Exercises and Lecture Notes, STK 9190, Spring 2018

Version 0.55, 12-iii-2018

Nils Lid Hjort

Department of Mathematics, University of Oslo

Abstract

Exercises and Lecture Notes collected here are indeed for the Bayesian Nonparametrics course STK 9190, given for the first time in the spring semester 2018. It is still useful to go through some prototype lower-dimensional Bayesian work, however, so a few exercises of that type are also included. This relates to clarifying concepts and principles, and also to Bayesian Nonparametrics constructions that use lower-dimensional pieces – as the famous interlocking versatile Lego bricks pieces.

1. Prior to posterior updating with Poisson data

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

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

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

Here a and b are positive parameters. Show that

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

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

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

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

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

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

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

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

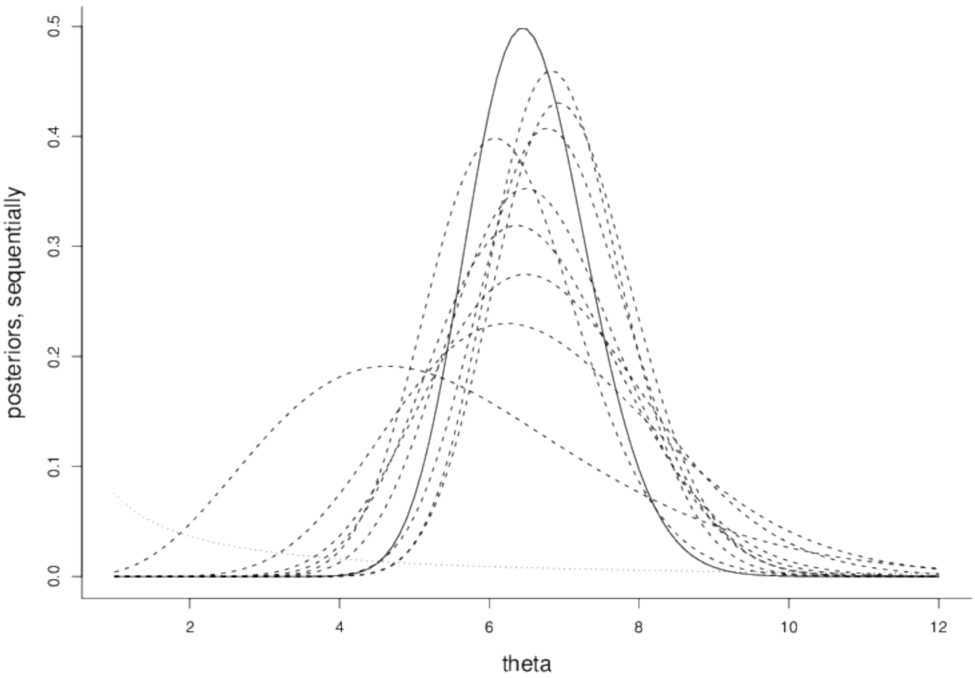

Figure 0.1: Eleven curves are displayed, corresponding to the Gamma$(0.1, 0.1)$ intial prior density for the Poisson parameter θ along with the ten updates following each of the observations 6, 8, 7, 6, 7, 4, 11, 8, 6, 3.

2. The Master Recipe for finding the Bayes solution

I decide to copy in this particular exercise from the lower-dimensional lower-ambition Bayes course, without changing the terms or the notation. The meta-exercise, however, is to understand that all of this still applies in the higher-level world of Bayesian Nonparametrics, partly at the price of the required higher-level mathematical abstraction level. Basically, where one for Bayesian Parametrics writes model likelihoods in terms of the famous generic θ, below, one needs for Bayesian Nonparametrics to think and write and work in terms of a very-high-dimensional or even infinite-dimensional parameter vector. This could be an unknown cumulative distribution function F, an unknown median regression function $m(x)$, an intensity function $\lambda(t)$, etc., rather than the prototypical θ. Often enough there are no clear-and-simple likelihood functions coming out of such constructions, however, as we shall see during the course. This does not stop us from trying to crunch our way from priors to posteriors.

Crucially and amazingly, the basic concepts of decision functions, prior and posterior, loss functions and risk functions, and the optimal Bayesian strategy, carry over. As long as the statistician has data y, a model in terms of some distribution P (i.e. rather than the ubiquitous θ), a clear (nonparametric) prior for this P, and a loss function $L(P,a)$ encountered for decision a if the truth is P – then there will be (a) a posterior $\pi(P\,|\,\mathrm{data})$; (b) a clear strategy for reaching the Bayes solution $\widehat{a}_B$; and (c) this strategy is unbeatable, the sole gold medal winner, in the Olympic competition against other strategies.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

(I follow a certain semi-classical convention here, regarding using the 'p' multipurposedly, and with each 'p' to be understood by the reader from the context.)

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

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

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

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

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

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

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

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

3. The Dirichlet-multinomial model

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

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

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

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

(b) Show that

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

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

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

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

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

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

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

(d) Infer from this that

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4. Gott würfelt nicht

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

(a) Use either of the priors (i) 'flat', $\mathrm{Dir}(1, 1, 1, 1, 1, 1)$; (ii) 'symmetric but more confident', $\mathrm{Dir}(3, 3, 3, 3, 3, 3)$; (iii) 'unwilling to guess', $\mathrm{Dir}(0.1, 0.1, 0.1, 0.1, 0.1, 0.1)$

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

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

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

(c) [xx something re the same ideas being generalisable to the fuller Dirichlet process case. xx]

5. The Dirichlet Process: definition, existence, constructions

Let $\mathcal{X}$ be some sample space, like the real line, with subsets A belonging to an appropriate sigma-algebra $\mathcal{A}$. Let P_0 be a fixed probability distribution on $\mathcal{X}$, and a a positive scalar. We say that P is a Dirichlet process on $\mathcal{X}$, with parameter aP_0, and write $P \sim \mathrm{Dir}(aP_0)$ to indicate this, if it is the case for each partition $(A_1, \ldots, A_m)$, we have

$$(p_1, \ldots, p_m) = \big(P(A_1), \ldots, P(A_m)\big) \sim \mathrm{Dir}(aP_0(A_1), \ldots, aP_0(A_m)).$$

This is required for any number m of elements in the partition.

(a) Show that the basic 'logic coherence' property is satisfied, that we may put some of the A_j together where the resulting distribution does not clash with the start definition. For example, with sets $A_1, \ldots, A_8$ in such a partition, deduce the distribution for

$$(P(A_1) + P(A_2), P(A_3) + P(A_4) + P(A_5), P(A_6), P(A_7) + P(A_8)),$$

and verify that this is as it should be (i.e. the same distribution as dictated from the start definition). This is the 'collapsibility property' for the Dirichlet distribution, cf. Exercise 3(g). Without this property, the start definition would not make sense, and there would be no Dirichlet process.

(b) The full existence of the $\mathrm{Dir}(aP_0)$ is not a trivial matter, however. There are several routes to proving that yes, lo & behold, it exists. Think a bit about the paths of proofs brief indicated below. If sufficiently curious (now or later), with enough time, go ad fontem and check the arguments.

(i) Check the original argument used by Ferguson (1973, Annals), appealing to Kolmogorov's consistency (or 'inherent coherence') theorem. Under a few natural and clearly necessary conditions, Kolmogorov proved that these are also sufficient; there will be no cognitive dissonance. Ferguson then verified the Kolmogorov dictated conditions. It is worth noting that in this fashion he 'only' got a random $P = \{P(A)\colon A \in \mathcal{A}\}$, with a certain well-defined probability distribution $\mathcal{P}$, in the enormous space $[0,1]^{\mathcal{A}}$ of all function P on the enormous space $\mathcal{A}$, with values $P(A)$ in $[0,1]$ for every A. He could then could go on to prove that $\mathcal{P}(\mathcal{M}) = 1$, where $\mathcal{M}$ is the space of all probability measures on $\mathcal{X}$. This is still not the same as having created a $\mathcal{P}$ working directly on $\mathcal{M}$. Several of the other Dirichlet process constructions are more direct than this, however.

(ii) Check also Ferguson (1974, Annals), where a representation in the form of $P = Z/Z(\mathcal{X})$ is worked through, with $Z(\cdot)$ a Gamma process.

(iii) Hjort (1976, last chapter) showed that the distribution $\mathcal{P}$ of a $P \sim \mathrm{Dir}(aP_0)$ can be reached as the well-defined limit in distribution of say $\mathcal{P}_m$, where $\mathcal{P}_m$ is an easier finite-dimensional construction, basically a Dirichlet process $aP_{0,m}$ for a simpler discrete $P_{0,m}$ concentrated in only finitely many positions (for which the Dirichlet process existence is immediate). With the $P_{0,m}$ sequence constructed to tend in distribution to the perhaps continuous P_0, Hjort showed that $\mathcal{P}_m$ is tight; that its finite-dimensional distributions converge; that it must have a unique limit; and this limit is identical to Ferguson's $\mathrm{Dir}(aP_0)$. Care needs to be exercised regarding the convergence of probability measures on a space of probability measures (yes, you heard that right). In other words, the complicatedness of the statement $\mathcal{P}_m \to_d \mathcal{P}$ needs to be examined carefully, as part of the construction.

'Det er å håpe at denne alternative konstruksjonen av en Dirichlet-prosess ikke bare er av teoretisk verdi. Konstruksjonen gir informasjon utover det tre år gamle faktum at Dirichlet-prosessen eksisterer.' (Hjort, 1976, last chapter.) Hjort's 1976 construction takes place directly on the subspace $\mathcal{M}_0$ of all *discrete* probability measures on $(\mathcal{X}, \mathcal{A})$, so Ferguson's non-trivial 1973 theorem that $\mathcal{P}$ with probability 1 selects a discrete probability measure is here automatic.

(iv) Tiwari and Sethuranam (1982, Purdue Symposium), and later Sethuraman (1994, Statistica Sinica), have given an intriguing explicit representation of a Dirichlet process, in the form of

$$P = \sum_{h=1}^{\infty} w_h \delta(\xi_h),$$

where the ξ_h are i.i.d. from P_0, and the random probability weights w_h constructed in a certain way, discussed in Exercise [xx ... xx] below. Here, $\delta(\xi_h)$ means the degenerate point-mass measure with value 1 at position ξ_h.

(v) Hjort (1990, Annals). [xx via the Beta process. xx]

(vi) Hjort (2003, HSSS book). [xx via the symmetric representation and then the limit. xx]

6. Some properties for the Dirichlet process

Let $P \sim \mathrm{Dir}(aP_0)$ on some space $\mathcal{X}$. Here are a few properties to go through, shedding light on the behaviour of the random P. Note that the Dirichlet process provides a model for random probability measures (hence also for random distribution functions, etc.), with independent or separate interest. The broader appeal lies however in its use as a prior for an unknown distribution, from which we then have observations, say $X_1, \ldots, X_n$. See exercises and notes below.

(a) With A a given set, show that

$$P(A) \sim \mathrm{Beta}(aP_0(A), aP_0(A^c)),$$

with mean and variance

$$\mathrm{E}\, P(A) = P_0(A) \quad \text{and} \quad \mathrm{Var}\, P(A) = \frac{P_0(A)\{1 - P_0(A)\}}{a+1}.$$

Thus P_0 is the mean of P, hence often called simply the prior mean. The a parameter indicates strength of belief in the prior guess; a large a means a tight distribution around P_0, and vice versa for a smaller a.

(b) Find the covariance and then correlation between $P(A)$ and $P(B)$, first for A and B disjoint, then with potential overlap.

(c) With $g\colon \mathcal{X} \to \mathcal{R}$ a function, consider the random mean

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

Show that

$$\mathrm{E}\, \theta = \theta_0 = \int g\, \mathrm{d}P_0,$$

so the mean of the random mean is the prior mean. Show also that

$$\mathrm{Var}\, \theta = \frac{\sigma_0^2}{a+1},$$

with $\sigma_0^2 = \int (g - \theta_0)^2\, \mathrm{d}P_0$ the prior variance.

(d) For two functions g_1, g_2, consider the two random means $\theta_1 = \int g_1\, \mathrm{d}P$ and $\theta_2 = \int g_2\, \mathrm{d}P$. Find expressions for the covariance and correlation between these two random neans.

7. The basic updating theorem for the Dirichlet process

Suppose $P \sim \mathrm{Dir}(aP_0)$, and that $X \,|\, P$ follows the P distribution:

$$\mathcal{P}\{X \in A \,|\, P\} = P(A) \quad \text{for all } A.$$

In yet other words, X is a sample of size $n = 1$ from the given P, where P is selected randomly from the $\mathrm{Dir}(aP_0)$ machine first.

(a) Show that X has distribution P_0. Start from

$$\mathrm{E}\{I(X \in A) \,|\, P\} = P(A)$$

and use double expectation.

(b) The task is then to deduce the distribution of P given $X = x$. Attempt to show that if $A_1, \ldots, A_m$ is a partition, where x happens to lie in say the first of these, then

$$(P(A_1), \ldots, P(A_m)) \,|\, (X = x) \sim \mathrm{Dir}(aP_0(A_1) + 1, aP_0(A_2), \ldots, aP_0(A_m)).$$

(c) This is an indication that P given x is actually itself a Dirichlet process, with updated parameter $aP_0 + \delta(x)$. This also fits nicely with the finite-dimensional situation, see Exercise 3(f). You may attempt to give a formal proof of this basic updating statement for the Dirichlet process. See Ferguson (1973, Annals) or Ghosal and van der Vaart (2017, CUP book, Ch. 4).

(d) Then consider a random sample $X_1, \ldots, X_n$ from the randomly selected P, with the defining property that

$$\mathcal{P}\{X_1 \in A_1, \ldots, X_n \in A_n \,|\, P\} = P(A_1) \cdots P(A_n)$$

for all $A_1, \ldots, A_n$. With P from the Dirichlet aP_0, this defines a joint probability measure for $(P, X_1, \ldots, X_n)$. Show, perhaps by induction, that

$$P \,|\, x_1, \ldots, x_n \sim \mathrm{Dir}(aP_0 + \sum_{i=1}^n \delta(x_i)).$$

This is really a wondrously and convenient convincing result, which matches the classical Dirichlet-multinomial situation examined in Exercise 3. Note that the parameter of the posterior Dirichlet process can be written

$$aP_0 + \sum_{i=1}^n \delta(x_i) = aP_0 + nP_n,$$

with $P_n = \sum_{i=1}^n (1/n)\delta(x_i)$ the empirical distribution for the n data points.

8. Simulating from the prior and posterior, for a Dirichlet process

We need to be able to simulate realisations from the prior and the posterior, and here, specifically, from a given Dirichlet process. There are indeed several recipes for accomplishing this, but the simplest and most direct is to cut the space into a high number of smaller boxes, and then use the ensuing finite-dimensional Dirichlet as a fully adequate approximation. To carry out such finite-dimensional simulation we may use the recipe implicit in Exercise 3(g), which here means simulating a long list of small Gamma pieces and then normalising in the end.

Suppose you observe the following data points on the unit interval:

$$0.103, 0.110, 0.140, 0.175, 0.186, 0.205, 0.219, 0.348, 0.511, 0.592.$$

I have actually generated these from another distribution, namely the Beta$(1, 2)$, but the statistician seeing and about to analyse the data does not know this. For the prior for the unknown cumulative distribution function (cdf) F, take $F \sim \mathrm{Dir}(aF_0)$, with F_0 the Beta$(2, 1)$.

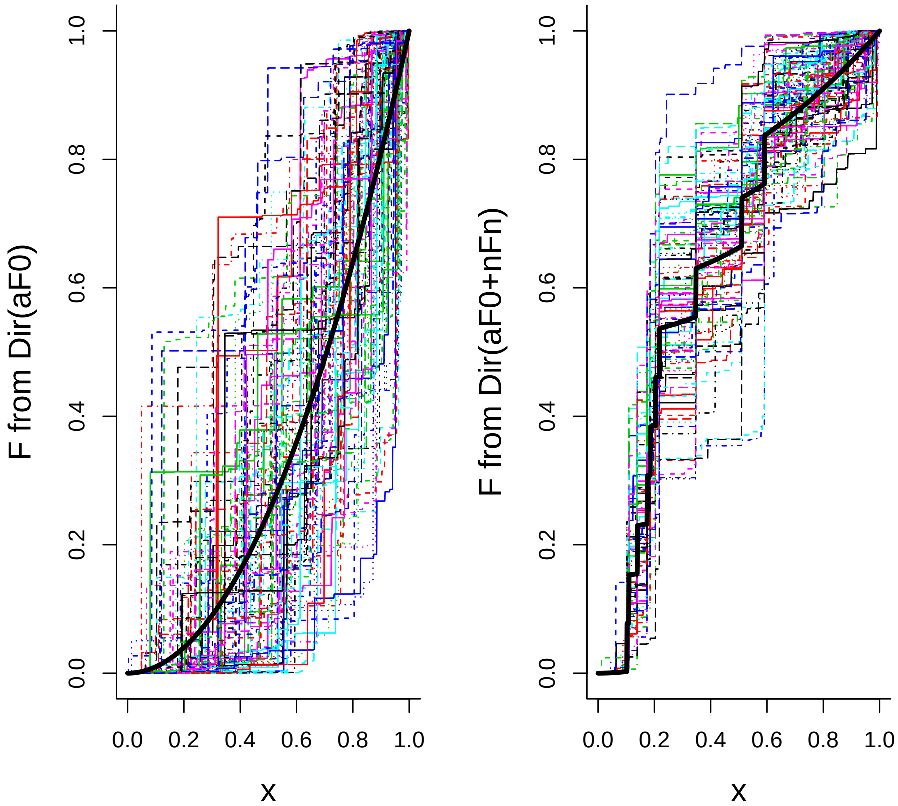

Figure 0.2: 100 simulations of F from the $\mathrm{Dir}(aF_0)$ prior (left); then 100 simulations of F from the $\mathrm{Dir}(aF_0 + nF_n)$ posterior (right), with the $n = 10$ data points of Exercise 8. The fat black curves are the prior mean and posterior mean, respectively.

(a) Simulate say 100 realisations $F = \{F(x) \colon x \in [0, 1]\}$ from the prior, using the 'lots of tiny boxes' scheme of things. See the left panel of Figure 0.2, where I've used $a = 3.333$.

(b) Then simulate say 100 realisations F from the posterior, where

$$F \,|\, \mathrm{data} \sim \mathrm{Dir}(aF_0 + nF_n),$$

with $nF_n = \sum_{i=1}^n \delta(x_i)$. See the right panel of Figure 0.2.

(c) Show that the Bayes estimator, under quadratic loss, is

$$\widehat{F}_B(x) = \mathrm{E}\,\{F(x)\,|\,\mathrm{data}\} = \frac{aF_0(x) + nF_n(x)}{a+n} = \frac{a}{a+n} F_0(x) + \frac{n}{a+n} F_n(x),$$

with F_n the empirical distribution function, i.e. the one having point-mass $1/n$ at each data point. Show furthermore that the posterior variance is

$$\widehat{\tau}^2(x) = \mathrm{Var}\,\{F(x)\,|\,\mathrm{data}\} = \frac{1}{n+a+1} \widehat{F}_B(x)\{1 - \widehat{F}_B(x)\}.$$

(d) Given realisations from F, these may be used to read off outcomes for parameters of interest, like $F(0.70) - F(0.60)$, the mean $\int_0^1 x \,\mathrm{d}F(x)$, or the median

$$\mu = \min\{x \colon F(x) \geq \tfrac{1}{2}\}.$$

Carry out analysis for this random median, by computing the $\mu = \mu(F)$ for each realisation of F, for the prior and the posterior. This leads to Figure 0.3, where I used 10^4 simulations.

(e) Play with your code a bit, to see the influence of a small a or a large a, and of the choice of the prior mean cdf F_0. You should also monitor what happens if you have say $n = 40$ data points from the underlying data generating mechanism, not only $n = 10$. You should get something similar to the right panel of Figure 0.2, but now with a slimmer and tighter spread around the Bayes estimator $\widehat{F}_B$.

(f) Then try $a = 0.0001$, a very tiny value, to see that happens with the posterior distribution of the median μ. You should learn that it has a distribution concentrated in the n data points. Try to find explicit formulae for these point masses,

$$\mathcal{P}(\mu = x_i \,|\, \text{data}), \quad \text{for } i = 1, \ldots, 10.$$

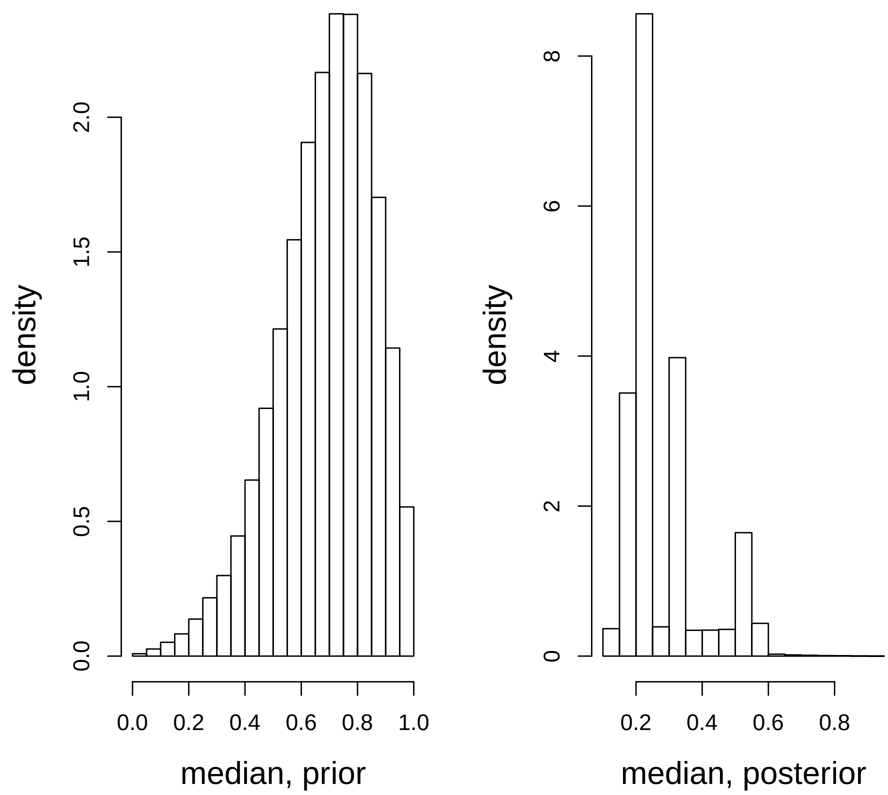

Figure 0.3: For the random median $\mu = \min\{x\colon F(x) \geq \frac{1}{2}\}$, I give histograms of its distribution, for the prior (left) and the posterior (right), based on 10^4 simulations, for each case.

9. War and peace, before and after Vietnam

Access the Tolstoyean `krigogfred-data` dataset on the course website and download it to your computer. It provides

$$(x_i, z_i) \quad \text{for } i = 1, \ldots, 95,$$

the 95 inter-state wars from 1823 to 2003 with at least 1000 battle deaths; here x_i is time of onset and z_i the number of battle deaths, for war i. Look through Hjort's FocuStat Blog Post (which apparently impressed Steven Pinker enough to cause an admiring tweet about it, to his 368,001 followers), and also the Cunen, Hjort, Nygård (2018) paper, to get a sense of the themes, the questions, the predictions for our common future, and the controversies.

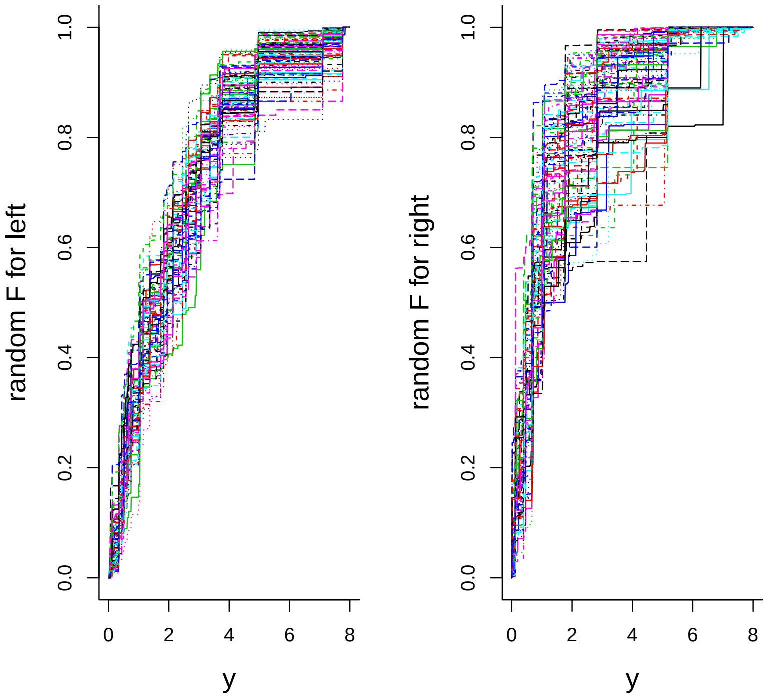

Figure 0.4: 100 simulated realisations of F_L, representing the past up to Vietnam (left), and 100 realisations of F_R, representing post-Vietnam period (right). The scale here is that of $y = \log(z/7061)$, for all wars with battle death counts at least 7061.

From these data, carry out the following two follow-up operations. First, limit attention to the 51 wars where $z_i \geq z_0$, with $z_0 = 7061$, a certain threshold value selected by A. Clauset, with the statistical intention that above this threshold, the density if proportional to $1/z^\alpha$, for an appropriate α. This is related to power laws and fat tails etc.; see again the Hjort blog post. Second, divide the remaining 51 value of (x_i, z_i) into a Left part, those 37 wars where $x_i \leq 1965.103$ (the onset-time for the Vietnam War), and a Right part, those 14 wars where $x_i > 1965.103$.

The statistical task is now to model and analyse the distributions of

$$y_i = \log(z_i/z_0) = \log z_i - \log 7061, \quad \text{for } i = 1, \ldots, 51,$$

divided into

$$\begin{aligned} y_1, \ldots, y_{37}, &\quad \text{with } x_i \text{ before and up to Vietnam,} \\ y_{38}, \ldots, y_{51}, &\quad \text{with } x_i \text{ after Vietnam.} \end{aligned}$$

Specifically, we take the 37 before and including Vietnam to be i.i.d. from some F_L, and the 14 after Vietnam to be i.i.d. from some F_R.

(a) Suppose Z has a density with the power law tail property that $f(z)$ is proportional to $1/z^\alpha$ for all z above some threshold z_0. Show that this is equivalent to saying that $Y = \log(Z/z_0)$ has an exponential tail, specifically that $\Pr\{Y \ge y \,|\, Y \ge y_0\} = \exp(-\theta(y - y_0))$ for $y \ge t_0$, with $\theta = \alpha - 1$. Power law tail behaviour for the z_i, the battle deaths counts, can therefore be examined and in terms of exponential tails for the $\log z_i$.

(b) It makes sense to take the same prior $\mathrm{Dir}(aF_0)$ for both F_L and F_R, since there is controversy in claiming that there is a difference between them at all; see Clauset's papers (2017, 2018). Take indeed $F_0(y) = 1 - \exp(-0.5\,y)$, and exponential, and $a = 3.333$ (later on you may tinker with that strength parameter). Work out the posterior distributions, and simulate say 100 realisations from each of them, as I have done to create Figure 0.4.

(c) Carry out the consequent Bayesian nonparametric inference for the difference function $\delta(y) = F_L(y) - F_R(y)$. Plot the Bayes estimate $\widehat{\delta}(y) = \mathrm{E}\,\{\delta(y) \,|\, \mathrm{data}\}$, along with a pointwise 90% credibility interval. The latter can be constructed accurately, via simulations, or via $\pm\, 1.645\,\kappa(y)$, where $\kappa(y)$ is the posterior standard deviation. Attempt both methods.

(d) [xx something more. inference for median of F_L minus median of F_R. xx]

10. The marginal distribution of a Dirichlet process sample

Suppose that $P \sim \mathrm{Dir}(aP_0)$, and that data points are subsequently drawn independently from that P. The defining property for a sample of size n, is again that

$$\mathcal{P}\{X_1 \in A_1, \ldots, X_n \in A_n \,|\, P\} = P(A_1) \cdots P(A_n),$$

for all sets $A_1, \ldots, A_n$. Here we look at a few properties.

(a) Let X be one of these points, say the first point. Show that its distribution is P_0; see also Exercise 7.

(b) Consider next (X_1, X_2), the two first data points. Show that their distribution can be expressed as

$$Q_2(A \times B) = \mathcal{P}\{X_1 \in A, X_2 \in B\} = \mathrm{E}\, P(A)P(B).$$

Then give formulae for this expression, (i) when A and B are disjoint; (ii) when they are identical; (iii) in the general case.

(c) Show that

$$Q_2 = \frac{a}{a+1} P_0 \times P_0 + \frac{1}{a+1} P_{0,12},$$

where $P_{0,12}(A \times B) = P_0(A \cap B)$. We may think about this latter probability component $P_{0,12}$ as a mechanism that first picks $X_1 \sim P_0$ and then automatically takes the X_2 equal to the first.

(d) Next study the joint distribution of three observations from a Dirichlet process. Note that X_1, X_2, X_3 are indeed i.i.d. given P, but the randomness in P makes the three dependent. Start from

$$Q_3(A \times B \times C) = \mathcal{P}_3\{X_1 \in A, X_2 \in B, X_3 \in X\} = \mathrm{E}\, P(A)P(B)P(C),$$

and give a formula for the case where A, B, C are disjoint.

(e) [xx then finish this, give clear representation of Q_3, find Hjort (1976). xx]

11. Ties and the slower stream of new guys from a Dirichlet

[xx about the process of seeing some old guys, occasionally a new guy, for Dirichlet samples. xx]

12. The number of discrete values in a Dirichlet sample

[xx to be written and polished. xx] we have $D_n = R_1 + \cdots + R_n$ representation. we find $D_n/\log n \to_{\rm pr} a$, and limiting normality from Nils 1976,

$$(\log n)^{1/2}(D_n/\log n - a) \to_d \mathrm{N}(0, a).$$

Also, the simple $D_n/\log n$ is large-sample equivalent to the maximum likelihood estimator.

13. A simple models for clusters in data

[xx to be written out and polished. xx] We consider a simple hierarchical model which in a natural fashion leads to clusters, or groups, in the data, and where the number of such clusters is not specified in advance. The setup can be described as a three-step machinery, as follows:

(i) A distribution P is taken from $\mathrm{Dir}(aP_0)$;

(ii) model parameters $\theta_1, \ldots, \theta_n$ are sampled from P (which in particular means various ties);

(iii) observations $y_1, \ldots, y_n$ are independent, given the $\theta_1, \ldots, \theta_n$, and $y_i \,|\, \theta_i \sim f(y_i \,|\, \theta_i)$.

The Bayesian task is to understand the posterior distribution of $P, \theta_1, \ldots, \theta_n$ given the observations $y_1, \ldots, y_n$.

To make this clear and understandable in a simple prototype setup, consider a case where the parameters θ_i form a sample from P, where $P \sim \mathrm{Dir}(aP_0)$, with $P_0 = \mathrm{N}(0, \sigma_0^2)$. We also take $y_i \sim \mathrm{N}(\theta_i, \sigma^2)$, with known σ. [xx more to come here. xx]

14. A clustering illustration

[xx a simple illustration here, with Dirichlet producing the model parameters, with lots of ties, etc. xx]

15. The Sethuraman stick-breaking representation

A somewhat surprising representation of the Dirichlet process, stemming from Sethuraman and Tiwari (1982, Purdue Symposium) and written out more fully in Sethuraman (1994, Sinica), is described here. With P_0 a probability measure, and a positive, we start with $B_1, B_2, B_3, \ldots$ being i.i.d. from $\mathrm{Beta}(1, a)$. From these we form weights $w_1, w_2, w_3, \ldots$, from

$$w_1 = B_1, \quad w_2 = (1 - B_1)B_2, \quad w_3 = (1 - B_1)(1 - B_2)B_3, \quad , w_h = (1 - B_1) \cdots (1 - B_{h-1})B_h.$$

In addition, we draw an infinite i.i.d. sequence $\xi_1, \xi_2, \ldots$ from P_0. The stick-breaking representation is

$$P = \sum_{h=1}^{\infty} w_h \delta(\xi_h),$$

with $\delta(\xi_h)$ the unit point-mass in position ξ_h.

(a) Show that

$$1 - w_1 - w_2 - w_3 = (1 - B_1)(1 - B_2)(1 - B_3),$$

with the immediate generalisation to $1 - w_1 - \cdots - w_n$. Show from this that $\sum_{h=1}^\infty w_h = 1$, with probability 1.

(b) For a fixed set A, consider the random probability $p = P(A)$, using the representation above. Show that p has mean $p_0 = P_0(A)$, and that

$$\mathrm{Var}\, p = \mathrm{E}\,(p - p_0)^2 = p_0(1 - p_0)/(a + 1).$$

(c) Attempt to prove that $p = P(A)$ is a $\mathrm{Beta}(ap_0, a(1 - p_0))$.

(d) For a given bounded function g, consider the random mean

$$\theta = \int g\,\mathrm{d}P = \sum_{h=1}^\infty w_h g(\xi_h).$$

Show that it has mean $\theta_0 = \int g\,\mathrm{d}P_0$ and variance $\sigma_0^2/(a + 1)$, with $\sigma_0^2 = \int (g - \theta_0)^2\,\mathrm{d}P_0$.

(e) Consider disjoint sets A and B, and work with

$$p = P(A) = \sum_{\xi_h \in A} w_h \quad \text{and} \quad q = P(B) = \sum_{\xi_h \in B} w_h.$$

Calculate the covariance between p and q, from the stick-breaking representation above, and verify that you get the correct answer, i.e. what we should have if P indeed is a $\mathrm{Dir}(aP_0)$.

(f) Then attempt to prove that the Sethuraman–Tiwari representation is correct, i.e. that P above with probability 1 becomes a $\mathrm{Dir}(aP_0)$. – You may check the Sethuraman (1994, Sinica) paper. A more concise proof is given in Ghosal and van der Vaart (2017, Ch. 4), but this needs certain other properties which must be established separately, including a distributional equation property that uniquely characterises the Dirichlet process. See also Hjort and Ongaro (2005, SISP).

16. Dependent Dirichlet processes, using stick-breaking representations

[xx something here. check nils's discussion contribution to Gelfand and Petrone. xx] the basic idea, for two Dirichlet processes, which now become dependent: suppose (ξ_h, ξ_h') are i.i.d. pairs, from some joint distribution, like the standardised binormal with correlation ρ. let P_0 and P_0' be the marginals of this joint distribution for pairs. then construct

$$P = \sum_{h=1}^\infty w_h \delta(\xi_h) \quad \text{and} \quad P' = \sum_{h=1}^\infty w_h \delta(\xi_h'),$$

with the same stick-breaking sequence of probabilities w_h as in Exercise [xx 15 xx]. by construction, $P \sim \mathrm{Dir}(aP_0)$ and $P' \sim \mathrm{Dir}(aP_0')$, and they are dependent. a quick illustration.

17. Quantile inference for the Dirichlet process

[xx something here. nils and sonia. cute formula

$$\widehat{Q}_0(y) = \sum_{i=1}^n \binom{n-1}{i-1} y^{i-1}(1 - y)^{n-i} x_{(i)}.$$

start from $F \sim \mathrm{Dir}(aF_0)$, and consider the random quantile function

$$Q(y) = \min\{x \colon F(x) \geq y\}.$$

find a clear expression for the distribution of $Q(y)$. check case of $a \to 0$ separately. also the resulting cute enough nonparametric automatic bandwidth-free density estimator $\widehat{f}_0(x)$. xx]

18. Quantile pyramids

[xx something here. Hjort and Walker (2009, Annals) and their quantile pyramids. first construction, then MCMC for posterior. xx]

19. Brownian motion via convergence of a partial-sum process

Here I briefly describe the construction of Brownian motion as the proper limit in distribution of an empirical partial-sum process. This is of interest in its own right, as it also gives a proof of the existence of the relevant Gaussian process. The point is also that similar constructions (where 'similar' could mean 'very similar' or 'somewhat similar' or 'long-distance part-time similar') in different straits are useful for Bayesian Nonparametrics, such as the Gamma and Beta processes down the road.

The Brownian motion, or Wiener process, say $W = \{W(t) \colon t \geq 0\}$, is a Gaussian process (all finite-dimensional distributions are Gaussian), with mean zero, independent increments, with $W(t) - W(s) \sim \mathrm{N}(0, t-s)$. The existence of such a process is a non-trivial delicate matter, but the construction I give below has 'yes, the Brownian motion process exists' as a by-product.

We start from an i.i.d. sequence $\varepsilon_1, \varepsilon_2, \ldots$, with mean zero and variances one, and then build the empirical process

$$Z_n(t) = (1/\sqrt{n}) \sum_{i/n \leq t} \varepsilon_i.$$

(a) Above we define Brownian motion via the property that the independent increments have $\mathrm{N}(0, t-s)$ distributions. Prove that if we somehow had started with $\mathrm{N}(0, |t-s|^\gamma)$ distributions instead, for some $\gamma \neq 1$ for the variances, then things would quickly backfire, turning the universe into massive cognitive dissonance. The Kolmogorov coherence theorem way of proving existence of Brownian motion indeed starts with checking that coherence matters are in order.

(b) Verify that $Z_n = \{Z_n(t) \colon t \geq 0\}$ has independent increments, mean zero, and variance $[nt]/n$, where $[nt]$ is the integer part of nt (so $[17] = 17$, $[17.01] = 7$, $[17.99] = 17$, etc.). Show also that

$$\mathrm{Var}\, \{Z_n(t) - Z_n(s)\} = (1/n)\, \mathrm{Var} \sum_{s < i/n \leq t} \varepsilon_i = [nt]/n - [ns]/n \to t - s,$$

for each $s < t$.

(c) Show that, for each t, we have $Z_n(t) \to_d \mathrm{N}(0, t)$. This is essentially the central limit theorem at work.

(d) For $t_1 < \cdots < t_k$, show that the vector of random differences

$$(Z_n(t_1), Z_n(t_2) - Z_n(t_1), \ldots, Z_n(t_k) - Z_n(t_{k-1}))$$

tends to the distribution of $(D_1, \ldots, D_k)$, where these are independent, with $D_j \sim \mathrm{N}(0, t_j - t_{j-1})$ (writing also $t_0 = 0$).

(e) Use this to verify that indeed

$$(Z_n(t_1), \ldots, Z_n(t_k)) \to_d (W(t_1), \ldots, W(t_k)),$$

for any $t_1 < \cdots < t_k$.

(f) The theory of convergence of probability measures, see e.g. the classic Billingsley (1968, Wiley), tells us that (d) is necessary, but sufficient, for properly proving that $Z_n \to_d W$, in the space $D[0, \tau]$ of all right-continuous functions $x\colon [0, \tau] \to \mathcal{R}$ with left-hand limits, and equipped with the Skorokhod topology. We do not go into the details here, but the added necessity factor is that of *tightness*, a condition that secures that the Z_n does not have any mass escaping away, or turning itself into too high oscillation with too high probability. A condition securing tightness, which again secures what we're after, namely full process convergence $Z_n \to_d W$, is

$$\mathrm{E}\,\{Z_n(t) - Z_n(s)\}^2\,\{Z_n(u) - Z_n(t)\}^2 \le \{K(u) - K(s)\}^2,$$

for all triples $s < t < u$, for a suitable monotone, continuous function K. Verify this condition here.

(g) Note that the above construction and reasoning hold, regardless of the actual distribution of the building blocks $\varepsilon_1, \varepsilon_2, \ldots$. In particular, we may take the two very different start distributions $\varepsilon_i \sim \mathrm{N}(0, 1)$, or ε_i equal to 1 or -1 with equal probability $\frac{1}{2}$, and have the same Brownian limit. Simulate some paths, from these two partial-sum processes, for say $n = 1000$, and check if you can tell the difference.

(h) When this cornerstone theorem is in place, there is a long list of implications and corollaries and new constructions. Let me mention the Brownian bridge, $W^0 = \{W^0(t)\colon 0 \le t \le 1\}$. It is Gaussian, zero mean, and covariance function $\mathrm{cov}\{W^0(s), W^0(t)\} = s(1 - t)$ for $s \le t$. It emerges in several ways, including starting from the Wiener process and then forming

$$W^0(t) = W(t) - tW(1) \quad \text{for } 0 \le t \le 1.$$

It is also the limit of the bridged version of the above empirical process,

$$Z_n^0(t) = Z_n(t) - tZ_n(1) \quad \text{for } 0 \le t \le 1.$$

(i) [xx perhaps a few illustrations of the 'invariance theorem' aspect of this Donsker theorem. xx]

20. A little lemma

We shall encounter situations involving long products of the type $a_n = \prod_{i \le n}(1 + z_{n,i})$, where there for each n is a well-defined sequence of $z_{n,i}$ for $i = 1, \ldots, n$. If these are small and their sum converges, the sequence of products will converge. Specifically, assume

(i) that $\sum_{i \le n} z_{n,i} \to z$;

(ii) that $\delta_n = \max_{i \le n} |z_{n,i}| \to 0$;

(iii) that $\sum_{i \le n} |z_{n,i}|$ remains bounded.

Show that then $a_n = \prod_{i \le n}(1 + z_{n,i}) \to a = \exp(z)$. It is helpful here to write

$$\log(1 + z) = z - \tfrac{1}{2}z^2 + z^2 K(z),$$

where $|K(z)| \le \frac{1}{2}$ for all $|z| \le \frac{1}{2}$.

Similar results also hold when the product is taken over suitable subsets of i/n, like

$$\prod_{s < i/n \le t} (1 + z_{n,i}) \to \exp(z_{s,t}),$$

if $\sum_{s<i/n\le t} z_{n,i} \to z_{s,t}$, etc.

21. The Gamma process

For a given monotone function $M(t)$, starting at $M(0) = 0$, we may define a Gamma process $Z = \{Z(t)\colon t \ge 0\}$ with the property that it has independent increments with $Z(t) - Z(s) \sim \mathrm{Gamma}(M(t) - M(s), 1)$. Existence of such a process is not entirely obvious, but one is of course helped by the fact that

$$\mathrm{Gamma}(M(t) - M(s), 1) + \mathrm{Gamma}(M(u) - M(t), 1) \sim \mathrm{Gamma}(M(u) - M(s), 1)$$

for $s < t < u$, with the two components on the left hand side being independent.

The purpose of this exercise is to work through some of the crucial details for the Gamma process, which also opens the door for more general constructions later on, like the extended Gamma process in the next exercise.

(a) Let $G \sim \mathrm{Gamma}(a, b)$, with density proportional to $x^{a-1}\exp(-bx)$. Show that its Laplace transform may be written as

$$\mathrm{E}\exp(-uG) = \frac{b^a}{\Gamma(a)}\frac{\Gamma(a)}{(b+u)^a} = \frac{1}{(1+u/b)^a} = \exp\{-a\log(1+u/b)\}.$$

(b) Use this to show that if $G_1, \ldots, G_m$ are independent Gamma distributed variables, with parameters $(a_1, b), \ldots, (a_m, b)$, then their sum is also Gamma distributed, with parameters $(\sum_{i=1}^m a_i, b)$.

(c) Show that the negative exponent in the Laplace transform can be expressed as

$$a\log(1 + u/b) = \int_0^\infty \{1 - \exp(-us)\}\,\mathrm{d}L(s),$$

with

$$\mathrm{d}L(s) = as^{-1}\exp(-bs)\,\mathrm{d}s.$$

(d) Suppose as above that $M(t)$ is monotone, with $M(0) = 0$; in various applications, it will be a cumulative intensity function and of the form $M(t) = \int_0^t a(s)\,\mathrm{d}s$, with an underlying nonnegative intensity function $a(s)$. Consider the process

$$Z_m(t) = \sum_{i/m \le t} G_{m,i} \quad \text{for } t \ge 0,$$

where the $G_{m,i}$ are independent, and $G_{m,i} \sim \mathrm{Gamma}(a_{m,i}, b)$, with $a_{m,i} = M(i/m) - M((i-1)/m)$ for $i \ge 1$. For the case of M being the integral of a, it is useful to think of $a_{m,i}$ as $a(i/m)(1/m)$. Show that the mean and variance converge properly,

$$\mathrm{E}\,Z_m(t) \to M(t)/b \quad \text{and} \quad \mathrm{Var}\,Z_m(t) \to M(t)/b^2.$$

(e) Show that the Laplace transform converges,

$$\mathrm{E}\exp\{-uZ_m(t)\} = \prod_{i/m\le t} \exp\{-a_{m,i}\log(1+u/b)\} \to \exp\{-M(t)\log(1+u/b)\}.$$

This establishes existence of the Gamma process with parameter $(M(\cdot), b)$, via process convergence arguments as in Exercise [xx ... xx].

(f) Show that the arguments above also work in the case where the underlying $M(\cdot)$ function is replaced by a $M_m(\cdot)$ function, which converges to a limit $M(\cdot)$. In particular, things go through for the case of $M_m(t) = \sum_{i/m\le t} a(i/m)(1/m)$, tending to $\int_0^t a(s)\,\mathrm{d}s$.

(g) [xx a bit more xx]

22. The extended gamma process

In the course of this exercise I build a more general process, which I term an extended Gamma process. [xx which has been worked with earlier, actually; find one or two references, from the two Walker students. xx] We start with independent and inherently small gammas,

$$G_{m,i} \sim \mathrm{Gamma}(a(i/m)(1/m), b(i/m)) \quad \text{for } i = 1, 2, \ldots,$$

where $a(s)$ and $b(s)$ are functions, taken positive and continuous, or at least piecewise continuous, and with $b(s)$ bounded above zero. From these we form the partial sum process

$$Z_m(t) = \sum_{i/m\le t} G_{m,i} \quad \text{for } t \ge 0.$$

(a) Show that the mean and variance converge,

$$\begin{aligned}\mathrm{E}\,Z_m(t) &= \sum_{i/m\le t} \frac{a(i/m)(1/m)}{b(i/m)} \to \int_0^t \frac{a(s)}{b(s)}\,\mathrm{d}s,\\ \mathrm{Var}\,Z_m(t) &= \sum_{i/m\le t} \frac{a(i/m)(1/m)}{b(i/m)^2} \to V(t) = \int_0^t \frac{a(s)}{b(s)^2}\,\mathrm{d}s.\end{aligned}$$

(b) Show that the Laplace transform converges properly:

$$\mathrm{E}\exp\{-uZ_m(t)\} = \prod_{i/m\le t} \mathrm{E}\exp(-uG_{m,i}) = \exp\Big[-\sum_{i/m\le t} a(i/m)(1/m)\log\{1+u/b(i/m)\}\Big],$$

which indeed tends to

$$\exp\Big[-\int_0^t a(s)\log\{1+u/b(s)\}\,\mathrm{d}s\Big].$$

(c) With $V(t)$ as above, show that

$$\mathrm{E}\,\{Z_m(t)-Z_m(s)\}^2\{Z_m(u)-Z_m(t)\}^2 \to \{V(t)-V(s)\}\{V(u)-V(t)\} \le \{V(u)-V(s)\}^2,$$

for $s < t < u$. Via general arguments and inequalities in Billingsley (1968, Section 15), see also Exercise [xx 18 xx], this can be seen to imply that the Z_m sequence is tight, in the space of right-continuous functions $z\colon [0,\tau] \to \mathcal{R}$ with left-hand limits, equipped with the Skorokhod metric. This establishes proper existence of the extended Gamma process.

(d) [xx a bit more. the Lévy representation of things. xx]

$$\mathrm{E}\,\exp\{-uZ(t)\} = \exp\Big[-\int_0^\infty \{1-\exp(-us)\}\,\mathrm{d}L_t(s)\Big].$$

23. The extended Gamma process with a Poisson process

[xx part of the nils-emil story. pointer to later exercise with covariates. pointer also to Beta process version of things. xx] I start with the time-discrete version of things. Consider a sequence of independent pairs $(\theta_{m,i}, z_{m,i})$, to be thought of as evolving over time points i/m, with

$$\theta_{m,i} \sim \mathrm{Gamma}(a(i/m)(1/m), b(i/m)) \quad \text{and} \quad z_{m,i}\,|\,\theta_{m,i} \sim \mathrm{Pois}(\theta_{m,i}).$$

In particular, $G_m(t) = \sum_{i/m\le t} \theta_{m,i}$ is the cumulative intensity process, and $Z_m(t) = \sum_{i/m\le t} z_{m,i}$ the cumulative Poisson count of events.

(a) Show that

$$\theta_{m,i}\,|\,\mathrm{data} \sim \mathrm{Gamma}(a(i/m)(1/m) + z_{m,i}, b(i/m)+1).$$

(b) Then consider the time-continuous version of this story, corresponding to letting $m\to\infty$ above. This leads to $G_m \to_d G$, an extended Gamma process, with parameter functions $a(s), b(s)$. Then, iven G, there is a limit $Z_m \to Z$, a nonhomogeneous observed Poisson process $Z = \{Z(t)\colon t\ge 0\}$, with cumulative intensity function G. Show that G given data is another extended Gamma process, with $b_{\mathrm{new}}(s) = b(s)+1$, and $A_{\mathrm{new}}(t) = \int_0^t a(s)\,\mathrm{d}s + Z(t)$. This translates to

$$\mathrm{d}G(s)\,|\,\mathrm{data} \sim \mathrm{Gamma}(a(s)\,\mathrm{d}s + \mathrm{d}Z(s), b(s)+1),$$

with $\mathrm{d}Z(s) = Z[s, s+\mathrm{d}s]$ the number of Poisson events observed in the small time window $[s, s+\mathrm{d}s]$.

(c) In particular, writing event times as $T_1 < T_2 < \cdots$, show that the posterior mean becomes

$$\widehat{G}(t) = \int_0^t \frac{a(s)\,\mathrm{d}s + \mathrm{d}Z(s)}{b(s)+1} = \int_{\text{no jumps}} \frac{a(s)}{b(s)+1}\,\mathrm{d}s + \sum_{\text{jumps }\le t} \frac{1}{b(T_j)+1}.$$

Find also an expression for the posterior variance.

(d) Suppose next that there are several observed nonhomogeneous Poisson processes, say $Z_1,\ldots,Z_k$, with the same underlying G. Show that G given the data is again an extended Gamma process, with

$$\mathrm{d}G(s)\,|\,\mathrm{data} \sim \mathrm{Gamma}(a(s)\,\mathrm{d}s + \sum_{j=1}^k \mathrm{d}Z_j(s), b(s)+k).$$

(e) With such observed processes, show that the Bayes estimator for the cumulative intensity process, i.e. the posterior mean, becomes

$$\widehat{G}(t) = \int_0^t \frac{a(s)\,\mathrm{d}s + \sum_{=1}^k \mathrm{d}Z_j(s)}{b(s)+k} = \int_{\text{no jumps}} \frac{a(s)}{b(s)+k}\,\mathrm{d}s + \sum_{\text{jumps }\le t} \frac{1}{b(T_j)+k},$$

now with 'jumps' referring to jumps in any of the k observed nonhomogeneous Poisson processes. Also, give a formula for the posterior variance.

24. The jumps of a Gamma process

[xx something from Hjort and Ongaro (2006, Metron). the density $g_1(x)$ for the biggest jump of a Gamma process over a time window $[0,\tau]$. then more generally a clear expression for the joint density $g_k(x_1,\cdots,x_k)$ for the biggest jump J_1, the next biggest jump J_2, etc., valid for $x_1>\cdots>x_k$. xx]

25. The Beta process

Hjort (1985, SJS, invited discussion contribution to the SJS paper by P.K. Andersen and Ø. Borgan on counting process models) introduced the Beta process, used as a prior process for cumulative hazard functions, and gave the crucial conjugacy property when used for survival data. A fuller account was then given in Hjort (1990, Annals). The present exercise indicates how the Beta process can be constructed from a limit operation for a partial-sum process involving small Beta components.

We start with a function $a_0(s)$, intended to be like a prior guess hazard function, with cumulative $A_0(t)=\int_0^t a_0(s)\,\mathrm{d}s$. For given m, let $B_{m,1},B_{m,2},\ldots$ be independent Beta random variables, with

$$B_{m,i}\sim \mathrm{Beta}\Big(c\Big(\frac{i}{m}\Big)a_0\Big(\frac{i}{m}\Big)\frac{1}{m},c\Big(\frac{i}{m}\Big)-c\Big(\frac{i}{m}\Big)a_0\Big(\frac{i}{m}\Big)\frac{1}{m}\Big).$$

Here $c(s)$ is a positive function, with at most finitely many discontinuities; it may e.g. be a constant. Our process is

$$A_m(t)=\sum_{i/m\le t}B_{m,i}\quad\text{for } t\ge 0.$$

(a) Show that

$$\mathrm{E}\,Z_m(t)=\sum_{i/m\le t}a_0(i/m)(1/m)\to A_0(t).$$

Show also that

$$\mathrm{Var}\,A_m(t)=\sum_{i/m\le t}\frac{a_0(i/m)(1/m)\{1-a_0(i/m)(1/m)\}}{c(i/m)+1}\to\int_0^t\frac{a_0(s)\,\mathrm{d}s}{c(s)+1}.$$

(b) Hjort (1985, 1990) proves that A_m really converges to a well-defined limit process $A=\{A(t)\colon t\ge 0\}$, with independent increments all inside $[0,1]$, and calls this the Beta process, with parameters (c,A_0). Proving convergence and existence of this limit process takes some care and tools from empirical processes. The crucial point here is that the Laplace transform has a well-defined limit, so let us work with

$$\mathrm{E}\,\exp\{-uA_m(t)\}=\prod_{i/m\le t}\mathrm{E}\,\exp(-uB_{m,i})=\prod_{i/m\le t}(1+z_{m,i}),$$

say. We must then work hard enough with the $z_{m,i}$ to be able to apply the Little Lemma of Exercise XX. Show via Beta moments that

$$\begin{aligned}\mathrm{E}\,\exp(-uB_{m,i})&=1+z_{m,i}\\&=1+\sum_{j=1}^\infty(-1)^j\frac{u^j}{j!}\frac{\Gamma(c(i/m))}{\Gamma(c(i/m)a_0(i/m)(1/m))}\frac{\Gamma(c(i/m)a_0(i/m)(1/m)+j)}{\Gamma(c(i/m)+j)}.\end{aligned}$$

(c) Then use

$$\Gamma(\varepsilon+j)/\Gamma(\varepsilon) = \varepsilon(\varepsilon+1)\cdots(\varepsilon+j-1) = (j-1)!\,\varepsilon + O(\varepsilon^2)$$

for small ε to deduce

$$\mathrm{E}\exp(-uB_{m,i}) = 1 + \sum_{j=1}^\infty (-1)^j \frac{u^j}{j}\frac{\Gamma(c(i/m))}{\Gamma(c(i/m)+j)} c(i/m)a_0(i/m)(1/m) + O(1/m^2).$$

(d) Show that this leads to

$$\mathrm{E}\exp\{-uA_m(t)\} \to \exp\Big\{-\int_0^t \sum_{j=1}^\infty (-1)^j \frac{u^j}{j!}\frac{\Gamma(c(z))\Gamma(j)}{\Gamma(c(z)+j)} a_0(z)\,\mathrm{d}z\Big\}.$$

(e) [xx more xx] Link to Lévy representation

$$\int_0^1 \{1-\exp(-us)\}\,\mathrm{d}L_t(s).$$

– The above establishes the existence of a Beta process, with parameters (c, A_0); for a fuller discussion, see Hjort (1990, Annals). It is a independent and nonnegative increments, and these are all in $[0,1]$. The intuitive interpretation for a Beta process is that

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

These tiny increments are not exactly Beta distributed, though; that distribution does not have any easy convolution properties, unlike e.g. the Gamma.

26. A time-discrete framework for survival analysis

Consider the following framework for life-times, now with time-discrete outcomes in $\{0,1,2,\ldots\}$, rather than the usual time-continuous setup of $[0,\infty)$. A random variable T then has probability masses

$$f_j = \Pr\{T=j\} \quad \text{for } j=0,1,2,\ldots,$$

with cumulative $F_j = \Pr\{T \le j\} = f_0 + f_1 + \cdots + f_j$. It is also very fruitful to work with the hazards

$$\alpha_j = \Pr\{T=j \,|\, T \ge j\} = f_j/(f_j + f_{j+1} + \cdots),$$

along with the cumulative hazards $A_j = \alpha_0 + \alpha_1 + \cdots + \alpha_j$.

Part of what I present in this exercise was also included in Hjort (1990, Annals, along with further results, extensions, and discussion). The framework and methods for the time-discrete setup have separate interest, and it inspired the invention of the Beta process, as a fine limit of the time-discrete grid.

(a) Show that

$$F_j = 1 - \prod_{k=0}^j (1-\alpha_k), \quad \text{for } j \ge 0.$$

(b) Show then that

$$f_j = (1-\alpha_0)(1-\alpha_1)\cdots(1-\alpha_{j-1})\alpha_j,$$

and give an interpretation of this identity.

– Assume now that we have observations (t_i, δ_i) for $i = 1, \ldots, n$, for different individuals, with $\delta_i = 1$ if the life-time is observed and $\delta_i = 0$ if there merely is censored information that the real life-time is larger than t_i. From these, define

$$Y_j = \sum_{i=1}^{n} I\{t_i \geq j\} \quad \text{and} \quad N_j = \sum_{i=1}^{n} I\{t_i = j, \delta_i = 1\},$$

the at-risk counter and the counting process of observed life-times. In particular, let ΔN_j be the jump of N at time point j, the number of the Y_j at risk who experience a transition at time j.

(d) Show, with the necessary efforts of bureaucratic book-keeping, that the likelihood information from such a dataset can be expressed as

$$L = \prod_{i\colon \delta_i=1} f_i \prod_{i\colon \delta_i=0} (1 - F_i) = \prod_{j=0}^{\infty} (1 - \alpha_j)^{Y_j - \Delta N_j} \alpha_j^{\Delta N_j}.$$

Discuss how or to what extent this can be interpreted as a succession of binomial trials, with $\Delta N_j \,|\, Y_j$ a binomial (Y_j, α_j).

(e) The representation above invites the idea of independent Beta priors for the hazards. Let in fact $\alpha_j \sim \mathrm{Beta}(c_j\alpha_{j,0}, c_j - c_j\alpha_{,0})$, for $j = 0, 1, 2, \ldots$, and deduce that these are independent and Beta distributed also given data, with updated parameters

$$\alpha_j \,|\, \mathrm{data} \sim \mathrm{Beta}(c_j\alpha_{j,0} + \Delta N_j, c_j - c_j\alpha_{0,j} + Y_j - \Delta N_j).$$

(f) Show that the Bayes estimator for the cumulative hazard function is

$$\widehat{A}_j = \mathrm{E}\,(A_j \,|\, \mathrm{data}) = \sum_{k=0}^{j} \frac{c_k\alpha_{k,0} + \Delta N_k}{c_k + Y_k}.$$

The noninformative case of the c_j becoming small leads to $\sum_{k=0}^{j} \Delta N_k / Y_k$, a time-discrete version of the Nelson–Aalen estimator (see Exercise [xx ... xx]).

(g) Then show that the Bayes estimator for the survival function $\Pr\{T > j\} = \prod_{k=0}^{j}(1 - \alpha_k)$ is

$$\widehat{S}_j = \prod_{k=0}^{j} \Big(1 - \frac{c_k\alpha_{k,0} + \Delta N_k}{c_k + Y_k}\Big).$$

For the noninformative case of the $c_j \to 0$, we find $\prod_{k=0}^{j}(1 - \Delta N_k / Y_k)$, a time-discrete version of the Kaplan–Meier estimator (see again Exercise [xx ... xx]).

27. The Beta process for survival data

[xx need polish. xx] Suppose a survival dataset of the usual form (t_i, δ_i) is available, to inform us about an underlying survival distribution F on $[0, \infty)$. As per tradition, $\delta_i = 1$ is an indicator for non-censoring, and means a fully observed life-time, whereas $\delta_i = 0$ means that the life-time involved is censored, but one knows that it is larger than t_i. The survival distribution is $S(t) = 1 - F(t) = \Pr\{T > t\}$, and the cumulative hazard rate is

$$A(t) = \int_0^t \mathrm{d}A(s), \quad \text{where} \quad \mathrm{d}A(s) = \frac{\mathrm{d}F(s)}{F[s, \infty)} = \Pr\{T \in [s, s + \mathrm{d}s] \,|\, T \geq s\}.$$

In Aalen–Borgan notation, consider the at-risk counter and the counting process of observed life-times,

$$Y(s) = \sum_{i=1}^n I\{t_i \ge t\} \quad \text{and} \quad N(t) = \sum_{i=1}^n I\{t_i \le t, \delta_i = 1\}.$$

In particular, $\mathrm{d}N(s)$ is 1 if a life-time has been observed in $[s, s+\mathrm{d}s]$, and 0 if not. The famous Nelson–Aalen and perhaps even more famous Kaplan–Meier estimator, for the cumulative hazard rate and the survival curve, are

$$\widehat{A}(t) = \int_0^t \frac{\mathrm{d}N(s)}{Y(s)} \quad \text{and} \quad \widehat{S}(t) = \prod_{[0,t]} \{1 - \mathrm{d}N(s)/Y(s)\}.$$

(a) Let $A \sim \mathrm{Beta}(c, A_0)$, with A_0 the prior mean and c the strength function. Try to show, perhaps using some intuitive arguments, based on the approximate prior distribution of $\mathrm{d}A(s)$, that

$$\mathrm{d}A(s) \,|\, \text{data} \approx_d \mathrm{Beta}\{c(s)\,\mathrm{d}A_0(s) + \mathrm{d}N(s), c(s) - c(s)\,\mathrm{d}A_0(s) + Y(s) - \mathrm{d}N(s)\},$$

and that these increments must be independent.

(b) Try, again perhaps using heuristic arguments, to show that this means that the posterior distribution of A is an updated Beta process,

$$A \,|\, \text{data} \sim \mathrm{Beta}(c + Y, \widehat{A}),$$

with posterior mean function

$$\widehat{A}(t) = \int_0^t \frac{c\,\mathrm{d}A_0 + \mathrm{d}N}{c + Y}.$$

This is the basic conjugacy property for the Beta process with survival data, proven in Hjort (1990, Annals) – involving, he says, 'heroic integrations'.

(c) Use the product integral representation

$$F(t) = 1 - \prod_{[0,t]} \{1 - \mathrm{d}A(s)\}$$

to find the posterior mean of the survival function,

$$\widehat{S}(t) = \prod_{[0,t]} \Big\{1 - \frac{\mathrm{d}N(s)}{Y(s)}\Big\}.$$

(d) Show that when the $c(s)$ function tends to zero, or if the data volume is relatively large compared to the $c(s)$, then we're back to the Nelson–Aalen and Kaplan–Meier estimators.

(e) Explain how one may simulate realisations of A and then S from the posterior distribution. This may then be used to read off what we might wish for from these, like the posterior median

$$\mu = \min\{t \colon F(t) \ge \tfrac{1}{2}\}.$$

28. Lifelengths in Roman Era Egypt

[xx this to be polished. xx] Access the `egypt-data` dataset from the course website, pertaining to the life-lengths of 82 men and 59 women from Roman Era Egypt, the 1st century b.C. This was a relatively peaceful society, without major wars, etc., and the life-lengths can be seen as having been sampled from the upper classes of that society. I've taken the data from the very first issue of Biometrika (1901), where Karl Pearson briefly discussed aspects of the life-lengths distribution, comparing this to Britain 1900.

Here we are interested in aspects of the underlying distributions F_w and F_m, for women and men, respectively, and, in particular, aspects where we might identify differences between the two. Let A_w and A_m be the cumulative hazard rate functions, along with survival curves

$$S_w(t) = \prod_{[0,t]} \{1 - \mathrm{d}A_w(s)\} \quad \text{and} \quad S_m(t) = \prod_{[0,t]} \{1 - \mathrm{d}A_m(s)\}. \tag{eg1}$$

We use Beta process priors for the cumulative hazard rates, $A_w \sim \mathrm{Beta}(c_w, A_{0,w})$ and $A_m \sim \mathrm{Beta}(c_m, A_{0,m})$.

(a) Assume for about two minutes that A_w and A_m are continuous functions. Then show from the product integrals that the familiar formulae

$$S_w(t) = \exp\{-A_w(t)\} \quad \text{and} \quad S_m(t) = \exp\{-A_m(t)\} \tag{eg2}$$

emerge. With the Beta process priors to be used, however, there are discrete components, and we prefer (eg1) over (eg2), in terms of setup, modelling, prior to posterior, analysis, and interpretation. See also the general discussion regarding this point in Hjort (1990, Annals).

(b) To make this concrete, choose the same Beta process prior for men and for women, with prior guess $A_0(t) = \int_0^t \alpha_0(s)\,\mathrm{d}s$ corresponding to a Gamma with mean 30.00 and standard deviation 20.00, and then your own $c(s)$ strength function. Simulate realisations from A_w, A_m, and by implication S_w, S_m, on your screen.

(c) Then update the Beta processes, given the data from the heroic Egyptian women and men, to say

$$A_w \,|\, \text{data} \sim \mathrm{Beta}(c_w + Y_w, \widehat{A}_w) \quad \text{and} \quad A_m \,|\, \text{data} \sim \mathrm{Beta}(c_m + Y_m, \widehat{A}_m).$$

In particular, compute and display both

$$\widehat{A}_w(t) = \int_0^t \frac{c_w\,\mathrm{d}A_0(s) + \mathrm{d}N_w(s)}{c_w(s) + Y_w(s)} \quad \text{and} \quad \widehat{A}_m(t) = \int_0^t \frac{c_m\,\mathrm{d}A_0(s) + \mathrm{d}N_m(s)}{c_m(s) + Y_m(s)},$$

and the survival curves

$$\widehat{S}_w(t) = \prod_{[0,t]} \Big\{1 - \frac{c_w(s)\,\mathrm{d}A_0(s) + \mathrm{d}N_w(s)}{c_w(s) + Y_w(s)}\Big\} \quad \text{and} \quad \widehat{S}_m(t) = \prod_{[0,t]} \Big\{1 - \frac{c_m(s)\,\mathrm{d}A_0(s) + \mathrm{d}N_m(s)}{c_m(s) + Y_m(s)}\Big\}.$$

(d) Compute and display also the standard deviation curves, say $\widehat{\kappa}_w(t)$ and $\widehat{\kappa}_m(t)$ for A_w and A_m, and $\widehat{\tau}_w(t)$ and $\widehat{\tau}_m(t)$ for S_w and S_m.

(e) Display the easy and simulation free approximate pointwise 90% confidence bands, of the type

$$\widehat{A}_w(t) \pm 1.645\,\widehat{\kappa}_w(t) \quad \text{and} \quad \widehat{A}_m(t) \pm 1.645\,\widehat{\kappa}_m(t),$$

and similarly for the survival curves. Crucially, in order to check the differences between the female and male populations, do this also for $A_w - A_m$ and $S_w - S_m$.

(f) Then re-do the above point, without formulae, but via simulations from the posterior Beta processes.

(g) This thing looks cool and relevant: Consider the survival curve ratio

$$\rho(t) = \frac{S_m(t)}{S_w(t)} = \prod_{[0,t]} \frac{1 - \mathrm{d}A_m(s)}{1 - \mathrm{d}A_w(s)}.$$

Find formulae for the prior and posterior mean of $\rho(t)$, and display the resulting $\widehat{\rho}(t)$. Supplement this with a pointwise 90% credibility band, from simulations, or from conditional variances.

(h) Summarise your findings properly. Yes, the women and the men of Roman Era Egypt had different life-length distributions. For which age interval is this most clear? And what could be the underlying mechanism or explanations?

(i) [xx nils then includes a couple of Old Egyptian plots here. xx]

29. The Bernoulli process and the Poisson process

We learn if not in kindergarten then perhaps in high school that a binomial (n, p) is close to a Poisson if n is big and p is small. This exercise exhibits generalisations of this basic result, leading also to a nonhomogeneous Poisson process limit of a suitably defined Bernoulli events process.

(a) For $y \sim \mathrm{Bin}(n, p)$, show that its Laplace transform is

$$L_n(u) = \mathrm{E}\,\exp(-uy) = \{\exp(-u)p + 1 - p\}^n.$$

(b) Show also that when n increases and p decreases in such a way that $np \to \theta$, then

$$L_n(u) = [1 - p\{1 - \exp(-u)\}]^n \to \exp[-\theta\{1 - \exp(-u)\}].$$

Verify that this limit is the Laplace transform of a $\mathrm{Pois}(\theta)$.

(c) Now study a Bernoulli event process, of the form

$$Z_m(t) = \sum_{i/m \le t} B_{m,i} \quad \text{for } t \ge 0,$$

with independent Bernoulli components $B_{m,i} \sim \mathrm{Bin}(1, a(i/m)(1/m))$. Here $a(s)$ is some nonnegative function, perhaps constant, perhaps evolving over time, and with cumulative intensity function $A(t) = \int_0^t a(s)\,\mathrm{d}s$. Show that Z_m has independent increments, and that its Laplace transform converges,

$$\mathrm{E}\,\exp\{-uZ_m(t)\} = \prod_{i/n \le t} [1 - a(i/m)(1/m)\{1 - \exp(-u)\}] \to \exp[-A(t)\{1 - \exp(-u)\}].$$

This means that the fine-grid Bernoulli process has properly converged, in the time-continuous limit, to a nonhomogeneous Poisson process, with $A(t) = \int_0^t a(s)\,\mathrm{d}s$ as cumulative intensity process.

30. The Beta process with a Bernoulli process

[xx to be written down. part of nils-emil story. xx] prior $A \sim \mathrm{Beta}(c, A_0)$ for the cumulative intensity of a Bernoulli process Z. then

$$A \mid \mathrm{data} \sim \mathrm{Beta}(c+1, \widehat{A}),$$

where

$$\widehat{A}(t) = \int_0^t \frac{c(s)\,\mathrm{d}A_0(s) + \mathrm{d}Z(s)}{c(s)+1}.$$

with variation: extended Gamma. perhaps with marks or covariates. cross-ref to other exercises here.

31. The Gamma process, with a Poisson process, with covariates

[xx to be polished. part of the nils-emil story with police tweets, now with covariates. xx] In Exercise [xx 23 xx] we worked with the extended Gamma process as a prior G for the cumulative intensity function of nonhomogeneous Poisson processes, say $Z_1, \ldots, Z_k$. The present exercise takes us through a certain statistically important generalisation, where covariate information is available for the Z_j event counting processes.

(a) As a start generalisation of the framework of Exercise [xx 23 xx], suppose that there is a sequence of independent pairs $(\theta_{m,i}, z_{m,i})$, where

$$\theta_{m,i} \sim \mathrm{Gamma}(a(i/m)(1/m), b(i/m)) \quad \text{and} \quad z_{m,i} \mid \theta_{m,i} \sim \mathrm{Pois}(w(i/m)\theta_{m,i}).$$

At the moment, the $w(s)$ is to be thought of as a given function, producing multiplicative Poisson intensity factors $w(i/m)$. Show that this leads to

$$\theta_{m,i} \mid \mathrm{data} \sim \mathrm{Gamma}(a(i/m)(1/m) + z_{m,i}, b(i/m) + w(i/m)).$$

(b) In the time-continuous limit, with $G_m(t) = \sum_{i/m \le t} \theta_{m,i}$ tending to a cumulative intensity process $G(t)$, and the nonhomogeneous Poisson counting process $Z_m(t) = \sum_{i/m \le t} z_{m,i}$ to a proper $Z(t)$, show that

$$\mathrm{d}G(s) \mid \mathrm{data} \sim \mathrm{Gamma}(a(s)\,\mathrm{d}s + \mathrm{d}Z(s), b(s) + w(s)).$$

(c) Suppose there are several Poisson event processes being observed, say $Z_1, \ldots, Z_k$, which are conditionally independent given G, and with

$$\mathrm{d}Z_j(s) \mid G \sim \mathrm{Pois}(w_j(s)\mathrm{d}G(s)) \quad \text{for } j = 1, \ldots, k,$$

where the $w_j(s)$ are multiplicative Poisson factor functions. Show that G given data again becomes an extended Gamma process, with

$$\mathrm{d}G(s) \mid \mathrm{data} \sim \mathrm{Gamma}\Big(a(s)\,\mathrm{d}s + \sum_{j=1}^k \mathrm{d}Z_j(s), b(s) + \sum_{j=1}^k w_j(s)\Big).$$

(d) Assume there are covariates $x_1, \ldots, x_k$ at work for the Poisson event counting processes $Z_1, \ldots, Z_k$; these may also depend on time, say with $x_j(s)$ related to the outcome $\mathrm{d}Z_j(s)$. Let $w_j = \exp(x_j^{\mathrm{t}}\beta)$, with a prior $\pi(\beta)$ for this regression parameter. The model at work then says (i) that β is drawn from the prior $\pi(\beta)$; (ii) that $G(\cdot)$ is an extended Gamma process, with parameters $(a(s), b(s))$; (iii) that the Poisson processes $Z_1, \ldots, Z_k$ have intensity functions $\exp(x_j^{\mathrm{t}}\beta)\,\mathrm{d}G(s)$, hence cumulative intensity functions $\int_0^t \exp(x_j^{\mathrm{t}}\beta)\,\mathrm{d}G(s)$. Show that G, given both the data and β, is another extended Gamma process, with parameters

$$\Big(a(s) + \sum_{j=1}^k \mathrm{d}Z_j(s), b(s) + \sum_{j=1}^k \exp(x_j^{\mathrm{t}}\beta)\Big).$$

(e) Show that

$$\widehat{G}(t \,|\, \beta) = \mathrm{E}\,\{G(t) \,|\, \mathrm{data}, \beta\} = \int_0^t \frac{a(s)\,\mathrm{d}s + \sum_{j=1}^k \mathrm{d}Z_j(s)}{b(s) + \sum_{j=1}^k \exp(x_j^{\mathrm{t}}\beta)},$$

which may also be written out as an integral over intervals with zero jumps plus the component summed over the precise jump times.

(f) Then work out an expression for the posterior density of β. It may be required to set up an MCMC scheme for simulation from this $\pi(\beta \,|\, \mathrm{data})$. This leads in particular to the Bayes estimator

$$\widehat{G}(t) = \int \widehat{G}(t \,|\, \beta)\pi(\beta \,|\, \mathrm{data})\,\mathrm{d}\beta.$$

(g) [xx just a little more here. point to similar story for Beta process. also link to Nils Beta process with Cox regression type data. xx]

32. The Gamma process, with a Poisson process, with a marks process

[xx something here. actuarial statistics is fond of compound Poisson processes, with total claim size

$$W(t) = \sum_{T_j \le t} \xi_j,$$

summed over claim times $T_1 < T_2 < \cdots$. may here build a Bayesian nonparametrics story, with a prior for the nonhomogeneous Poisson process etc. xx]

XX. Bernshteĭn–von Mises theorems

[xx to be written down xx] first for Dirichlet, with fairly clear details. but it takes the Donsker and Kolmogorov thing. then for Beta processes.

XX. The Bayesian bootstrap

well

XX. Hjort's informative Bayesian bootstrap

well

XX. Simulating realisations of a Gaussian process

[xx to be written down and polished. xx] We say that $Z = \{Z(x)\colon x \in [a, b]\}$ is a Gaussian process if all its finite-dimensional distributions are Gaussian. In particular, $Z(x)$ is normal, say $\mathrm{N}(m(x), \sigma^2(x))$, and $(Z(x), Z(x'))$ is binormal, with correlation say $\rho(x, x')$.

(a) Explain why giving the mean function $m(x)$, the standard deviation function $\sigma(x)$, and the correlation function $\rho(x, x')$, is actually sufficient to determine the full distribution of Z.

– For some Gaussian processes there are specialised techniques making it easier-than-brute-force to simulate realisations. In general, however, we can't do much better than brute-force, which means simulating $Z^* = (Z(x_1), \ldots, Z(x_n))$, for a fine enough grid $x_1, \ldots, x_n$. The implied distribution is multinormal,

$$Z^* \sim \mathrm{N}_n(\xi, \Sigma),$$

with ξ having components $m(x_i)$ and Σ of size $n \times n$ and with components $\sigma(x_i)\sigma(x_j)\rho(x_i, x_j)$. Thus simulating from Z becomes practically the same as being able to simulate from a general multinormal $\mathrm{N}_n(0, \Sigma)$.

(c) The R algorithm `rmvnorm` may be used, for simulating from a given multinormal, but my impression is that it might not work well for higher n. A general technique that can be used here is as follows. First, find a unitary matrix P such that

$$P \Sigma P^{\mathrm{t}} = D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n).$$

A unitary or orthonormal matrix Q is one having the property that $QQ^{\mathrm{t}} = I = Q^{\mathrm{t}}Q$. Finding such a P, for given Σ, can be achieved via the `eigen` algorithm in R. Then define, compute, and store the root-matrix

$$\Sigma^{1/2} = P D^{1/2} P^{\mathrm{t}}, \quad \text{with } D^{1/2} = \mathrm{diag}(\lambda_1^{1/2}, \ldots, \lambda_n^{1/2}).$$

Verify that $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$. Then use

$$z = \Sigma^{1/2}\varepsilon, \quad \text{where } \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{t}} \sim \mathrm{N}_n(0, I_n),$$

i.e. these are independent standard normals. Verify that z then has the desired multinormal distribution. Check that the following R code works:

```
squareroot <- function(Sigma)
{rootL <- 0*Sigma
 diag(rootL) <- sqrt(eigen(Sigma, symmetric = T)$values)
 P <- eigen(Sigma, symmetric = T)$vectors
 P %*% rootL %*% t(P)}
```

(d) Consider an Ornstein–Uhlenbeck process Z on $[0, 10]$, with mean zero and covariance function $\mathrm{cov}\{Z(x), Z(x')\} = \exp(-a|x - x'|)$, say with $a = 1.3579$. Simulate and plot 50 realisations of the Z process.

XX. Bayesian Kriging

[xx to be written out and polished. xx] Suppose there is a continuous process $Z(x)$ on $[0, 1]$, which we have observed only in a small number of locations. How can we estimate $Z(x)$ where we have not seen it, along with a measure of precision? This translates to 'spatial interpolation' and so on, and with Kriging one of its names (from the Master Thesis of Danie Gerhardus Krige, 1919–2013, a South African geostatistician).

Suppose $Z(x)$ is Gaussian, with constant mean function a, and covariance function

$$\mathrm{cov}\{Z(x), Z(x')\} = \sigma^2 K_0(|x - x'|),$$

where $K_0(r)$ is the correlation function. This means a stationary setup, where $Z(x)$ and $Z(x + r)$ have a correlation independent of position x.

(a) Use $a = 1.3579$ and $K_0(r) = \exp(-\lambda r)$, with $\lambda = 2.222$. Simulate realisations of $Z(x)$, for $x \in [0, 1]$. Take $\sigma = 1$ here (but later on we may tinker with this precision parameter).

(b) Assume now that the scientific team has come back from their expedition and report that for positions $0.11, 0.22, 0.33, 0.77, 0.88$, they found that $Z(x)$ is equal to $0.99, 1.33, 1.66, 1.22, 1.11$ (yes, I'm inventing this, and will search for a real application later on). Find expressions giving the posterior distribution of $Z = \{Z(x)\colon x \in [0, 1]\}$.

(c) Find in particular an expression for

$$\widehat{Z}(x) = \mathrm{E}\,\{Z(x) \,|\, \mathrm{data}\},$$

and plot that curve.

(d) Find also a formula for

$$\widehat{\kappa}(x)^2 = \mathrm{Var}\,\{Z(x) \,|\, \mathrm{data}\},$$

and plot the 90% prediction confidence band

$$\widehat{Z}(x) \pm 1.645\,\widehat{\kappa}(x).$$

(e) Simulate say 50 realisations from the distribution of $Z = \{Z(x)\colon x \in [0, 1]\}$ given the data, and plot them.

XX. Bayesian nonparametric regression

[xx to be written out and polished. xx] model is

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

where the ε_i are i.i.d. from $N(0, \sigma^2)$. Suppose $m(x)$ is Gaussian, with mean function $m_0(x)$ and covariance function for the form $\sigma_0^2 K_0(|x - x'|)$, with a given correlation function $K_0(r)$.

Then find expressions for the conditional mean, the conditional variance, and conditional covariance, of the process $m(x)$, given the data (x_i, y_i).

XX. A nonparametric minimax estimator for an unknown mean

[xx to be polished. xx] Suppose i.i.d. observations $X_1, \ldots, X_n$ are available from an unknown distribution P on the unit interval $[0, 1]$. We only know that $P \in \mathcal{M}$, the set of all distribution functions on $[0, 1]$. We wish to estimate the mean $\theta = \int x \,\mathrm{d}P(x)$, with quadratic loss function $(\widehat{\theta} - \theta)^2$. Below I exhibit a minimax estimator (mm) for this problem. Lehmann (1951, Mimeographed Lecture Notes on the Theory of Point Estimation from Berkeley) did this, with similar arguments.

Lehmann also claimed in these Lecture Notes that the estimator given in (mm) is admissible – but his argument was not correct, as it turns out. Nils Lid Hjort's perhaps First Theorem was to prove that the (mm) estimator *is* nonparametrically admissible (in an exam project for Erik N. Torgersen on decision theory, 1975, which consisted in reading, comprehending, and presenting the Ferguson 1973 paper for the exam marker).

(a) Work out the risk function for the direct sample average $\bar{X}$:

$$R(\bar{X}, P) = (1/n)\sigma(P)^2, \quad \text{with} \quad \sigma(P)^2 = \int \{x - \theta(P)\}^2 \, \mathrm{d}P(x).$$

(b) Show that the variance $\sigma(P)^2$ is maximal, over all distributions on $[0, 1]$, when P is concentrated in the end-points 0 and 1, with equal probabilities $\frac{1}{2}$, $\frac{1}{2}$. Hence

$$\max\{R(\bar{X}, P) \colon P \in \mathcal{M}\} = (1/n)(1/4).$$

(c) Then consider the cool enough estimator

$$\widehat{\theta} = \frac{1}{\sqrt{n}+1}\tfrac{1}{2} + \frac{\sqrt{n}}{\sqrt{n}+1}\bar{X}. \qquad \text{(mm)}$$

Show that its risk function can be written

$$\begin{aligned} R(\widehat{\theta}, P) &= \Big(\frac{\sqrt{n}}{\sqrt{n}+1}\Big)^2 \frac{\sigma(P)^2}{n} + \Big\{\frac{\sqrt{n}}{\sqrt{n}+1}\theta(P) + \frac{1}{\sqrt{n}+1}\tfrac{1}{2} - \theta(P)\Big\}^2 \\ &= \frac{1}{(\sqrt{n}+1)^2}\big[\sigma(P)^2 + \{\tfrac{1}{2} - \theta(P)\}^2\big]. \end{aligned}$$

(d) Show that the max risk for the (mm) estimator is

$$\max\{R(\widehat{\theta}, P) \colon P \in \mathcal{P}\} = \frac{1/4}{(\sqrt{n}+1)^2}.$$

(e) Then show that it is minimax (Lehmann 1951, Berkeley Notes, precursor to the Theory of Point Estimation book). [xx fill in, not too hard. xx]

(f) Then show that it is actually also admissible; Lehmann made an error here, in these 1951 Berkeley Notes, but Nils 1976 has several proofs. [xx i fill in one of these, perhaps as a separate exercise; this is considerably harder than proving minimaxity. xx]

References

Billingsley, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

Clauset, A. (2017). The enduring threat of a large interstate war. Technical Report, One Earth Foundation.

Clauset, A. (2018). Trends and fluctuations in the severity of interstate wars. *Science Advances* **4**, xx-xx.

Cunen, C., Hjort, N.L., and Nygård, H. (2018). Statistical sightings of better angels. To be submitted for publication within 2-iv-2018.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209–230.

Ferguson, T.S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics* **2**, 615–629.

Ferguson, T.S. and Klass, M.J. (1972). A representation of independent increment processes without Gaussian components. *Annals of Mathematical Statistics* **43**, 1634–1643.

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference.* Cambridge University Press, Cambridge.

Hjort, N.L. (1976). *The Dirichlet Process Applied to Some Nonparametric Problems.* Cand. real. thesis [in Norwegian], Department of Mathematics, Nordlysobservatoriet, University of Tromsø.

Hjort, N.L. (1985). Discussion contribution to P.K. Andersen and Ø. Borgan's 'Counting process models for life history data: A review'. *Scandinavian Journal of Statistics* **12**, xx–xx.

Hjort, N.L. (1985). An informative Bayesian bootstrap. Technical Report, Department of Statistics, Stanford University.

Hjort, N.L. (1986). Discussion contribution to P. Diaconis and D. Freedman's paper 'On the consistency of Bayes estimators'. *Annals of Statistics* **14**, 49–55.

Hjort, N.L. (1990). Nonparametric Bayes estimators based on Beta processes in models for life history data. *Annals of Statistics* **18**, 1259–1294.

Hjort, N.L. (1991). Bayesian and empirical Bayesian bootstrapping. Statistical Research Report, Department of Mathematics, University of Oslo.

Hjort, N.L. (2003). Topics in nonparametric Bayesian statistics [with discussion]. In *Highly Structured Stochastic Systems* (eds. P.J. Green, N.L. Hjort, S. Richardson). Oxford University Press, Oxford.

Hjort, N.L. (2018). Towards a More Peaceful World [Insert '!' or '?' Here]. FocuStat Blog Post.

Hjort, N.L. (2010). An invitation to Bayesian nonparametrics. In *Bayesian Nonparametrics* (by Hjort, N.L., Holmes, C.C., Müeller, P., and Walker, S.G.), 1–21.

Hjort, N.L., Holmes, C.C., Müller, P., and Walker, S.G. (2010). *Bayesian Nonparametrics.* Cambridge University Press, Cambridge.

Hjort, N.L. and Ongaro. A. (2005). Exact inference for random Dirichlet means. *Statistical Inference for Stochastic Processes* **8**, 227–254.

Hjort, N.L. and Ongaro. A. (2006). On the distribution of random Dirichlet jumps. *Metron* **LXIV**, 61–92.

Hjort, N.L. and Petrone, S. Nonparametric quantile inference using Dirichlet processes. In *Festschrift for Kjell Doksum* (ed. V. Nair).

Hjort, N.L. and Walker, S.G. (2009). Quantile pyramids for Bayesian nonparametrics. *Annals of Statistics* **37**, 105–131.

Lehmann, E.L. (1951). Notes on the Theory of Point Estimation. (Mimeographed by C. Blyth.) Department of Statistics, University of Berkeley, California.

Müller, O., Quintana, F.A., Jara, A., and Hanson, T. (2015). *Bayesian Nonparametric Data Analysis.* Springer-Verlag, Berlin.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.

Sethuraman, J. and Tiwari, R. (1982). Convergence of Dirichlet measures and the interpretation of their parameter. In: *Proceedings of the Third Purdue Symposium on Statistical Decision Theory and Related Topics* (eds. S.S. Gupta and J. Berger), 305–315. Academic Press, New York.

Wolpert, R.L. and Ickstadt, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika* **85**, 251–267.