STK 4050 – Statistical Simulation
Autumn 2007

Course Notes and Exercises
by Nils Lid Hjort

– This version: as of 10 November 2007 –

1. Approximating integrals

Let

$$I = \int_0^1 \int_0^1 \exp\{\sin(\sqrt{|xy|})\exp(|y|^{3/2})\}\,\mathrm{d}x\,\mathrm{d}y.$$

(a) Approximate the value of I using a sample of i.i.d. uniform draws $(X_1, Y_1), \ldots, (X_m, Y_m)$ from the unit square, for cases m equal to $10^2, 10^3, 10^4, 10^5, 10^6$. For each of these cases, supply also a 99% confidence interval for I.

(b) In an attempt at improving on the precision reached with the direct method of (a), implement a weighted importance sampling method of the form

$$\widehat{I}_m = \frac{1}{m}\sum_{j=1}^{m} \frac{g(X_j, Y_j)}{\pi(X_j, Y_j)},$$

where the (X_j, Y_j) pairs are drawn from the density $\pi(x, y)$ over the unit square. Use for example product of Beta densities, i.e.

$$\pi(x, y) = \mathrm{be}(x; a_1, b_1)\mathrm{be}(y; a_2, b_2),$$

e.g. with $(a_1, b_1) = (3, 1)$ and $(a_2, b_2) = (4, 1)$. Again, produce estimates and 99% confidence intervals for I for m equal to $10^2, 10^3, 10^4, 10^5, 10^6$. Experiment a bit for other weight densities.

(c) What is the best possible weight function π?

(d) Find the exact value of I via two-dimensional numerical integration.

2. Null distribution of test statistics

Let $X_1, \ldots, X_n$ be i.i.d. from some density, and assume one wishes to test the null hypothesis that the distribution is normal. Of course there are a great many different goodness-of-fit tests that may be used for this purposes.

(a) Argue that under the normality $\mathrm{N}(\mu, \sigma^2)$ assumption, we ought to have

$$X_{(i)} \approx \mu + \sigma\Phi^{-1}\Big(\frac{i}{n+1}\Big) \quad \text{for } i = 1, \ldots, n,$$

where $X_{(1)} < \cdots < X_{(n)}$ are the order statistics of the sample and Φ is the standard normal distribution function.

(b) In light of (a), argue that

$$D_n = \sqrt{n} \max_{\varepsilon \le i/(n+1) \le 1-\varepsilon} \Big| \frac{X_{(i)} - \bar{X}}{s} - \Phi^{-1}\Big(\frac{i}{n+1}\Big) \Big|$$

is a natural test statistic. Here ε is some small and fixed number, like 0.05, and $\bar{X}$ and s are the usual empirical mean and standard deviation numbers. Show that D_n has a null distribution independent of (μ, σ). (The '$\sqrt{n}$' is not important here, but is there to indicate that there is actually a well-defined limit distribution of D_n as n grows to infinity.)

(c) One rejects normality when D_n is large enough. For $n = 100$ and $n = 1000$, simulate the null distribution of D_n to find the rejection point that gives a test with significance level 0.05.

(d) Suppose the real distribution of data is not a normal, but rather of the form

$$f(x) = t_\nu\Big(\frac{x-\mu}{\sigma}\Big)\frac{1}{\sigma},$$

where t_ν is the t density with ν degrees of freedom. Use simulations to find the power of the D_n test, for $n = 100$ and $n = 1000$, against these alternatives, for different ν values. You may also draw power curves as a function of ν.

3. Estimating standard deviation with parametric bootstrapping

We are to study the Weibull distribution, where the distribution function takes the form

$$F(t) = 1 - \exp\Big\{-\Big(\frac{t}{\theta}\Big)^\alpha\Big\} \quad \text{for } t \ge 0.$$

(a) Show that the median of this distribution is equal to

$$m = \theta(\log 2)^{1/\alpha}.$$

(b) Simulate a data set from the Weibull distribution with parameters $(\theta_0, \alpha_0) = (2.22, 3.33)$, of size $n = 250$. Write an expression for the log-likelihood of the data and find the maximum likelihood estimates $(\widehat{\theta}, \widehat{\alpha})$. Compute also the resultant maximum likelihood estimate $\widehat{m}$ for the median.

(c) How can we estimate the standard deviation of $\widehat{m}$? Let

$$h = h(\theta, \alpha) = \mathrm{sd}_{\theta,\alpha}(\widehat{m}),$$

the standard deviation of $\widehat{m}$ provided the parameters are precisely (θ, α), for the fixed sample size $n = 250$ under consideration. The *parametric bootstrapping* approach is to compute h via stochastic simulation, at the estimated position $(\widehat{\theta}, \widehat{\alpha})$:

$$\widehat{h} = h(\widehat{\theta}, \widehat{\alpha}) \doteq \Big\{\frac{1}{B-1}\sum_{j=1}^{B}(\widehat{m}_j^* - \bar{m}^*)^2\Big\}^{1/2},$$

where $\widehat{m}_1^*, \ldots, \widehat{m}_B^*$ is a suitably large number of estimates formed from B quasi-samples of Weibull data $X_1^*, \ldots, X_n^*$, using the estimated position $(\widehat{\theta}, \widehat{\alpha})$ in the parameter space. Also, $\bar{m}^*$ is the average of these bootstrap estimates. – Now execute this idea for your data set. How large should B be?

(d) Follow the ideas above to also estimate *the bias* of $\widehat{m}$, as an estimator of the true m, for your data set.

4. Level and power of the t test

Among the most frequently applied statistical tests on this planet is the so-called t test: if one has data points $X_1, \ldots, X_n$ from a distribution with mean μ, compute the quantity

$$t = \sqrt{n}\bar{X}/\widehat{\sigma},$$

where $\widehat{\sigma}$ is the familiar empirical standard deviation. Then one claims that $\mu \neq 0$ provided $|t| > t_0$, the upper 2.5% quantile in the t distribution with $n-1$ degrees of freedom.

(a) Under which conditions does this lead to a test with exact significance level 0.05?

(b) Show that the limit level of the test, when n grows, will in fact always be 0.05 (under certain mild conditions).

(c) To check how the test behaves outside strict model conditions, assume that $n = 25$ and that the real underlying density of the X_is is the double exponential one, with $f(x) = \frac{1}{2}\exp(-|x|)$. Compute the level of the test, via simulations.

(d) Then use simulations to compute the power of the t-test, under conditions of the X_is being $\mathrm{N}(\mu, \sigma^2)$ (where the level of the test described above is guaranteed to be exactly and not only approximately equal to 0.05), as a function of μ/σ, for $n = 25$. Compare with the exact results.

(e) Finally use simulations to compute the power of the t-test, for the case of alternatives taking the double exponential form of $f(x) = \frac{1}{2}\exp(-|x-\mu|)$. Display the two power curves (for the normal case and the double exponential case) in the same diagram (both adjusted to have the same null level of 0.05), for $n = 25$, as functions of μ/σ, where σ in both cases is the standard deviation.

5. Testing uniformity of random generators

Most software packages have in-built algorithms for simulating from the most popular statistical distributions.

(a) Use the `x <- runif(1000)` of **R** to simulate 1000 data points, say $x_1, \ldots, x_n$, from the uniform distribution over the unit interval. In particular, one ought

to have

$$p_j = \Pr\{X_i \in ((j-1)/10, j/10]\} = 1/10 \quad \text{for } j = 1, \ldots, 10.$$

Use the Karl Pearson test to test this particular underlying aspect of the hypothesis that data really follow the uniform distribution.

(b) If this course has m students, and each of them carries out the homework assignment detailed in (a), finding their Pearson test statistics values $P_1, \ldots, P_m$, how would these be distributed? How many of the students shall be moderately surprised to find that their tests *reject* the hypothesis of a perfect uniform distribution?

(c) Another aspect of the underlying distributional hypothesis is that X_{i-1} and X_i ought to be stochastically independent, i.e. there is no serial dependence. Propose a test to check for this aspect of the hypothesis of a well-working software algorithm, and execute the test.

(d) Discuss a couple of further essential aspects of the underlying i.i.d. uniformity assumption, and give methods for checking that these aspects are in fact respected by the data delivered by **R**.

6. Some simulation tricks

(a) Let F be a continuous and strictly increasing distribution function. Show that if U is a uniform on the unit interval, then $X = F^{-1}(U)$ has in fact distribution F.

(b) Generate $n = 1000$ data points from the distribution with density $f(x) = 4x^3$ over $[0, 1]$. Check, using a Pearson test, that the number of data points in the ten boxes $[0.0, 0.1], \ldots, [0.9, 1.0]$ behave as they ought to.

(c) Let $F(t) = 1 - \exp\{-A(t)\}$ be a distribution function over $[0, \infty)$, where $A(t)$ starts at 0 and increases continuously to infinity. If E is standard exponentially distributed, show that $X = A^{-1}(E)$ has distribution F.

(d) The so-called Weibull distribution has the form $F(t) = 1 - \exp\{-(\theta t)^\alpha\}$, where θ and α are positive parameters. Explain how one may simulate n data points from any given Weibull distribution.

7. Rejection sampling

(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?

8. A ratio method

Let h be a nonnegative function with finite integral over $\mathbb{R}$, and consider the region

$$A_h = \{(u, v) : 0 \le u \le h(v/u)^{1/2}\}.$$

(a) For the case of $h(x) = e^{-x}$, make a drawing of the region A_h.

(b) Show that A_h has finite area $\frac{1}{2}\int h(x)\,\mathrm{d}x$.

(c) Let (U, V) be uniform over the region A_h. Show that the distribution of $X = V/U$ has density $f(x) = h(x)/\int h(x)\,\mathrm{d}x$.

(d) Use this method to generate 1000 independent values from the density $f(x) = e^{-x} I\{x > 0\}$.

9. Estimating standard deviation via nonparametric bootstrapping

Let $X_1, \ldots, X_n$ be independent data from some unknown distribution F. Among various robust measures for the spread of the distribution is the quantity

$$\theta = \theta(F) = \left\{\mathrm{E}_F |X - \mathrm{E}_F X|^{3/2}\right\}^{2/3}.$$

(a) Show that if X is normal with standard deviation σ, then $\theta = 0.9044\,\sigma$.

(b) Show that that natural nonparametric estimator $\widehat{\theta} = \theta(\widehat{F})$, where $\widehat{F}$ is the empirical distribution (defined by assigning probability mass $1/n$ to each of the data points) takes the form

$$\widehat{\theta} = \left\{\frac{1}{n}\sum_{i=1}^{n} |X_i - \bar{X}|^{3/2}\right\}^{2/3}.$$

(c) Simulate $n = 50$ data points via $X_i = \exp(aY_i)$, where $a = 0.333$ and the Y_is are standard normal. (One says that the X_is are log-normally distributed.) Estimate θ for these data. In addition find, or approximate, the value of the underlying θ for this situation.

(d) Generate BOOT = 1000 bootstrap estimates $\widehat{\theta}^*$. Estimate the standard deviation for $\widehat{\theta}$, and give a confidence interval for θ with level approximately equal to 90%. Compare with the interval that uses a normal approximation.

10. Rejection sampling (cont.)

(a) 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 $g(x) \leq Kh(x)$ for all x, for some K; and (ii) that we keep this candidate X with probability $g(x)/\{Kh(x)\}$. Verify that the probability density of a surviving X is really $f(x)$. – The importance of this variation on the rejection sampling recipe of Exercise 7 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.

(b) 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 $[0, 1] \times [0, 1]$, studied also in Exercise 1. Make fine histograms of the two marginal distributions, and find means, standard deviations, and the correlation, numerically.

(c) 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.

(d) 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.

11. Transformation tricks

Sometimes one manages to sample from some given target distribution by cleverly transforming simulation output from a simpler distribution. Note the following from standard probability theory: If Y comes from density $g(y)$, and $X = h(Y)$ is a smooth 1–1 transformation, then X follows the density

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

where $y = h^{-1}(x)$ is the inverse transformation to $x = h(y)$ and the factor on the right is the absolute value of the Jacobi matrix.

(a) Suppose (X, Y) is a pair of independent standard normals. Find the simultaneous density of (R, θ), where these are the polar coordinates, as in

$$X = R\cos\theta \quad \text{and} \quad Y = R\sin\theta.$$

(b) You are on a desert island with your semi-modern computer, which is set up with a random generator for the uniform $(0, 1)$ distribution and with ordinary functions. You are in desperate need of simulations from the standard normal, however, which is not part of your computer's standard setup. How can you survive?

12. Markov chains

Let $X_1, \ldots, X_n$ be a Markov chain over states 1,2,3, with transition matrix

$$\mathbf{P} = \begin{pmatrix} a & b & c \\ b & c & a \\ c & a & b \end{pmatrix},$$

where you may put e.g. $(0.5, 0.3, 0.2)$ for (a, b, c).

(a) Make a routine that simulates such a chain, for say $n = 100$. Let X_1 be drawn from 1,2,3 with equal probabilities. Show, by the way, that the stationary distribution for the chain is precisely $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$.

(b) Characterise the most probable and the least probable sequenes, of all 3^{500} possibilities.

(c) One is interested in the events

$$A = \text{at least once, there are five consecutive 1's,}$$
$$B = \text{at least once, there are three consecutive 3's.}$$

It is hard to compute the probabilities of these events exactly. But you may approximate these via simulations! Do so. Find also $P(A \cap B)$.

(d) Let Y be the number of times the following event occurs in the life of the $X_1, \ldots, X_n$ sequence: three consecutive X_is have three distinct values. Estimate the probability distribution for Y. Investigate briefly the influence of the chain length n for this probability distribution. Find also $\mathrm{E}\, Y$ via an explicit formula, and compare to your simulation output.

13. 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 rae 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.

14. 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.

15. 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 modetrate 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.

16. 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 $\bar{x} = (x_1 + \ldots + x_{12})/12$. 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 achieveing 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) Consctruct 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

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

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

17. 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 17, 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 101, 111, 121, 131, etc. for final analysis. Discuss aspects of such schemes.

18. The multinormal distribution

'Multivariate statistics' is broadly speaking the area of statistical modelling and analysis where data exhibit dependencies. The most important multivarite 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)?

19. 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

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

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.

20. Gaußian processes

Let $Y = \{Y(t): t \geq 0\}$ be a stochastic (or random) process. One may in general terms prove that the full probability distribution of such a process is completely specified via all its finite-dimensional distributions. In other words, if Y and Z are two random processes such that the distributions of $(Y(t_1), \ldots, Y(t_k))$ and $(Z(t_1), \ldots, Z(t_k))$ are identical, for all finite subsets $\{t_1, \ldots, t_k\}$, then Y and Z are probabilistically equivalent, with $\Pr\{Y \in A\} = \Pr\{Z \in A\}$ for all measurable sets A. When defining a stochastic process in such a way, via its finite-dimensional distributions, one must also check certain coherence criteria ('Kolmogorov's consistency conditions'), but we will not go into this here.

(a) We say that a stochastic process is *normal*, or *Gaußian*, if all its finite-dimensional distributions are multinormal. Show that it then suffices to define its *mean function* $m(t) = \mathrm{E}\,Y(t)$ and *covariance function* $k(s,t) = \mathrm{cov}\{Y(s), Y(t)\}$. Show that $k(s,t)$ must satisfy the *nonnegative definite condition*, which is that

$$a^{\mathrm{t}}\Sigma a = \sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j k(t_i, t_j) \geq 0$$

for all $t_1, \ldots, t_m$ and $a_1, \ldots, a_m$, and all finite m. (Here Σ denotes the $m \times m$ matrix of all $k(t_i, t_j)$.) Conversely one may show that a given function $k(s,t)$ satisfying this condition is really a covariance function for a Gaußian process.

(b) Let Y be a Gaußian process over say $[0, 10]$ with mean function $m(t) = 0$ and some given covariance function $k(s,t)$ – in particular, the standard deviation of $Y(t)$ is $k(t,t)^{1/2}$. Give a general recipe for simulating paths from Y, via values across a grid.

(c) Consider the particular Gaußian process Y defined over the unit interval $[0, 1]$ that has mean zero and covariance function $k(s,t) = \min(s,t)$. Simulate say ten realisations of Y, and display them in the same diagram. Use a grid of type $0, 1/m, 2/m, \ldots, m/m$, perhaps with $m = 100$ or more, and explore where the matrix square root operation (described in Exercise 19) might have its current pain limit. (In the **R** version included with my 2007 laptop, I manage to use this recipe with grid size up to say $m = 1000$ without real problems, but the computations slow down with increasing grid size m, as these involve eigenvalues and squarerooting of general symmetric $m \times m$ matrices. For

$m = 100$ computations take 1 second; for $m = 500$ they need about 10 seconds; for $m = 1000$ they need about 40 seconds; for $m = 2000$ they take about three minutes. These computations are what is needed to compute the square root matrix via eigenvalues decomposition. When this matrix is stored, the following computations required to simulate a large number of process paths take very little extra time.)

21. Brownian Motion

The most important Gaußian process is the *Brownian Motion*, defined as follows: $W = \{W(t): t \geq 0\}$ is a standard Brownian motion (or Wiener process, or Einstein process) if (i) $W(0) = 0$; (ii) all increments $W(t) - W(s)$ for $s < t$ are independent; and (iii) if $W(t) - W(s) \sim \mathrm{N}(0, t - s)$ for all $s < t$. In particular, $W(t)$ is zero-mean normal with standard deviation $\sqrt{t}$. The t is often interpreted as 'time', though this is not necessary.

(a) Show that the covariance function for the Brownian motion process is $k(s,t) = \min(s,t)$, i.e. as for Exercise 20(c). Also show that the definition is 'logicall coherent' in that the distributions of

$$W(u) - W(s) \quad \text{and} \quad (W(t) - W(s)) + (W(u) - W(t))$$

agree, for $s < t < u$. In this sense, if one tries to define a process V with independent increments and $V(t) - V(s) \sim \mathrm{N}(0, |t - s|^{\alpha})$, then the only valid choice is $\alpha = 1$; all other choices lead to illogical consequences.

(b) To simulate paths from the W process, using

$$W(i/m) = D_1 + D_2 + \cdots + D_i \quad \text{for } i = 1, 2, 3, \ldots,$$

where $D_1, D_2, \ldots$ are independent and $\mathrm{N}(0, 1/m)$, is much more computationally efficient (and mathematically elegant) than the 'general brute force machinery' of Exercise 20.

(c) Simulate ten Brownian motion paths over $[0, 5]$, and display them in the same diagram.

(d) For a Brownian motion over the time interval $[0, 5]$, consider the event A that the maximum value of W exceeds 3.333 while its minimum value lands below -2.222. Compute the probability of A via stochastic simulations.

22. The Brownian Bridge

The *Brownian bridge* $W^0 = \{W^0(t): 0 \leq t \leq 1\}$ may be characterised as the Brownian motion process $W = \{W(t): 0 \leq t \leq 1\}$ conditional on the event $W(1) = 0$. This process is important in various probabilistic analyses of phenomena related to Wiener processes and empirical processes.

(a) For $0 < s < t < 1$, show that

$$\begin{pmatrix} W^0(s) \\ W^0(t) \end{pmatrix} \sim \mathrm{N}_2(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} s(1-s) & s(1-t) \\ s(1-t) & t(1-t) \end{pmatrix}).$$

In particular, the standard deviation of $W^0(t)$ is $\sqrt{t(1-t)}$, smaller than $\sqrt{t}$, the standard deviation of the non-constrained $W(t)$.

(b) Use the result above to infer that the covariance function of a Brownian bridge is $k(s,t) = \min(s,t)\{1 - \max(s,t)\}$, and use this to simulate ten bridges. Display these in the same diagram.

(c) Show that the process $Z(t) = W(t) - tW(1)$ has the very same covariance function as that of (b), and hence infer that W^0 and Z are equivalent processes. Simulating paths of W^0 is therefore more easily accomplished via paths of W. Repeat the task of point (b) using this alternative view.

(d) The Kolmogorov–Smirnov goodness-of-fit test statistic for testing whether the distribution F underlying a given data set $X_1, \ldots, X_n$ is equal to some specified F_0, say the standard normal, is $D_n = \max_t |F_n(t) - F_0(t)|$, where $F_n(t) = (1/n)\sum_{i=1}^n I\{X_i \le t\}$ is the empirical distribution function of the data set. Empirical processes theory may be used to prove that

$$\sqrt{n} D_n \to_d D = \max_{t \in [0,1]} |W^0(t)|.$$

Simulate the distribution of D using e.g. 10,000 bridge paths, and estimate in particular the upper 0.05 quantile point, say c. Rejecting $F = F_0$ when $D_n \ge c/\sqrt{n}$ is then a test with asymptotic significance level 0.05.

23. The Ornstein–Uhlenbeck process

The *Ornstein–Uhlenbeck process* may be defined in various ways, but its main purpose is to portray a normal process with constant variability level (whereas e.g. the Wiener process has standard deviation growing as the square root of elapsed time). For our purposes, say that $Y = \{Y(t) : t \ge 0\}$ is an Ornstein–Uhlenbeck process if is Gaußian, with mean zero, and covariance function

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

Here $\rho = \exp(-a)$ is the correlation between random process points positioned 1 time unit away from each other.

(a) Since Y is Gaußian with a given covariance function, the 'brute force' method of Exercise 20(c) may be used to simulate paths across a given grid, say along points with inter-distance $1/m$ for $m = 100$ or $m = 1000$. Use this to simulate paths of Y over the time interval $[1, 4]$.

(b) There is however an alternative definition or representation of Y that makes simulation rather easier (and more mathematically transparent), as a normalised Wiener process. Define

$$Z(t) = \frac{W(\exp(2at))}{\exp(at)} \quad \text{for } t \in I\!R.$$

Show that this is actually an Ornstein–Uhlenbeck process. Use this to simulate say 10,000 paths over the time interval $[1, 4]$, and compute mean and standard deviations for $U = \min_t Z(t)$, $V = \max_t Z(t)$, and their inter-correlation.

(c) Consider the process

$$Z(t) = (1 - c)\, Y(t) + c\, W(t) \quad \text{for } 0 \le t \le 4,$$

where Y is an Ornstein–Uhlenbeck process independent of the Brownian motion process W, and where c is a given constant, perhaps small or zero. Simulate paths of this process. Determine the distribution of $\max_{t\in[0,4]} |Z(t)|$, for $c = 0$ and for other positive values of c. Suggest ways in which to test the statistical null hypothesis that 'the world is stable' (i.e. $c = 0$) versus the alternative hypothesis that 'the world is changing' (i.e. $c > 0$).

24. The Dirichlet distribution and multiomial data

A very classical Bayesian calculations is the following, dating (in one of its particular forms) all the way back to the 1764 publication *An Essay Towards Solving a Problem in the Doctrine of Chances* by Rev. Thomas Bayes (1702–1761): if X is binomial (n, p), and the unknown p is given a Beta(a, b) prior distribution, then it may be *updated*, and the posterior distribution is a Beta$(a + x, b + n - x)$. – This exercise relates to the natural generalisations of this classic result to situations with more than two cells. The multinomial generalises the binomial, and the Dirichlet generalises the Beta. Let $(p, q, 1 - p - q)$ have a Dirichlet distribution with parameters (a, b, c), i.e. the density of (p, q) is of the form

$$f(p, q) = \frac{\Gamma(a + b + c)}{\Gamma(a)\Gamma(b)\Gamma(c)} p^{a-1} q^{b-1} (1 - p - q)^{c-1} \quad \text{for } p > 0,\ q > 0,\ p + q < 1.$$

(a) Show that $p \sim \text{Beta}(a, b + c)$ and that $q \sim \text{Beta}(b, a + c)$.

(b) Find the distribution for p given q and for q given p. Express your answers in terms of Beta distributions so that **R** routines may be used to generate realisations from $p \mid q$ and $q \mid p$.

(c) One is interested in the probabilities p_1 and p_6 for achieving a '1' and '6' respectively for a suspicious-looking die ('Gott würfelt nicht'). As a prior distribution for these I choose for simplicity the uniform one, over the triangle,

i.e. $\pi(p_1, p_6) = 2$ for $p_1 + p_6 < 1$. (This is the Dirichlet with parameters $(1, 1, 1)$.) The die is cast $n = 100$ times, and one observes a 1 $X_1 = 10$ times and a 6 $X_6 = 20$ times. Set up a Gibbs sampler that produces realisations $(p_{1,i}, p_{6,i})$ from the posterior density. Display a density estimate for the distribution of $\theta = p_6/p_1$, estimate θ, and compute a 90% confidence interval (or credibility interval) for this parameter.

(d) Here one may actually find the exact mean and variance of p_6/p_1 given data. Do this, and compare with the estimates from the Gibbs sampler.

(e) There is an alternative method for simulating Dirichlet distributed vectors that does not require iterative Gibbsian chains. Show that if G_1, G_2, G_3 are independent and Gamma distributed with parameters respectively $(a, 1)$, $(b, 1)$, $(c, 1)$, then

$$(X, Y, Z) = \Big(\frac{G_1}{G_1 + G_2 + G_3}, \frac{G_2}{G_1 + G_2 + G_3}, \frac{G_3}{G_1 + G_2 + G_3}\Big)$$

has a Dirichlet distribution with parameters (a, b, c). Use this in the case above to make an alternative sequence of θ values.

(f) Finally one may also find the exact probability density for $\theta = p_6/p_1$ given data. Do this, and compare with the density estimate based on an observed simulation chain from the Gibbs sampler. – Here Point 1 is that the Gibbs sampler just as easily produces answers for other parameters of arbitrary complexity, where analytical calculations become too complicated. Point 2 is that analogous Gibbs samplers may be set up also for more complicated choices of the prior than the one used above for (p_1, p_6), and where again exact calculations quickly become too complicated.

25. Gibbs sampler for a two-dimensional density

This exercise provides a concrete illustration of the Gibbs sampler in a two-dimensional situation.

(a) Consider the density

$$f(x) = \begin{cases} c\exp(x-2) & \text{for } 0 \le x \le 1, \\ c\exp(-x) & \text{for } x \ge 1. \end{cases}$$

Find the cumulative distribution function. Construct a method that manages to simulate say 1000 independent realisations from f. Use these to estimate the mean value and the standard deviation for this distribution. (You should also find exact numerical values for these.)

(b) Then consider the probability density

$$f(x, y) = c(\theta)\exp\{-|x| - |y| - \theta|x - y|\} \quad \text{for } x, y \in I\!R.$$

What is the proper parameter region for the θ parameter? Implement a rejection sampling method that produces 10,000 independent pairs (X_i, Y_i) from this density, and do this for $\theta = 0.333$. Plot these simulated data, and estimate means, standard deviations, and correlation.

(c) Then construct a Metropolis algorithm of the type

$$(X_i, Y_i) = \begin{cases} (X_i + \delta_i, Y_i + \varepsilon_i) & \text{with probability } p_i, \\ (X_i, Y_i) & \text{with probability } 1 - p_i, \end{cases}$$

where $(\delta_i, \varepsilon_i)$ is drawn from a symmetric distribution around origo, with small reach. Again, get hold of 10,000 independent pairs, for the case of $\theta = 0.333$, and answer questions raised in the previous point.

(d) Find the densities $f_2(y \,|\, x)$ for Y given x and $f_1(x \,|\, y)$ for X given y. Simulate 10,000 independent Ys from its distribution given $x = 1.111$, again with $\theta = 0.333$. Estimate the density $f_2(y \,|\, x = 1.111)$ from these, and compare with the exact curve.

(e) Set up and implement a Gibbs sampler (X_i, Y_i) that utilises

$$X_i \sim f_1(x \,|\, Y_{i-1}) \quad \text{and} \quad Y_i \sim f_2(y \,|\, X_i);$$

the start value X_1 is arbitrary. How can this be used to get hold of 10,000 independent pairs from f? Do this, and again answer questions raised in point (b).

26. Computing risk functions via simulation

Assume that X is normal $(\theta, 1)$, where θ is unknown and needs to be estimated. When $\widehat{\theta} = \widehat{\theta}(X)$ is an estimator of θ there is a long tradition of examining its properties via its so-called *risk function* (under quadratic loss),

$$R(\theta) = \mathrm{E}_\theta(\widehat{\theta} - \theta)^2 \qquad \text{for } \theta \in I\!R.$$

In various cases this function may be evaluated analytically, but for many estimators this is too complicated, and one needs to compute the risk via numerical integration or simulation. The point here is to illustrate and discuss the simulation approach. The reader is asked to realise that the methods are rather general, and apply to much more complicated situations than the present $\mathrm{N}(\theta, 1)$ case.

(a) The classic estimator in the normal situation is of course $\widehat{\theta} = X$ itself. Find its risk function.

(b) An alternative estimator is

$$\widetilde{\theta} = \frac{X^2}{1 + X^2}\, X.$$

There are several different lines of arguments that lead to this estimator, including empirical Bayes and shrinkage constructions. Your task is to use simulations to compute its risk function $R(\theta)$ to a good precision level.

(c) Estimate the R function for each of the θ values $-4.0, -3.9, \ldots, 3.9, 4.0$ (for example) by for each simulating a large number of X from the $N(\theta, 1)$ and then recording the average value of $(\widetilde{\theta} - \theta)^2$. Plot the result in a diagram. (This is what may be termed 'the direct method'.)

(d) Alternatively one may use that an X from $N(\theta, 1)$ may be represented as $\theta + Z$, where $Z \sim N(0, 1)$, which implies

$$R(\theta) = \mathrm{E}\, H(Z, \theta) = \mathrm{E}\Big\{ \frac{(\theta + Z)^2}{1 + (\theta + Z)^2} (\theta + Z) - \theta \Big\}^2.$$

Estimate the $R(\theta)$ function by simulating a large number of Z_is and then computing the average of $H(Z_i, \theta)$. This may be termed 'the contextual method'. Plot also this estimated $R(\theta)$ curve in a diagram and compare with the result from the direct method.

(e) Discuss differences and pros and cons between these methods, based on your experiments, and also via analysis of the precision of risk function estimates. Comment also on the differences in risk between the classic estimator $\widehat{\theta}$ and the empirical Bayes estimator $\widetilde{\theta}$.

27. One-dimensional image restoration

This exercise develops an example of a simulation based one-dimensional image restoration method. Suitable generalisations of the models and methods presented here have broad applications within modern image analysis, where the images in question may be of dimension one, two, three, or even four (time and space).

(a) Consider the simple Markov model for a chain $(X_1, \ldots, X_n)$ that takes on values 0–1, that has

$$\mathbf{P} = \begin{pmatrix} 1 - a, & a \\ a, & 1 - a \end{pmatrix}$$

as its transition probability matrix. The equilibrium distribution is $(\frac{1}{2}, \frac{1}{2})$. Show that the so-caled *local characteristics*, namely $\Pr\{X_i = x_i \,|\, \text{rest}\}$, only depend on the two nearest neighbours:

$$p_i(x_i \,|\, x_{\partial i}) = \Pr\{X_i = x_i \,|\, \text{rest}\} = p_i(x_i \,|\, x_{i-1}, x_{i+1}).$$

Find these explicitly, for each of the cases $(0,0), (0,1), (1,0), (1,1)$ for (x_{i-1}, x_{i+1}).

(b) Simulating realisations from the X chain directly is easily done, by letting X_i be generated after X_{i-1} (cf. Exercise 12). But it will be worth the trouble to work out how a Metropolis algorithm may be used to solve the task, since this is a more powerful tool in the situation to be considered later, where the

X chain is not fully observed but needs to be modelled given a noisy image thereof. – From a realised chain $(x_1, \ldots, x_n)$, select an arbitrary position i, and consider the possibility of changing x_i to $x_i' = 1 - x_i$. Show that the Metropolis method accepts this suggestion with probability $\text{pr}_i = \min(1, S_i)$, der

$$S_i = \frac{p_i(x_i' \,|\, x_{i-1}, x_{i+1})}{p_i(x_i \,|\, x_{i-1}, x_{i+1})} = \frac{P_{x_{i-1},x_i'} P_{x_i',x_{i+1}}}{P_{x_{i-1},x_i} P_{x_i,x_{i+1}}}.$$

Show further that this gives rise to the following table of S_is, for the possible neighbour-dependent transitions from x_i to x_i', where we write $\rho = a/(1-a)$:

$$\begin{aligned}
&\text{from } (0,0,0) \text{ to } (0,1,0) \qquad \text{with } S_i = \rho^2,\\
&\text{from } (0,0,1) \text{ to } (0,1,1) \qquad \text{with } S_i = 1,\\
&\text{from } (0,1,0) \text{ to } (0,0,0) \qquad \text{with } S_i = 1/\rho^2,\\
&\text{from } (0,1,1) \text{ to } (0,0,1) \qquad \text{with } S_i = 1,\\
&\text{from } (1,0,0) \text{ to } (1,1,0) \qquad \text{with } S_i = 1,\\
&\text{from } (1,0,1) \text{ to } (1,1,1) \qquad \text{with } S_i = 1/\rho^2,\\
&\text{from } (1,1,0) \text{ to } (1,0,0) \qquad \text{with } S_i = 1,\\
&\text{from } (1,1,1) \text{ to } (1,0,1) \qquad \text{with } S_i = \rho^2.
\end{aligned}$$

(c) Assume now there is a *true image* $x^0 = (x_1^0, \ldots, x_n^0)$, which one however only is able to observe with additional noise, or blurring:

$$y_i = x_i^0 + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

where the ε_is are independent and normal $(0, \sigma^2)$. The noise level σ may most often be estimated separately with good precision and is considered known. The statistical challenge is to somehow estimate (or restore) the true image x^0 from its noisy version $y = x^0 + \varepsilon$. – A simple start estimate $\widetilde{x}^0$ for x^0 emerges by letting $\widetilde{x}_i^0$ be 1 if $y_i > \frac{1}{2}$ and 0 if $y_i \leq \frac{1}{2}$. We shall investigate hos this works when $n = 100$, in a situation where the true image is

```
0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 1 1 1 1 1 1
1 1 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0
0 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0
0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 1 1
```

This sequence may be concisely generated in **R** as

```
x0 <- round( (sin(b*list))^2 ),
```

where `list` is $(1, 2, 3, \ldots, n)$ and $b = 0.20$. (One may later on experiment with other images with different degrees of 'context', via other values of b, etc.) – Now make an y image by adding simulated noise, with $\sigma = 0.60$. Illustrate the situation via

```
matplot(list, cbind(x0, y), type = "l"),
```

and compute the error rate $n^{-1} \sum_{i=1}^n I\{\widetilde{x}_i^0 \neq x_i^0\}$ for the simple (and non-contextual) $\widetilde{x}^0$ method.

– The idea is now to put a Metropolis chain of some thousands of x chains in motion, that have $p(x \,|\, y)$ as its equilibrium distribution (in their sample space of all 2^{100} chains of length $n = 100$). We let the chain of chains start in position $\widetilde{x}^0$, the non-contextual reconstruction of the image. As in point (b), let the proposal for change from a given $x = (x_1, \ldots, x_i, \ldots, x_n)$ be the simple $x' = (x_1, \ldots, x'_i, \ldots, x_n)$, where i is selected uniformly across the indexes in question and where $x'_i = 1 - x_i$. For simplicity we let these indexes be merely $2, 3, \ldots, n-1$, without $i = 1$ og $i = n$, in that we let $x_1 = \widetilde{x}^0_1$ and $x_n = \widetilde{x}^0_n$ for all iterations. (These outer positions do not have two neighbours, so certain special rules would have to be devised for these.)

(d) Show that the Metropolis algorithm is to accept the proposed change $x'_i = 1 - x_i$, in the selected position i, with probability $\mathrm{pr}_i = \min(1, S_i T_i)$, where S_i is as above and where

$$T_i = \frac{f(y_i \,|\, x'_i)}{f(y_i \,|\, x_i)} = \exp\Big[\tfrac{1}{2}\frac{1}{\sigma^2}\{(y_i - x_i)^2 - (y_i - x'_i)^2\}\Big].$$

Show that this also may be expressed as $\exp\{(1/\sigma^2)(y_i - \frac{1}{2})\}$ for the situation where $x_i = 0$ perhaps is to be replaced by $x'_i = 1$, and as $\exp\{-(1/\sigma^2)(y_i - \frac{1}{2})\}$ for the situation where $x_i = 1$ perhaps is to be replaced by $x_i = 0$. This may also be formulated as $T_i = \exp\{(1/\sigma^2)(1 - 2x_i)(y_i - \frac{1}{2})$.

(e) Implement this method, and use it for generating a suitably large number of chains of chains x. Then attempt to reconstruct the original image x^0, using the estimate $\widehat{x}^0 = (\widehat{x}^0_1, \ldots, \widehat{x}^0_n)$, where $\widehat{x}^0_i$ is the most popular state at position i, i.e. the one among the the two states 0 and 1 that has been seen most frequently at position i in the course of the simulations. Compute the error rate for $\widehat{x}^0$ og discuss the result. This operation is to be executed a suitably larte number of chains of chains x. Then attempt to reconstruct the original image x^0, using the estimate $\widehat{x}^0 = (\widehat{x}^0_1, \ldots, \widehat{x}^0_n)$, where $\widehat{x}^0_i$ is the most popular state at position i, i.e. the one among the the two states 0 and 1 that has been seen most frequently at position i in the course of the simulations. Compute the error rate for $\widehat{x}^0$ and discuss the result. This operation is to be executed for $a = 0.15$ and $\sigma = 0.60$. (You should also experiment with other values of b, a, σ.)

(f) The method developed above has as its aim to choose the final estimate $\widehat{x}^0$ such that

$$\Pr\{X_i = \widehat{x}^0_i \,|\, y\} \quad \text{is maximal, for each } i.$$

Show that this is the same as minimising the expected error rate given data. The underlying *loss function* is in other words

$$L(x^0, \widehat{x}^0) = n^{-1} \sum_{i=1}^{n} I\{\widehat{x}^0_i \neq x^0_i\}.$$

A different loss function that often is used (or that one often attempts to use) is

$$L(x^0, \widehat{x}^0) = I\{\widehat{x}^0 \neq x^0\},$$

with full score if one has done everything perfectly and zero award if one does one more more errors. Show that the optimal estimator then becomes the sequence $x^* = (x_1^*, \ldots, x_n^*)$, among all 2^{100} sequences, that has the maximal probability given data, i.e. the one chain that maximises $p(x \,|\, y)$. Do you have any suggestions or ideas for finding this *maximum a posteriori probability image*?

28. Finding the mode

For a given probability distribution on some sample space, where is the most probable location? Or the most probable state for a distribution over a large number of states? The problem is rather simple in situations with a small or a moderate number of outcomes, but much more demanding when the multitude of possible outcomes exceeds ordinary bounds – which is often the case! (The number of different chess games far exceeds the number of atoms in our galaxy.) This 'finding the mode' problem is central in many classical and modern statistical applications. We shall see here that stochastic simulations may be used to search for the solution.

Let $f(i) = \pi_i$ be some given probability distribution over positions $i = 1, \ldots, M$. These positions may e.g. be the 2^{100} different possibilities for an 0–1 chain $(X_1, \ldots, X_{100})$. It is not vital that they are numbered from 1 to M, and the generality of the situation encompasses cases of vectors or processes.

(a) Assume that i_0 is the unique position at which $f(i)$ is maximal, i.e.

$$\pi_{i_0} > \pi_i \quad \text{for all } i \neq i_0.$$

Now we apply heat!, and consider the heated-up probability distribution

$$f_t(i) = \text{const.}\, \pi_i^t = \frac{\pi_i^t}{\pi_1^t + \cdots + \pi_M^t} \quad \text{for } i = 1, \ldots, M.$$

One terms t 'the temperature'. Show that as the temperature increases towards the one found at the surface of the Sun, then f_t will concentrate at i_0 alone.

(b) Assume then that there are two equally worth winners in the competition of maximal probability, i.e. that for two indexes i_1 and i_2 one has $\pi_{i_1} = \pi_{i_2}$ bigger than all other π_i. Show that the heated-up $f_t(i)$ in the end concentrate with equal probabilities $\frac{1}{2}$ and $\frac{1}{2}$ at the two positions i_1 and i_2. Generalise.

(c) Let there be a unique maximand i_0 of the f distribution. Assume you succeed in generating one or more X_t from the f_t distribution, for any selected t. Let $t_1 < t_2 < t_3 < \cdots$ be an increasing sequence of temperatures growing towards infinity, and let X_{t_j} be simulated from f_{t_j}. Show that X_{t_j} converges towards i_0 in probability. – Hence this is a simulation based recipe for searching for i_0.

(d) Generalise the preceding results to probability distributions with infinitely many states.

(e) Let $f(i) = \pi_i$ be the Poisson distribution with parameter 100.5. Find i_0. Then try to search out the value of i_0 (that you now already know) via simulation. Use `sample` in **R**, using sample region `list <- 50:150`, for example, in that you ignore outcomes falling outside this region. – Here it is fruitful to express the probabilities via the exponential function, such that

$$f_t(x) = \text{const. } \exp(-t\theta + t\theta \log x - t \log \Gamma(x+1)).$$

(One ought to be aware of **R**'s understandable aversion to, frustration, weltangst and insecurity when faced with the very tiny and the very large exponents, and therefore plan programmes with the right modicum of care to avoid these symptoms.)

Rather often one is however not able to simulate directly or easily from f_t. Then one may attempt to use MCMC schemes, say Metropolis–Hastings or the Gibbs sampler, for steadily increasing temperatures. In the end one hopes that outcomes concentrate around i_0. Here one ought to expect many competing specialisations of these ideas, depending also on the computational cost and how to spread this cost across different temperatures. Probability theory arguments may be furnished to demonstrate that one should not let the temperature increase too quickly.

(f) Attempt to run such a programme in the situation with the Poisson distribution with parameter 100.5. Let

$$f_t(x) = \text{const.} \left(e^{-\theta} x^\theta / x!\right)^t \quad \text{for } x = 0, 1, 2, \ldots$$

(cf. the note made in point (e)), and make a Metropolis algorithm that produces a chain $X_1, X_2, \ldots$ that has the f_t as its equilibrium distribution. Put a collection of such into motion, for steadily increasing temperatures, and check whether you actually succeed in simulating your way towards the correct answer i_0 (that you in this particular situation knew in advance, but that in most situations of interest will not be known pre simulations and analysis).

(g) Apply this machinery for the case of a Poisson with parameter 100, and comment on the results.

29. Finding the mode: continuous case

The previous exercise has parallels to problems with continuous distributions. Let $f(x)$ be a continuous probability density for $x \in \mathbb{R}$ and assume it has a unique maximiser $x = x_0$. The task is to ascertain this x_0, from simulations. (The method we give here works in higher dimensions, where the simulation techniques are more needed than in dimension one.)

(a) Consider the heated distribution

$$f_t(x) = \text{const.}\, f(x)^t = f(x)^t \Big/ \int f(u)^t \, \mathrm{d}u.$$

Show that when t grows towards infinity, then f_t will concentrate more and more around x_0. More formally, X_t, drawn from f_t, will converge in probability to x_0.

(b) Illustrate this in the exceedingly simple situation where f is the standard normal density. What is the mean value and standard deviation for X_t, drawn from const. $f(x)^t$?

(c) Let now $f = 0.67\, g_1 + 0.33\, g_2$, where g_1 is $\mathrm{N}(0,1)$ and g_2 is $\mathrm{N}(2, 0.5^2)$. Display the density in a diagram and find x_0 with three correct decimals.

(d) For given temperature t, set up and implement a MCMC scheme that has f_t as its stationary distribution. Roll chain.

(e) Try to find x_0 via simulations, via increased temperatures. Sum up your findings.

(f) Carry out experiments of the previous type for the situation where

$$f = 0.22\, \mathrm{N}(-2, 0.3^2) + 0.33\, \mathrm{N}(0, 0.7^2) + 0.45\, \mathrm{N}(2, 0.4^2).$$

Try out different start values and proposal distributions for the Metropolis chain.

30. Finding the most probable image

We shall soon direct our attention back to the problem raised at the end of Exercise 27, where the precise question is: which one, among the multitude of 2^n chains $x = (x_1, \ldots, x_n)$ has the highest a posteriori probability $p(x \,|\, \text{data})$? But we start more carefully and simplistically, in a situation without y measurements.

(a) Let again the Markov chain have transition probability matrix

$$\mathbf{P} = \begin{pmatrix} 1-a, & a \\ a, & 1-a \end{pmatrix},$$

where we assume $a < \frac{1}{2}$ (more context and comradery than repulsion and shiftnyness). Let

$$N_{u,v} = \sum_{i=2}^{n} I\{x_{i-1} = u, x_i = v\}$$

count the number of transitions from state u to state v. Show that the simultaneous probability distribution may be written

$$p(x) = p(x_1, \ldots, x_n) = a^M (1-a)^{n-M},$$

where

$$M = N_{0,1} + N_{1,0} = \sum_{i=2}^{n} \{x_i \neq x_{i-1}\}$$

counts the number of shifts $0 \to 1$ and $1 \to 0$.

(b) Find the most probable outcomes of chains $x = (x_1, \ldots, x_n)$. There are acutally several different x chains with identical and highest probability. You may also find those chains that have the lowest probability, and compare these lowest probabilities with the highest probabilities. Generalise these findings to the case where the transition probability matrix is any

$$\mathbf{P} = \begin{pmatrix} 1-a, & a \\ b, & 1-b \end{pmatrix}.$$

– In these particular rather simple situations there is no real need to simulate oneself forwards via heated temperatures to find appropriate answers. It is nevertheless fruitful to attempt to do so, in that the much more important and much more difficult challenge, where x is a hidden rather than an openly observed Markov chain, which does not have any explicitly findable solution, may be attacked using similar ideas and techniques.

(c) The recipe of Exercises 28 and 29 is to study and then to simulate from the distribution

$$\begin{aligned} p_t(x) = p_t(x_1, \ldots, x_n) &= \text{const.}\, \{a^M (1-a)^{n-M}\}^t \\ &= \text{const.}\, \exp\{tM \log a + t(n-M)\log(1-a)\} \\ &= \text{const.}\, \exp\{-tM \log((1-a)/a)\}. \end{aligned}$$

We see that the lower M is, the higher are the probabilities; the pedagogically instructive exercise is to see whether temperature increase methods with Metropolis succeeds in learning this by itself! Note that p_t is no longer a Markov chain, so one can not simulate x chains along the ordinary Markovian road. – As in Exercise 27 one selects and index i uniformly, between 2 and $n-1$, and asks whether the shift from x_i to $x_i' = 1 - x_i$ is worthwhile. Show

that this shift is to be taken with acceptance probability $\text{pr}_i = \min(1, S_i)$, where

$$S_i = \exp\{-t(M' - M)\log((1-a)/a)\},$$

where $M' = M(x_1, \ldots, x_i', \ldots, x_n)$ and $M = M(x_1, \ldots, x_n)$. Show also that this may be simplified to

$$\begin{aligned} M' - M &= 2I\{x_i = x_{i-1}\} + 2I\{x_i = x_{i+1}\} - 2 \\ &= \begin{cases} 2 & \text{for } (0,0,0) \text{ og } (1,1,1); \\ 0 & \text{for } (0,1,1),\ (1,1,0),\ (0,0,1) \text{ og } (1,0,0); \\ -2 & \text{for } (0,1,0) \text{ og } (1,0,1). \end{cases} \end{aligned}$$

(d) Implement this Metropolis method, in a situation with $n = 100$ and $a = 0.15$, for example, and run qua temperature master a sensible scheme that leads to simulated chains coming close to the most probable ones.

31. Image restoration: finding the MAP

We shall now attack the more serious problem of finding the most probable image x based on a blurred or noisy image $y = x + \varepsilon$. The probability distribution for the image x given the observed y is of the form

$$p(x \,|\, y) = \text{const.}\, p(x) f(y \,|\, x) = \text{const.}\, p(x) \prod_{j=1}^{n} f(y_j \,|\, x_j),$$

with x following a Markov chain. The (ambitious) task is to use simulations to find the so-called *MAP solution* (the maximum a posteriori probability solution), defined as the x^* (possibly among several equally worthy candidates) that has highest $p(x^* \,|\, y)$.

(a) When the noise is white, i.e. the ε_is are independent and normal $(0, \sigma^2)$ for some σ, show that the MAP solution is identical to the chain x^* that minimises

$$\text{crit}(x) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2 + \frac{2\sigma^2}{n} M \log\frac{1-a}{a},$$

where again $M = M(x_1, \ldots, x_n) = N_{0,1} + N_{1,0}$ is the number of paradigm shifts for the x chain. Explain what this leads to for very low σ and for very high σ.

(b) Let as in Exercise 27 the truth be

```
x0 <- round( (sin(b*list))^2 ),
```

where `list` is $(1, 2, 3, \ldots, n)$ and $b = 0.20$, and simulate $y_i = x_i + \varepsilon_i$ with noise level $\sigma = 0.60$. We are aiming for the MAP estimate for the underlying

x. The idea is to simulate e.g. `sim` = 100 chains, for each of a number of increasing temperatures t, from the probability distribution

$$p_t(x\,|\,y) = \text{const.}\, p(x)^t f(y\,|\,x)^t$$
$$= \text{const.}\, \{a^M(1-a)^{n-M}\}^t \prod_{j=1}^n \exp\Big\{-\tfrac12 \frac{t}{\sigma^2}(y_i - x_i)^2\Big\},$$

via the Metropolis algorithm. The temperature is to increase slowly. Show that Metropolis (and the famous Tellers) accept the move from x_i to $x_i' = 1 - x_i$ with probability $\text{pr}_i = \min(1, S_i T_i)$, where S_i is as in Exercise 27 and where

$$T_i = \Big(\frac{f(y_i\,|\,x_i')}{f(y_i\,|\,x_i)}\Big)^t = \exp\Big[\tfrac12 \frac{t}{\sigma^2}\{(y_i - x_i)^2 - (y_i - x_i')^2\}\Big].$$

This quantity is as in Exercise xx, but with $\sigma/\sqrt{t}$ replacing σ.

(c) Let the MAP solution be your message to García. For each round with a new temperature t and let us say 100 Metropolis rounds, compute the criterion $\text{crit}(x)$. It is hoped and intended that the criterion's level is sinking with increased temperature.

32. Point processes

Here we shall consider a probability model for the placing of points in a planar region. For the sake of concreteness we let this region be the unit square $[0,1] \times [0,1]$. The model specifies that for given number of n points $\mathbf{x} = \{x_1, \ldots, x_n\}$ inside the unit square the probability density of their geographical location is

$$f(\{x_1, \ldots, x_n\}) = c(\rho)\rho^M, \quad \text{where } M = M(\mathbf{x}) = \sum_{i<j} I\{\|x_j - x_i\| \le r\}.$$

Here ρ is a parameter i $(0,1]$, while M counts the number of neighbour pairs, in the sense that two points are neighbours when their inter-distance is at most r.

Implement and test the Metropolis–Hasting method for simulating realisations from f. The starting point may e.g. be a regular point pattern or a pattern drawn from the uniform distribution. The algorithm shall within one iteration first first select one of the n points randomly, say $x_i = (u_i, v_i)$, and then judge whether x_i is to be moved to a new candidate position, say $x_i' = (u_i', v_i')$.

You may e.g. test your methods for $n = 50$, $r = 0.15$, and some values of ρ. When following the iterations of the MCMC scheme, you ought to follow the number of times points are moving and also the number of neighbouring points M. You may simulate e.g. 100 iterations at a time, and after each such 100-job evaluate how many point translations there has been and whether M has shifted in value. You are to try out two versions of this idea. In both cases your implementation should contain '`plot(u,v)`' at the end of the programming loop, so that you may seen the points jump & dance on your screen.

(a) The first version should use the proposal distribution where $q(x' \mid x)$ is the uniform over the unit square.

(b) The second version should utilise $q(x_i' \mid x_i)$ equal to the inform over $B(x_i)$, where $B(x_i)$ is the largest box around the midpoint x_i that still fits inside the unit square; that is,

$$B(x) = B(u_i, v_i) = [u_i - d_i, u_i + d_i] \times [v_i - e_i, v_i + e_i],$$

where $d_i = \min(u_i, 1 - u_i)$ and $e_i = \min(v_i, 1 - v_i)$.

(c) Assume I give you a point pattern with 50 points in the unit square, during an exam or another rainy day, and tell you that these have been drawn from the model above, for $r = 0.15$ and for a suitable value of the parameter ρ. Can you then estimate the value of ρ that I have used?