UNIVERSITETET I OSLO

Matematisk Institutt

Exam in:	**STK 4190/9190:**
	Bayesian Nonparametrics
	Part I of two parts: The project
With:	**Nils Lid Hjort**
Time for exam:	**4.–16.xii.2019**

This is the exam project set for STK 4190/9190, autumn semester 2019. It is made available on the course website as of *Wednesday 4 December 10:01*, and candidates must submit their written reports by *Monday 16 December 10:59* (or earlier), to the reception office at the Department of Mathematics, *in duplicate*. The supplementary oral examination part takes place *Thursday December 19* (practical details concerning this are provided elsewhere). Reports may be written in nynorsk, bokmål, riksmål, English or Latin, and should preferably be text-processed (TeX, LaTeX). Give your name (and your student-web identification number) on the first page; the markers need to couple project reports with the oral examinations. Write concisely (in der Beschränkung zeigt sich erst der Meister; brevity is the soul of wit; краткость – сестра таланта). Relevant figures need to be included in the report. Copies of relevant parts of machine programmes used (in `R`, or `matlab`, or similar) are also to be included, perhaps as an appendix to the report. Candidates are required to work on their own (i.e. without cooperation with any others). They are graciously allowed not to despair should they not manage to answer all questions well.

Importantly, each student needs to submit *two special extra pages* with her or his report. *The first* (page A) is the 'erklæring' (self-declaration form), properly signed; it is available at the webpage as 'Exam Project, page A, self-declaration form'. *The second* (page B) is the student's one-page summary of the exam project report, which should also contain a brief self-assessment of its quality.

This exam set contains four exercises and comprises nine pages.

Exercise 1

Norsk bistandspolitikk er som en betasuppe som har stått og kokt i lengre tid (says a certain professor of economics from the University of Bergen). Here we shall employ Beta processes from a different soup, for analysing two groups of Australian drug users, associated with two different clinics. The question is how long time a drug user spends with the clinic, before he or she leaves. There's an appropriate technical definition of 'leaving the clinic', involving transition back to normal society, etc., but which does not need to concern us here.

The data are available at the course website, as `heroin2-data`, with six columns, say `id, tt, delta, x1, x2, x3`, with `tt` the time spent in the clinic, in years; `delta` is 1 if

that time is observed and 0 if it is censored; `x1` is methadone dose; `x2` is an indicator for the user to have spent time in prison or not; and `x3` is 1 for clinic 1 and 2 for clinic 2. Here we shall ignore covariates `x1` and `x2`. You may access the data in `R` using

```
heroin <- matrix(scan("heroin2-data",skip=7),byrow=T,ncol=6)
```

after which you may set up data (t_i, δ_i) for clinics 1 and 2 using

```
heroin1 <- heroin[x3 == 1, c(2,3)]
heroin2 <- heroin[x3 == 2, c(2,3)]
```

Consider the cumulative hazard functions $A_1(t)$ and $A_2(t)$, for the processes involved in getting users of clinics 1 and 2 to leave. The interpretation is that

$$\mathrm{d}A_j(s) = \Pr\{\text{leaves in } [s, s+\mathrm{d}s] \,|\, \text{still there at time } s\}, \quad \text{for } j = 1, 2.$$

These are associated with survival curves

$$S_j(t) = \Pr\{T_j \geq t\} = \prod_{[0,t]} \{1 - \mathrm{d}A_j(s)\} \quad \text{for } j = 1, 2,$$

where 'survival' here means 'have not yet left the clinics'. The usual nonparametric estimators for the A_j and the S_j are the Nelson–Aalen and Kaplan–Meier estimators. Figure A shows such Kaplan–Meier estimates, for clinic 1 (black) and clinic 2 (red), along with parametrically fitted curves, where I have used the gamma distribution. The point of the present exercise is to use Beta processes for carrying out Bayesian analysis.

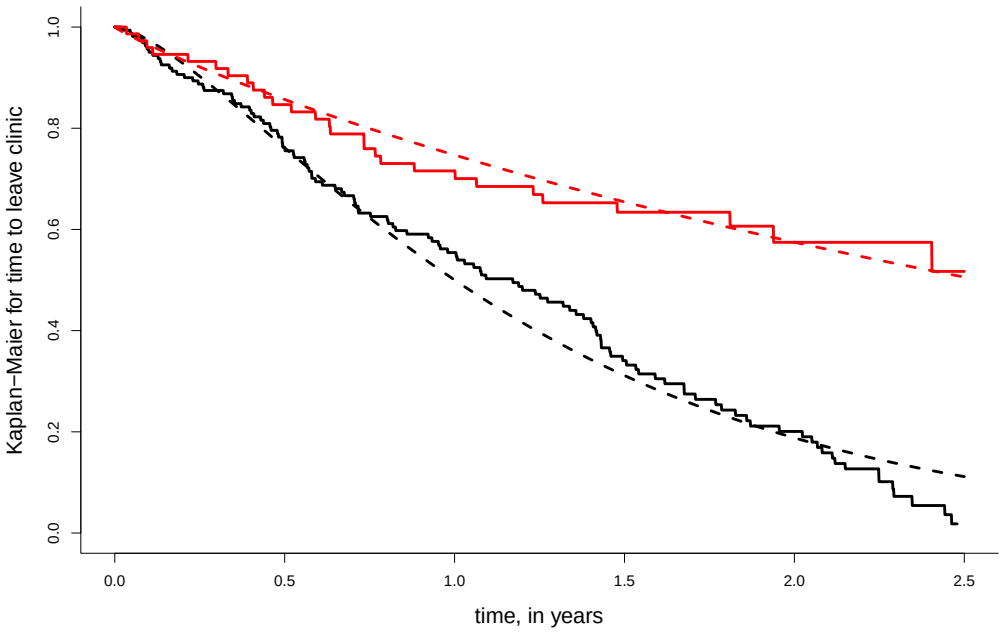

Figure A: Estimated survival curves for Australian drug users in clinics 1 (black) and 2 (red), where survival means not yet having left the clinic. The rugged full curves are Kaplan–Meier estimates and the smooth dashed curves are fitted gamma distributions.

(a) Since part of the motivation for analysing the leaving-the-clinic data is to give weight to claims of their being different, it makes sense to use the same prior for the two A_j; in case one finds differences one can then not be criticised for having started out with a bias. Let therefore both A_1 and A_2 have Beta process priors, with independent increments

$$\mathrm{d}A_j(s) \approx_d \mathrm{Beta}(c(s)\,\mathrm{d}A_0(s), c(s)\,\{1-\mathrm{d}A_0(s)\}).$$

For A_0, take $A_0(t) = \lambda_0 t$, with $\lambda_0 = 0.462$; show that this entails that the prior median is 1.50 years. For $c(s)$, take the constant 3.333. Simulate 25 realisations of $A_j = \{A_j(t)\colon t \in [0, 2.50]\}$, and transform these to 25 realisations of $S_j = \{S_j(t)\colon t \in [0, 2.50]\}$.

(b) The Beta process was introduced in Hjort (1990, Annals of Statistics), and among the central results, covered in class, is that the A_j given the data is still a Beta process. Specifically,

$$\mathrm{d}A_j(s)\,|\,\mathrm{data} \approx_d \mathrm{Beta}(c(s)\,\mathrm{d}A_0(s) + \mathrm{d}N_j(s), c(s)\{1-\mathrm{d}A_0(s)\} + Y_j(s) - \mathrm{d}N_j(s)\},$$

where $Y_j(t)$ is the at-risk process, counting the number of individuals still present at time t, and $N_j(t)$ the counting process, counting individuals observed to make the transition from 'in clinic' to 'leaving the clinic' in $[0, t]$. In particular, $\mathrm{d}N_j(s)$ is counting the number of individuals observed to leave clinic j in the small time interval $[s, s+\mathrm{d}s]$. Simulate say 25 realisations of the A_1 process and equally many from the A_2 process, given data (perhaps in the same diagram, with two different colours, or if you judge it better in two different diagrams).

(c) Give formulae for the Bayes estimators $\widehat{A}_j(t) = \mathrm{E}\,\{A_j(t)\,|\,\mathrm{data}\}$ and also for the conditional variances. Give a plot with the two Bayes estimators, along with pointwise approximate 90% credibility bands.

(d) Let $p_j = \Pr\{T_j \geq 1.50\,|\,T_j \geq 1.00\}$ for the two clinics, the probability that a drug user still in the clinic system after one year will also be there half a year later. Simulate perhaps 1000 realisations of p_1 and p_2, from the appropriate distributions given data, and give point estimates and 90% credibility intervals for p_1, p_2, and the ratio $\rho = p_2/p_1$.

(e) Find an explicit formula for the Bayes estimate $\widehat{\rho} = \mathrm{E}\,(\rho\,|\,\mathrm{data})$, and compute it.

Exercise 2

There are two types of people in the world: those who can extrapolate from incomplete data. This exercise concerns estimating the values of a prior Gaußian process, along with uncertainty assessment, based on the process having been observed in a small number of locations.

(a) Assume $Z = \{Z(x)\colon 0 \leq x \leq 10.00\}$ is a Gaußian process, with

$$\mathrm{E}\,Z(x) = m(x) = a + bx \quad \text{and} \quad \mathrm{cov}\{Z(x), Z(x')\} = \sigma^2 K_0(|x-x'|),$$

where $K_0(u) = \exp(-\lambda|u|^q)$, for suitable $\lambda > 0$ and $q \in (0,2)$. In particular, the variance function is a constant, equal to σ^2. Suppose that this Z process has been observed in locations $x_1 < \cdots < x_n$, say $Z(x_i) = z_{\text{obs}.i}$ for $i = 1, \ldots, n$. Show that with x any value in $[0, 10.00]$,

$$\begin{pmatrix} Z(x) \\ z_{\text{obs}} \end{pmatrix} \sim \mathrm{N}_{n+1}(\begin{pmatrix} m(x) \\ m_{\text{obs}} \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & k(x)^{\mathrm{t}} \\ k(x) & \Sigma \end{pmatrix}),$$

in which m_{obs} is the $n \times 1$ vector of $m(x_i)$, $k(x)$ is the $n \times 1$ vector of $K_0(x - x_i)$, and Σ the $n \times n$ matrix of $K_0(|x_i - x_j|)$.

(b) Use properties of conditional distributions for multinormals to put up clear formulae for $\widehat{Z}(x) = \mathrm{E}\{Z(x) \,|\, \text{data}\}$ and $\widehat{\kappa}(x)^2 = \mathrm{Var}\{Z(x) \,|\, \text{data}\}$.

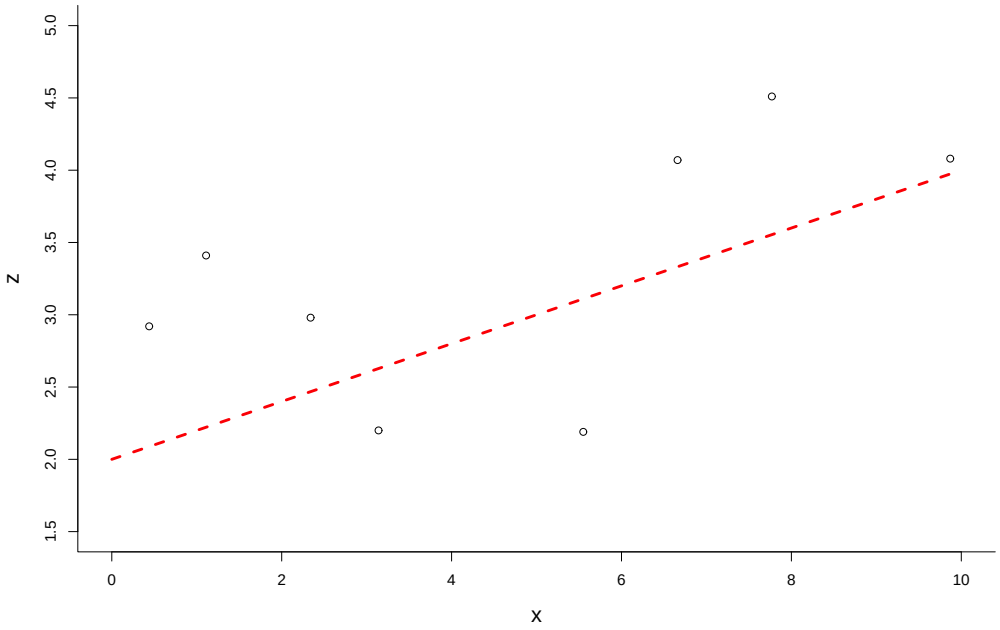

Figure B: The values of $Z(x)$ at the observed positions $x_1, \ldots, x_n$, along with the prior mean function $a + bx$. The task is to estimate the full $Z(x)$ curves, along with assessment of precision.

(c) Suppose now that the prior process parameters are specified as $(a, b) = (2.00, 0.20)$, $\sigma = 0.50$, $\lambda = 0.66$, $q = 1.25$. What is the probability that $Z(x)$ and $Z(x')$ shall be closer to each other than 0.50, if $|x - x'| = 1.00$? Assume that the Z process has been observed in the eight locations x_i, with values of $Z(x_i)$, as follows:

```
xobs:  0.44 1.11 2.34 3.14 5.55 6.66 7.77 9.87
zobs:  2.92 3.41 2.98 2.20 2.19 4.07 4.51 4.08
```

Compute and display the Bayes estimate $\widehat{Z}(x)$, with a 90 percent pointwise credibility band.

(d) Generate and display say 25 realisations, in a suitable figure, from the Z process given the data.

(e) The analysis above relies on the assumption that at positions x_i, one really observes the underlying $Z(x_i)$. In many cases, when applying this type of modelling and methodology for spatial and temporal interpolation and extrapolation, one only observes $Z(x_i)$ with an extra measurement error, say $Z^*(x_i) = Z(x_i) + \varepsilon_i$ for $i = 1, \ldots, n$, where the ε_i are independent and $\mathrm{N}(0, \tau^2)$. Explain how this 'changes the game'. Use the data above, take $\tau = 0.20$, and for this new situation compute and display the Bayes estimate function $\mathrm{E}\{Z(x) \,|\, \text{data}\}$, along with a 90 percent pointwise credibility band.

Exercise 3

The luscious clusters of the vine / Upon my mouth do crush their wine, and in this exercise we exploit the discreteness of the Dirichlet process to create clusters and mixtures. For the questions below you're free to use the following formulae. First, if X is a $\mathrm{Beta}(ap, a(1-p))$, for some positive a and $p \in (0,1)$, then

$$\mathrm{E}\, X = p, \quad \mathrm{E}\, X^2 = p\frac{ap+1}{a+1}, \quad \mathrm{Var}\, X = \frac{p(1-p)}{a+1}.$$

Second, if $(X_1, \ldots, X_k) \sim \mathrm{Dir}(ap_1, \ldots, ap_k)$, with $(p_1, \ldots, p_k)$ a probability vector, then

$$\mathrm{E}\, X_i = p_i, \quad \mathrm{Var}\, X_i = \frac{p_i(1-p_i)}{a+1}, \quad \mathrm{cov}(X_i, X_j) = -\frac{p_i p_j}{a+1}$$

for $i \neq j$.

(a) Let G_0 be any continuous distribution on the real line, like the standard normal, and let $G \sim \mathrm{Dir}(aG_0)$, a Dirichlet process with basis measure aG_0, for some fixed and positive a. Show that

$$\mathrm{E}\, G(A) = G_0(A), \quad \mathrm{E}\, G(A)G(B) = \frac{a}{a+1}G_0(A)G_0(B) + \frac{1}{a+1}G_0(A \cap B).$$

(b) Consider first a single θ drawn from this random distribution G, i.e. $\Pr\{\theta \in A \,|\, G\} = G(A)$ for each A. Show that $\Pr\{\theta \in A\} = G_0(A)$, so the random θ from the random G has distribution equal to that of the prior mean, i.e. G_0.

(c) Consider next the case of θ_1, θ_2 drawn i.i.d. from the random G, i.e.

$$\Pr\{\theta_1 \in A_1, \theta_2 \in A_2 \,|\, G\} = G(A_1)G(A_2) \quad \text{for all } A_1, A_2.$$

Show that the marginal distribution of these two may be expressed as

$$Q_2 = \frac{a}{a+1}G_0 \times G_0 + \frac{1}{a+1}G_{0,12},$$

where $(G_0 \times G_0)(A_1 \times A_2) = G_0(A_1)G_0(A_2)$ corresponds to two independent draws from G_0 and $G_{0,12}(A_1 \times A_2) = G_0(A_1 \cap A_2)$ corresponds to enforcing $\theta_1 = \theta_2$, a single value, drawn from G_0. In other words, with probability $a/(a+1)$ the two are different and independent, both from G_0, and with probability $1/(a+1)$ they are equal, a single value drawn from G_0.

(d) One may similarly show, with further efforts, that the marginal distribution of a triple $(\theta_1, \theta_2, \theta_3)$ drawn i.i.d. from the random G, has a marginal distribution which can be expressed as

$$\begin{aligned} Q_3 = &\frac{a^2}{(a+1)(a+2)} G_0 \times G_0 \times G_0 \\ &+ \frac{a}{(a+1)(a+2)} (G_{0,12} \times G_{0,3} + G_{0,13} \times G_{0,2} + G_{0,23} \times G_{0,1}) \\ &+ \frac{2}{(a+1)(a+2)} G_{0,123}, \end{aligned}$$

necessitating a bit of work to get the notation right etc. Here

$$G_{0,13} \times G_{0,2}(A_1 \times A_2 \times A_3) = G_0(A_1 \cap A_3) G_0(A_2),$$

which means that $\theta_1 = \theta_3$ is from G_0, independent of θ_2 from G_0, etc. With D_3 the number of distinct values of $\theta_1, \theta_2, \theta_3$, write down the probabilities of having $D_3 = 3$, $D_3 = 2$, $D_3 = 1$. Comment briefly on the cases a very large and a very small.

(e) Assume that $\theta_1, \ldots, \theta_i$ have been drawn from the random G. What is the posterior distribution of G, given these first i draws? And what is the distribution of the next θ_{i+1}, given $\theta_1, \ldots, \theta_i$? Find the probability that θ_{i+1} is a new value, i.e. not in $\{\theta_1, \ldots, \theta_i\}$.

(f) With D_n the number of distinct values among $\theta_1, \ldots, \theta_n$, from the random $G \sim \text{Dir}(aG_0)$, one may demonstrate, via arguments used above, that $D_n = R_1 + \cdots + R_n$, where the R_i are independent 0-1 variables, and with $\Pr\{R_i = 1\} = a/(a+i-1)$. Show that $D_n / \log n$ converges in probability to a.

– And now, finally, for the cluster modelling. The general idea is to take

(i) $G \sim \text{Dir}(aG_0)$;

(ii) parameters $\theta_1, \ldots, \theta_n$ as i.i.d. from that random G, and these underlying parameters are not visible to the statistician;

(iii) observed data points $y_1, \ldots, y_n$ conditional on $\theta_1, \ldots, \theta_n$ as independent from densities $f(y_1 \,|\, \theta_1), \ldots, f(y_n \,|\, \theta_n)$.

Thus there are approximately $a \log n$ different parameter values. By proper simulation from $(\theta_1, \ldots, \theta_n)$ given data $(y_1, \ldots, y_n)$ we hope to learn both about the number of clusters and where these are in the parameter space.

– The setup and methodology can be worked with using different techniques, and can also be extended in several directions. Here we shall however be content to illustrate the general principles in a kindergarten setup with only $n = 3$ data points, and with simple models for both G_0 and $f(y_i \,|\, x_i)$. Indeed we take $G_0 = \text{N}(0,1)$ and $y_i \,|\, x_i \sim \text{N}(\theta_i, 1)$. Write $g_0(y) = \phi(y)$ for the $\text{N}(0,1)$ density and $\phi(y_i - \theta_i)$ for the $\text{N}(\theta_i, 1)$ density of $y_i \,|\, \theta_i$. For the following you may use the fact that

$$g_0(\theta) \prod_{i=1}^m \phi(y_i - \theta) = g_m(\theta \,|\, y_1, \ldots, y_m) \bar{f}_m(y_1, \ldots, y_m),$$

where

$$g_m(\theta \,|\, y_1,\ldots,y_m) = \frac{1}{(2\pi)^{1/2}}(m+1)^{1/2}\exp\Big\{-\tfrac12(m+1)\Big(\theta-\frac{m}{m+1}\bar y_m\Big)^2\Big\},$$
$$\bar f_m(y_1,\ldots,y_m) = \frac{1}{(2\pi)^{m/2}}\frac{1}{(m+1)^{1/2}}\exp\Big(-\tfrac12 z_m - \tfrac12\frac{m}{m+1}\bar y_m^2\Big),$$

in terms of $\bar y_m = (1/m)\sum_{i=1}^m y_i$ and $z_m = \sum_{i=1}^m (y_i-\bar y_m)^2$. In particular,

$$g_0(\theta)\phi(y-\theta) = \mathrm{N}\Big(\frac12 y,\frac12\Big)(\theta)\frac{1}{(2\pi)^{1/2}}\frac{1}{2^{1/2}}\exp\Big(-\frac12\frac12 y^2\Big),$$
$$g_0(\theta)\prod_{i=1}^2\phi(y_i-\theta) = \mathrm{N}\Big(\frac23\bar y_2,\frac13\Big)(\theta)\frac{1}{(2\pi)^{2/2}}\frac{1}{3^{1/2}}\exp\Big(-\frac12 z_2 - \frac12\frac23\bar y_2^2\Big),$$
$$g_0(\theta)\prod_{i=1}^3\phi(y_i-\theta) = \mathrm{N}\Big(\frac34\bar y_3,\frac14\Big)(\theta)\frac{1}{(2\pi)^{3/2}}\frac{1}{4^{1/2}}\exp\Big(-\frac12 z_3 - \frac12\frac34\bar y_3^2\Big).$$

You don't need to show these particular identities here, for the exam; they are essentially algebraic and I'm just being helpful having done it for you. The g_m part is the usual posterior given the data, but on this occasion we need the marginal densities $\bar f_m$ too.

(g) Show from the above efforts that the joint density of parameters and data points, say $\pi(\theta,y) = \pi(\theta_1,\theta_2,\theta_3,y_1,y_2,y_3)$, can be expressed as follows:

. with probability $a^2/\{(a+1)(a+2)\}$, the three θ_i are different, and

$$\begin{aligned}\pi(\theta,y) &= g_0(\theta_1)g_0(\theta_2)g_0(\theta_3)\phi(y_1-\theta_1)\phi(y_2-\theta_2)\phi(y_3-\theta_3)\\ &= g_1(\theta_1\,|\,y_1)g_1(\theta_2\,|\,y_2)g_1(\theta_3\,|\,y_3)\bar f_1(y_1)\bar f_1(y_2)\bar f_1(y_3);\end{aligned}$$

. with probability $a/\{(a+1)(a+2)\}$, nos. 1 and 2 are equal, say θ_{12}, different from θ_3, and

$$\begin{aligned}\pi(\theta,y) &= g_0(\theta_{12})\phi(y_1-\theta_{12})\phi(y_2-\theta_{12})g_0(\theta_3)\phi(y_3-\theta_3)\\ &= g_2(\theta_{12}\,|\,y_1,y_2)g_1(\theta_3\,|\,y_3)\bar f_2(y_1,y_2)\bar f_1(y_3);\end{aligned}$$

. with probability $a/\{(a+1)(a+2)\}$, nos. 1 and 3 are equal, say θ_{13}, different from θ_2, and

$$\begin{aligned}\pi(\theta,y) &= g_0(\theta_{13})\phi(y_1-\theta_{13})\phi(y_3-\theta_{12})g_0(\theta_2)\phi(y_2-\theta_2)\\ &= g_2(\theta_{13}\,|\,y_1,y_3)g_1(\theta_2\,|\,y_2)\bar f_2(y_1,y_3)\bar f_1(y_2);\end{aligned}$$

. with probability $a/\{(a+1)(a+2)\}$, nos. 2 and 3 are equal, say θ_{23}, different from θ_1, and

$$\begin{aligned}\pi(\theta,y) &= g_0(\theta_{23})\phi(y_2-\theta_{23})\phi(y_3-\theta_{23})g_0(\theta_1)\phi(y_1-\theta_1)\\ &= g_2(\theta_{23}\,|\,y_2,y_3)g_1(\theta_1\,|\,y_1)\bar f_2(y_2,y_3)\bar f_1(y_1);\end{aligned}$$

. with probability $2/\{(a+1)(a+2)\}$, the three θ_i are equal, say θ_{123}, and

$$\begin{aligned}\pi(\theta,y) &= g_0(\theta_{123})\phi(y_1-\theta_{123})\phi(y_2-\theta_{123})\phi(y_3-\theta_{123})\\ &= g_3(\theta_{123}\,|\,y_1,y_2,y_3)\bar f_3(y_1,y_2,y_3).\end{aligned}$$

(h) To illustrate the above calculations and results, take $a = 2.00$. First, what are the prior chances of having 1, 2, 3 clusters (i.e. the number of distinct values among $\theta_1, \theta_2, \theta_3$)? Now study two cases of data outcomes for y_1, y_2, y_3. Case A has $1.50, 1.55, 1.60$; Case B has $-2.75, 0.03, 2.78$. For these cases, compute the posterior probabilities of there being 1, 2, 3 clusters.

(i) Again, use $a = 2.00$. Describe the posterior densities for the three mean parameters, say $\pi(\theta_i \,|\, \text{data})$ for $i = 1, 2, 3$. Display these, in one diagram for Case A and another diagram for Case B. You may use simulation if you find that more convenient.

(j) Regardless of how successful you might have been for tackling the questions of (g), (h), (i), write up a paragraph or two regarding what you see as the potential fruitfulness of the general approach taking in this exam Exercise 3.

Exercise 4

I Egypten bodde egypterne / for lenge, lenge siden. Here we shall revisit our distantly admired $n = 141$ acquaintances, from Roman era Egypt, about a century b.C. Lifelengths for 82 men and 59 women from that population have been recorded and are available at the course website, as `egypt_data`. Earlier in the course we've analysed the two distributions nonparametrically, via Beta processes, but presently we shall use our Bayesian Nonparametrics toolkit in the modus of 'from Processes to Models'. In the following, let $H(x, b_1, b_2)$ be the cumulative distribution function for a Beta(b_1, b_2); it is available in `R` as `pbeta(x,b1,b2)`.

(a) Let $G \sim \text{Dir}(aG_0)$, a Dirichlet process on $[0, 100]$ with an appropriate prior mean cumulative distribution function G_0 and a positive strength parameter a. Consider the survival model where death occurs as soon as the random $G(t)$ crosses threshold c. Show that the consequent survival curve can be written

$$S(t) = \Pr\{T \geq t\} = H(c, aG_0(t), a\bar{G}_0(t)) \quad \text{for } t \geq 0,$$

where $\bar{G}_0(t) = 1 - G_0(t)$.

(b) We now fit such Dirichlet process threshold crossing models, one for the men and one for the women. Here we take $G_0(t) = t/100$ on $[0, 100]$, i.e. the simple uniform on that interval. Explain how you can compute and programme log-likelihood functions, say $\ell_m(a_m, c_m)$ and $\ell_w(a_w, c_w)$ for the men and for the women, via numerical derivation.

(c) You are not required to do such programming and numerical optimisation here, but I have carried out such efforts and find that the maximum likelihood estimates are $(1.629, 0.216)$ for the men and $(3.681, 0.184)$ for the women. With this information, reconstruct a version of Figure C. Make also a figure featuring the estimated hazard rates, say $\widehat{\alpha}_m(t)$ and $\widehat{\alpha}_w(t)$, with $\alpha(t) = f(t)/S(t)$ in terms of density and survival.

(d) Construct a figure, or if you prefer two, showing ten realisations of the Dirichlet process G_m for men, and ten for the G_w for women, and where you also draw horizontal lines for the thresholds $\widehat{c}_m$ and $\widehat{c}_w$.

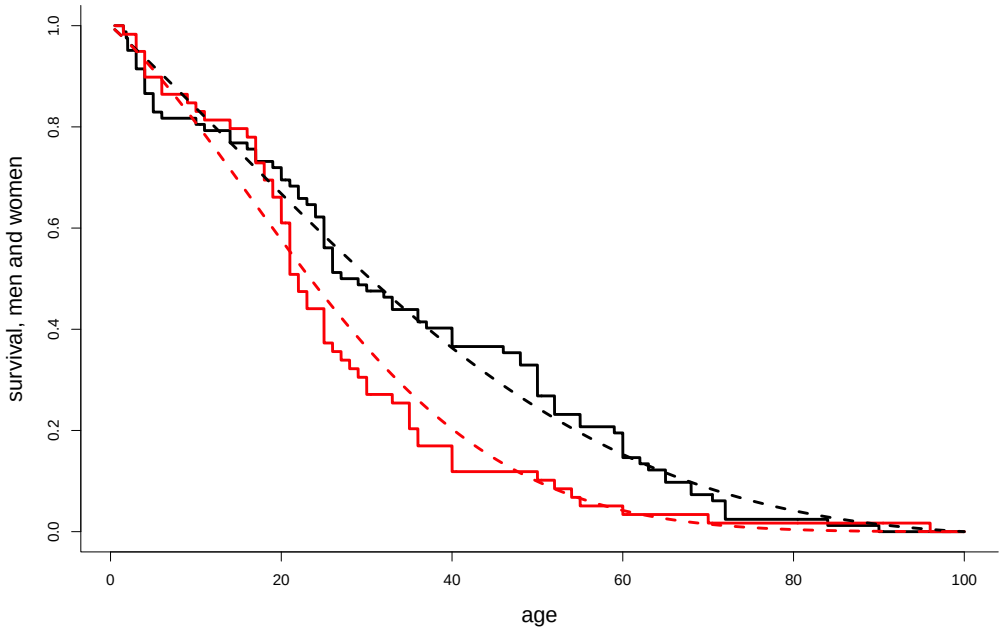

Figure C: Estimated survival curves for the men (black curves) and women (red curves) of Roman era ancient Egypt; the ragged full curves are Kaplan–Meier curves whereas the smoothed dashed curves are from the Dirichlet process threshold crossing model.

(e) The two-parameter Dirichlet threshold crossing models fit the data decently well, with G_0 taken as the uniform, as seen in Figure C. There might be better versions, however, in particular to reflect what has been discussed in the course via Beta process models, that young women might have had a harder time surviving than young men (for this period of relative peace in Roman era Egypt). Indicate how better models could perhaps be constructed, and try them out, if you have time.