Exercises and Lecture Notes, STK 9190, Spring 2018

Version 0.31, 9-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 $Z \sim \text{Gamma}(a, b)$ if its density is

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

Here a and b are positive parameters. Show that

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

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

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

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

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

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

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

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

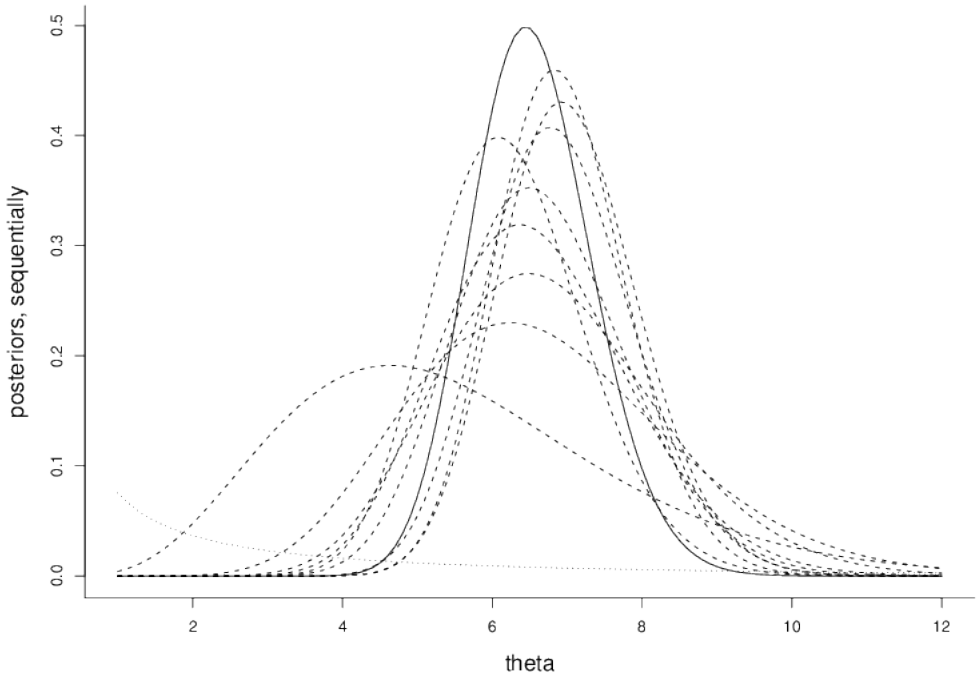

Figure 0.1: Eleven curves are displayed, corresponding to the Gamma$(0.1, 0.1)$ intial prior density for the Poisson parameter θ along with the ten updates following each of the observations 6, 8, 7, 6, 7, 4, 11, 8, 6, 3.

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 $\pi(P \,|\, \text{data})$; (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 $\{\text{reject}, \text{accept}\}$ if a hypothesis test is being carried out; or the set of all intervals, if the statistician needs a confidence interval.

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

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

(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

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

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

(b) Show that

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

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

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

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

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

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

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

(d) Infer from this that

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 $\rho, \alpha, \beta, \gamma$.

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 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 $P \sim \text{Dir}(aP_0)$ 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 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 $P \sim \text{Dir}(aP_0)$ 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

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

with mean and variance

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

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

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

Show that

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

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

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

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

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

7. The basic updating theorem for the Dirichlet process

Suppose $P \sim \text{Dir}(aP_0)$, 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 $\text{Dir}(aP_0)$ 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

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

(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 normalisting 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$(2, 1)$.

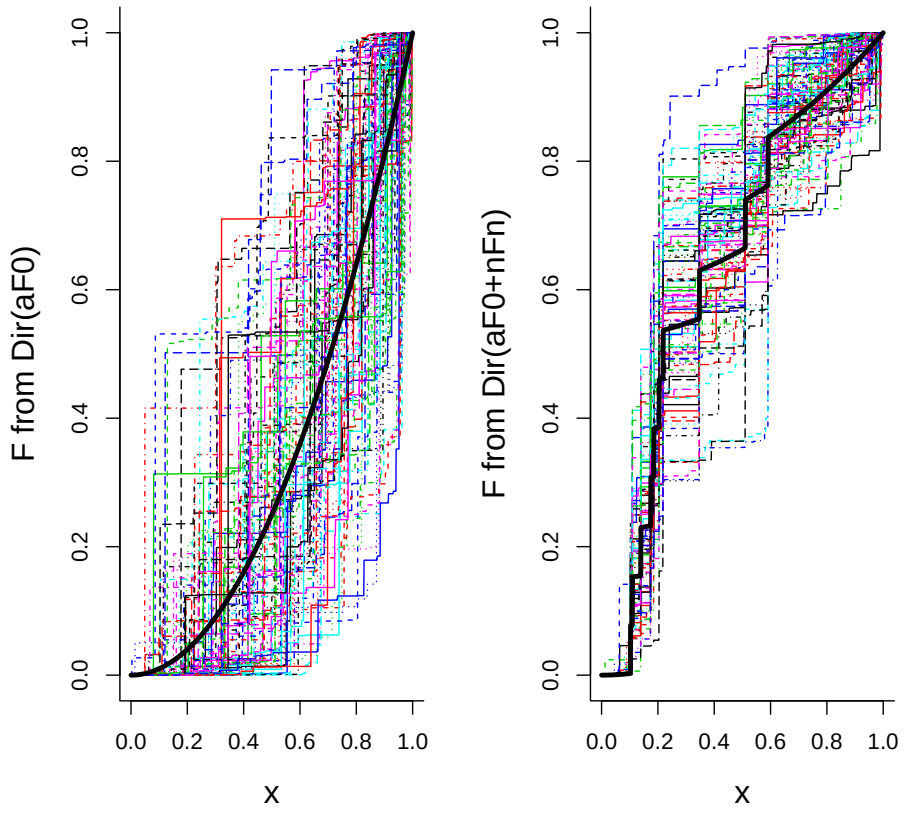

Figure 0.2: 100 simulations of F from the Dir(aF_0) prior (left); then 100 simulations of F from the Dir$(aF_0 + nF_n)$ posterior (right), with the $n = 10$ data points of Exercise 8. The fat black curves are the prior mean and posterior mean, respectively.

(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

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

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

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

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

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

(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 $\int_0^1 x\,\text{d}F(x)$, 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,

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

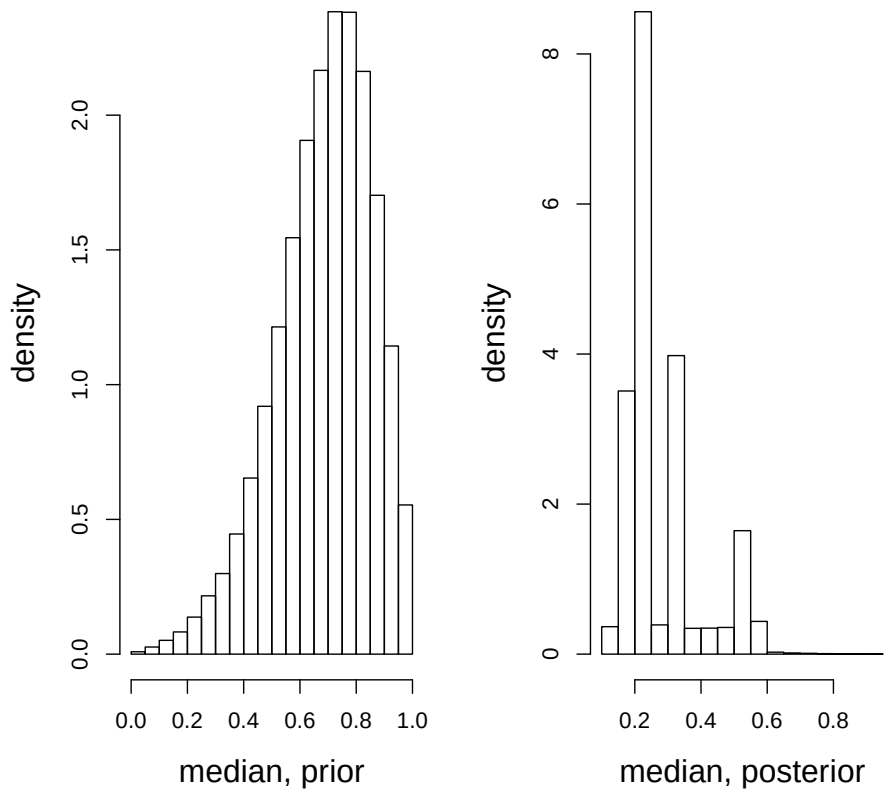

Figure 0.3: For the random median $\mu = \min\{x\colon F(x) \geq \frac{1}{2}\}$, I give histograms of its distribution, for the prior (left) and the posterior (right), based on 10^4 simulations, for each case.

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 of our common future, and the controversies.

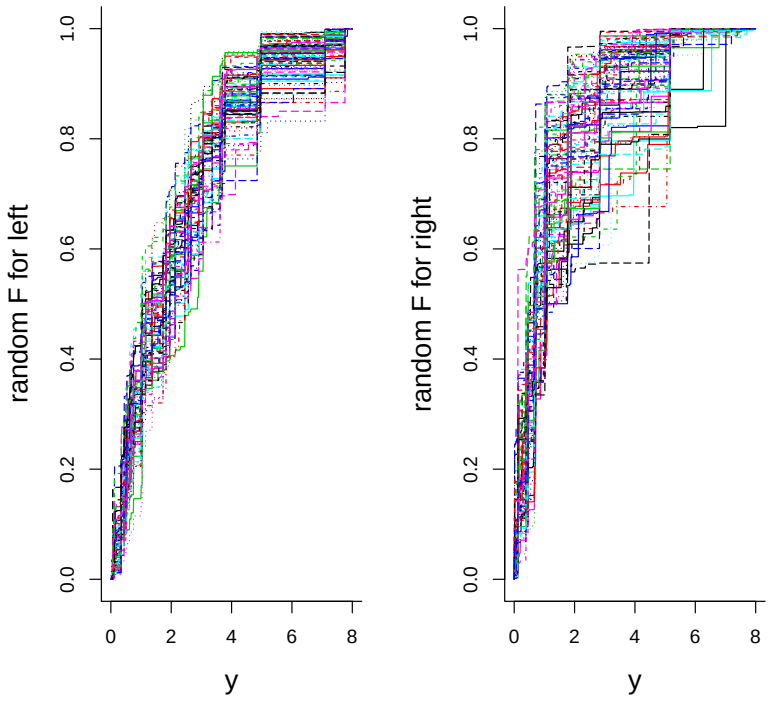

Figure 0.4: 100 simulated realisations of F_L, representing the past up to Vietnam (left), and 100 realisations of F_R, representing post-Vietnam period (right). The scale here is that of $y = \log(z/7061)$, for all wars with battle death counts at least 7061.

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 the power law tail property that

(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 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. 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 $D_n/\log n \to a$, 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.

12. 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]

13. 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) 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 $\sigma_0^2 = \int (g-\theta_0)^2 \, \mathrm{d}P_0$.

(d) well

(e) well

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

well

XX. 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.

XX. 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 $Z(t) - Z(s) \sim \text{Gamma}(M(t) - M(s), 1)$. Existence of such a process is not entirely obvious, but one is of course helped by the fact that

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

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

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

(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)

(e)

(f)

XX. The extended gamma process

[xx In the course of this exercise I build a more general process, which I term an extended gamma process. xx] We start with independent and inherently small gammas,

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

and from these form the partial sum process

$$Z_m(t) = \sum_{i/m \le t} G_{m,i}.$$

Show that the Laplace transform converges properly:

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

which indeed tends to

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

XX. The extended gamma process with a Poisson process

well

XX. The jumps of a gamma process

[xx something from Hjort and Ongaro (2006, Metron). xx]

XX. The Beta process

Hjort (1985, SJS) 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 $A_0(t) = \int_0^t a_0(s)\,dds$. For given m, let $B_{m,1}, B_{m,2}, \ldots$ be independent Beta random variables, with

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

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

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

Show also that

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

(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

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

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

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

(c)

(d)

(e)

XX. The Beta process for survival data

[xx write down and polish. xx] with conjugacy property and updating. the Bayes estimator for the cumulative hazard is

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

with link to the Nelson–Aalen estimator. also, the Bayes estimator for the survival function is

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

with link to Kaplan–Meier. Can simulate from A and S given data, and read off what we might wish for from these, like the posterior median

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

XX. 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 them 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 distributions. 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, $A_w \sim \text{Beta}(c_w, A_{0,w})$ and $A_m \sim \text{Beta}(c_m, A_{0,m})$.

(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 (eq1) over (eq2), in terms of setup, modelling, prior to posterior, analysi, 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

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

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

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

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?

XX. The Bernoulli process and the Poisson process

[xx to be written down. xx] showing that a Bernoulli construction becomes a Poisson process.

XX. The Beta process with a Bernoulli process

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

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

where

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

to become a Nils-Emil story, with the Oslo Police tweets. with variation: extended Gamma. perhaps with marks or covariates.

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

well

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

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

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.

(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 written out and polished. should be cool. xx] observations $x_1, \ldots, x_n$ are i.i.d. from some F on the unit interval. wish to estimate $\theta = \int x \,\mathrm{d}F(x)$, with quadratic loss function $(\widehat{\theta} - \theta)^2$.

risk function for the direct sample average $\bar{x}$:

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

find the max-risk. Then the cool enough

$$\widehat{\theta} = \frac{1}{\sqrt{n}+1}\tfrac{1}{2} + \frac{\sqrt{n}}{\sqrt{n}+1}\bar{x}.$$

find the risk function and its max value. Then show that it is minimax (Lehmann 1951, Berkeley Notes, precursor to the Theory of Point Estimation book). Then show that it is actually also admissible; Lehmann made an error her, in these 1951 Berkeley Notes, but Nils 1976 has several proofs.

References

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. (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). [xx intro chapter to HHMW. xx]

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.

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.