Part III: Bayesian Nonparametrics

focustat
FOCUS DRIVEN STATISTICAL
INFERENCE WITH COMPLEX DATA

Nils Lid Hjort

Department of Mathematics, University of Oslo

Geilo Winter School, January 2017

[Note: This is the pdf version of the 2 x 45 minutes Nils Talk III I gave at the Geilo Winter School, January 2017. In my actual presentation I of course did both of (a) saying quite a bit more than is on the page and (b) skidding semi-quickly over chunks of the material, including parts of the mathematics, complete with the usual mixture of hand-waving, glossing over technicalities, and swiping of details under imaginary carpetry. The pdf notes themselves are meant to be decently coherent, though, and may be suitable for study.]

Traditional Bayesian analysis: with data y from model with likelihood $L(\theta)$, the Bayes formula takes the pre-data prior $\pi(\theta)$ to the post-data posterior

$$\pi(\theta \,|\, \mathrm{data}) = \frac{\pi(\theta)L(\theta)}{\int \pi(\theta')L(\theta')\,\mathrm{d}\theta'}.$$

This requires well-defined densities (w.r.t. suitable measures), and in essence that θ is finite-dimensional.

Bayesian nonparametrics: about attempts to carry out such schemes, from pre-data to post-data, in infinite- or very high-dimensional models.

This is a tall order: conceptually; elicitation-wise; mathematically (distributions over very big spaces); operationally (there is no direct Bayes theorem); computationally (MCMC with 1000 parameters?).

Priors and posteriors for densities, regression functions, hazard rates, big hierarchical models, etc.

Plan & outline

A What is it?

B From the Dirichlet distribution to the Dirichlet process

C The Beta process (e.g. for survival analysis)

D Bernshteĭn–von Mises theorems (sometimes not)

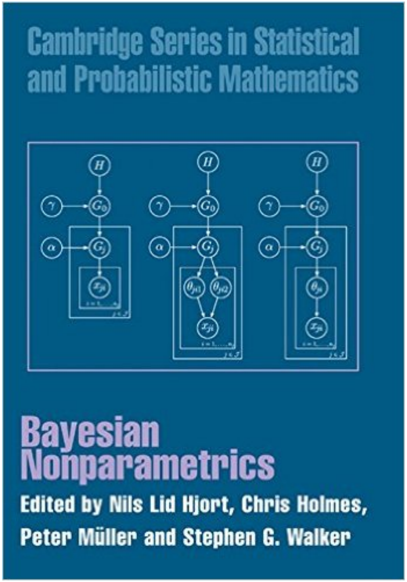
E Stationary time series

F Bayesian nonparametrics for quantile analysis

G Clustering models

H Other issues & concluding remarks

Hjort, Holmes, Müller, Walker, 2010



A: What is it?

Well, what is it^c? Theorem: If the world is frequentist or Bayes, and parametric or nonparametric, then

$$\mathrm{IV} = (\mathrm{I} \cup \mathrm{II} \cup \mathrm{III})^c.$$

	frequentist	Bayes
parametric	I	II
nonparametric	III	IV

I: Smallish finite models, estimation and inference for aspects of θ.

II: Smallish finite models, estimation and posterior inference, via prior $\pi(\theta)$ (this was all of Bayes inference, from c. 1774 to c. 1973).

III: Bigger models, density estimation, nonparametric regression, confidence bands, etc.

IV: Priors and posteriors for random functions, bigger structures, hierarchies of hierachies, ...

Bayesian nonparametrics invites constructions for 'approximately normal', 'approximately linear regression', etc.

With $\psi_1, \psi_2, \ldots$ orthogonal functions on $[0, 1]$, like $\psi_j(u) = \sqrt{2}\cos(j\pi u)$, try

$$f(y) = f(y, \theta)\exp\Big\{\sum_{j=1}^{100} a_j\psi_j(F(y, \theta))\Big\}\Big/ c_{100}(a_1, \ldots, a_{100}),$$

with a prior on θ along with $a_j \sim \mathrm{N}(0, \tau^2/j^2)$. Data will tell us (via lots o' MCMC) how close the real f is to the parametric start.

Approximately linear regression with approximately normal errors:

$$y_i = a + bx_i + \sum_{j=1}^{100} \gamma_j h_j(x_i) + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

with perhaps $\gamma_j \sim \mathrm{N}(0, \kappa^2/j^2)$, where the ε_i are from $f \approx \mathrm{N}(0, \sigma^2)$.

Fascinating and promising – but raises a long list of questions.

A de Finetti theorem

Why should we go for Bayesian nonparametrics? – Apart from 'it works, and can solve big problems', there's a mathematical-probabilistic argument:

Consider $y_1, y_2, \ldots$, a sequence of observations, with the exchangeability property:

$$(y_1, y_2, y_3, y_4, y_5) \sim (y_3, y_1, y_5, y_4, y_2),$$

i.e. have the same distribution – and similarly for all permutations, and all lengths. Then there is a de Finetti measure π, on the set of all distributions P, such that

$$\Pr\{y_1 \in A_1, \ldots, y_n \in A_n\} = \int P(A_1) \cdots P(A_n)\, \pi(\mathrm{d}P)$$

for all $A_1, \ldots, A_n$, and all n.

So there is a (nonparametric) prior behind what you see (whether you knew or not), and $y_1, y_2, \ldots$ are i.i.d. given P.

B: The Dirichlet process: from finite to infinite

I begin with two boxes (1-or-0 measurements): With $y \sim \mathrm{Bin}(n, p)$,

$$f(y \mid p) \propto p^y (1-p)^{n-y},$$

and with $p \sim \mathrm{Beta}(a, b)$,

$$p \mid y \propto p^{a-1}(1-p)^{b-1} p^y (1-p)^{n-y} = p^{a+y-1}(1-p)^{b+n-y-1},$$

which means $p \mid \mathrm{data} \sim \mathrm{Beta}(a+y, b+n-y)$:

$$\widehat{p} = \mathrm{E}(p \mid y) = \frac{a+y}{a+b+n} = \frac{a+b}{a+b+n} p_0 + \frac{n}{a+b+n} \frac{y}{n},$$

$$\mathrm{Var}\,(p \mid y) = \frac{1}{a+b+n} \widehat{p}(1-\widehat{p}).$$

Thomas Bayes did this, with $(a, b) = (1, 1)$, i.e. a uniform prior – not in *Divine Benevolence, or an Attempt to Prove That the Principal End of the Divine Providence and Government is the Happiness of His Creatures* (1731), but in the other one (1763).

Then k boxes: from binomial to multinomial. With $(y_1, \ldots, y_k)$ the number of cases of types $1, \ldots, k$, the likelihood is

$$\frac{n!}{y_1! \cdots y_k!} p_1^{y_1} \cdots p_{k-1}^{y_{k-1}} (1 - p_1 - \cdots - p_{k-1})^{y_k},$$

if the n trials are independent with the same probabilities $p_1, \ldots, p_k$ each time.

This calls on the Dirichlet distribution, $\mathrm{Dir}(a_1, \ldots, a_k)$:

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

on the simplex of $(p_1, \ldots, p_{k-1})$.

Multiplying prior and likelihood:

$$(p_1, \ldots, p_k) \,|\, (y_1, \ldots, y_k) \sim \mathrm{Dir}(a_1 + y_1, \ldots, a_k + y_k).$$

So we can pass from prior $\mathrm{Dir}(a_1, \ldots, a_k)$ to posterior $\mathrm{Dir}(a_1 + y_1, \ldots, a_k + y_k)$ by simply adding the observed counts for the k boxes. We have

$$\widehat{p}_j = \mathrm{E}(p_j \mid \mathrm{data}) = \frac{a_j + y_j}{a + n},$$
$$\widehat{\sigma}_j^2 = \mathrm{Var}\,(p_j \,|\, \mathrm{data}) = \frac{1}{a + n + 1}\widehat{p}_j(1 - \widehat{p}_j),$$

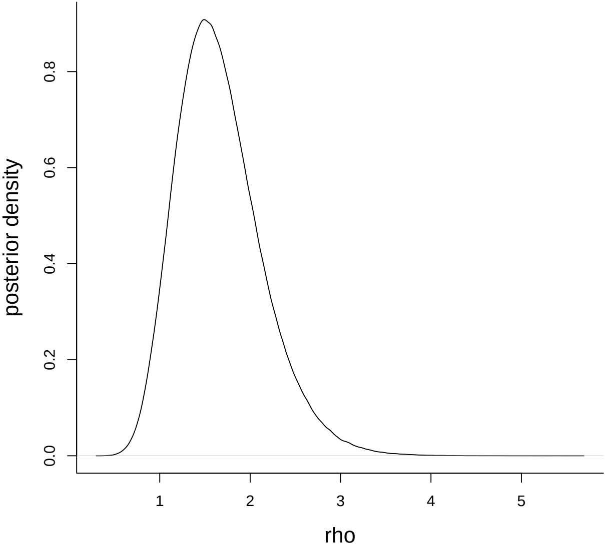
with $a = a_1 + \cdots + a_k$.

Easy to use, via simulation:

$$(p_1, \ldots, p_k) = \Big(\frac{G_1}{G_1 + \cdots + G_k}, \ldots, \frac{G_k}{G_1 + \cdots + G_k}\Big),$$

with $G_j \sim \mathrm{Gamma}(a_j, 1)$ (for the prior), or $G_j \sim \mathrm{Gamma}(a_j + y_j, 1)$ (for the posterior).

Example: I throw my die 60 times and get $8, 8, 7, 13, 9, 15$. Is p_6 bigger than it should be? With prior $\mathrm{Dir}(2, 2, 2, 2, 2, 2)$ this is the posterior for $\rho = p_6/(p_1 \cdots p_5)^{1/5}$, and $\Pr(\rho > 1 \mid \mathrm{data}) = 0.947$:



Then from k boxes to the infinite full-space process: We're helped by this collapsibility lemma: If

$$(p_1,\ldots,p_{10}) \sim \mathrm{Dir}(a_1,\ldots,a_{10}),$$

then

$$\begin{aligned}(p_1+p_2, p_3+p_4+p_5, p_6, p_7+p_8+p_9+p_{10})\\ \sim \mathrm{Dir}(a_1+a_2, a_3+a_4+a_5, a_6, a_7+a_8+a_p+a_{10}),\end{aligned}$$

etc. With P_0 a distribution on the sample space S, we say that $P \sim \mathrm{Dir}(aP_0)$, a Dirichlet process with parameter aP_0, if

$$(P(A_1),\ldots,P(A_k)) \sim \mathrm{Dir}(aP_0(A_1),\ldots,aP_0(A_k))$$

for each partition $A_1,\ldots,A_k$.

Existence is non-trivial (Ferguson, 1973, Doksum 1974). We have

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

The Dirichlet process has various uses as a probabilistic model for a random distribution. It is also well-suited for inference after observations from unknown distribution. Master lemma says that if (i) $P \sim \mathrm{Dir}(aP_0)$ and (ii) $y_1, \ldots, y_n \,|\, P$ are i.i.d. from P, then

$$P \,|\, \mathrm{data} \sim \mathrm{Dir}(aP_0 + \delta(y_1) + \cdots + \delta(y_n)),$$

i.e. with posterior measure $aP_0 + nP_n$, with $P_n = n^{-1} \sum_{i=1}^n \delta(y_i)$, the empirical measure with point mass $1/n$ in each data point.

In particular:

$$\widehat{P}(A) = \mathrm{E}\,\{P(A) \,|\, \mathrm{data}\} = \frac{a}{a+n} P_0(A) + \frac{n}{a+n} P_n(A),$$

$$\mathrm{Var}\,\{P(A) \,|\, \mathrm{data}\} = \frac{1}{a+n+1} \widehat{P}(A)\{1 - \widehat{P}(A)\}.$$

May also form confidence bands for $P(A)$, and may e.g. simulate 1000 realisations from $P \,|\, \mathrm{data}$, from which we can read off posterior for any $\theta(P)$.

How to simulate a $\mathrm{Dir}(aP_0)$, over the sample space S?

(i) May discretise, sample space cut into tiny pieces, and use a finite-dimensional Dirichlet $(aP_0(A_1), \ldots, aP_0(A_{1000}))$.

(ii) May write $P(A) = G(A)/G(S)$, with G a Gamma process with independent pieces over disjoint sets, $G(A) \sim \mathrm{Gamma}(aP_0(A))$.

(iii) Via the Tiwari–Sethuraman representation theorem:

$$P = \sum_{j=1}^{\infty} p_j \delta(\theta_j),$$

where the random locations $\theta_1, \theta_2, \ldots$ are i.i.d. from P_0, and where the random stick-breaking probabilities are

$$p_1 = B_1, p_2 = (1 - B_1)B_2, p_3 = (1 - B_1)(1 - B_2)B_3, \ldots,$$

with $B_1, B_2, B_3, \ldots$ i.i.d. $\mathrm{Beta}(1, a)$.

Note that the random distribution P is discrete.

C: The Beta process

For survival analysis, consider life-times T, with distribution F, and cumulative hazard rate function A:

$$\mathrm{d}A(s) = \frac{\mathrm{d}F(s)}{F[s,\infty)} = \Pr\{\text{die in } [s, s+\mathrm{d}s] \,|\, \text{still alive at } s\}.$$

The survival curve is

$$S(t) = \Pr\{T \geq t\} = \prod_{[0,t]} \{1 - \mathrm{d}A(s)\}.$$

Hjort (1985, 1990) constructs a Beta process, with independent increments: With $A_0(\cdot)$ the prior mean function, and $c(\cdot)$ a prior strength function,

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

The existence is non-trivial, since a sum of Betas is not a Beta; a fine-limit-argument is needed (cf. Lévy processes).

As for the Dirichlet process, also the Beta process has various probabilistical uses in studies of random transitions phenomena, and as the de Finetti measure of the Indian Buffet Processes.

They are particularly well-suited for survival analysis. Survival data (t_i, δ_i), with $t_i = \min(t_i^0, z_i)$ and $\delta_i = I\{t_i^0 \le z_i\}$ the indicator for non-censoring: The classical nonparametric estimators for cumulative hazard and survival are

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

the Nelson–Aalen and Kaplan–Meier estimators. Here $Y(s)$ is the number at risk at time s and $\mathrm{d}N(s)$ the number of those dying in $[s, s + \mathrm{d}s]$.

With the Beta process, we reach Bayesian nonparametric extensions of these.

If $A \sim \mathrm{Beta}(c, A_0)$, then

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

with

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

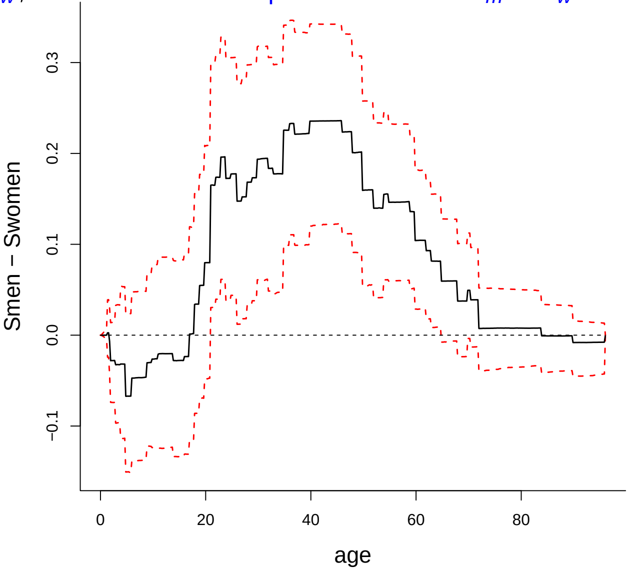
The Bayes estimator for survival is

$$\widehat{S}(t) = \mathrm{E}\{S(t) \,|\, \mathrm{data}\} = \prod_{[0,t]} \Big\{1 - \frac{c(s)\,\mathrm{d}A_0(s) + \mathrm{d}N(s)}{c(s) + Y(s)}\Big\}.$$

With $c(s) \to 0$ we do not trust the prior, and we get the Nelson–Aalen and Kaplan–Meier estimators.

May also simulate 1000 posterior realisations from the distributions $A \,|\, \mathrm{data}$ and $S \,|\, \mathrm{data}$, and read off relevant features and probabilities.

Example: Analyse life-lengths from ancient Egypt, for 82 men and 59 women [see Nils talk I], via posterior distribution for $\Pr(T_m \geq t) - \Pr(T_w \geq t)$. I start with Beta process priors for A_m and A_w, and simulate 5000 posterior curves $S_m - S_w$.



Sir David Cox (b. 1924) is an Eternal Guru of Statistics (the first ever winner of the International Prize in Statistics, 2017). His most important invention (from 1972) is the hazard rate regression model

$$\alpha_i(s) = \alpha(s) \exp(x_i^{\mathrm{t}} \beta)$$

along with deep methodology for handling such and similar models, starting with survival data (t_i, δ_i, x_i).

The canonical semiparametric Bayesian extension of this method (Hjort, 1990) starts with

$$1 - \mathrm{d}A_i(s) = \{1 - \mathrm{d}A(s)\}^{\exp(x_i^{\mathrm{t}} \beta)},$$

a prior for β, and $A \sim \mathrm{Beta}(c, A_0)$.

There is a long list of further generalisations and uses of the Beta process in models with transtions over time (medicine, demography, biology, event history analysis, etc.).

D: Bernshteĭn–von Mises theorems (do not always hold)

For ordinary parametric inference, there's a comforting general theorem saying frequentist and Bayesian inferences 'agree in the end', with enough data.

Suppose n data points or vectors have been observed, from a model $f(y, \theta)$, with $\widehat{\theta}$ the maximum likelihood estimator. First,

$$\sqrt{n}(\widehat{\theta} - \theta_0) \to_d \mathrm{N}_p(0, J^{-1}).$$

Second, having started with any prior, the posterior $\pi(\theta \,|\, \mathrm{data})$ is such that

$$\sqrt{n}(\theta - \widehat{\theta}) \,|\, \mathrm{data} \to_d \mathrm{N}(0, J^{-1}).$$

So frequentist and (every) Bayesian inference tend to agree, with $\widehat{\mu} \pm 1.96\,\widehat{\kappa}/\sqrt{n}$ as the 95% confidence or credibility interval, etc.

The nonparametric world is bigger and scarier (?), however. Various reasonable-looking nonparametric Bayesian schemes don't work – lack of consistency, wrong coverage, etc.

E: Stationary time series

Bayesian nonparametrics for covariance functions with application to time series – the following reports briefly on joint work with Gudmund Hermansen.

- Covariance and correlation functions: via spectral measure F
- Prior on $F \Rightarrow$ prior on covariances and correlations
- F a Dirichlet: $C(h) = \int_0^\pi \cos(h\omega)\,\mathrm{d}F(\omega)$ is a valid correlation sequence
- Stationary time series: full nonparametric Bayes inference
- Other spatial and spatial-temporal models: prior, posterior, Bayesian inference ok; but fewer hard results

Plan:

- 1. Priors for stationary Gaussian time series – Spectral representation: F first, then C
- 2. Frequentist analysis – Periodogramme, cumulative, Brownian motion
- 3. Exact and approximate Bayesian updating – Whittle approximation, MCMC
- 4. Limit theorems and Bernshteĭn–von Mises
- 5. [Illustration: sun spots, etc.; not here (!)]

E1. Priors for stationary Gaußian time series

Let $Y_1, Y_2, \ldots$ be a zero-mean stationary Gaußian time series with unknown covariance function

$$C(h) = C(|i-j|) = \mathrm{cov}(Y_i, Y_j) \quad \text{for } |j-i| = h.$$

Wish to use Bayesian nonparametrics for modelling $C(\cdot)$ – and hence any function of $C(\cdot)$, i.e. any function of the $n \times n$ covariance matrix.

If we manage, this leads to full Bayesian inference for a time series with 'uncertain covariance function'. Can then also answer predictive questions, like the Geiloesque

$$\alpha = \Pr\{Y_{n+1} \geq y_0, Y_{n+2} \geq y_0, Y_{n+3} \geq y_0 \,|\, \mathrm{data}\}.$$

Would also wish to centre the random $C(\cdot)$ as some given $C_0(\cdot)$, say $C_0(h) = \sigma^2 \rho^h$ for $\mathrm{AR}(1)$.

Not easy to do it 'directly' – placing a random band around ρ^h quickly produces outcomes that are non-valid, i.e. the associated covariance matrices may be negative definite. We need

- the random $C(\cdot)$ is positive definite;
- clear interpretation of prior;
- big (or full) prior support;
- simulations (or approximations) to the posterior;
- posterior consistency;
- perhaps more, e.g. Bernshteĭn–von Mises.

General approach (but not the only one): modelling $C(\cdot)$ via spectral measure $F(\cdot)$ on $[0, \pi]$. Wold's theorem:

$$C(h) = 2\int_0^\pi \cos(hu)\,\mathrm{d}F(u) \quad \text{for } h = 0, 1, 2, \ldots,$$

with F nondecreasing and finite; in particular $C(0) = 2F(\pi) = \sigma^2 = \operatorname{Var} Y_i$. Hence also for correlation function:

$$\operatorname{corr}(h) = \int_0^\pi \cos(hu)\frac{\mathrm{d}F(u)}{F(\pi)} \quad \text{with } \frac{F}{F(\pi)} \text{ random cdf.}$$

Main idea & programme:

- model for $F(\cdot)$
- $\Rightarrow$ model for $C(\cdot)$
- $\Rightarrow$ model for full covariance matrix and all related quantities
- $\Rightarrow$ full posterior distribution of 'everything', when coupled with data likelihood, which is

$$L_n \propto \exp(-\tfrac{1}{2}\log|\Sigma_n| - \tfrac{1}{2}y^{\mathrm{t}}\Sigma_n^{-1}y).$$

To understand F models from C properties (and vice versa, they are a 'Fourier couple'):

$$C(h) = 2\int_0^\pi \cos(hu)\,\mathrm{d}F(u),$$

$$f(u) = \frac{1}{2\pi}\sum_{h=-\infty}^{\infty}\exp(-ihu)C(u) = \frac{\sigma^2}{2\pi} + \frac{1}{\pi}\sum_{h=1}^{\infty}\cos(hu)C(h).$$

Prior for F should match prior knowledge for $C(\cdot)$. May centre F at F_0 that matches some C_0.

Example: AR(1). Here $C_0(h) = \sigma^2 \rho^h$ and spectral density becomes

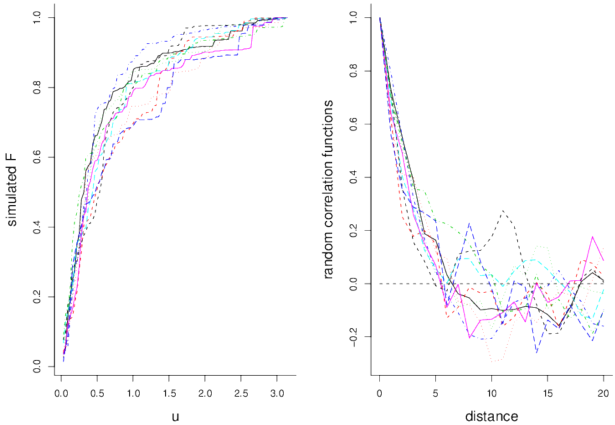
$$f_0(u) = \frac{\sigma^2}{2\pi} \frac{1-\rho^2}{1 - 2\rho\cos u + \rho^2} \quad \text{for } u \in [0, \pi].$$

May e.g. choose a prior for F with prior mean matching $F_0(u) = \int_0^u f_0(v)\,\mathrm{d}v$, and with uncertainty band matching prior uncertainty about $C(h)$ around $C_0(h)$.

Many possibilities: any random finite measure F gives rise to a random C etc.

Take J Bayesian nonparametrics papers for random finite measures $\Rightarrow$ new papers $\mathrm{paper}_1, \ldots, \mathrm{paper}_J$ for Bayesian nonparametric analysis of stationary time series.

Good tool: F a Gamma process $(aF_0, 1) \Rightarrow F/F(\pi)$ a Dirichlet (cF_0) on $[0, \pi]$, etc.



Left: simulated Dirichlet processes, $F/F(\pi) \sim \mathrm{Dir}(cF_0)$ *on* $[0,\pi]$*;*
right: the accompanying random correlation functions
$C(h) = \int_0^\pi \cos(hu)\,\mathrm{d}F(u)/F(\pi)$.

E2. Frequentist analysis

Time series analysis is nearly always parametric (typically also with model selection issues etc.), though nonparametric analysis is also possible – for spectral density f, its cdf F, and hence the covariance function. Consider the periodogramme (Schuster, 1898)

$$I_n(u) = \frac{1}{n}\frac{1}{2\pi}\Big|\sum_{k=1}^{n}\exp(-iku)y_k\Big|^2 \quad \text{for } u \in [0,\pi].$$

One has $I_n(u_{n,j}) \approx f_{\rm true}(u)\,{\rm Expo}(1)$ when $u_{n,j} \to u$, and asymptotic independence between these limits when $u_{n,j} \doteq \pi j/n$.

There's a big literature on smoothed periodogrammes etc., for estimating $f_{\rm true}$, but here we are more interested in F than f.

We may use either of

$$F_n(u) = \int_0^u I_n(v)\,\mathrm{d}v \quad \text{and} \quad \widehat{F}_n(u) = \frac{\pi}{n} \sum_{\pi j/n \le u} I_n(\pi j/n),$$

and have process convergence:

$$Z_n(u) = \sqrt{n}\{\widehat{F}_n(u) - F_{\rm true}(u)\} \to_d W\Big(2\pi \int_0^u f_{\rm true}(v)^2\,\mathrm{d}v\Big).$$

The associated estimators of covariances $C(h)$ are $\widehat{C}_n(h) = 2\int_0^\pi \cos(uh)\,\mathrm{d}\widehat{F}_n(u)$. From $Z_n = \sqrt{n}(\widehat{F}_n - F_{\rm true}) \to_d Z$, a time-transformed Brownian motion, follows

$$\sqrt{n}\{\widehat{C}_n(h) - C_{\rm true}(h)\} \to_d A_h = 2\int_0^\pi \cos(uh)\,\mathrm{d}Z(u),$$

and variances and covariances may be written down and estimated consistently. We have also good large-sample nonparametric control over all other smooth functions of the Σ_n matrix, and can put down normal approximations and confidence intervals etc.

E3. Exact and approximate Bayesian updating

Attractive class of priors: let F have independent increments, and split the spectral domain into windows,

$$[0,\pi] = W_1 \cup \cdots \cup W_m,$$

perhaps of equal width, $W_j = (\pi(j-1)/m, \pi j/m]$. The log-likelihood also almost splits into m components, across windows:

$$\ell_n = -\tfrac{1}{2}\log|\Sigma_n| - \tfrac{1}{2}y^{\mathrm{t}}\Sigma_n^{-1}y + \mathrm{const},$$
$$\widetilde{\ell}_n = -\tfrac{1}{2}n\frac{1}{\pi}\int_0^\pi \Big\{\log f(u) + \frac{I_n(u)}{f(u)}\Big\}\,\mathrm{d}u + \mathrm{const} = \sum_{j=1}^m \widetilde{\ell}_{n,j}.$$

Here $\widetilde{\ell}_n$ is the Whittle approximation to the exact ℓ_n, and $I_n(u)$ the periodogramme. We need F to have a density f.

Hence Bayesian updating can be undertaken 'window by window'.

Special prior: locally constant spectral density,

$$f(u) = f_j \quad \text{for } u \in \text{window } W_j,\ j = 1, \ldots, m,$$

with priors $\pi_1(f_1), \ldots, \pi_m(f_m)$ for these constants. The random F is continuous and piecewise linear.

Exact posterior distribution

$$\pi(f_1, \ldots, f_m \,|\, \text{data}) \propto \pi_1(f_1) \cdots \pi_m(f_m) \exp\{\ell_n(f_1, \ldots, f_m)\},$$

can be worked with, both practically (MCMC) and theoretically. For growing n (and windows not too small) it is close enough to its easier Whittle approximation:

$$\propto \pi_1(f_1) \exp\{\widetilde{\ell}_{n,1}(f_1)\} \cdots \pi_m(f_m) \exp\{\widetilde{\ell}_{n,m}(f_m)\},$$

where

$$\widetilde{\ell}_{n,j} = -\tfrac{1}{2} n \frac{1}{\pi} \int_{W_j} \Big\{\log f_j + \frac{I_n(u)}{f_j}\Big\} \,\mathrm{d}u = -\tfrac{1}{2} n \frac{1}{\pi} \Big(w_j \log f_j + \frac{v_{n,j}}{f_j} \Big)$$

for window W_j, with w_j length of W_j and $v_{n,j} = \int_{W_j} I_n(u) \,\mathrm{d}u$.

Full Bayesian analysis may now be carried out, from a given number of windows and given priors for the spectral heights $f_1, \ldots, f_m$.

We may use exact posterior via MCMC, or approximate posterior via Whittle and independence,

$$\pi(f_j \,|\, \mathrm{data}) \propto \pi_j(f_j) \exp\{-\tfrac{1}{2} n (1/\pi)(w_j \log f_j + v_{n,j}/f_j)\}$$

for $j = 1, \ldots, m$, with $v_{n,j} = \int_{W_j} I_n(u)\,\mathrm{d}u$ and w_j the width of W_j.

Can use inverse gamma priors for the local constants (convenient updating), but there are reasons for preferring gamma priors, say $f_j \sim \mathrm{Gam}(af_{0,j}, a)$.

We may compute posterior mean and variance directly (involving Bessel functions etc.), and also draw samples from $\pi(f_j \,|\, \mathrm{data})$.

E4. Large-sample results

We have proven nice large-sample theorems that in a Bernshteĭn–von Mises fashion mirror the frequentist results. The essential conditions are $m \to \infty$ and $m/\sqrt{n} \to 0$, 'more and more windows, but not too many'. Then:

- the Whittle approximation becomes good enough (same limit with exact and with Whittle);
- the parametric BvM theorem has time to kick in, for each window:

$$f_j \,|\, \mathrm{data} \approx \mathrm{N}\Big(w_j^{-1} \int_{W_j} \mathrm{d}\widehat{F}_n(u), 2\pi \int_{W_j} f_{\mathrm{true}}(v)^2 \,\mathrm{d}v/n\Big),$$

  with x_j midpoint of W_j;
- nonparametric BvM process convergence:

$$\sqrt{n}(F - \widehat{F}_n) \,|\, \mathrm{data} \to_d W\Big(2\pi \int_0^u f_{\mathrm{true}}(v)^2 \,\mathrm{d}v\Big);$$

  with 'invariance theorem' consequences for $C(h)$ etc.

F: Nonparametric quantile inference

Suppose $x_1, \ldots, x_n$ are i.i.d. from a distribution F, with quantile function

$$Q(y) = F^{-1}(y) = \inf\{t\colon F(t) \ge y\}.$$

So $Q(\frac{1}{2})$ is the median; $Q(\frac{1}{4})$ and $Q(\frac{3}{4})$ the two quartiles, etc.

A quantile pyramid is constructed in this fashion:

1 give a prior for $Q(\frac{1}{2})$;

2 give priors for $Q(\frac{1}{4})$ and $Q(\frac{3}{4})$, given $Q(\frac{1}{2})$;

3 give priors for $Q(\frac{1}{8}), Q(\frac{3}{8}), Q(\frac{5}{8}), Q(\frac{7}{8})$, given $Q(\frac{1}{4}), Q(\frac{2}{4}), Q(\frac{3}{4})$;

– &cetera, &cetera.

Under some conditions, this pans out well (Hjort and Walker, Annals, 2009) – the full $Q = \{Q(y)\colon y \in (0,1)\}$ exists; there is a characterisation of $Q \mid \mathrm{data}$; this may be computed and simulated from.

The class of Quantile Pyramids is very large. It may be used for purely probabilistic analyses of different types of phenomena, and for statistical quantile inference. A broad model is

$$z_i = m(x_i) + \varepsilon_i, \quad \text{where } \varepsilon_i \text{ has quantile process } Q,$$

with a prior process for $m(x)$. May then reach inference for

$$(Q(0.05 \,|\, x), Q(0.50 \,|\, x), Q(0.95 \,|\, x),$$

presented as bands in x, etc.

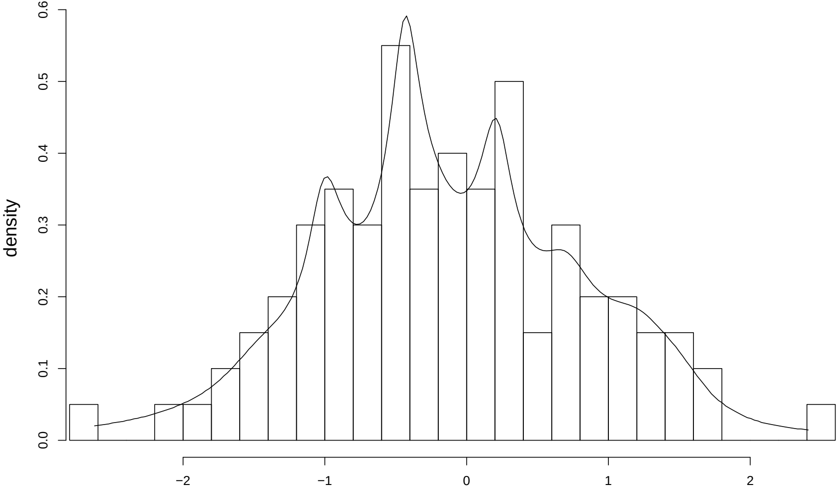
A special case of the Quantile Pyramid $Q = \{Q(y)\colon y \in (0,1)\}$ corresponds to $F = \{F(x)\colon x \in R\}$ being a $\mathrm{Dir}(aF_0)$. Cute quantile estimator:

$$\widehat{Q}(y) = \sum_{i=1}^{n} \binom{n-1}{i-1} y^{i-1}(1-y)^{n-i} x_{(i)} \quad \text{for } 0 \le y \le 1.$$

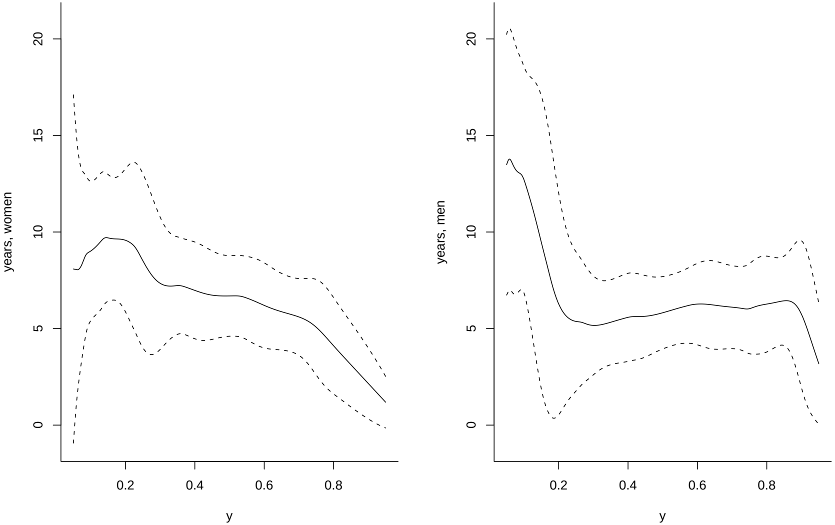
It has $\widehat{Q}(0) = x_{(1)}$ and $\widehat{Q}(1) = x_{(n)}$.

A fully automatic density estimator (Hjort and Petrone, 2007): solve $\widehat{Q}(y) = x$ to identify $y = \widehat{F}(x)$, and then

$$\widehat{f}(x) = \Big[\sum_{i=1}^{n-1} (x_{(i+1)} - x_{(i)})\,\mathrm{beta}(\widehat{F}(x); i, n-i)\Big]^{-1}.$$



Quantile difference function $G^{-1}(y) - F^{-1}(y)$, with bands, for F age at hospitalisation and G age at death, for women and for men. (I've re-analysed data from Laake, Laake, Aaberge, 1985.)



G: Models and methods for clusters

A simple prototype setup: Data points $y_1, \ldots, y_n$ are to be clustered, say as belonging to $\mathrm{N}(\xi_j, 1)$, with cluster centres ξ_j, and we do not know the number of clusters in advance (so this is not k-means or similar).

- 1 Let $P \sim \mathrm{Dir}(aP_0)$.
- 2 Let $\mu_1, \ldots, \mu_n$ i.i.d. P – but only D_n of these n will be distinct.
- 3 Let $y_i \sim \mathrm{N}(\mu_i, 1)$ for $i = 1, \ldots, n$.

May then set up a posterior scheme simulating from $(\mu_1, \ldots, \mu_n)$. From this one reads off both $\pi(D_n \,|\, \mathrm{data})$ and the positions of cluster centres.

$\exists$ hundreds of variations – many in heavy use. Note that a influences size of D_n: $D_n \approx a \log n$. (Can also have a prior on a.)

H: Concluding remarks

Bayesian nonparametrics has grown drastically, from c. 1973 to now – in horizon size, ambition level, flexibility, convenience, popularity (!), computational feasibility, applicability, maturity. It's close friends with branches of probability theory and applications and with machine learners and with all uses of Big Hierarchical Constructions.

Its uses include more flexibility around stricter models (nonparametric envelopes around parametric models).

It links with machine learning for nonparametric regression and classification; for hierarchical structures ('Dirichlet process of Dirichlet processes'); and for clustering and allocation processes (Chinese Restaurant Process, Indian Buffet Process).

FocuStat Workshop May 2015: CDs and Related Fields
FocuStat Workshop May 2016: FICology
FocuStat Workshop May 23–25 2017: Building Bridges, ‘from parametrics to nonparametrics’, including Bayesian nonparametrics.