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:	**1.–12.vi.2018**

This is the exam project set for STK 4190/9190, spring semester 2018. It is made available on the course website as of *Friday 1 June 10:01*, and candidates must submit their written reports by *Tuesday 12 June 11:59* (or earlier), to the reception office at the Department of Mathematics, *in duplicate*. The supplementary oral examination part takes place *Friday June 15* (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, Word), but may also be hand-processed. Give your name (and your student-web identification number) on the first page; the markers need to couple project with the oral examination. 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, 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 eight pages.

Exercise 1

The only limits we have are the limits we believe (perhaps). This exercise looks into a certain construction of the Dirichlet process as a limit of simpler probability measures. Consider a sample space $\mathcal{X}$, e.g. the real line, with a fixed probability measure P_0, like the standard normal, and a positive strength parameter a. Let now

$$P_m = \sum_{i=1}^{m} D_i \delta(\xi_i), \tag{1}$$

where $\xi_1, \xi_2, \ldots$ are i.i.d. from P_0, and where $(D_1, \ldots, D_m)$, independently, is a Dirichlet distributed vector with symmetric parameters $(a/m, \ldots, a/m)$. Also, $\delta(\xi_i)$ means unit

point-mass at position ξ_i, so that, in particular,

$$P_m(A) = \sum_{i=1}^m D_i I(\xi_i \in A) = \sum_{i:\xi_i \in A} D_i, \tag{2}$$

which is a random sum of random probabilities.

Below you may find use for some or more of the following Dirichlet distribution formulae (and which you do not need to prove here). If $(D_1, \ldots, D_m) \sim \text{Dir}(ap_1, \ldots, ap_m)$, with the p_i positive and summing to one, then

$$\text{E}\, D_i = p_i, \quad \text{E}\, D_i^2 = \frac{ap_i(ap_i+1)}{a(a+1)}, \quad \text{Var}\, D_i = \frac{p_i(1-p_i)}{a+1},$$

for each i, and when $i \neq j$,

$$\text{E}\, D_j D_j = \frac{ap_i\, ap_j}{a(a+1)}, \quad \text{cov}(D_i, D_j) = -\frac{p_i p_j}{a+1}.$$

Also, the 'summing over cells' property is known, that if $(D_1, \ldots, D_m) \sim \text{Dir}(ap_1, \ldots, ap_m)$, then a shorter vector formed by summing over cells is another Dirichlet vector, with parameters obtained by summing over the appropriate cells. In particular, with I a set of indexes, $\sum_{i\in I} D_i \sim \text{Beta}(aq, a(1-q))$, where $q = \sum_{i\in I} p_i$, &c.

(a) For $P_m(A)$, obtained in (2), show that

$$P_m(A) \,|\, (\xi_1, \ldots, \xi_m) \sim \text{Beta}\big(a\widehat{P}_m(A), a\{1-\widehat{P}_m(A)\}\big),$$

where $\widehat{P}_m(A) = (1/m)\sum_{i=1}^m I(\xi_i \in A)$ is the empirical proportion of points in A. Show from this that with probability 1,

$$P_m(A) \to_d \text{Beta}(aP_0(A), a\{1-P_0(A)\}).$$

(b) Show more generally that if $A_1, \ldots, A_k$ is some partition of the sample space $\mathcal{X}$, then $(P_m(A_1), \ldots, P_m(A_k))$ converges in distribution, with probability 1, to the distribution of $(P(A_1), \ldots, P(A_k))$, where $P \sim \text{Dir}(aP_0)$. – With a few extra arguments, depending on the level of mathematical precision, one may demonstrate that P_m of (1) converges in distribution, with probability 1, to the $P \sim \text{Dir}(aP_0)$, in the space of all distributions on $\mathcal{X}$, with an appropriate topology of convergence.

(c) With $P \sim \text{Dir}(aP_0)$, consider the random Dirichlet process mean

$$\theta = \theta(P) = \int x\, \text{d}P(x).$$

With $\theta_m = \theta(P_m)$, and with notation as at the start of this exercise, show that

$$\theta_m = \sum_{i=1}^m D_i \xi_i.$$

Assuming P_0 having finite mean $\theta_0 = \int x\, \text{d}P_0(x)$ and variance $\sigma_0^2 = \int (x-\theta_0)^2\, \text{d}P_0(x)$, show that $\text{E}\,\theta_m = \theta_0$, and explain why this implies $\text{E}\,\theta = \theta_0$.

(d) Then demonstrate that

$$\operatorname{Var}\theta_m \to \frac{\sigma_0^2}{a+1},$$

and explain why this is also a formula for $\operatorname{Var}\theta$. (These formulae have also been demonstrated during Hjort's lectures, but then using different arguments.)

(e) The distribution of $\theta = \int x\,\mathrm{d}P(x)$ is typically quite complicated. Consider the case of $P_0 \sim \mathrm{N}(0,1)$. Show that

$$\theta_m \,|\, (D_1,\ldots,D_m) \sim \mathrm{N}(0, V_m),$$

where $V_m = \sum_{i=1}^m D_i^2$. One may show that V_m converges in distribution to a certain V, with an appropriate (but complicated) density $h(v)$. Find the mean of V. Use this to show that the distribution of θ must be a normal mixture, with density

$$f(\theta) = \int_0^\infty \phi\Big(\frac{\theta}{v}\Big)\frac{1}{v}h(v)\,\mathrm{d}v.$$

As usual, ϕ is the standard normal density.

(f) Use finally the above to simulate 1000 independent realisations of θ, say with $a = 3.14159$, and display a histogram of these. Comment briefly on other possible simulation schemes.

Exercise 2

Incidentally, did you know that using non-linear regression is currently out of line? Consider the data portrayed in Figure 1, consisting of the for a segment of the Oslo population drastically important number of skiing days per year, at the location Bjørnholt, an hour's trasking upwards from Frognerseteren along Historias Kraftlinjer. The dataset, with such skiing days numbers from 1897 to 2012, but with a hole in the series from 1938 to 1954, is available at the course website (and a skiing day is defined as there being 25 cm or more snow on the ground).

Though various generalisations are relevant and doable, we shall start out considering the data in the following somewhat simple fashion. At year x_i, we observe the number of skiing days

$$Z(x_i) = m(x_i) + \varepsilon_i \quad \text{for } i = 1,\ldots,n, \tag{3}$$

for the appropriate years $x_1,\ldots,x_n$, and with $n = 99$. Here $m(x)$ is seen as the underlying trend function, the 'signal', and with the ε_i as 'noise', the random variations around the trend function, and here, for simplicity, taken as i.i.d. from the $\mathrm{N}(0,\sigma^2)$.

This becomes a Bayesian nonparametrics model when a prior process is used for the $m(x)$ function. Here we take it as a Gaussian, with mean and covariance function of the form

$$\mathrm{E}\,m(x) = m_0(x) \quad \text{and} \quad \operatorname{cov}\{m(x), m(x')\} = \sigma_0^2 K_0(x-x'),$$

for a covariance function K_0, with $K_0(0) = 1$. In other words, the variance of $m(x)$ is σ_0^2, and the correlation between two points of $m(x)$ being a distance d from each other is $K_0(d)$.

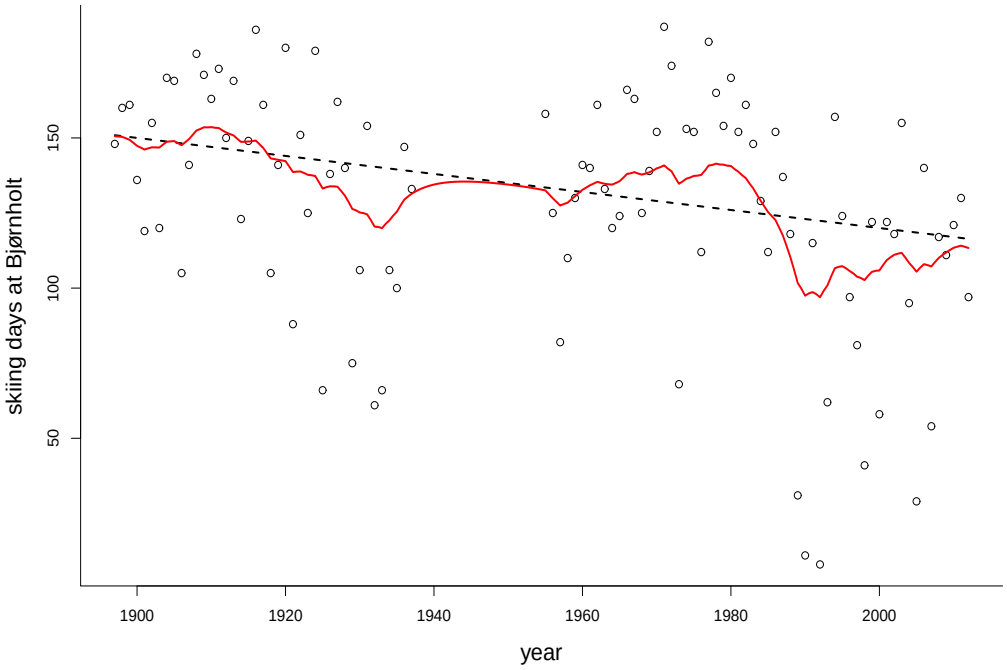

Figure 1: The number of skiing days at Bjørnholt per year, from 1897 to 2012 but with missing data from 1938 to 1954. The red curve is a Bayesian nonparametrics posterior mean, with a certain Gaussian prior process. See Cunen, Hermansen, Hjort (JSPI, 2018) or Schweder and Hjort's CLP book (2016, Cambridge) for more information and analysis.

(a) Consider first the observed vector $Z_{\rm obs} = (Z(x_1), \ldots, Z(x_n))^{\rm t}$. Show that its marginal distribution is

$$Z_{\rm obs} \sim {\rm N}_n(m_{0,{\rm obs}}, \sigma_0^2 \Sigma_0 + \sigma^2 I_n),$$

with $m_{0,{\rm obs}}$ the vector of $m_0(x_i)$, where I_n is the identity matrix of size $n \times n$, and Σ_0 is the matrix consisting of all $K_0(x_i - x_j)$.

(b) In extension of this show that for a given position x_0,

$$\begin{pmatrix} m(x_0) \\ Z_{\rm obs} \end{pmatrix} \sim {\rm N}_{n+1}\Big(\begin{pmatrix} m_0(x_0) \\ m_{0,{\rm obs}} \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_0^2 k(x_0)^{\rm t} \\ \sigma_0^2 k(x_0) & \sigma_0^2 \Sigma_0 + \sigma^2 I_n \end{pmatrix} \Big).$$

Here $k(x_0)$ is the column vector of $K_0(x_0 - x_i)$.

(c) Deduce that $m(x_0)$, given the data, has a mean value which can be expressed as

$$\widehat{m}(x_0) = m_0(x_0) + \sigma_0^2 k(x_0)^{\rm t} (\sigma_0^2 \Sigma_0 + \sigma^2 I_n)^{-1} (Z_{\rm obs} - m_{0,{\rm obs}}).$$

Find also a formula for the conditional variance of m at position x_0.

(d) Now implement such a concrete scheme, for the Bjørnholt data. For the prior, use $m_0(x) = 150 - 0.20\,x$, on the scale of $x = \text{year} - 1900$, and furthermore let $\sigma_0 = 12.50$, $K_0(d) = \rho_0^d$ with $\rho_0 = 0.75$. For the data given the trend function $m(x)$, take $\sigma = 33.33$. Compute the Bayes estimate curve, and make a figure similar to Figure 1 (it will not be entirely equal to this figure, though, as I used somewhat different prior parameters there).

(e) In addition to the posterior mean, compute also the posterior standard deviation curve, and display a pointwise 90% posterior credibility band.

(f) Simulate and display say 25 curves from $m(x)$ given the data.

(g) The above setup has used the simplifying assumption in (3) that the ε_i are i.i.d. Explain how the apparatus and results are changed if we model also the $Z(x)$ given the trend function $m(x)$ as an autocorrelated time series, with $\text{cov}(\varepsilon_i, \varepsilon_j) = \sigma^2 \rho^{|x_i - x_j|}$, say with $\rho = 0.333$.

(h) There are many natural generalisations of the setup above. Briefly consider one of these, namely the step from taking a known σ (above taken as 33.33) to an unknown parameter, with say a Gamma distribution prior for $1/\sigma^2$. Explain how one may now compute the posterior mean and posterior standard deviation for $m(x)$.

Exercise 3

"The mystic chords of memory will swell when again touched, as surely they will be, by the better angels of our nature." Well, let's see. Figure 2 below tells a dramatic and gruesome story for mankind. For the 95 inter-state wars, from 1823 to the present, where the number of battle deaths has been 1000 or more, the figure displays the log of these sad numbers, i.e. $(x_i, \log z_i)$ for $i = 1, \ldots, 95$, where (x_i, z_i) denote the onset time and number of battle deaths for war i. The lower limit in the figure is hence $\log 1000 = 6.908$. The red points and black points correspond to before and after time point 1950.83, which corresponds to the change-point found in Cunen, Hjort, Nygård's 'Statistical Sightings of Better Angels' article (May 2018).

The dataset of (x_i, z_i) is available at the course website. For the purposes of this exam exercise it is convenient to read the data into your computer as follows. It involves a little cosmetic trick of setting the nine numbers recorded as '1000' in the pre-Korea list (though these are clearly meant as rough approximations) to $1002, 1003, \ldots, 1010$, in order to avoid certain artificial numerical issues with some of the estimation procedures. The $z_i = 1001$ for the Falklands war is however taken as accurate and kept as it is.

```
krig <- matrix(scan("krigogfred-data",skip=4),byrow=T,ncol=2)
xx <- krig[ ,1]
zz <- krig[ ,2]
x0 <- 1950.825 # Korea
xxL <- xx[xx <= x0]
xxR <- xx[xx > x0]
```

```
zzL <- zz[xx <= x0]
zzR <- zz[xx > x0]
nnL <- length(zzL) # 60
nnR <- length(zzR) # 35
# pushing nine values of "1000" for zz left to 1002, ..., 1010
where <- (1:nnL)[zzL==1000]
zzL[where] <- 1002:1010
yyL <- log(zzL/1000)
yyR <- log(zzR/1000)
```

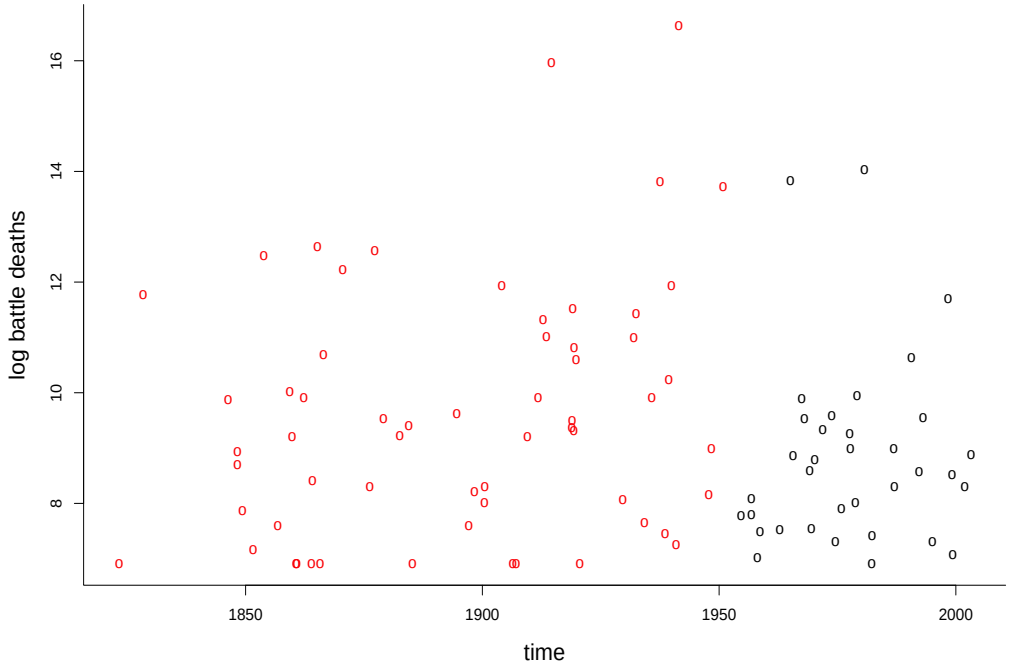

Figure 2: The log number of battle deaths, for all 95 major wars from 1823 to 2003 where the battle death number is above 1000. Red and black circles correspond to before and after the Korean War, cf. the breakpoint found in Cunen, Hjort, Nygård's 'Statistical Sightings of Better Angels' (2018).

Your task, in this exercise, involves analysing the distribution of $y_i = \log(z_i/1000)$, with $n_L = 60$ to the left and $n_R = 35$ to the right of the potential regime-shift, that of the Korean War. It is easier to model and carry out analysis on this log-scale, and then revert back to the full-drama scale of $z_i = 1000 \exp(y_i)$ afterwars.

(a) For the y_i distributions F_L and F_R to the left and right of Korea, give them Dirichlet process priors, with parameters $a_L F_{0,L}$ and $a_R F_{0,R}$. Since we wish to let data speak for themselves, regarding any potential differences between before and after 1950, use the same $F_0(y) = 1 - \exp(-\theta_0 y)$ for both prior mean functions, and the same $a_L = a_R$. For concreteness, take $a_L = a_R = 3.33$ and $\theta_0 = 0.50$. Simulate a few F_L and F_R from these priors, and discuss briefly how 'reasonable' they seem to be.

(b) First compute and display the empirical distribution functions,

$$\widehat{F}_L(y) = (1/n_L)\sum_{i=1}^{n_L} I(y_{L,i} \le y) \quad \text{and} \quad \widehat{F}_R(y) = (1/n_T)\sum_{i=1}^{n_T} I(y_{R,i} \le y).$$

Comment on whether and to which degree they look different.

(c) Then do the Bayesian updating, and explain the posterior distributions for F_L and F_R. Simulate say 50 curves for each, displayed to the left and to the right in a figure. Also compute and display the posterior means, along with a 90% pointwise quantile band, where you for the latter ought to use more than 1000 simulations for good precision.

(d) Simulate say 1000 (or more) realisations of F_L and of F_R, reading off their medians

$$\text{med}_L = \inf\{y: F_L(y) \ge \tfrac{1}{2}\} \quad \text{and} \quad \text{med}_R = \inf\{y: F_R(y) \ge \tfrac{1}{2}\}.$$

Show histograms of $\text{med}_L - \text{med}_R$ and of $\text{med}_L/\text{med}_R$, and comment on what you learn from this.

(e) We all hope the wars will go down, in both frequency and volume. But assume, for the sake of statistical imagination and communication, that there will be (God forbid) future inter-state wars, following the same distribution as implied by the F_R distribution. Simulate say 1000 y, from such future wars, and give the 0.10, 0.50, 0.90 quantiles of the appropriate predictive battle deaths distribution.

Exercise 4

Countings one's blessings involves recording these in repeated rounds, presumably, and where the underlying intensity measures may change over time. Here we look at a certain framework for modelling and analysing count data.

(a) Suppose that $y_1, \ldots, y_m$ are independent Poisson counts, recorded over m different and perhaps small time intervals. Assume next that the underlying parameters, say $\theta_1, \ldots, \theta_m$, are independent Gamma variables, with parameters (a_i, b_i), i.e. densities proportional to $\theta_i^{a_i-1}\exp(-b_i\theta_i)$. Show that

$$\theta_i \,|\, \text{data} \sim \text{Gam}(a_i + y_i, b_i + 1),$$

and that these are independent.

(b) A time-continuous version of the above is as follows, via fine limits. First, let $Z_m(t) = \sum_{i/m \le t} \theta_{m,i}$, where the individual and independent intensity components are

$$\theta_{m,i} \sim \text{Gam}(a(i/m)(1/m), b(i/m)) \quad \text{for } i = 1, 2, 3, \ldots.$$

Here $a(s)$ and $b(s)$ are given smooth positive functions. Show that the Z_m process has a clear distribution limit Z, as m grows. Show also that

$$\text{E}\, Z(t) = \int_0^t \frac{a(s)}{b(s)}\,\text{d}s \quad \text{and} \quad \text{Var}\, Z(t) = \int_0^t \frac{a(s)}{b(s)^2}\,\text{d}s.$$

Discuss briefly the special case of $b(s)$ being equal to a constant b.

(c) In addition to the Gamma increments $\theta_{m,i}$, consider $Y_m(t) = \sum_{i/m \le t} y_{m,i}$, with the $y_{m,i}$ being independent and Poisson, with the $\theta_{m,i}$ as parameters. Show that the Y_m process has a clear distributional limit Y, which for the given parameters is Poisson with cumulative intensity function the $Z(t)$ above.

(d) We have now defined time-continuous processes, with $Z(\cdot)$ as cumulative extended Gamma process and $Y(\cdot)$ given $Z(\cdot)$ a Poisson process. Show that $Z(\cdot)$ given the observed counts $Y(\cdot)$ is another cumulative extended Gamma process, with cumulative intensity function

$$M(t) = \mathrm{E}\{Z(t) \,|\, \text{data}\} = \int_0^t \frac{a(s)\,\mathrm{d}s + \mathrm{d}Z(s)}{b(s)+1}.$$

Find also an expression for the conditional variance of $Z(t)$, given data.

(e) Suggest further variations and extensions of this initial Bayesian nonparametrics setup with Gamma increments and counts, and indicate something which could be or become a statistical application.