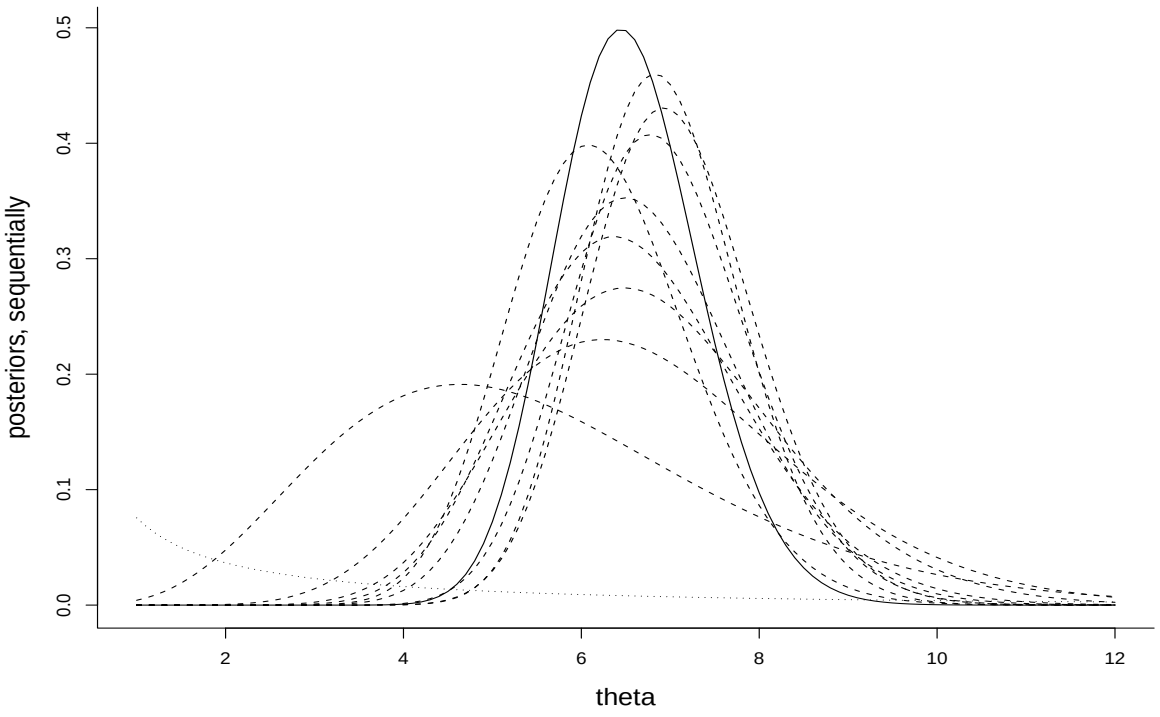

Figure 1: Eleven curves are displayed, corresponding to the $\text{Gamma}(0.1, 0.1)$ *intial prior density for the Poisson parameter* θ *along with the ten updates following each of the observations 6, 8, 7, 6, 7, 4, 11, 8, 6, 3.*

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

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

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

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 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 $p(\theta \,|\, \text{data}_j)$, where data_j is $y_1, \ldots, y_j$, along with the prior density. Also compute the ten Bayes estimates $\widehat{\theta}_j = \text{E}(\theta \,|\, \text{data}_j)$ 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 $p(y \,|\, \theta)$ 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 $\{\text{reject}, \text{accept}\}$ 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:

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

(In particular, in this expectation operation the random element is y, having its $p(y \,|\, \theta)$ 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

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

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

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

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

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

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 $\theta \sim \text{Beta}(k\theta_0, k(1-\theta_0))$. 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 $\text{N}(\theta, \sigma^2)$, and let θ have the $\text{N}(0, \tau^2)$ prior. Find the marginal distribution of y.

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

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

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 $\theta \sim \text{Gamma}(a, b)$. 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

$$\begin{aligned} f(y) &= L_n(\widehat{\theta}) \int \exp\{\ell_n(\theta) - \ell_n(\widehat{\theta})\}\pi(\theta)\,\text{d}\theta \\ &\doteq \exp(\ell_{n,\max}) \int \exp\{-\tfrac{1}{2}(\theta - \widehat{\theta})^{\text{t}} \widehat{J}(\theta - \widehat{\theta})\}\pi(\theta)\,\text{d}\theta, \end{aligned}$$

with $\widehat{J}$ the Hessian matrix $-\partial^2 \ell_n(\widehat{\theta})/\partial\theta\partial\theta^{\text{t}}$, 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

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

(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 $J_n^* = -n^{-1}\partial^2\ell_n(\widehat{\theta})/\partial\theta\partial\theta^{\mathrm{t}}$ 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,

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

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

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

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.

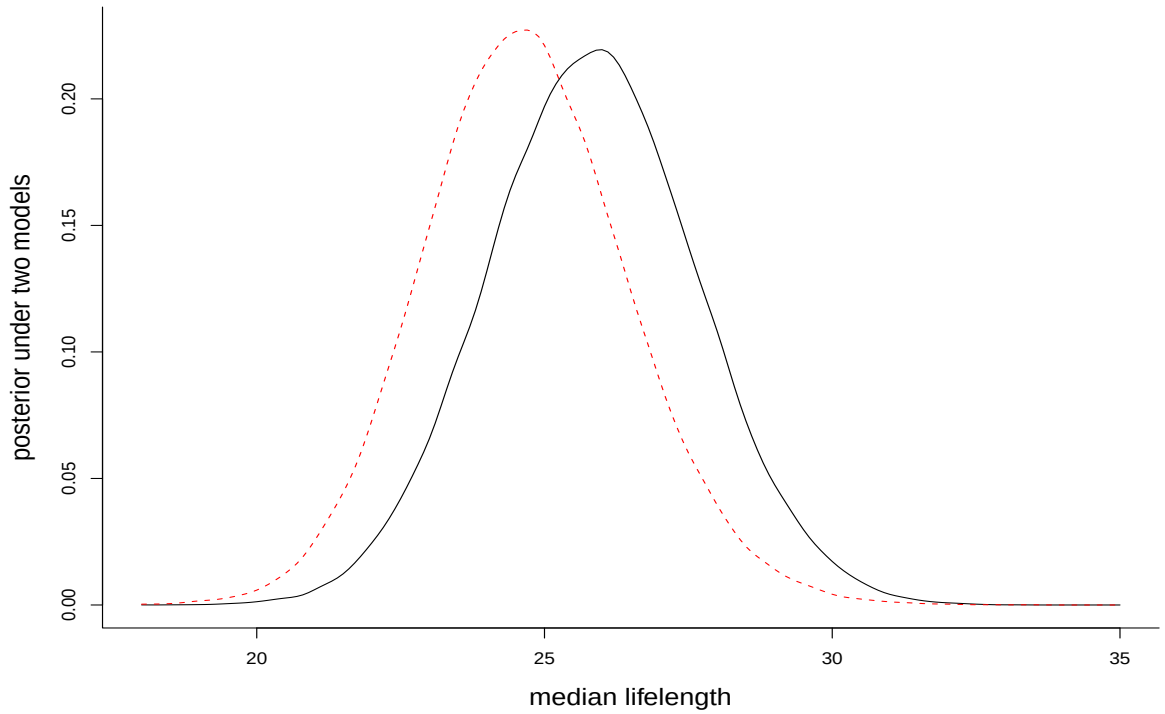

Figure 2: Posterior density for the median life-length in Roman era Egypt, based on respectively the Weibull model (full line) and the Gamma model (dotted line). The posterior model probabilities are respectively 0.825 and 0.175.

(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 $X = (X_1, \ldots, X_k)^{\mathrm{t}}$ 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

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

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

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

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

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

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

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

justifying the semantic terms used above.

(c) Show that X is multinormal if and only if all linear combinations are normal. In particular, if $X \sim \mathrm{N}_k(\xi, \Sigma)$, then $a^{\mathrm{t}} X = a_1 X_1 + \cdots + a_k X_k$ is $\mathrm{N}(a^{\mathrm{t}}\xi, a^{\mathrm{t}}\Sigma a)$. – We will also allow saying '$X \sim \mathrm{N}_k(\xi, \Sigma)$' 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

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

then

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

(e) How tall is Professor Hjort? Assume that the heights of Norwegian men above the age of twenty follows the normal distribution $\mathrm{N}(\xi, \sigma^2)$, 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

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

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

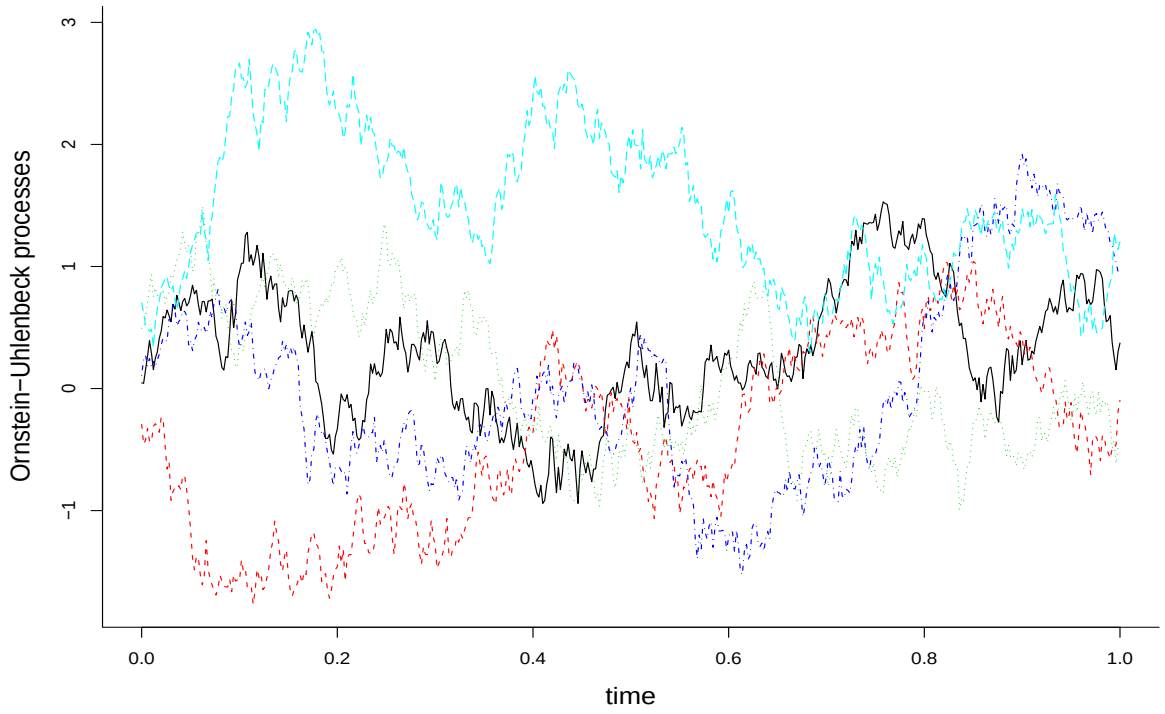

Figure 3: Five simulated Ornstein–Uhlenbeck processes, with dependence parameter $\rho = \exp(-3.00) = 0.0498$. *The grid used for this figure has fineness* $1/m$ *with* $m = 500$.

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

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

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 $Z_{\text{grid}} = \{Z(i/m): 0 \le i \le m\}$; (iii) then simulating and plotting such Z_{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

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

(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

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

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 $(1, \ldots, 1)$ and prior density $(m-1)!$ over the simplex, find the means, variances, covariances.

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

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

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 $p_1, \ldots, p_m$ given observed counts $y_1, \ldots, y_m$, one needs a recipe for simulating from the Dirichlet distribution. One such is as follows: Let $X_1, \ldots, X_m$ be independent with $X_j \sim \text{Gamma}(a_j, 1)$ for $j = 1, \ldots, m$. Then the ratios

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

are in fact $\text{Dir}(a_1, \ldots, a_m)$. 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 $(Z_1, \ldots, Z_{m-1}, S)$, 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 $(p_1, \ldots, p_8)$ is $\text{Dir}(a_1, \ldots, a_8)$, show that then the collapsed vector $(p_1 + p_2, p_3 + p_4 + p_5, p_6, p_7 + p_8)$ is $\text{Dir}(a_1 + a_2, a_3 + a_4 + a_5, a_6, a_7 + a_8)$.

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.

(a) Use either of the priors
- 'flat', $\text{Dir}(1, 1, 1, 1, 1, 1)$,
- 'symmetric but more confident', $\text{Dir}(3, 3, 3, 3, 3, 3)$,
- 'unwilling to guess', $\text{Dir}(0.1, 0.1, 0.1, 0.1, 0.1, 0.1)$

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

$$\rho = p_6/p_1,$$
$$\alpha = (1/6)\sum_{j=1}^{6}(p_j - 1/6)^2,$$
$$\beta = (1/6)\sum_{j=1}^{6}|p_j - 1/6|,$$
$$\gamma = (p_4p_5p_6)^{1/3}/(p_1p_2p_3)^{1/3}.$$

(b) The above priors are slightly artificial in this context, since they do not allow the explicit possibility that the die in question is plain boring utterly simply a correct one, i.e. that $p = p_0 = (1/6, \ldots, 1/6)$. The priors used hence do not give us the possibility to admit that ok, then, perhaps $\rho = 1$, $\alpha = 0$, $\beta = 0$, $\gamma = 1$, after all. This motivates using a mixture prior which allows a positive chance for $p = p_0$. Please therefore redo the Bayesian analysis above, with the same $(2, 5, 3, 7, 5, 8)$ data, for the prior $\frac{1}{2}\,\delta(p_0) + \frac{1}{2}\,\mathrm{Dir}(1, 1, 1, 1, 1, 1)$. Here $\delta(p_0)$ is the 'degenerate prior' that puts unit point mass at position p_0. Compute in particular the posterior probability that $p = p_0$, and display the posterior distributions of $\rho, \alpha, \beta, \gamma$.

15. Rejection-acceptance sampling

This exercise provides the basics of the so-called rejection-acceptance sampling strategy. It it presented and exemplified here in the general framework of random variables on certain sample spaces, but applies for this course particularly fruitfully in situations where the target density is the posterior distribution (say $\pi(\theta \,|\, \text{data})$, rather than the generic density $f(x)$ used in this particular exercise).

(a) Let Y come from some density $g(y)$, and assume that we choose to keep the Y with probability $h(y)$; otherwise we throw it away and go on to the next round. Show that an accepted Y then follows the density

$$f(x) = g(x)h(x) \Big/ \int g(x)h(x)\,\mathrm{d}x.$$

(b) Suppose we wish to draw Xs from some density $f(x)$ but that it appears difficult to do so 'directly'. Assume that $f(x) \le Mg(x)$ for all x, where g is an easier job to draw samples from. Show that the two-step algorithm that first draws Y from g, and then keeps this value with probability $f(y)/\{Mg(y)\}$, succeeds in its aim, i.e. being a sample from f. – What is the frequency of rejected Y values, i.e. of 'wasted efforts'?

(c) Let

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1} \quad \text{for } 0 < x < 1,$$

i.e. the Beta distribution with parameters (a, b). Show that f is unlimited if a or b is smaller than 1, and finds its maximum value M_0 for the case $a \ge 1$, $b \ge 1$.

(d) Let $a = 1.33$ and $b = 1.67$. Draw $n = 1000$ samples from the Beta distribution with these parameters, using the rejection algorithm that starts with uniforms. How many Ys did you need to make, in order to produce 1000 Xs?

(e) Suppose in general terms that we wish to sample from a density of the form $f(x) = g(x)/I$, where g is nonnegative over a certain region and $I = \int g \, \mathrm{d}x$. Assume (i) that we sample X from a (simpler) start-density $h(x)$, where $g(x) \leq Mh(x)$ for all x, for some K; and (ii) that we keep this candidate X with probability $g(x)/\{Mh(x)\}$. Verify that the probability density of a surviving X is really $f(x)$. – The importance of this variation on the rejection sampling recipe of point (a) above lies in the fact that we do not need to know the number I, i.e. it is sufficient to know the target density up to an (unknown) factor.

(f) Set up a rejection sampling regime to get hold of say 100,000 samples (X_i, Y_i) from the density

$$f(x, y) = g(x, y)/I, \quad \text{where } g(x, y) = \exp\{\sin(\sqrt{|xy|}) \exp(|y|^{3/2})\},$$

and I is its integral over $[0, 1] \times [0, 1]$. Make fine histograms of the two marginal distributions, and find means, standard deviations, and the correlation, numerically.

(g) Consider the following idiosyncratic recipe for creating $\mathrm{N}(0, 1)$ variables: sample X from the $\mathrm{N}(0, 2)$ (standard deviation $\sqrt{2}$), and keep with probability $\exp(-\frac{1}{4}X^2)$. Verify that the recipe works. Simulate 100,000 samples in this way, and set up a Pearson test with 1000 cells to test statistically that the recipe works.

(h) The binormal density, for the case of means equal to zero and standard deviations equal to one, is of the form

$$f(x, y) = \frac{1}{2\pi} \frac{1}{(1 - \rho^2)^{1/2}} \exp\Big\{-\tfrac{1}{2} \frac{1}{1 - \rho^2}(x^2 + y^2 - 2\rho xy)\Big\}.$$

Use rejection sampling to generate 10,000 pairs (X, Y) from this binormal distribution, for a couple of values of the correlation parameter ρ. Plot the pairs and make some empirical checks that your algorithm works properly.

16. The Metropolis and Metropolis–Hastings algorithm

Let (π_i) be a probability distribution over some large sample space. The task is to simulate realisations from this distribution.

(a) Let $X_1, X_2, X_3, \ldots$ come from a Markov chain with transition probability matrix $P_{i,j} = \Pr\{X_{n+1} = j \,|\, X_n = i\}$. Show that if these are constructed such that

$$\pi_i P_{i,j} = \pi_j P_{j,i} \quad \text{for all } i, j,$$

and also that the chain is irreducible with period 1, then the stationary (or equilibrium) distribution for the chain is actually (π_i).

(b) There ought to be quite some elbow room for many different $P_{i,j}$ constructions that obey the conditions of (a). The Metropolis method of 1953 uses

$$P_{i,j} = Q_{i,j} \min\Big(1, \frac{\pi_j}{\pi_i}\Big) \quad \text{for } j \neq i,$$

where $Q_{i,j} = \Pr\{X' = j \,|\, X = i\}$ is the so-called proposal distribution, assumed here to be symmetric ($Q_{i,j} = Q_{j,i}$). Show that the condition of (a) really is in force with such $P_{i,j}$ constructions.

(c) Sometimes it is however practical, or even necessary, to employ $Q_{i,j}$ that are not symmetric in (i, j). Let $Q_{i,j}$ be a potentially non-symmetric proposal distribution that from the present i proposes a j. Attempt using an accept probability of the type $\min(1, S_{i,j}\pi_j/\pi_i)$, i.e.

$$P_{i,j} = Q_{i,j} \min\Big(1, S_{i,j} \frac{\pi_j}{\pi_i}\Big).$$

Show that this really works, provided $S_{i,j} = Q_{j,i}/Q_{i,j}$! This amounts to Hastings's 1970 generalisation of the Metropolis algorithm: propose j from a symmetric or non-symmetric $Q_{i,j}$, and accept with probability

$$\min\Big(1, \frac{Q_{j,i}}{Q_{i,j}} \frac{\pi_j}{\pi_i}\Big).$$

(d) Comment specifically on the special cases where $Q_{i,j}$ is symmetric and where $Q_{i,j} = q_j$ is independent of i.

17. The continuous space Metropolis algorithm

Methods and results from the previous exercise have analogues in the continuous world. The task and challenge is to simulate samples from a given continuous density $f(x)$. The methods we develop now are meant to be able to work even in high dimension. If judged instructive you may prefer to think in terms of a given $f(x)$ that is too difficult to attack with more direct means. Let $q(y \,|\, x)$ be a proposal distribution that for a given x proposes a y.

(a) The Metropolis–Hastings method consists in generating $X_0, X_1, X_2, \ldots$, by giving X_0 some start value and by letting

$$X_{i+1} = \begin{cases} Y_i & \text{with probabiloity } \mathrm{pr}_i, \\ X_i & \text{with probability } 1 - \mathrm{pr}_i, \end{cases}$$

where Y_i is drawn from $q(y \,|\, X_i)$, and where

$$\mathrm{pr}_i = \min\Big(1, \frac{q(X_i \,|\, Y_i)}{q(Y_i \,|\, X_i)} \frac{f(Y_i)}{f(X_i)}\Big).$$

Show, heuristically if needed, that the Markov process $X_1, X_2, X_3, \ldots$ indeed has $f(x)$ as its equilibrium distribution.

(b) Explain which conditions that ought to be met in order for the simulation strategy just described being practically effective.

(c) Study and comment on the special cases where $q(y \,|\, x) = q(x \,|\, y)$ and where $q(y \,|\, x) = q_0(y)$ is independent of x.

(d) You are to simulate 10000 data points from the density

$$f(x) = \frac{1}{\Gamma(\frac{3}{2})}\, x^{1/2}\, e^{-x},$$

i.e. the Gamma density with parameters $(\frac{3}{2}, 1)$. This is easily done in **R**, but the task is to achieve this via the Metropolis–Hastings algorithm, with proposal distribution $q(y \,|\, x)$ equal to the uniform on $[\frac{1}{3}x, 3x]$. Compute the mean and standard deviation for the 10000 points you generate, and compare with the theoretical values.

18. Using Metropolis for a steep distribution

One would like to simulate independent realisations $X_1, \ldots, X_n$ from the probability distribution $\pi_j = j/c_M$ over the set $\{1, \ldots, M\}$, where $c_M = j(j+1)/2$. This is an easy task for low and moderate M, and an **R** routine is at your disposal. If however M is large the problem is more difficult, and Markov Chain Monte Carlo methods may become necessary. In the following points, let first $M = 20$, for the sake of easy illustrataion; the MCMC machinery is then not necessary, but it is a good exercise to solve the problem using these tools.

(a) Run for free in **R**: use the command

```
x0 <- sample(list, sim, replace=T, prob)
```

to simulate say 1000 data points $X_{0,i}$ from the distribution π_j. Check that the data points really appear to come from the wished-for distribution over $\{1, \ldots, 20\}$, by checking the histogram, and by using the Pearson test statistic.

(b) Then try the Metropolis method. The challenge is to simulate a Markov chain $Y_1, Y_2, \ldots$ over $\{1, \ldots, 20\}$ that has $(\pi_1, \ldots, \pi_{20})$ as its equilibrium distribution, and that only uses very simply transitions mechanisms. Implement the Metropolis algorithm for this purpose, where you use as proposal that Y_i moves up one step or down one step, from its previous value Y_{i-1}, with equal probability $\frac{1}{2}$. Then the proposal is accepted with probability $\min(1, \pi(Y_i)/\pi(Y_{i-1}))$ (where I write $\pi(j)$ for π_j). Here 'up' and 'down' is meant as with 'clock addition modulo M'; up one step from M means 1, down one step from 1 means M.

(c) Who invented the so-called H bomb?

(d) Let the chain run for a long while, say $Y_1, \ldots, Y_{5000}$. Check if the Y_i can be seen as a (correlated) sample from the π_j-fordelingen.

(e) Take out each 100th Y from the chain, and check if the sub-chain $Y_{100}, Y_{200}, Y_{300}, \ldots$ can be seen as making up an independent sample from the π_j distribution.

(f) The method described above runs into certain problems when M is large, say $M = 5000$. What kind of problems, and Что делать (as Lenin said)? Discuss some alternative proposal mechanisms (i.e. the choice of symmetric $Q_{i,j}$ matrix) inside the MCMC chain above. Implement and try out.

19. Metropolis for a distribution for telephone numbers

Consider a probability distribution across all natural numbers from zero up to a million million, defined by

$$\pi(x_1, \ldots, x_{12}) = \frac{1}{Z(\lambda)} \exp\Big\{-\lambda \sum_{j=1}^{12} (x_j - \bar{x})^2\Big\} \quad \text{for } (x_1, \ldots, x_{12}) \in \{0, \ldots, 9\}^{12}.$$

Here $Z(\lambda)$ is the required summation constant, that perhaps not even HAL could manage to compute accurately, and $\bar{x} = (x_1 + \cdots + x_{12})/12$. We may think of an outcome $x = (x_1, \ldots, x_{12})$ as a random telephone number, in a country employing 12-digit telephone numbers.

(a) What kind of numbers will be preferred by this distribution, i.e. what type of x are likely and what type less likely? Describe some aspects of outcomes, for situations where λ is respectively negative, close to zero, moderate, and large.

(b) How can one manage to sample say 10,000 random telephone numbers from this distribution, for a given λ? Set up and implement a Metropolis algorithm for achieving this, and discuss how well it works.

(c) Let

$$\xi(\lambda) = \mathrm{E}_\lambda U(X) = \mathrm{E}_\lambda \sum_{j=1}^{12} (X_j - \bar{X})^2,$$

the mean of the random variance, as a function of the underlying λ. Set up simulations in a loop across λ values from -3 to 3, to find numerical approximations for $\xi(\lambda)$, and plot the resulting curve. Check directly that $\xi(0) = 90.75$.

(d) I have got hold of $n = 200$ telephone numbers from the country in question, and computed $U(x) = \sum_{j=1}^{12} (x_j - \bar{x})^2$ for each of these. Their average value turns out to be $\bar{U} = 11.111$. Estimate the λ parameter. (Answer: show that maximum likelihood estimation is equivalent to solving $\xi(\lambda) = 11.111$, and use simulation to show that its solution is $\widehat{\lambda} \doteq 0.488$.)

(e) How can you supplement the $\widehat{\lambda}$ parameter estimate you found in (d) with a confidence interval, or a standard deviation estimate?

(f) Construct also a Gibbs Sampler to solve the simulation problem, implement it, and test its efficiency vs. the direct Metropolis method above. Here you will need

$$\begin{aligned} \pi(x_i \,|\, \text{rest}) &= \Pr\{X_i = x_i \,|\, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{12}) \\ &= \frac{g_i(x_i \,|\, \text{rest})}{\sum_{y=0}^{9} g_i(y \,|\, \text{rest})}, \end{aligned}$$

where g_i ought to be made as simple (and easily interpretable) as possible.

20. Autocorrelation in simulation output

Situations with independence tend to be much easier to analyse than for cases with dependence. This comments also applies to simulation output; if such output stems from MCMC then one must expect positive correlations between neighbouring realisations, with consequences for precision of estimates etc. This exercise briefly considers the phenomenon of autocorrelation and some of its implications.

(a) Suppose $X_1, \ldots, X_n$ are independent with the same distribution, with mean $\mu = \mathrm{E}\, X_i$. Then, famously, $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$ has mean μ and variance σ^2/n, where σ is the standard deviation of X_i. Verify this, and show that

$$\mathrm{CI}_n = \bar{X} \pm 1.96\,\widehat{\sigma}/\sqrt{n}$$

is a confidence interval that captures the μ parameter with probability tending to 0.95. The sole condition securing this statement is that the standard deviation is finite.

(b) Assume now that the X_is are again from the same distribution, with mean μ and standard deviation σ, but that they are dependent, with

$$\mathrm{cov}(X_i, X_j) = \sigma^2 \rho^{|j-i|}, \quad \text{or} \quad \mathrm{corr}(X_i, X_j) = \rho^{|j-i|},$$

for an appropriate autocorrelation parameter ρ. Typically, ρ is in $(0,1)$, but may in certain special cases also be negative. Show that

$$\mathrm{Var}\,\bar{X}_n = \frac{\sigma^2}{n}\Big\{1 + \frac{2\rho}{1-\rho}\Big(1 - \frac{1}{n}\frac{1-\rho^n}{1-\rho}\Big)\Big\} \doteq \frac{\sigma^2}{n}\frac{1+\rho}{1-\rho}.$$

Under various mild conditions on the exact nature of the dependence one may prove that

$$\sqrt{n}(\bar{X}_n - \mu) \to_d \mathrm{N}(0, \sigma^2\frac{1+\rho}{1-\rho}).$$

(c) The consequence for estimation of means based on autocorrelated simulation output is that *the variances are inflated*. In particular, the confidence interval of (a) is now too naive, is too narrow, and undershoots its intended level of confidence. Show that the real coverage probability of that confidence interval tends to

$$p = \Pr\Big\{|\mathrm{N}(0,1)| \le \sqrt{\frac{1-\rho}{1+\rho}}1.96\Big\}.$$

With $\rho = 0.90$, for example, which may be a typical value for various MCMC schemes, one finds that the real confidence level is around 0.347 rather than the intended 0.95.

(d) A better confidence interval, under autocorrelation conditions, is

$$\mathrm{CI}_n^* = \bar{X}_n \pm 1.96\frac{\widehat{\sigma}}{\sqrt{n}}\sqrt{\frac{1+\widehat{\rho}}{1-\widehat{\rho}}},$$

for a suitable estimate of ρ. One such estimate is

$$\widehat{\rho} = \frac{1}{n-1} \sum_{i=2}^{n} \frac{X_i - \bar{X}}{\widehat{\sigma}_0} \frac{X_{i-1} - \bar{X}}{\widehat{\sigma}_0},$$

where $\widehat{\sigma}_0$ is an estimate of the standard deviation (not identical to the usual empirical standard deviation). Discuss versions of such a $\widehat{\sigma}_0$.

(e) For the random telephone numbers model of Exercise 19, use the described Metropolis Markov chain $X_1, X_2, \ldots$ that converges in distribution to the target distribution, and use `acf` in **R** to assess the degree of autocorrelation in the chain of $U(X_1), U(X_2), \ldots$. Concretely, `acf(Usim)` produces an autocorrelation plot of the simulated $U(X_i)$ values, and `acf(Usim)$acf` gives the estimated correlation values for pairs of points 1 position apart, 2 positions apart, 3 positions apart, etc.

(f) For the telephone numbers model, construct a diagram that displays (i) the estimated $\widehat{\xi}(\lambda)$ curve, for values $0 \leq \lambda \leq 2$ and (ii) pointwise 95% confidence intervals, qua upper and lower curves:

$$\Pr\{a(\lambda) \leq \xi(\lambda) \leq b(\lambda)\} \doteq 0.95 \quad \text{for each } \lambda.$$

(g) A simple trick for avoiding too high autocorrelation is to 'skip data', keeping e.g. only simulated values corresponding to positions 1001, 1051, 1101, 1151, etc. for final analysis. Discuss aspects of such schemes.

21. Two Metropolis–Hastings exercises

This exercise provides two reasonably simple illustrations of uses of the Metropolis type algorithm. It is useful to use these two and similar simpler situations as 'playing grounds', both for investigating aspects of different tuning parameters, start values, etc., and to get a sense for the type of computer programmes necessary in bigger and more complex problems.

(a) Consider the distribution with probability function

$$f(x) = c \exp(-\lambda |x|^{\alpha}) \quad \text{for } x = 0, \pm 1, \pm 2, \ldots,$$

with $c = c(\lambda, \alpha)$ the summation constant. Show that this indeed defines a distribution on the integers provided λ and α are positive.

(b) For given values of λ, α, set up a Metropolis algorithm for creating a Markov chain $X_1, X_2, \ldots$ with proposal $X_{n+1} = X_n \pm 1$, using equal probabilities for $X_n + 1$ and $X_n - 1$. Implement the procedure, run the chain for a coupl' o' values of (λ, α), and demonstrate that it really converges in distribution to the target distribution f (even if you start the chain far out of the main probability domain). For $\lambda = 1$, throw in another programme loop to compute and draw the curve $\mathrm{sd}(\alpha) = \mathrm{sd}_{\alpha}(X)$, portraying the standard deviation as a function of α.

(c) A perhaps rather silly but nevertheless worthwhile method of simulating from the standard normal density is by means of a Metropolis scheme with proposals created by uniform perturbations. Specifically, set up a Markov chain $X_1, X_2, \ldots$ with proposals $X_{i+1,\text{prop}} = X_i + \delta U_i$, where the U_i are i.i.d. uniform on $[-1, 1]$ and δ a parameter signalling whether the proposed changes are big or small. Accept these proposals in the Metropolis fashion, with the standard normal as target. Run the chain and demonstrate that its equilibrium distribution is indeed the standard normal. Add another loop to your programme to monitor the acceptance rate as a function of δ. For which δ is the acceptance rate equal to the quasi-magical value 0.234 (which is the optimal balancing value, according to some criteria)?

22. Gamma-normal conjugate inference for the normal model

Let data $y_1, \ldots, y_n$ for given parameters μ and σ be i.i.d. $\mathrm{N}(\mu, \sigma^2)$. We know that when σ may be taken as a known quantity, then the canonical class of priors for μ is the normal one. When both parameters are unknown, however, as in most practical encounters, a more elaborate analysis is called for.

(a) Show that the likelihood function may be written as being proportional to

$$L_n(\mu, \sigma) = \exp\Big[-n \log \sigma - \tfrac{1}{2} \frac{1}{\sigma^2} \{Q_0 + n(\mu - \bar{y})^2\Big],$$

where $\bar{y} = (1/n) \sum_{i=1}^n y_i$ and $Q_0 = \sum_{i=1}^n (y_i - \bar{y})^2$.

(b) With *any* given prior $p(\mu, \sigma)$, explain how you may set up a Metropolis type MCMC to draw samples from the posterior distribution. Try this out in practice, using the prior that takes μ and $\log \sigma$ independent and uniform on say $[5, 5]$ and $[10, 10]$, with data that you simulate for the occasion from a $\mathrm{N}(2.345, 1.234^2)$, with $n = 25$. Note that this approach does not need more mathematical algebra as such, apart from the likelihood function above.

(c) There is however a popular and convenient conjugate class of priors for which posterior distributions become particularly clear, with the appropriate algebraic efforts. These in particular involve placing a Gamma prior on the inverse variance $\lambda = 1/\sigma^2$. Say that (λ, μ) has the gamma-normal distribution with parameters (a, b, μ_0, v), and write this as

$$(\lambda, \mu) \sim \mathrm{GN}(a, b, \mu_0, v),$$

provided

$$\lambda = 1/\sigma^2 \sim \mathrm{Gamma}(a, b) \quad \text{and} \quad \mu \,|\, \sigma \sim \mathrm{N}(\mu_0, \sigma^2/v).$$

Show that the prior can be expressed as

$$p(\lambda, \mu) \propto \lambda^{a-1} \lambda^{1/2} \exp\big[-\lambda\{b + \tfrac{1}{2}(\mu - \mu_0)^2 v\}\big].$$

What is the unconditional prior variance of μ?

(d) Prove first the convenient formula

$$v(\,-\mu_0)^2 + n(\mu - \bar{y})^2 = (v+n)(\mu - \mu^*)^2 + d_n(\bar{y} - \mu_0)^2,$$

where

$$\mu^* = \frac{v\mu_0 + n\bar{y}}{v+n} \quad \text{and} \quad d_n = \frac{vn}{v+n} = (v^{-1} + n^{-1})^{-1}.$$

Then show that if the prior is $(\lambda, \mu) \sim \mathrm{GN}(a, b, \mu_0, v)$, then

$$(\lambda, \mu) \,|\, \text{data} \sim \mathrm{GN}(a + \tfrac{1}{2}n, b + \tfrac{1}{2}Q_0 + \tfrac{1}{2}d_n(\bar{y} - \mu_0)^2, \mu^*, v+n).$$

(e) The special case of a 'flat prior' for μ, corresponding to letting $v \to 0$ above, is particularly easy to deal with. Show that then

$$(\lambda, \mu) \,|\, \text{data} \sim \mathrm{GN}(a + \tfrac{1}{2}, b + \tfrac{1}{2}Q_0, \bar{y}, n).$$

Find the posterior mean of σ^2 under this prior.

(f) To illustrate, go to `lib.stat.cmu.edu/DASL/Stories/cigcancer.html` and create the data vector y of 44 death rates per 100,000 inhabitants of lung cancer (for 43 American states plus the District of Colombia), assuming these data to represent an i.i.d. normal sample. First carry out a Bayesian analysis of (μ, σ) using an informative prior, namely one identified by (i) using a gamma prior for $1/\sigma^2$ such that the 0.10 and 0.90 prior quantiles of σ are respectively 2.2 and 7.7, and (ii) using a $\mathrm{N}(15.0, (3.3\,\sigma)^2)$ for μ given σ. Find 95% credibility intervals for μ, for σ, and for the probability that $y \geq 25.0$.

(g) Re-do the Bayesian analysis above, but now with the simpler and less informative prior that takes a flat prior for μ. Also compare with 95% confidence intervals arrived at via classical frequentist analysis.

23. Gamma-normal conjugate inference for the linear regression model

The aim of the present exercise is to generalise the Gamma-Normal conjugate prior class above to the linear-normal regression model. The model is the very classical one where

$$y_i = x_{i,1}\beta_1 + \cdots + x_{i,k}\beta_k + \varepsilon_i = x_i^{\mathrm{t}}\beta + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

with the ε_i taken i.i.d. $\mathrm{N}(0, \sigma^2)$. Write X for the $n \times k$ matrix of covariates (explanatory variables), with $x_i = (x_{i,1}, \ldots, x_{i,k})$ as its ith row, and use y and ε to indicate the vectors of y_i and ε_i. Then

$$y = X\beta + \varepsilon \sim \mathrm{N}_n(X\beta, \sigma^2 I_n)$$

is a concise way to write the full model.

(a) Show that the likelihood function may be written as being proportional to

$$L_n(\beta,\sigma) = \sigma^{-n}\exp\Big\{-\tfrac{1}{2}\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - x_i^{\mathrm{t}}\beta)^2\Big\}$$
$$= \sigma^{-n}\exp\Big[-\tfrac{1}{2}\frac{1}{\sigma^2}\{Q_0 + n(\beta-\widehat{\beta})^{\mathrm{t}}M_n(\beta-\widehat{\beta})\}\Big],$$

in which

$$M_n = (1/n)X^{\mathrm{t}}X = n^{-1}\sum_{i=1}^{n}x_i x_i^{\mathrm{t}} \quad \text{and} \quad \widehat{\beta} = (X^{\mathrm{t}}X)^{-1}X^{\mathrm{t}}y = M_n^{-1}n^{-1}\sum_{i=1}^{n}x_i y_i.$$

Also,

$$Q(\beta) = \|y - X\beta\|^2 = Q_0 + n(\beta-\widehat{\beta})^{\mathrm{t}}M_n(\beta-\widehat{\beta}),$$

with $Q_0 = \sum_{i=1}^{n}(y_i - x_i^{\mathrm{t}}\widehat{\beta})^2$ the minimum value of Q over all β. Note that $\widehat{\beta}$ is the classical least squares estimator (and the ML estimator), which in the frequentist framework is unbiased with variance matrix equal to $\sigma^2(X^{\mathrm{t}}X)^{-1} = (\sigma^2/n)M_n^{-1}$. This is the basis of all classical methods related to the widely popular linear regression model.

(b) Let $p(\beta,\sigma$ be *any* prior for the $(k+1)$-dimensional parameter of the model. Set up formulae for a Metropolis type MCMC algorithm for drawing samples from the posterior distribution of (β,σ).

(c) In spite of the possibility of solving problems via MCMC (or perhaps acceptance-rejection sampling), as with the previous exercise it is very much worthwhile setting up explicit formulae for the case of a certain canonical prior class. Write

$$(\lambda,\beta) \sim \mathrm{GN}_k(a,b,\beta_0,M_0)$$

to indicate the gamma-normal prior where

$$\lambda = 1/\sigma^2 \sim \mathrm{Gamma}(a,b) \quad \text{and} \quad \beta \sim \sigma \sim \mathrm{N}_k(\beta_0,\sigma^2 M_0^{-1}).$$

Show that this prior may be expressed as

$$p(\lambda,\beta) \propto \lambda^{a-1}\lambda^{k/2}\exp\big[-\lambda\{b + \tfrac{1}{2}(\beta-\beta_0)^{\mathrm{t}}M_0(\beta-\beta_0)\}\big].$$

(d) When multiplying the prior with the likelihood it is convenient to use the following linear algebra identity about quadratic forms, which you should prove first. For symmetric and invertible matrices A and B, and for any vectors a,b,x of the appropriate dimension,

$$(x-a)^{\mathrm{t}}A(x-a) + (x-b)^{\mathrm{t}}B(x-b) = (x-\xi)^{\mathrm{t}}(A+B)(x-\xi) + (b-a)^{\mathrm{t}}D(b-a),$$

where $\xi = (A+B)^{-1}(Aa+Bb)$ (a weighted average of a and b) and D is a matrix for which several equivalent formulae may be used:

$$\begin{aligned} D &= A(A+B)^{-1}B = B(A+B)^{-1}A \\ &= AA(A+B)^{-1}A = BB(A+B)^{-1}B = (A^{-1}+B^{-1})^{-1}. \end{aligned}$$

(e) Prove that if (λ, β) has the $\mathrm{GN}_k(a, b, \beta_0, M_0)$ prior, then

$$(\lambda, \beta) \,|\, \text{data} \sim \mathrm{GN}_k(a + \tfrac{1}{2}n, b + \tfrac{1}{2}Q_o + \tfrac{1}{2}(\widehat{\beta} - \beta_0)^{\mathrm{t}} D_n (\widehat{\beta} - \beta_0), \beta^*, M_0 + nM_n),$$

where

$$\beta = (M_0 + nM_n)^{-1}(M_0\beta_0 + nM_n\widehat{\beta}) \quad \text{and} \quad D_n = M_0(M_0 + nM_n)^{-1}nM_n.$$

This characterisation makes it easy to simulate a large number of (β, σ) from the posterior distribution and hence to carry out Bayesian inference for any parameter of quantity of interest.

(f) Note the algebraic simplifications that result when the M_0 in the prior is chosen as being proportional to the covariate sample variance matrix, i.e. $M_0 = c_0 M_n$. Show that then

$$\beta^* = \frac{c_0\beta_0 + n\widehat{\beta}}{c_0 + n} \quad \text{and} \quad D_n = \frac{c_0 n}{c_0 + n}.$$

In this connection c_0 has a natural interpretation as 'prior sample size'.

(g) A special case of the above, leading to simpler results, is that where β has a flat, non-informative prior, corresponding to very large prior variances, i.e. to $M_0 \to 0$. Show that with such a prior,

$$(\lambda, \beta) \,|\, \text{data} \sim \mathrm{GN}_k(a + \tfrac{1}{2}n, b + \tfrac{1}{2}Q_0, \widehat{\beta}, nM_n).$$

The prior is improper (infinite integral), but the posterior is proper as long as $\widehat{\beta}$ exists, which requires $X^{\mathrm{t}}X$ to have full rank, which again means at least k linearly independent covariate vectors, and, in particular, $n \geq k$.

(h) Go again to `lib.stat.cmu.edu/DASL/Datafiles/cigcancerdat.html`, for illustration and for flexing your operational muscles. For y use the lung cancer column of deaths per 100,000 inhabitants and for x use the number of cigarettes sold per capita. Your task is to carry out Bayesian analysis within the linear regression model

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad \text{for } i = 1, \ldots, 44,$$

with ε_i taken i.i.d. $\mathrm{N}(0, \sigma^2)$. Specifically, we wish point estimates along with 95% credibility intervals for (i) each of the three parameters α, β, σ; (ii) the probability that $y \geq 25.0$, for a country with cigarette consumption $x = 35.0$; (iii) the lung

cancer death rates y_{45} and y_{46}, per 100,000 inhabitants, for countries with cigarette consumption rates $x_{45} = 10.0$ (low) and $x_{46} = 50.0$ (high). You are to carry out such inference with two priors:

- First, the informative one which takes $1/\sigma^2$ a gamma with 0.10 and 0.90 quantiles for σ equal to 1.0 and 5.0, and α and β as independent normals $(15.0, (2.0\,\sigma)^2)$ and $(0.0, (2.0\,\sigma)^2)$, given σ.
- Then, the simpler and partly non-informative one that takes a flat prior for (α, β) and the less informative one for σ that uses 0.10 and 0.90 prior quantiles 0.5 and 10.0.
- Finally, compare your results from those arrived at using classical frequentist methods.

24. The Stein effect and empirical Bayes

Suppose there is an ensemble of parameters $\theta_1, \ldots, \theta_k$ to estimate, where these are thought to be not unreasonably dissimilar, and where it may make sense to think about them as having arisen from a distribution of parameter values. In such cases various empirical Bayes constructions will often be successful, in the sense that they lead to 'joint estimation' procedures that typically perform better than using 'separate estimation'. What *is and remains* surprising is that for certain situations of the above type, there are empirical Bayes methods that *always and uniformly* improve upon the 'separate estimation' procedures, i.e. even when the underlying parameters are widely dissimilar. This phenomenon is loosely referred to as 'the Stein effect', or even 'the Stein paradox', from influential papers by Charles Stein in 1956 and later. Even *The Scientific American* have had papers on this for a wider audience. The paradox in question is that when needing to estimate apples, oranges, bananas, then it is counterintuitively possible to do better by calling in information about oranges and bananas to estimate apples, etc.

The present exercise looks into one of these models where reasonably clean proofs may be given for the type of universal risk dominance of certain procedures over the standard ones. Let $Y_i \sim \mathrm{N}(\theta_i, 1)$ be independent for $i = 1, \ldots, k$, where the aim is to estimate each of the θ_i with a combined loss function

$$L(\theta, \widehat{\theta}) = k^{-1} \sum_{i=1}^{k} (\widehat{\theta}_i - \theta_i)^2.$$

The ensuing risk for the $\widehat{\theta}$ procedure is

$$R(\widehat{\theta}, \theta) = \mathrm{E}_\theta L(\theta, \widehat{\theta}) = k^{-1} \sum_{i=1}^{k} \mathrm{E}_\theta (\widehat{\theta}_i - \theta_i)^2.$$

This may again be represented as the average variance plus the average squared bias (as a function of the position in parameter space). Note that $\widehat{\theta}_i$ for θ_i ought to be allowed to depend on all the data, not merely Y_i.

(a) The standard estimator here is simply using Y_i for θ_i, for $i = 1, \ldots, k$; Y_i is after all the least squares estimator, the maximum likelihood estimator, the best unbiased estimator, it is admissible, etc. Show that its risk function is simply 1, constant across the parameter space. The challenge is to find an estimator which has risk function smaller than 1 everywhere in the parameter space.

(b) For a single $Y \sim \mathrm{N}(\theta, 1)$, show that under very mild conditions on the function $b(y)$, one has

$$\mathrm{E}_\theta(Y - \theta)b(Y) = \mathrm{E}_\theta b'(Y).$$

(Use 'partial integration'.) Check with e.g. $b(Y) = Y$ and $b(Y) = Y^2$ to get a feeling for how the identity works.

(c) Using the same technique, generalise the above to

$$\mathrm{E}_\theta(Y_i - \theta_i)b_i(Y) = \mathrm{E}_\theta b'_{i,i}(Y),$$

where $b'_{i,i}(y) = \partial b_i(y)/\partial y_i$.

(d) Consider a general competitor to Y of the form $\widehat{\theta}_i = Y_i - b_i(Y)$. Show that

$$\mathrm{E}_\theta\{(Y_i - b_i(Y) - \theta_i)^2 - (Y_i - \theta_i)^2\} = \mathrm{E}_\theta\{b_i(Y)^2 - 2b_{i,i}(Y)\}$$

and hence that

$$R(\widehat{\theta}, \theta) = R(Y, \theta) + E_\theta D(Y) = 1 + \mathrm{E}_\theta D(Y),$$

where

$$D(y) = k^{-1} \sum_{i=1}^{k} \{b_i(y)^2 - 2b_{i,i}(y)\}.$$

If in particular we manage to find $b_i(y)$ functions for which $D(y) < 0$ for all y, then $\widehat{\theta}$ is a uniform improvement over the standard estimator Y. It turns out to be impossible to find such functions for $k = 1$ or $k = 2$, but indeed possible for $k \geq 3$.

(e) Try $b_i(y) = cy_i/\|y\|^2$, with $\|y\|^2$ being the squared Euclidean norm $\sum_{i=1}^k y_i^2$, corresponding to

$$\widehat{\theta} = y - b(y) = \Big(1 - \frac{c}{\|y\|^2}\Big)y.$$

Show that

$$D(y) = \frac{1}{k}\frac{1}{\|y\|^2}\{c^2 - 2c(k-2)\},$$

and that this is indeed negative for an interval of c values, provided the dimension is $k \geq 3$. Indeed demonstrate that the best value is $c_0 = k - 2$ and that the consequent risk function can be expressed as

$$R(\widehat{\theta}, \theta) = 1 - \frac{(k-2)^2}{k}\mathrm{E}_\theta \frac{1}{\|Y\|^2} = 1 - \frac{k-2}{k}\mathrm{E}\frac{k-2}{\chi^2_k(\|\theta\|^2)}.$$

Here $\chi^2_k(\lambda)$ is the excentric chi-squared distribution with k degrees of freedom and excentre parameter λ.

(f) The arguments above led to the estimator

$$\widehat{\theta}_i = \Big(1 - \frac{k-2}{\|y\|^2}\Big) y_i \quad \text{for } i = 1, \ldots, k,$$

which is a version of the Stein estimator. A useful modification is to truncate the shrinking factor $1 - (k-2)/\|y\|^2$ to zero in the case of this being negative, i.e. $\|y\|^2 \leq k-2$. We write this as

$$\widehat{\theta}_{\text{Stein}} = \Big(1 - \frac{k-2}{\|y\|^2}\Big)_+ y, \quad \text{where } x_+ = \max(0, x).$$

Prove that this modification actually improves the performance further. (It remains easier to work directly with $\widehat{\theta}$, though, e.g. regarding risk functions.)

(g) Show that the greatest risk reduction for $\widehat{\theta}$ takes place at zero, with $R(\widehat{\theta}, 0) = 2/k$. For a few values of k, say $k = 5, 10, 100$, compute and display the risk functions for Y and $\widehat{\theta}$, as functions of $\|\theta\|$. Do the same with the $\widehat{\theta}_{\text{Stein}}$ estimator (for which you may use simulations to compute the risk).

(h) Now make the empirical Bayes connection, as follows. Start with the prior that takes $\theta_1, \ldots, \theta_k$ independent from the $\text{N}(0, \tau^2)$, and show that the Bayes estimator takes the form

$$\theta_i^* = \theta_i^*(\rho) = \rho y_i, \quad \text{with } \rho = \frac{\tau^2}{\tau^2 + 1}.$$

Show that the marginal distribution of $y_1, \ldots, y_k$ is that of independent $\text{N}(0, 1 + \tau^2)$ components, with maximum likelihood estimate $\widehat{\tau}^2 = (W-1)_+$, where $W = k^{-1} \sum_{i=1}^k y_i^2$. This invites

$$\widehat{\rho} = \frac{(W-1)_+}{W} = \Big(1 - \frac{k}{\|y\|^2}\Big)_+,$$

or versions close to this, for the empirical Bayes estimator

$$\widehat{\theta}_{i,\text{EB}} = \theta_i^*(\widehat{\rho}) = \widehat{\rho} y_i.$$

The Stein type estimator above can accordingly be viewed as an empirical Bayes construction. Note that $\widehat{\theta}_{i.\text{EB}}$ can be motivated and constructed without any direct concern or calculations for the risk functions per se.

25. Shrinking towards means and regression lines

The construction of Exercise #24 shrinks the standard estimator $\widetilde{\theta} = Y$ towards zero. In the framework of the empirical Bayes setup this can be traced to our using the zero-mean prior $\text{N}(0, \tau^2)$ for the θ_i. It is a simple matter to shrink towards any given θ_0 instead, leading to

$$\widehat{\theta}_i = \Big(1 - \frac{k-2}{\|y - \theta_0\|^2}\Big) y_i + \frac{k-2}{\|y - \theta_0\|^2} \theta_0 \quad \text{for } i = 1, \ldots, k.$$

The present exercise looks into further useful generalisations of the Steinean leitmotif, involving the shrinking of standard estimators towards estimated means and even estimated regression curves. The perspectives pertain to both direct risk performance calculations and empirical Bayes considerations.

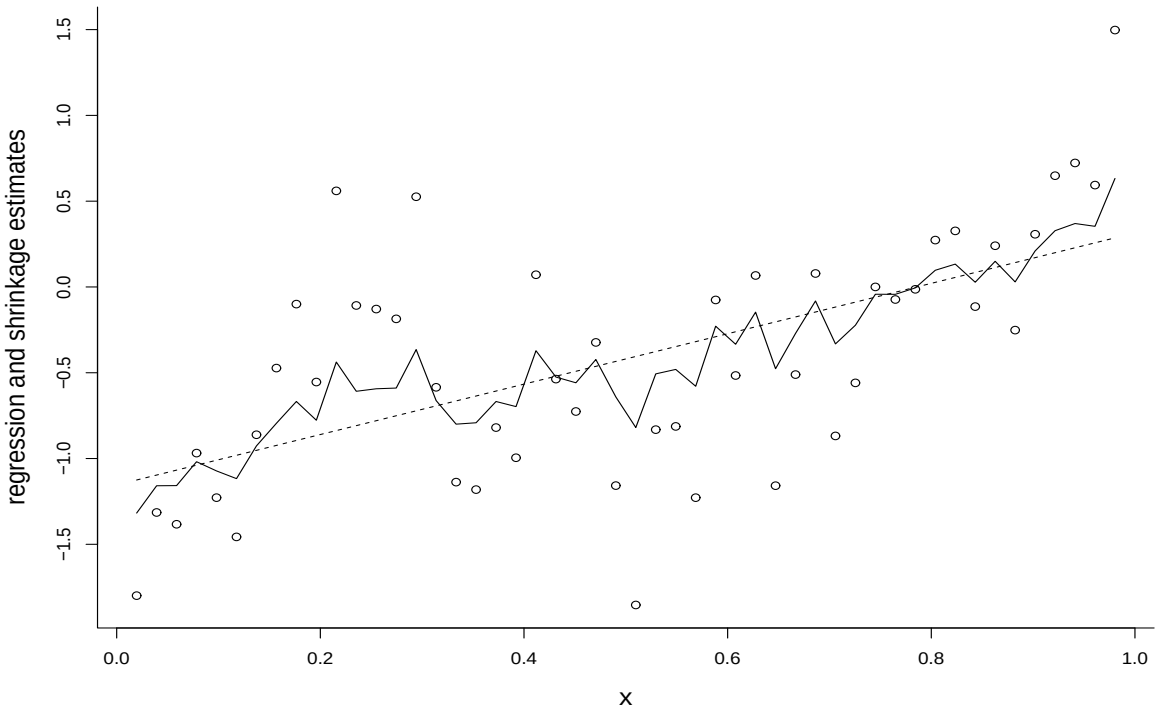

Figure 4: Shrinkage of raw estimates y_i towards the regression line, with $\widehat{\theta}_i = \widehat{\rho} y_i + (1-\widehat{\rho})(\widehat{\beta}_0 + \widehat{\beta}_1 x_i)$. This estimation procedure does not demand that the true θ_i points actually follow a regression line, but nevertheless uniformly dominates that of using the raw data.

(a) In the spirit of empirical Bayes one would also like to insert a data based estimate for θ_0 in the formula above. Since $\bar{y}$ is the natural candidate for such an estimate of a common centre value, we are led to

$$\widehat{\theta}_{i,\mathrm{EB}} = \Big(1 - \frac{c}{\|y-\bar{y}\|^2}\Big) y_i + \frac{c}{\|y-\theta_0\|^2}\bar{y} \quad \text{for } i = 1, \ldots, k.$$

The reasons given are already sufficiently clear and natural in order for us to accept this construction as a clever one, for a suitably fine-tuned c, but we have of course not yet studied its performance. To do so, employ the risk difference representation apparatus of the previous exercise for

$$b_i(y) = c\frac{y_i - \bar{y}}{\|y-\bar{y}\|^2} = c\frac{y_i - \bar{y}}{z} \quad \text{for } i = 1, \ldots, k,$$

where $z = \sum_{i=1}^k (y_i - \bar{y})^2$. Show that

$$D(y) = \frac{1}{k}\frac{1}{z}\{c^2 - 2c(k-3)\}$$

and that there are c values making this $D(y)$ entirely negative as long as the dimension is $k \geq 4$. Show that the best values if $c_0 = k-3$ and that this leads to

$$\widehat{\theta}_{\mathrm{EB}} = \Big(1 - \frac{k-3}{\|y-\bar{y}\|^2}\Big) y + \frac{k-3}{\|y-\theta_0\|^2}\bar{y}$$

with

$$R(\widehat{\theta}_{\rm EB},\theta) = 1 - \frac{k-3}{k}{\rm E}\frac{k-3}{\chi^2_{k-1}(\|\theta-\bar\theta\|^2)}.$$

Show in particular that the risk reduction is largest when the θ_i are close to each other, and that $R(\widehat{\theta}_{\rm EB},\theta) = 3/k$ when θ has all components equal to each other. Though not entirely necessary, in practice we are again thresholding the weights here to lie inside $[0,1]$, yielding

$$\widehat{\theta}_{i,{\rm EB}} = \widehat{\rho}y_i + (1-\widehat{\rho})\bar y \quad \text{for } i = 1,\ldots,k,$$

with $\widehat{\rho} = \{1-(k-3)/z\}_+$.

(b) Though the arguments and calculations for question (a) come from direct risk analysis, so to speak, also provide clear empirical Bayes arguments leading to the estimator exhibited there.

(c) Now attempt to generalise the above constructions in the direction of 'shrinking towards a general linear subspace'. For concreteness, suppose there are data pairs $(x_1,y_1),\ldots,(x_n,y_n)$, with x_i a covariate for y_i, but where one is unwilling to go as far as assuming the traditional linear regression structure, though without entirely abandoning the idea. Write therefore $y_i = \theta_i + \varepsilon_i$, with the ε_i being i.i.d. and zero-mean normal, with a standard deviation we here take to be known, where the idea is to shrink the raw data y_i towards the regression line, and letting the data dictate the degree of shrinking. Make an empirical Bayes construction of the form

$$\widehat{\theta}_{i,{\rm EB}} = \Big(1-\frac{c}{z}\Big)y_i + \frac{c}{z}(\widehat{\beta}_0+\widehat{\beta}_1x_i) = y_i - (c/z)(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i) \quad \text{for } i=1,\ldots,n$$

with $z = \sum_{i=1}^n (y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i)^2$ the residual sum of squares. Again using the techniques and formulae of Exercise #24, show that for this case,

$$D(y) = \frac{1}{n}\frac{1}{z}\{c^2 - 2c(n-4)\}$$

with best value $c_0 = n-4$, and in particular that there is *uniform risk improvement* by shrinking to the regression line provided $k \geq 5$. Also, demonstrate that the risk improvement over the raw data estimator is greatest near the regression line, i.e. where the θ_i are close to $\beta_0+\beta_1x_i$, with maximal risk reduction, from 1 to $4/n$, when the θ_i actually follow the line.

(d) Your task now is to duplicate a version of Figure 4, with suitable variations. It provides an illustration of the general 'shrinking towards a regression line even if it does not fit data particular well' procedure. The raw data y_i have been sampled from a model that does *not* take their means θ_i to come from any regression line, but rather from the setup where the x_i are uniformly spread on $(0,1)$ and with

$$y_i \sim {\rm N}(\beta_0+\beta_1x_i+\gamma\sin(2\pi x_i),\sigma_0^2) \quad \text{for } i=1,\ldots,n,$$

with $n = 50$. Do this, for a few values of γ, and monitor the amount of 'smoothing' that takes place with the

$$\widehat{\theta}_i = \widehat{\rho} y_i + (1 - \widehat{\rho})(\widehat{\beta}_0 + \widehat{\beta}_1 x_i)$$

operation.

(e) Let us generalise further, to the case of shrinking raw data estimates y_i in the direction of a general regression structure $x_i^{\rm t}\widehat{\beta}$, where

$$\widehat{\beta} = M^{-1} \sum_{i=1}^n x_i y_i = \Big(\sum_{i=1}^n x_i x_i^{\rm t} \Big)^{-1} \sum_{i=1}^n x_i y_i$$

is the least squares estimate. We are assuming here that the M matrix is of full rank p, the dimension of β. Work with

$$\widehat{\theta}_i = y_i - b_i(y) = y_i - \frac{c(y_i - x_i^{\rm t}\widehat{\beta})}{z} = \Big(1 - \frac{c}{z}\Big) y_i + \frac{c}{z} x_i^{\rm t} \widehat{\beta},$$

where $z = \sum_{i=1}^n (y_i - x_i^{\rm t}\widehat{\beta})^2 = \sum_{i=1}^n \widehat{r}_i^2$ is the sum of squared residuals. In order to work out an expression for $D(y)$ of Exercise #24(d) you need first to show that $\partial\widehat{\beta}/\partial y_i = M^{-1}x_i$ and $\partial\widehat{\beta}^{\rm t} M \widehat{\beta}/\partial y_i = 2x_i^{\rm t}\widehat{\beta}$, which conspire to lead you to

$$b_{i,i}(y) = c\{(1 - x_i^{\rm t} M^{-1} x_i) z - 2(y_i - x_i^{\rm t}\widehat{\beta})^2\}/z^2$$

and then to

$$D(y) = (nz)^{-1}\{c^2 - 2c(n - p - 2)\}.$$

For $n \geq p+3$ we have the Stein phenomenon at work, with uniform risk improvement and risk $(p+2)/n$ when the regression model being shrunk to is actually correct. As before, make sure you can argue for this procedure from both risk function investigation and empirical Bayes perspectives.

(f) To extend our repertoire we need to modestly and patiently generalise some of the above work to the case of the observations having standard deviation say σ instead of merely $\sigma = 1$. Show that Exercise #24(c) generalises to

$$\mathrm{E}_\theta (Y_i - \theta_i) b_i(Y) = \sigma^2 \, \mathrm{E}_\theta b_{i,i}(Y),$$

and that the risk difference identity becomes

$$R(Y - b(Y), \theta) = \sigma^2 - \mathrm{E}_\theta D(Y), \quad \text{with } D(y) = k^{-1} \sum_{i=1}^k \{b_i(y)^2 - 2\sigma^2 b_{i,i}(y)\}.$$

Show how earlier results of this and the previous exercise generalise, e.g. to the following general 'shrinking to regression' recipe, for the model with independent observations $Y_i \sim \mathrm{N}(\theta_i, \sigma^2)$:

$$\widehat{\theta}_{i,\mathrm{EB}} = \widehat{\rho} y_i + (1 - \widehat{\rho}) x_i^{\rm t} \widehat{\beta},$$

with

$$\widehat{\rho} = \Big(1 - \frac{\sigma^2}{\widehat{\kappa}^2}\Big)_+ \quad \text{and} \quad \widehat{\kappa}^2 = \frac{1}{n - p - 2} \sum_{i=1}^n (y_i - x_i^{\rm t}\widehat{\beta})^2.$$

One needs a bit of care when interpreting and implementing these ideas and methods. Here we view $\widehat{\kappa}^2$ as estimating the overall variability around the regression line, whereas σ^2 remains the variance of Y_i around its expected value θ_i. In an empirical Bayes setup of the above we would have $\kappa^2 = \sigma^2 + \tau^2$ with τ^2 signalling the variance of the θ_i around the regression line $x_i^{\mathrm{t}}\beta$.

(g) In realistic cases the σ parameter above cannot be taken as known. Attempt to build a suitable empirical Bayes framework involving a prior for σ, leading, under some conditions, to a fully automatic 'data smoother', shrinking the raw data towards a regression structure with a shrinkage factor dictated by the data.