Spring 2004 | **STA 294: Stochastic Processes & Bayesian Nonparametrics** | **R. Wolpert**

PARAMETRIC AND NONPARAMETRIC STATISTICS

Parametric Models

There is little to learn from observing some random quantities $X_1, \cdots, X_n$ all drawn independently from a *known* probability distribution $\mu(A) = \mathsf{P}[X_i \in A]$; if we already know the measure $\mu(dx)$ (possibly by knowing its density function, if $\mu(A) = \int_A f(x)\,dx$ for some nonnegative integrable density function $f(x)$) then no *statistical* problems remain.

If we aren't certain about $f(x)$ (or, equivalently, about $\mu(dx)$), if we only know it to be some member $f_\theta(x)$ from a family $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ labeled by some parameter θ from an index set Θ, then we hope to learn something by observing $\mathbf{X}_n = (X_1, \cdots, X_n)$— perhaps to learn the index θ (e.g. to estimate the mean and variance $\theta = (\mu, \sigma^2)$, in a normal-distribution example), or learn more about the likely values of a future observation X_{n+1}, or about some other quantity entirely that somehow depends on the unknown $\theta \in \Theta$.

If we're very unimaginative about our uncertainty, and are only willing to consider a p-dimensional family of densities $f_\theta(x)$ such as the usual parametric families (normal, Poisson, binomial, t, etc.), then our course is clear: we'd start by calculating the *likelihood function*

$$L(\theta) = L(\theta|\mathbf{X}_n) \equiv \prod_{i=1}^{n} f_\theta(X_i).$$

How we proceed from here depends on our inferential paradigm (i.e. statistical religion) and on the problem we're trying to solve. Classical statisticians trying to estimate θ would (usually) use the value $\hat{\theta}$ that maximizes $L(\theta)$, for example, and would then try to compute "standard errors" or derive "confidence sets" to give a measure of how precisely θ is determined by the observation of $\mathbf{X}_n$; Bayesians faced with the same problem would specify some prior probability distribution $\pi(d\theta)$ on Θ or on $\mathcal{F}$ (technically a simple matter, since $\Theta \subset \mathbb{R}^p$ and we have the usual machinery of probability density functions, etc. available on any subset of Euclidean space), then base inference on the posterior distribution

$$\pi(d\theta|\mathbf{X}_n) = \frac{L(\theta|\mathbf{X}_n)\pi(d\theta)}{\int L(\theta'|\mathbf{X}_n)\pi(d\theta')};$$

the mean $\bar{\theta} = \int \theta\, \pi(d\theta|\mathbf{X}_n)$ of this distribution might then serve as a point estimate for θ, for example, and its variance $\mathsf{V}[\theta|\mathbf{X}_n] = \int |\theta - \bar{\theta}|^2\, \pi(d\theta|\mathbf{X}_n)$ as a measure of precision.

But what if we are *not* willing to postulate a "small" parametric family of probability distributions $\{f_\theta(x)\,dx \mid \theta \in \Theta\}$ for X_i? No parametric model can be expected to describe *exactly* the chance mechanism generating an observed data set, and unrealistic features of some common models (for example, the thin tails of the normal distribution when compared to the distribution of observed data) can lead us to make unsatisfactory inferences. Is there a way we can let the *data* suggest the distribution of the X_i, without constraining the choice to a tiny part (*i.e.*, a low-dimensional manifold) in $\mathcal{F}$?

Nonparametric Models

The formal description above didn't make any use of the fact that Θ was "small"— and, indeed, nonparametric statistics (both Classical and Bayesian) can be approached as routine statistical analysis with a *very large* index set Θ for the possible distributions of X— like

1. "all" distributions $\mu(dx)$, or
2. all distributions with density functions $f(x)$, or
3. all distributions with continuous density functions $f(x)$, or
4. all distributions with symmetric unimodal density functions $f(x)$, *etc.*

Formally we can write a nonparametric "likelihood" function in these cases in the form

$$L(f|\mathbf{X}_n) \equiv \prod_{i=1}^{n} f(X_i), \qquad *$$

now with the density function f as an argument instead of the parameter θ; we can denote by $\mathcal{F}$ the space of all possible density functions we're considering ($L^1(\mathbb{R})$ and $\mathcal{C}(\mathbb{R})$ for cases 2. and 3. above, for example). For continuously distributed observations X maximum-likelihood methods are now almost useless for the *estimation* problem, in the sense that $*$ increases to infinity as we take a sequence of densities $f(x)$ approaching the empirical distribution function, $1/n$ times the sum of unit point masses at each of the observed values X_i. For example, if $\delta_\epsilon(x)$ is the approximate identity $\delta_\epsilon(x) = 1/2\epsilon$ for $|x| < \epsilon$, and zero for $|x| \geq \epsilon$, then the approximate empirical density

$$f_\epsilon(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_\epsilon(x - X_i)$$

satisfies $L(f_\epsilon|\mathbf{X}_n) \equiv \prod_{i=1}^{n} f_\epsilon(X_i) \geq (2n\epsilon)^{-n} \to \infty$ as $\epsilon \to 0$, for any fixed $n \in \mathbb{N}$. Thus the maximum likelihood is not attained for any distribution with a density function, and in some sense the MLE of the distribution of X is simply the empirical distribution

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - X_i),$$

where we use the unsubscripted Dirac delta notation $\delta(x - X_i)$ for the unit point mass at X_i. This is also the classical predictive distribution— after observing a sample of size n, we predict that the next observation is equally likely to take any of the previous values X_i and that it cannot take any new value. Clearly this seems nonsensical, for continuously-distributed variables X_i (where repeated values are impossible), except possibly as an approximation in case n is very large.

A Brief Hope for an Easy Way Out

In the case of discrete distributions (say, for integer-valued random variables X_i) the empirical distribution is still the MLE and the predictive distribution, but it is no longer degenerate or particularly ill-behaved. In general nonparametric methods work splendidly when there isn't "too much freedom" for the uncertain distributions. Perhaps we can exploit that idea and simplify our models and our analysis by deliberately reducing the level of detail with which we model the distribution of X_i; for example, we might divide $\mathbb{R}$ up into a small number J (say, five) of disjoint "bins" $\mathbb{R} = \bigcup R_j$ and only concern ourselves with the J probabilities $p_j \equiv \mathsf{P}[X_i \in R_j]$, and only observe the occupancy counts $N_j \equiv \sum_i 1_{R_j}(X_i)$ of the J bins. These counts always have the multinomial distribution, and there are any number of ways of putting a prior probability distribution on the J bin probabilities p_j (subject to the usual constraints of $p_j \geq 0$ and $\Sigma_{j\leq J} p_j = 1$).

While this *can* be a useful nonparametric mode of statistical analysis, it can be criticised on several points:

1. The bins R_j must be chosen before observing the data, and the choice is critical— we will learn very little from the data if the bin probabilities are too close to zero or one, and whatever we learn will be of little use if the bins are too wide or don't fit in well with the decision or estimation problem we face. If we look at the data before selecting bins, the multinomial distribution no longer governs our bin counts.
2. Except for a few special probability distributions (e.g. the dirichlet), it is not possible to reconcile prior distributions for two different bin selections— so our inference can never be independent of the apparently arbitrary choice of bins, and with most families of probability distributions on the $\{p_j\}$ we cannot even refine our bin choice and stay within the family.
3. It ignores some features of the data (by reducing the observations to bin counts) and hence must lose some efficiency.
4. It's not very elegent or appealing.

Now let's look at an alternative— doing nonparametric Bayesian analysis by using the tools of stochastic processes to study random *functions* $f_\theta(x)$ and *measures* $\mu_\theta(dx)$ as prior distributions. While historically stochastic processes were introduced and studied in order to model time-dependent random phenomena (stock prices, gamblers' fortunes, population sizes, projectile trajectories, etc.) there is no reason we can't use the same theory in a new arena— Bayesian nonparametric statistical analysis.

NONPARAMETRIC BAYESIAN STATISTICS

Uncertain Functions

We all know that Bayesians express uncertainty about a finite-dimensional parameter $\theta \in \mathbb{R}^p$ using a "prior" probability measure $\pi(d\theta)$, often one with a density function $\pi(\theta)$ for Lebesgue measure $d\theta$. But how can Bayesian statisticians express uncertainty about a *function* of one or more variables? The problem comes up all the time— for example,

1. Estimation

When we observe *iid* random variables X_i we're uncertain about their density function

$$X_i \sim f(x)$$

(though we might believe or pretend that they come from some standard distribution like the normal or exponential, so that we're only uncertain about a finite-dimensional parameter $\theta = (\mu, \sigma^2)$ or $\theta = 1/\lambda$).

2. Regression

When we observe regression data (X_i, Y_i) we're uncertain about the regression function

$$\mathsf{E}[Y_i | X_i] = G(X_i),$$

though we might believe or pretend that $G(x)$ (or its derivative $g(x) = G'(x)$) has some standard form like $G(x) = \alpha + x\beta$ (or $g(x) = \beta$), for normal data $Y_i \sim N\big[G(X_i), \sigma^2\big]$, or, for (normalized) binomial data $Y_i \sim Bi\big[N_i, G(X_i)\big]/N_i$, the probit $G(x) = \Phi(\alpha + x\beta)$ (or $g(x) = \beta\phi(\alpha + x\beta)$) or logit $G(x) = \frac{e^{\alpha+x\beta}}{1+e^{\alpha+x\beta}}$ functions with finite-dimensional parameters $\theta = (\alpha, \beta)$).

3. Survival

When we observe survival data T_i and explanatory data X_i we're uncertain about the survival function $S(t, x) = \mathsf{P}[T_i > t | X_i]$ or its negative logarithm the cumulative Hazard function $H(t, x) = -\log S(t, x)$, so

$$\mathsf{P}[T_i > t | X_i] = e^{-H(t, X_i)},$$

though we might believe or pretend that the hazard has some simple parametric form like the proportional-hazard exponential model $H(t, x) = t[\alpha + x\beta]$, so we're only uncertain about a parameter vector θ.

When we *are* willing to use a specific parametric model and limit our attention to functions in some finite-dimensional set $f \in \mathcal{F} = \{f_\theta : \theta \in \Theta\}$, we can use the usual statistical methods (density functions, etc.) to construct prior probability distributions on the finite-dimensional set Θ.

What if we're *not* willing to use such a model? Most experimental data feature more outliers and more widely-distributed observations than would be consistent with a normal-distribution model; if we're uncertain how much "fatter" the tails must be, then perhaps we should use a model in which the form of the CDF $F(x)$ or PDF $f(x) = F'(x)$ isn't specified at all. Seldom are there good reasons for prefering a probit model to a logit, or visa-versa; perhaps we should "let the data determine the model" without specifying a specific functional form for the binomial regression function $G(X_i)$. Perhaps cumulative hazard function $H(x,t)$ (or its derivative $h(x,t) = \frac{\partial}{\partial t} H(x,t)$, the so-called *instantaneous hazard function*) should be unspecified too. In any of these cases we're left with the problem of specifying probability distributions for a *random function* $f(x)$, $g(x)$, or $h(x,t)$.

Stochastic Processes

A **stochastic process** is just a collection of random variables, all defined on the same probability space $(\Omega, \mathcal{F}, \mathsf{P})$. If we index the collection by elements t of some set $\mathcal{T}$, and denote the process by $X = \{X_t(\omega) : t \in \mathcal{T}\}$, we can regard the process either as

1. a family of real-valued random variables $\omega \to X_t = X_t(\omega)$, indexed by $t \in \mathcal{T}$, or
2. a random function $t \to X_t = X_t(\omega) \in \mathcal{F}$, indexed by $\omega \in \Omega$.

With the first interpretation we can think of a stochastic process as a family of random variables whose probability distributions are somehow related; with the second, we can think of it as a single random variable taking values in an infinite-dimensional function space $\mathcal{F}$— so Stochastic Processes becomes simply the study of infinite-dimensional probability theory, and *Bayesian Nonparametrics* becomes simply ordinary Bayesian analysis with an infinite-dimensional parameter space Θ.

A Bayesian nonparametric approach to the three statistical problems introduced above (estimation, regression, and survival analysis) can begin with a description of the prior probability distribution for the stochastic processes $f(x)$, $g(x)$, or $h(x,t)$, respectively, or (more-or-less equivalently) for their integrated forms $F(x)$, $G(x)$, or $H(x,t)$.

We've already begun to look at the first problem (estimation) using Dirichlet Process prior distribution for $F(x)$. This approach made computations very simple, since the distribution is conjugate for binomial sampling, but had at least two disadvantages: the distribution $F(x)$ was almost surely singular (so the density function $f(x) = F'(x)$ doesn't even exist, except as a sum of point masses or Dirac "delta functions" $\delta(x - \tau_i)$), and the increments $[F(x+\epsilon) - F(x)]$ were all negatively correlated, making it impossible to express prior belief that the distribution was unimodal or was "close" to a distribution with continuous density function. It's possible to use the Dirichlet Process priors (or close relatives) for the other two problems, too, and it's possible to modify the approach in ad-hoc ways to avoid these two problems— for example, to pass to mixtures of Dirichlets.

On the other hand, maybe we can take advantage of the huge amount of work that others have already done constructing and studying classes of stochastic processes. Their work may have been motivated by completely different problems (usually modeling time-varying phenomena, like stock prices or electrical voltages) but some of the tools they've developed can be useful starting places for us as we explore nonparametric Bayesian statistics, using stochastic processes for prior distributions.

Some Examples of Stochastic Processes

A few familiar examples of stochastic processes include

1. Random Walks, $X_{t+1} = X_t \pm 1$ with probabilities p and $1-p$, respectively;
2. Markov chains, with $\mathsf{P}[X_{t+1} = j | X_t = i] = P_{ij}$ for some fixed matrix P_{ij}, conditionally independent of $\{X_u : u < t\}$;
3. Poisson Processes, with $[X_{t+s} - X_t] \sim \text{Poi}(\lambda s)$ and independent of $\{X_u : u \le t\}$;
4. Gaussian processes, with specified mean and covariance functions $\mu_t = \mathsf{E}[X_t]$ and $\gamma_{st} = \mathsf{E}[(X_s - \mu_s)(X_t - \mu_t)]$;
5. Brownian motion with drift $X_t = \alpha + \beta t + \sigma\omega_t$, with the continuous-path Wiener process ω_t satisfying $\omega_0 = 0$ and $[\omega_{t+s} - \omega_t] \sim \text{N}(0, s)$ independent of $\{\omega_u : u \le t\}$;
6. Stationary Independent-Increment (SII) Processes, with characteristic functions (by the Lévy-Khinchine formula) $\mathsf{E}[e^{i\lambda X_t}] = e^{i\lambda x_0 + it\lambda m - t\frac{\sigma^2\lambda^2}{2} + t\int_{\mathbb{R}} \left(e^{i\lambda u} - 1\right)\nu(du)}$ for some initial value x_0, drift m, diffusion constant σ^2, and jump rate ("Lévy measure") ν;
7. Diffusion processes, with $X_{t+\epsilon} = X_t + \beta_t(X_t)\,\epsilon + \sigma_t(X_t)[\omega_{t+\epsilon} - \omega_t] + o(\epsilon)$.

Positive Processes, Increasing Processes, etc.

This list is neither exclusive (Brownian motion, for example, is also a Gaussian process, an SII process, and a diffusion) nor exhaustive, but it suggests a few classes of processes that might prove useful to us, either in themselves or as building blocks. For example, regression functions are often taken to be monotonic (higher toxic doses seldom lead to *lower* probabilities of death) and PDF's and CDF's are always taken to be non-negative and increasing, respectively, so some of the processes listed above— Gaussian processes (which necessarily take on any values from $-\infty$ to $+\infty$, and which are typically *not* monotonic), for example— can't be used directly for either of those purposes. Of course $e^{X_t} > 0$ if $X_t \in \mathbb{R}$, however, and both $Y_t = \int_0^t e^{X_s}\,ds$ and $Z_t \equiv Y_t/(1 + Y_t)$ are increasing, so we can still construct suitable processes from any of the classes listed above.

First let's exploit what the traditional approach to Stochastic Processes has to say about the Dirichlet process and its variants, and gain further insight into their use in nonparametric Bayesian statistics.

Gamma Process

Preliminaries: Gamma, Beta, and Dirichlet Distributions

If $\alpha_i \geq 0$ and $X_i \sim \text{Ga}(\alpha_i, 1)$ for $i = 1, 2, ..., k$, then a routine computation shows that $S \equiv X_1 + \cdots + X_k$ and $\mathbf{Y} = [Y_1, \ldots, Y_k]$ with $Y_i \equiv X_i/S$ are independent random variables with the $\text{Ga}(\Sigma_i \alpha_i, 1)$ and $\text{Dir}(\alpha_1, \ldots, \alpha_k)$ distributions, respectively; in particular, for $k = 2$, the conditional distribution of X_1, given $S \equiv X_1 + X_2$, is that of S times an independent $\text{Be}(\alpha_1, \alpha_2)$ variable. We will need this for $\alpha_1 = \alpha_2 = 1/2^n$.

First Construction of the Gamma and Dirichlet Processes

In this section we will construct both Gamma Process and Dirichlet Process prior distributions. We will use the mnemonic notation G_α and D_α (respectively) both for the Gamma and Dirichlet random measures $G_\alpha(E)$ and $D_\alpha(E)$ and for the Stieltjes functions $G_\alpha(x) = G_\alpha\big((0, x]\big)$ and $D_\alpha(x) = D_\alpha\big((-\infty, x]\big)$ that determine the random measures; note that the Gamma process is well-defined even for infinite measures while the Dirichlet process is not, so by convention the Stieltjes functions are normalized so that $G_\alpha(0) = 0 = D_\alpha(-\infty)$. Neither measure has a density function (the random CDF's $G_\alpha(x)$ and $D_\alpha(x)$ are nowhere differentiable) so we are forced to consider random *measures* or CDF's, and not PDF's. First we construct the Gamma process.

Let α be any σ-finite positive measure on the space $(\mathcal{X}, \mathcal{B})$ (maybe the real line) and let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space; the *Gamma Process with mean* α is a random measure $G_\alpha : \mathcal{B} \times \Omega \to \mathbb{R}$ which assigns independent Gamma random variables $G_\alpha(\Lambda_i) \sim \Gamma(\alpha_i, 1)$ to disjoint sets $\Lambda_i \in \mathcal{B}$ whose measure $\alpha_i = \alpha(\Lambda_i)$ is finite. Here is an explicit construction of G_α for $\mathcal{X} = \mathbb{R}$, initially for Lebesgue measure and then for any σ-finite α. We begin by defining X_t for integers $t \in \mathbb{Z}$, then proceed inductively for dyadic rationals of the form $t = i/2^n$ (separately for odd and even i), and finally define X_t for irrational t by right-continuity.

Let U_i^n be a doubly-indexed independent family of random variables with the Beta distribution $\text{Be}(\frac{1}{2^n}, \frac{1}{2^n})$, for $n \in \mathbb{N}$ and $i \in \mathbb{Z}$; the U_i^0 are independent with uniform distributions (so each $-\log(U_j^0)$ will have the exponential or $\text{Ga}(1, 1)$ distribution), the U_i^1 have the $\text{Be}(1/2, 1/2)$, *etc.* Define a random sequence x_i^0 for all (not just positive) integers $i \in \mathbb{Z}$ by

$$x_i^0 = \begin{cases} \sum_{j=1}^{i} -\log(U_j^0) & \text{for } 0 < i \in \mathbb{Z}, \\ 0 & \text{for } 0 = i \in \mathbb{Z}, \text{ and} \\ \sum_{j=i}^{-1} \quad \log(U_j^0) & \text{for } 0 > i \in \mathbb{Z}. \end{cases}$$

For successive $n \geq 0$ define sequences x_i^{n+1} by:

$$x_i^{n+1} = \begin{cases} x_j^n & \text{if } i = 2j \text{ is even, and} \\ x_j^n + \big(x_{j+1}^n - x_j^n\big) U_i^{n+1} & \text{if } i = 2j+1 \text{ is odd.} \end{cases}$$

Setting $X_t \equiv x_i^n$ for $t = i/2^n$ defines a process X_t for all dyadic rational t; by induction, all the increments $[X_t - X_s]$ are independent with the $\mathrm{Ga}(t-s, 1)$ distribution. The process X_t is nonnegative and nondecreasing, so we can extend the definition to all of $\mathbb{R}$ by right-continuity: $X_t \equiv \inf[x_i^n : t \le i/2^n]$. For both rational and irrational $s < t$, the increments $[X_t - X_s]$ are independent with the Gamma $\Gamma\big((t-s), 1\big)$ distributions, and hence with finite means $\mathsf{E}[X_t - X_s] = (t-s)$ and variances $\mathsf{V}[X_t - X_s] = (t-s)$. Since X_t has stationary, independent increments, the discussion of SII processes starting on page 6 will show that the process X_t does *not* have continuous sample paths— in fact, it has infinitely many jumps in every open interval $(t, t+\epsilon)$ almost surely!

Given any σ-finite measure α on $\mathbb{R}$, define a right-continuous function by

$$\alpha_t = \begin{cases} \;\;\alpha\big((0,t]\big) & \text{if } t > 0, \\ 0 & \text{if } t = 0, \text{ and} \\ -\alpha\big((t,0]\big) & \text{if } t < 0, \end{cases}$$

and a random measure G_α by setting

$$G_\alpha\big((s,t]\big) = X_{\alpha_t} - X_{\alpha_s}$$

on intervals, for the standard Gamma process X_t defined above. Extend by additivity to the field generated by the half-open intervals $(s,t]$, and by continuity to all Borel sets with finite α-measure, upon noting that

$$\begin{aligned} \mathsf{E}G_\alpha\big((s,t]\big) &= [\alpha_t - \alpha_s] = \alpha\big((s,t]\big) \\ \mathsf{V}G_\alpha\big((s,t]\big) &= [\alpha_t - \alpha_s] = \alpha\big((s,t]\big), \end{aligned}$$

so (by L^2 continuity) $\mathsf{E}[G_\alpha(B)] = \mathsf{V}[G_\alpha(B)] = \alpha(B)$ for all $B \in \mathcal{B}$ with $\alpha(B) < \infty$. The SII section will show that, almost surely, G_α is a discrete measure concentrated on a (random) countable set of points $\tau_i(\omega)$.

The Dirichlet Process

Now let α be a *finite* nonnegative measure on $(\mathcal{X}, \mathcal{B})$ and let $G_\alpha(dx)$ be a Gamma process random measure with mean $\mathsf{E}[G_\alpha(dx)] = \alpha(dx)$; since $\alpha(\mathbb{R}) < \infty$, $G_\alpha(\mathbb{R})$ is a well-defined random variable and we can construct

$$D_\alpha(A) = \frac{G_\alpha(A)}{G_\alpha(\mathbb{R})}$$

for all $A \in \mathcal{B}$. Each random variable $D_\alpha(A)$ has a Beta $\mathrm{Be}\big(\alpha(A), \alpha(A^c)\big)$ distribution, and for any partition Λ_i of $\mathcal{X}$ into n disjoint measurable sets, the n-variate random variables $X_i = D_\alpha(\Lambda_i)$ have the Dirichlet $\mathrm{Dir}(\alpha_1, \cdots, \alpha_n)$ distribution with parameters $\alpha_i = \alpha(\Lambda_i)$. Just as the Gamma process G_α was almost-surely concentrated on a countable set of points τ_i, so too is the Dirichlet process D_α... in fact, it is the *same* set $\{\tau_i\}$! The Dirichlet too process is, almost surely, a discrete distribution.

The Dirichlet Process is an important example, because of its use in nonparametric Bayesian statistics. The principal result is this:

Theorem. *Let $D \sim \mathrm{DP}(\alpha_o)$ for some finite measure α_o and let $X_1, X_2, \ldots, X_n$ be independent observations all with distribution $D(\omega)$. Then, conditional on $X_1, \ldots, X_n$, $D \sim \mathrm{DP}(\alpha_n)$ for the measure $\alpha_n(dx) = \alpha_o(dx) + \sum_{i=1}^n \delta(x - X_i)\, dx$ equal to $\alpha_o(dx)$ plus a unit point mass at each observed X_i.*

Corollary. *Under the same conditions, the predictive distribution for X_{n+1} assigns mass $\frac{1}{n+\alpha(\mathbb{R})}$ to $x = X_i$ for each $1 \le i \le n$ and the rest of the mass $\frac{\alpha(\mathbb{R})}{n+\alpha(\mathbb{R})}$ to the prior mean, $\frac{\alpha(dx)}{\alpha(\mathbb{R})}$.*

Note that, from the corollary, the probability of a *tie* among the first n variables is at least

$$1 - \prod_{i=1}^{n} \left(\frac{\alpha(\mathbb{R})}{\alpha(\mathbb{R}) + i - 1} \right) = 1 - \frac{\alpha(\mathbb{R})^n \Gamma\big(\alpha(\mathbb{R})\big)}{\Gamma\big(\alpha(\mathbb{R}) + n\big)},$$

arbitrarily close to 1 for large enough n (it exceeds $\frac{n}{n+\alpha(\mathbb{R})}$ for $n \ge 3$); if the X_i "really" come from any continuous distribution, and are observed without round-off, no ties will be observed no matter how large n might be. This fact has an alarming consequence: if $f(x)$ is *any density at all* and $\epsilon > 0$, the posterior for a Bayesian prior giving probability ϵ to $f(x)\, dx$ and $1{-}\epsilon$ to $D_\alpha \sim \mathrm{DP}(\alpha_o)$ will eventually be concentrated on $f(x)$. This proves that Bayesian analysis can be inconsistent. See me for more references if you're interested in this point.

Path Discontinuity and SII Processes

The Gamma process G_t with Lebesgue mean measure $\alpha(dt) = dt$ has independent increments $[G_t - G_s]$ whose distribution $(\mathrm{Ga}(t{-}s, 1))$ depends only on $t{-}s$; such processes are said to have Stationary Independent Increments, or to be SII processes. A celebrated theorem of Lévy and Khinchine asserts that every SII process X_t has a characteristic function of the form:

$$\mathsf{E}[e^{i\lambda X_t}] = e^{i\lambda x_0 + it\lambda m - t\frac{\sigma^2\lambda^2}{2} + t\int_{\mathbb{R}} \left(e^{i\lambda u} - 1\right) \nu(du)}$$

for some initial value x_0, drift m, diffusion constant σ^2, and jump rate ("Lévy measure") ν. The theorem is important because it shows exactly what the sample paths of every SII process looks like:

1. a linear function $x_0 + mt$, plus
2. a Brownian motion $\sigma\omega_t$, plus
3. a generalized Poisson process that takes jumps of sizes $[X_s{-}X_{s-}] = u \in E$ at rate $\nu(E)$ for each measurable set $E \subset \mathbb{R}$.

The total jump rate $\nu(\mathbb{R})$ need not be *finite*, if ν satisfies $\int_{\mathbb{R}} [1 \wedge |u|] \nu(du) < \infty$; it's OK to have infinitely many tiny jumps so long as they're small enough to (almost surely) have a finite sum. If $\nu(\mathbb{R}) < \infty$ then we can interpret the process as one with exponentially distributed waiting times (with means $\frac{1}{\nu(\mathbb{R})}$) between successive jumps, whose sizes are random with probability distribution $\nu(du)/\nu(\mathbb{R})$. With a little more work it's possible to make sense of processes with jump measures satisfying only the weaker condition $\int [1 \wedge u^2]\nu(du) < \infty$, but the argument gets more subtle. Ask if you'd like details or references.

Brownian motion with drift $X_t = \alpha + \beta t + \sigma \omega_t$, for example, has characteristic function

$$\begin{aligned} \mathsf{E}[e^{i\lambda X_t}] &= \int_{\mathbb{R}} e^{i\lambda(\alpha+\beta t+\sigma x)} \frac{e^{-x^2/2t}}{\sqrt{2\pi t}}\, dx \\ &= e^{i\lambda\alpha + it\lambda\beta - t\sigma^2\lambda^2/2}, \end{aligned}$$

corresponding to $x_0 = \alpha$, $m = \beta$, $\sigma^2 = \sigma^2$, and $\nu(du) \equiv 0$ so $X(t)$ has continuous paths with no jumps at all; a standard Poisson process $N(t)$ has characteristic function

$$\begin{aligned} \mathsf{E}[e^{i\lambda N_t}] &= \sum_{k=0}^{\infty} e^{i\lambda k} \frac{e^{-t}\ t^k}{k!} \\ &= e^{t(e^{i\lambda}-1)}, \end{aligned}$$

corresponding to $x_0 = 0$, $m = 0$, $\sigma^2 = 0$, and $\nu(du) = \delta(u-1)\,du$, so $N(t)$ has no drift or "Brownian" part but has jumps of size $u = 1$ at rate $\nu(\{1\}) = 1$. Any SII process can be written as the sum of Brownian motion with drift and an independent generalized Poisson Process. In particular, its paths will be continuous if and only if it is Brownian (*i.e.*, $\nu \equiv 0$), and otherwise its only path discontinuities will be jumps whose rate we can find by identifying the Lévy measure ν.

For example, the characteristic function of the standard Gamma Process $G_t \sim \mathsf{Ga}(t, 1)$ is $\mathsf{E}[e^{i\lambda G_t}] = (1 - i\lambda)^{-t}$; it has no drift or diffusion part, and has Lévy measure $\nu(du)$ satisfying

$$-t\log(1 - i\lambda) = t \int_{\mathbb{R}} \left(e^{iu\lambda} - 1\right) \nu(du)$$

or, after differentiating with respect to λ and dividing by it,

$$\frac{1}{1 - i\lambda} = \int_{\mathbb{R}} u e^{iu\lambda}\, \nu(du)$$

But $\int_0^\infty e^{-u(1-i\lambda)} du = (1 - i\lambda)^{-1}$, so $e^{-u(1-i\lambda)} 1_{(0,\infty)}(u)\, du$ must be $u\, e^{iu\lambda}\, \nu(du)$ and we find

$$\nu(du) = 1_{(0,\infty)}(u) u^{-1} e^{-u}\, du.$$

This is not a finite measure, so the Gamma process jumps infinitely often in every time interval; the rate of jumps bigger than ϵ is $E_1(\epsilon) = \int_\epsilon^\infty e^{-u}/u\, du$, finite for every $\epsilon > 0$

(it's called the "exponential integral function"; see Abramowitz and Stegun p.218), and the mean sum of all jumps in time t is $t\int_0^\infty ue^{-u}/u\,du = t$, while the *number* of such jumps in time t that exceed ϵ in size has the Poisson $\mathrm{Poi}(E_1(\epsilon)\,t)$ distribution. Thus the Gamma process and Dirichlet process priors both are concentrated on discrete measures supported on countably many points. It's not hard to find the joint distribution of the sizes and locations of *all* the jumps of the Dirichlet process X_t in any time interval, leading to another explicit construction; here's a sketch, you can ask me for details, more hints, or references.

Another Construction

Let α be a finite measure on $\mathbb{R}$ and let $\{U_i\}$ and $\{\tau_i\}$ be independent families of IID random variables with the $\mathrm{Be}\big(1, \alpha(\mathbb{R})\big)$ and $\alpha(dx)/\alpha(\mathbb{R})$ distributions, respectively, and let $S \sim \mathrm{Ga}\big(\alpha(\mathbb{R}), 1\big)$ have a standard Gamma distribution. For $n \in \mathbb{N}$, set $J_1 = U_1$ and $J_n = U_n{\cdot}(1{-}U_{n-1})\cdots(1{-}U_1) = U_n[1{-}\sum_{i<n} J_i]$. Then one can show that $1 \equiv \sum_{n<\infty} J_n$ and that

$$D_\alpha(dx) = \sum_{n=1}^{\infty} J_n \delta(x-\tau_n)\,dx \qquad \text{and} \qquad G_\alpha(dx) = \sum_{n=1}^{\infty} S\,J_n \delta(x-\tau_n)\,dx$$

determine Dirichlet Process and Gamma Process priors, respectively, both with parameter α. The "jumps" J_n are the sizes of the Dirichlet's point masses at sites X_n. Note that they are not given in decreasing order by this construction (due to Sethuraman and Tiwari), although another (more complicated) construction will give the largest jumps first— for example, the biggest jump $S\,J^t_{(1)}$ during time $(0,t]$ for the Gamma process has CDF

$$\mathsf{P}[S\,J^t_{(1)} \le x] = e^{-tE_1(x)}$$

and consequently PDF $S\,J^t_{(1)} \sim f^t_{(1)}(x) = \frac{t}{x} e^{-x-tE_1(x)}$ for $0 < x < \infty$.

Examples

Need nice examples from estimation, regression, and survival, using variations on Gaussian processes, Dirichlet processes and their mixtures, and diffusions. Volunteers?