7/24/89

An Elementary Continuous-Type Nonparametric Distribution Estimate

by
James M. Dickey, Paul H. Garthwaite, and Guorui Bian*

University of Minnesota, School of Statistics
Technical Report No. 525
Revised and expanded July 1989

*Dickey's and Bian's research were supported in part by National Science Foundation Grants DMS-8614793 and DMS-8911548.

An Elementary Continuous-Type Nonparametric Distribution Estimate

JAMES M. DICKEY, PAUL H. GARTHWAITE, and GUORUI BIAN*

Abstract. The theory of filtered-variate distributions can be used to construct a continuous-type distribution estimate directly from sample data that has interesting properties. The distribution estimate is obtained synthetically as the distribution of a linear transformation of a Dirichlet-distributed vector. The matrix of the transformation is a rescaled version of the matrix of n (fixed) sample points, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ $(k \times n)$, so our estimate of the population distribution is identical to a generalized and rescaled form of Rubin's (1981) Bayesian bootstrap of the posterior distribution of the population mean. The distribution can be constructed to have its first and second order moments match the empirical moments of the sample. Given a fixed sample outcome, limiting forms of this continuous-type distribution will include the discrete empirical distribution of the data, and at another extreme, a normal distribution parameterized by the sample mean and sample variance. Thus the method includes the two usual statistical practices as special cases. The method can produce a distribution from data of any dimension k, and the method is invariant under marginalizations. Representations and computational and simulation methods for the distribution estimate are given. The distribution is interpreted as, and generalized to, the distribution of a random mean of a Dirichlet process, enabling the study of limiting forms for increasing sample size. Dirichlet parameter measures with continuous-type components are introduced, including "prior" components to utilize pre-data information. The distribution estimate is not strictly a Bayesian function of the data, but the prior component and the data combine according to sample pooling.

---

*Dickey's and Bian's research were supported in part by National Science Foundation Grants DMS-8614793 and DMS-8911548. The authors are grateful for helpful discussions with Morris Eaton, David Lane, and Ron Pruitt.

1. INTRODUCTION

Nonparametric inference problems were considered by L. J. Savage (1964) to be an embarrassment for Bayesian statistics. Since the population distribution in these problems would not be known to within only a few parameters that could be given a prior distribution in the Bayesian approach, then a whole prior random process for the unknown distribution would need to be constructed. The often considered Dirichlet prior random process (Ferguson 1973) is generally unsuited for such use, primarily because it is prejudiced in favor of highly nonsmooth distributions, in the important sense that the prior correlations between the unknown probabilities of disjoint neighboring interval events are all small and slightly negative. This drawback, however, together with the problem of discrete realizations (Blackwell 1973), has been circumvented in various apparently successful Bayesian uses of the process. For an elegant review of Bayesian work based on the Dirichlet process, see Ferguson, Phadia, and Tiwari (1989). The Bayesian literature on nonparametric problems is extensive and could be said to include also Good (1950,1965), Whittle (1958), Hill (1968,1987ab), Dickey (1968ab,1969), Good and Gaskins (1971,1980), Leonard (1973,1978), Lo (1984, 1987), Lenk (1984,1988), Rubin (1981), Titterington (1986), Olkin and Spiegelman (1987), and Weng (1989).

We propose here an ostensibly nonBayesian elementary nonparametric distribution estimate from sample data for a continuous population. The distribution estimate, itself, is a continuous-type distribution and even has a continuous density function. However, the distribution estimate will not be defined in terms of its density. In general, the density function can be expensive to compute, and so the method may not be useful in some density-estimation contexts. The distribution will be defined, rather, by synthetically defining its random variable or random vector as a linear function, a "filter," of a finite dimensional Dirichlet random vector. (It will not be a "mixture of Dirichlet distributions," because only one Dirichlet distribution will be involved.) Since it is easy to simulate sampling from a Dirichlet distribution, the method can be used readily in problems where it is desired to draw monte carlo samples from the distribution estimate. This can be valuable in applied simulation studies when a population distribution is not fully known but sample data is available. Although the method is not strictly Bayesian, a Bayesian-like aspect is that modified versions of the proposed distribution can be interpreted as a predictive distribution for further data.

The moments of the distribution are easily calculated. Indeed, the distribution can be constructed to have its low-order moments identical to the empirical sample moments. Given the sample data, the resulting

continuous-type distribution is simply parameterized and its limiting forms are, at one extreme of its parameter, the so-called sample distribution, the discrete empirical distribution of the data, itself, and at the other extreme, a normal distribution parameterized by the sample mean and sample variance. This implies that the two most common and contrasting statistical practices are included as special cases of the method. The method produces a distribution from a sample of n data vectors of any dimension k, and a marginalization property is satisfied. Applying the method directly to k_1 coordinates of the n data vectors, $k_1 \leq k$, yields a distribution estimate that is identical to the k_1-dimensional marginal distribution from the full k-dimensional distribution estimate.

Following a dilation or expansion of the data points away from their sample mean, the distribution estimate will be defined as the distribution of the Dirichlet-weighted average of the fixed dilated data points. This is identical to a generalization of the Bayesian bootstrap (Rubin 1981) in the form that was designed for simulating the posterior distribution of the population mean. The present suggestion is for an estimate of the population distribution. This may have its own simulation uses in generalized bootstrapping or cross-validation. The distribution estimate can be interpreted as the random mean function of a simple case of a Dirichlet random process. This interpretation will permit investigations of the limiting distribution with increasing sample size, and also a generalization of the method to the use of Dirichlet processes with more general parameter measures than a dilated-sample empirical measure. This will permit distribution estimates having further-smoothed density functions and the use of parameter measures having components modeling "prior" information. A Bayesian aspect is that the prior components and the sample data combine similarly to sample pooling.

After setting the stage with filtered-variate distributions and the Bayesian bootstrap in Section 2, we will establish the required second-order moment properties in Section 3, and define the basic distribution estimate in Section 4. Sections 5, 6, and 7 give representations and properties of the distribution bearing on its computation and use. The interpretation in terms of a special Dirichlet process is given in Section 8, and a subsequent generalization to the random mean of a Dirichlet process. Methods of simulating such a general process are outlined in Section 9, and uses of parameter measures with continuous-type components are discussed in Section 10.

2. A FIRST ATTEMPT: THE BAYESIAN BOOTSTRAP

The simple idea involved in constructing the proposed distribution is that a weighted average of sample data column vectors, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $k \times n$,

$$\tilde{\mathbf{x}} = \mathbf{X}\,\mathbf{w} = w_1 \mathbf{x}_1 + \ldots + w_n \mathbf{x}_n \,, \tag{2.1}$$

can be given a synthetic post-sampling distribution with controllable properties by taking the data vectors as fixed and the vector of weights $\mathbf{w}$ to be random with a parameterized distribution. Such a random vector $\tilde{\mathbf{x}}$ can also be described as a linear function of random $\mathbf{w}$, and we will call the distribution of $\tilde{\mathbf{x}}$ a **filtered-variate** transformation of the distribution of $\mathbf{w}$. If the distribution of $\mathbf{w}$ is of continuous type over a region, Support($\mathbf{w}$) in R^n, then $\tilde{\mathbf{x}}$ will have a continuous distribution, too, over a linear transformation of the region, Support($\tilde{\mathbf{x}}$) = $\mathbf{X}$ Support($\mathbf{w}$), in R^k. In particular, $\mathbf{w}$ may have a Dirichlet distribution,

$$\mathbf{w} \sim D(\boldsymbol{\beta}) \tag{2.2}$$

on the probability simplex in R^n, $\Delta^{(n)} = \{\mathbf{w} : \text{each } w_j \geq 0 \text{ and } w_1 + \ldots + w_n = 1\}$. If each $0 < \beta_j < \infty$, $j = 1, \ldots, n$, the density of any $n-1$ coordinates, say $w_1, \ldots, w_{n-1}$, is

$$p(\mathbf{w}) = B(\boldsymbol{\beta})^{-1} \prod_{j=1}^{n} w_j^{\beta_j - 1}, \tag{2.3}$$

for $\mathbf{w}$ in the interior of $\Delta^{(n)}$, where $B(\boldsymbol{\beta}) = [\prod \Gamma(\beta_j)] / \Gamma(\beta_{tot})$, $\beta_{tot} = \sum \beta_j$. The distribution of $\tilde{\mathbf{x}}$, if $\mathbf{w}$ has a Dirichlet distribution (including limiting forms), will be called a **filtered-variate Dirichlet distribution**. In the case of a positive density (2.2) on the interior of $\Delta^{(n)}$, the closed region of support of $\tilde{\mathbf{x}}$ is the convex hull of the sample points, Support($\tilde{\mathbf{x}}$) = CHull($\mathbf{x}_1, \ldots, \mathbf{x}_n$). Under the usual assumption of full-rank data, $n \geq k+1$ and rank($\mathbf{X} - \mathbf{x}_n \mathbf{1}_n^T$) = k, where $\mathbf{1}_n = (1, \ldots, 1)^T$, $n \times 1$, this support is a k-dimensional solid polytope whose vertices are a subset of the sample points. In one dimension (k = 1), this gives a distribution that has a positive density function over the interior of the observed range interval of the data, (min x_j, max x_j).

The random variable $\tilde{\mathbf{x}}$ is the extension, to a more general parameter $\boldsymbol{\beta}$, of the Bayesian-bootstrap random variable that was defined to simulate the posterior distribution of the unknown population mean. For the Bayesian bootstrap, all $\beta_j \equiv 1$, $j = 1, \ldots, n$, so $\beta_{tot} = n$. The Bayesian bootstrap,

introduced by Rubin (1981), was studied by Lo (1987), Weng (1989), and others.

Known properties of the Dirichlet distribution establish properties of the tentative distribution estimate, (2.1), (2.2). For simplicity, assume an exchangeable or symmetric Dirichlet distribution with all $\beta_j \equiv \beta$, $j = 1, \ldots, n$, $\beta_{tot} = n\beta$, and thus the mean vector $\boldsymbol{\beta}/\beta_{tot} = n^{-1}\mathbf{1}_n$. As β_{tot} grows without bound, for fixed n, the Dirichlet degenerates to a one-point distribution concentrated at its mean point, $P[\mathbf{w} = n^{-1}\mathbf{1}_n] = 1$, and then our random vector $\tilde{\mathbf{x}}$ concentrates right at the sample average, $P[\tilde{\mathbf{x}} = \bar{\mathbf{x}}] = 1$, where $\bar{\mathbf{x}} = n^{-1}\sum \mathbf{x}_j$. In the other extreme, as the common parameter value β shrinks to zero, for fixed n, the distribution of $\mathbf{w}$ becomes a discrete uniform distribution on the set of unit coordinate vectors $P[\mathbf{w} = \boldsymbol{\delta}_{(j)}] = 1/n$, $j = 1, \ldots, n$, where $\boldsymbol{\delta}_{(j)} = (\delta_{(j)1}, \ldots, \delta_{(j)n})^T$, $\delta_{(j)j}=1$, and $\delta_{(j)i}=0$ for $i \neq j$. And so in this case, the vector $\tilde{\mathbf{x}}$ will have the discrete uniform distribution over the set of sample points, the empirical distribution of the sample itself: $P[\tilde{\mathbf{x}} = \mathbf{x}_j] = 1/n$, $j = 1, \ldots, n$. More generally, as we shall see below, whenever $0 \leq \beta \leq \infty$, the low order moments for $\tilde{\mathbf{x}}$ are simply related to the sample moments,

$$E\,\tilde{\mathbf{x}} = \bar{\mathbf{x}} \qquad \text{and} \qquad \text{Var}\,\tilde{\mathbf{x}} = c\,\mathbf{S}_{\mathbf{x}}, \tag{2.4}$$

where $c = (n\beta + 1)^{-1}$ and $\mathbf{S}_{\mathbf{x}} = n^{-1}\sum (\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})^T$.

So what we have in the tentative distribution estimate, (2.1) with (2.2) and $\beta_j \equiv \beta$, is a continuous-type distribution (if $0 < \beta < \infty$) with interesting limiting forms, whose mean is the sample mean and whose variance matrix is proportional to the sample variance matrix. Note, however, that the constant of proportionality in the variance (2.4) cannot be greater than unity, $0 \leq c \leq 1$, and its upper bound, $c = 1$ for $\beta = 0$, would be achieved at the expense of degeneracy of the distribution to the discrete sample empirical distribution. This, together with the limitation that the support cannot extend beyond the observed data points, leads us to speculate that both these restrictions can be removed by a variance-dilation or proportional extension of the data outward from its mean point. We shall see that this can lead to an improved distribution estimate, following a discovery that the filtered-variate Dirichlet family is not unique for our purposes.

3. VARIATE FILTERS OF DISTRIBUTIONS HAVING A MEAN-STRUCTURED VARIANCE

By the form of the Dirichlet density (2.3), the Dirichlet moments are obvious ratios of products of gamma functions. These give the Dirichlet mean and variance as

$$E\,\mathbf{w} = \boldsymbol{\omega}\,, \quad \text{and} \quad \text{Var}\,\mathbf{w} = c\,[\text{Diag}(\boldsymbol{\omega}) - \boldsymbol{\omega}\,\boldsymbol{\omega}^T] \tag{3.1}$$

where the mean vector $\boldsymbol{\omega} = \boldsymbol{\beta}/\beta_{tot}$ and the proportionality factor $c = (\beta_{tot} + 1)^{-1}$. Let us say that any distribution on $\Delta^{(n)}$ satisfying (3.1) for a mean vector $\boldsymbol{\omega}$ and some proportionality value c, where $0 \le c \le 1$, has a **mean-structured variance (MSV)**. Denote this by

$$\mathbf{w} \sim \text{MSV}(\boldsymbol{\omega}, c). \tag{3.2}$$

Another familiar MSV distribution is the normalized multinomial: if $\mathbf{n} \sim$ multinomial(N, $\boldsymbol{\omega}$) and $\mathbf{w} = \mathbf{n}/N$, then $\mathbf{w} \sim \text{MSV}(\boldsymbol{\omega}, c)$ with $c = 1/N$. Many other families of distributions on the probability simplex are MSV. Indeed, we find that the MSV property is preserved under: (i) mixing over the mean by a mixing distribution that is itself MSV; (ii) any mixing over the proportionality factor; (iii) partitioning and grouping of coordinate indices and corresponding summation of coordinates; and (iv) weighted averaging between independent (or uncorrelated) MSV vectors. The following theorem giving the details is easily proved by calculating the moments of the derived distributions.

Theorem 3.1

(i) If $\mathbf{y}\,|\mathbf{x} \sim \text{MSV}(\mathbf{x}, d)$ and $\mathbf{x} \sim \text{MSV}(\boldsymbol{\mu}, c)$, then $\mathbf{y} \sim \text{MSV}(\boldsymbol{\mu}, c + d - cd)$;

(ii) If $\mathbf{x}\,|c \sim \text{MSV}(\boldsymbol{\mu}, c)$, for random c and fixed $\boldsymbol{\mu}$, then $\mathbf{x} \sim \text{MSV}(\boldsymbol{\mu}, E(c))$.

(iii) If $\mathbf{x} \sim \text{MSV}(\boldsymbol{\mu}, c)$ and $\mathbf{y} = \mathbf{G}\mathbf{x}$, where matrix $\mathbf{G}$ has entries 0 and 1 with orthogonal rows, then $y \sim \text{MSV}(\mathbf{G}\boldsymbol{\mu}, c)$;

(iv) If $\mathbf{x} \sim \text{MSV}(\boldsymbol{\mu}, c)$, $\mathbf{y} \sim \text{MSV}(\boldsymbol{\mu}, d)$, $\text{cov}(\mathbf{x}, \mathbf{y}) = 0$ and $0 \le u \le 1$, then $u\mathbf{x} + (1-u)\mathbf{y} \sim \text{MSV}(\boldsymbol{\mu}, u^2c + (1-u)^2d)$.

For example, (i) implies that the normalized form of the Dirichlet-multinomial, the conjugate Bayesian predictive distribution for multinomial sampling, is MSV with $c = (N + \beta_{tot}) / [N(\beta_{tot} + 1)]$, in an obvious choice of notation. Bar-Lev and Enis (1986), who studied moment

structures related to MSV in one dimension, referred to an analog of property (iv), with unweighted averaging, as "reproducibility".

The following two theorems generalize limiting properties of the Dirichlet to MSV distributions.

Theorem 3.2 As $c \downarrow 0$, an MSV distribution $\mathbf{w} \sim \text{MSV}(\boldsymbol{\omega}, c)$ on $\Delta^{(n)}$ approaches the singular distribution with unit mass at $\mathbf{w} = \boldsymbol{\omega}$. As $c \uparrow 1$, the limiting distribution is the discrete distribution, $\mathbf{w} \sim \mathbf{w}^*$, where the random vector $\mathbf{w}^*$ is supported on the extreme points of $\Delta^{(n)}$ with each $P[\mathbf{w}^* = \boldsymbol{\delta}_{(j)}] = \omega_j, \quad j = 1, \ldots, n.$

Proof. The first part of the theorem is immediate and the second part is a consequence of the following lemma.

Lemma 3.1 If the (scalar) random quantity w is supported on the unit interval, $0 \le w \le 1$, and $Ew = \omega$, then the variance of w is maximized for fixed ω by $P(w = 1) = \omega$, $P(w = 0) = 1 - \omega$.

Proof of Lemma. By a standard linear optimization argument, the extreme measure under two linear constraints (total measure unity and mean fixed) is a two-point distribution, $P(w = u_i) = p_i$, $i = 1, 2$. Then $\text{Var}(w) = p_1 p_2 (u_1 - u_2)^2 = (u_2 - \omega)(\omega - u_1)$, which is maximized by $u_1 = 0$, $u_2 = 1$.

Now, consider any **filtered-variate MSV distribution** with random vector $\mathbf{y} = \mathbf{Z}\,\mathbf{w}$, for some matrix $\mathbf{Z}$, in which $\mathbf{w} \sim \text{MSV}(\boldsymbol{\omega}, c)$. The limiting distributions of $\mathbf{y}$ with c, as follows, are immediate from the limiting forms of $\text{MSV}(\boldsymbol{\omega}, c)$. Considered as points in R^k, the n column vectors $\mathbf{z}_j$ of $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ will be called **filter points**.

Theorem 3.3 As $c \downarrow 0$ (with Z held fixed), the induced filtered-variate distribution of $\mathbf{y} = \mathbf{Z}\,\mathbf{w}$, where $\mathbf{w} \sim \text{MSV}(\boldsymbol{\omega}, c)$, approaches the singular distribution with unit mass at $\mathbf{y} = \mathbf{Z}\,\boldsymbol{\omega}$. As $c \uparrow 1$, the limiting distribution is the finitely supported distribution, $\mathbf{y} \sim \mathbf{y}^*$, of the random vector $\mathbf{y}^*$ on the set of n filter points $\mathbf{z}_j$, with each $P[\mathbf{y}^* = \mathbf{z}_j] = \omega_j, \quad j = 1, \ldots, n.$

The moments of $\mathbf{y} = \mathbf{Z}\,\mathbf{w}$ must be the usual linear transform of the moments of $\mathbf{w}$,

$$E\,\mathbf{y} = \mathbf{Z}\,(E\,\mathbf{w}), \qquad \text{Var}\,\mathbf{y} = \mathbf{Z}\,(\text{Var}\,\mathbf{w})\,\mathbf{Z}^T. \tag{3.3}$$

Equations (3.1) and (3.3) yield the following important representation for these first two moments in terms of empirical moments of the filter points.

Theorem 3.4 If $\mathbf{w} \sim \mathrm{MSV}(\boldsymbol{\omega}, c)$, the low-order moments of $\mathbf{y} = \mathbf{Z}\,\mathbf{w}$ are the following simple functions of the moments of $\mathbf{y}^*$, the weighted empirical moments of the filter points,

$$E\,\mathbf{y} = E\,\mathbf{y}^* = \Sigma\,\omega_j\,\mathbf{z}_j \tag{3.4a}$$

$$\begin{aligned}\mathrm{Var}\,\mathbf{y} &= c\,\mathrm{Var}\,\mathbf{y}^* \\ &= c\,\Sigma\,\omega_j\,(\mathbf{z}_j - E\,\mathbf{y}^*)\,(\mathbf{z}_j - E\,\mathbf{y}^*)^T.\end{aligned} \tag{3.4b}$$

This yields the following result, in the symmetric case $\omega_j \equiv 1/n$.

Corollary 3.1 The low-order moments of a filtered-variate symmetric MSV distribution are simply expressed in terms of the (unweighted) empirical moments of the filter points. If $\mathbf{y} = \mathbf{Z}\,\mathbf{w}$ and $\mathbf{w} \sim \mathrm{MSV}(\boldsymbol{\omega}, c)$ with $\boldsymbol{\omega} = n^{-1}\mathbf{1}_n$, then

$$E\,\mathbf{y} = \bar{\mathbf{z}} \qquad \text{and} \qquad \mathrm{Var}\,\mathbf{y} = c\,\mathbf{S}_z, \tag{3.5}$$

where $\bar{\mathbf{z}} = n^{-1}\,\Sigma\,\mathbf{z}_j$ and $\mathbf{S}_z = n^{-1}\Sigma\,(\mathbf{z}_j - \bar{\mathbf{z}})\,(\mathbf{z}_j - \bar{\mathbf{z}})^T$.

More can be said about the moments of all orders in the case of Dirichlet-distributed weights.

Theorem 3.5 Suppose $\mathbf{w} \sim D(\boldsymbol{\beta})$. Then the moments of the scaler $y = w_1 z_1 + \ldots + w_n z_n$ $(k=1)$ are equal to B. C. Carlson's (1977, p. 91) symmetrized multiple hypergeometric function R,

$$E\,y^r = R_r(\boldsymbol{\beta}, \mathbf{z}), \qquad \mathbf{z} = (z_1, \ldots, z_n)^T. \tag{3.6}$$

More generally, for a $k \times n$ matrix $\mathbf{Z}$, the vector $\mathbf{y} = \mathbf{Z}\,\mathbf{w}$ has mixed moments equal to Carlson's (1971) matrix-variable function, script $\mathcal{R}$,

$$E\,\Pi_1^k\,y_i^{r_i} = \mathcal{R}_{r_{tot}}(\boldsymbol{\beta}, \mathbf{Z}, -\mathbf{r}), \qquad r_{tot} = \Sigma_1^k\,r_i. \tag{3.7}$$

(See Dickey 1983 for an introduction to R and $\mathcal{R}$ for statisticians.)

Theorems 3.4 and 3.5 yield obvious and apparently new forms of Carlson functions for $r_{tot} = 1, 2$.

4. THE DISTRIBUTION ESTIMATE

Given a sample $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $k \times n$, and a proportionality constant $d \geq 1$, we define the set $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ of outward proportionally extended points,

$$\mathbf{z}_j = d(\mathbf{x}_j - \bar{\mathbf{x}}) + \bar{\mathbf{x}}, \tag{4.1}$$

$j = 1, \ldots, n$. These will then satisfy

$$\bar{\mathbf{z}} = \bar{\mathbf{x}} \qquad \text{and} \qquad \mathbf{S}_\mathbf{z} = d^2 \mathbf{S}_\mathbf{x} . \tag{4.2}$$

Since the inverse transformation is a convex combination, $\mathbf{x}_j = d^{-1}\mathbf{z}_j + (1 - d^{-1})\bar{\mathbf{z}}$, $j = 1, \ldots, n$, the original data points $\mathbf{x}_j$ are all contained within the convex hull of the new points, $\text{CHull}(\mathbf{z}_1, \ldots, \mathbf{z}_n)$. We let $\mathbf{w}$ have a symmetric mean-structured variance, $\mathbf{w} \sim \text{MSV}(n^{-1}\mathbf{1}_n, c)$, and then we use $\mathbf{w}$ and the new points to define the filtered-variate distribution with random vector $\mathbf{y}$,

$$\begin{aligned} \mathbf{y} &= \mathbf{Z}\,\mathbf{w} \\ &= w_1 \mathbf{z}_1 + \ldots + w_n \mathbf{z}_n, \end{aligned} \tag{4.3a}$$

or, equivalently,

$$\begin{aligned} \mathbf{y} &= d(\mathbf{X} - \bar{\mathbf{x}}\mathbf{1}_n^T)\,\mathbf{w} + \bar{\mathbf{x}} \\ &= w_1 d(\mathbf{x}_1 - \bar{\mathbf{x}}) + \ldots + w_n d(\mathbf{x}_n - \bar{\mathbf{x}}) + \bar{\mathbf{x}}, \end{aligned} \tag{4.3b}$$

Then in terms of $\tilde{\mathbf{x}}$ (2.1), $\mathbf{y} = d(\tilde{\mathbf{x}} - \bar{\mathbf{x}}) + \bar{\mathbf{x}}$.

The distribution of $\mathbf{y}$ (4.3) replaces that of $\tilde{\mathbf{x}}$ as our distribution estimate. The distribution of $\mathbf{y}$ is restricted to $\text{CHull}(\mathbf{z}_1, \ldots, \mathbf{z}_n)$, and by (3.5) and (4.2) it has the moments in terms of the sample moments of the data $\mathbf{X}$,

$$E\,\mathbf{y} = \bar{\mathbf{x}} \qquad \text{and} \qquad \text{Var}\,\mathbf{y} = c\,d^2\,\mathbf{S}_\mathbf{x} . \tag{4.4}$$

Interesting choices of parameters satisfy

$$d = c^{-1/2}, \tag{4.5}$$

for which $c\,d^2 = 1$, and so the variance of our constructed distribution matches the sample variance, $\text{Var}\,\mathbf{y} = \mathbf{S}_\mathbf{x}$. This leaves only one parameter to tune in the distribution estimate, one extreme of which, $c \uparrow 1$ ($d \downarrow 1$), gives again the sample empirical distribution. In the other extreme,

$c \downarrow 0$ ($d \uparrow \infty$), our distribution will remain a distribution of continuous type in the case of Dirichlet-distributed $\mathbf{w}$, and by the following lemma, if $\text{rank}(\mathbf{X} - \mathbf{x}_n \mathbf{1}_n^T) = k$, this limiting distribution will have the whole space R^k as its support.

Lemma 4.1 (Grunbaum 1967, p 3.) The dimension in R^k of the region $\text{CHull}(\mathbf{z}_1, \ldots, \mathbf{z}_n)$ is equal to $\text{rank}(\mathbf{z}_1 - \mathbf{t}, \ldots, \mathbf{z}_n - \mathbf{t})$ where $\mathbf{t}$ is any point in the region.

In the case of the exchangeable Dirichlet distribution for the random weights,

$$\mathbf{w} \sim D(\beta \mathbf{1}_n), \tag{4.7}$$

we have

$$c = (n\beta + 1)^{-1}, \qquad \text{so} \qquad d = (n\beta + 1)^{1/2}. \tag{4.8}$$

Then the constructed distribution will be of continuous type supported on $\text{CHull}(\mathbf{z}_1, \ldots, \mathbf{z}_n)$ if $0 < \beta < \infty$. The following theorem shows that in this case our revised distribution estimate becomes a normal distribution as $\beta \uparrow \infty$ ($c \downarrow 0$, $d \uparrow \infty$).

Lemma 4.2 The Dirichlet distribution $\mathbf{w} \sim D(\boldsymbol{\beta})$ is asymptotically normal as $\beta_{tot} \uparrow \infty$ for fixed $\boldsymbol{\omega} = \boldsymbol{\beta} / \beta_{tot}$.

The standardization in Lemma 4.2 involves dividing the vector $(w_1, \ldots, w_{n-1})$ by a matrix proportional to $c^{1/2} = (\beta_{tot} + 1)^{-1/2}$, which is essentially what we do when we form $\mathbf{y}$ (4.3b) with the choice of dilator, $d = c^{-1/2} = (n\beta + 1)^{1/2}$ (4.8). So we obtain, as our two extreme forms, the two usual contrasting methods of statistical practice, as follows.

Theorem 4.1 The distribution estimate based on the random vector $\mathbf{y} = \mathbf{Z}\,\mathbf{w}$, where $\mathbf{w} \sim D(\beta \mathbf{1}_n)$ and where $\mathbf{Z}$ dilates the data by (4.1) with $d = (n\beta + 1)^{1/2}$, is of continuous-type for $0 < \beta < \infty$, is supported on $\text{CHull}(\mathbf{Z})$, and has the two extreme limiting forms in β, for fixed sample size n:

(i) $\mathbf{y}$ has the sample empirical distribution of the data $\mathbf{X}$ for $\beta \downarrow 0$.

(ii) $\mathbf{y} \sim \text{Normal}(\bar{\mathbf{x}}, \mathbf{S}_\mathbf{x})$ for $\beta \uparrow \infty$.

Figure1 illustrates the method in the setting of Theorem 4.1 for moderate β. The raw data consists of $n=20$ scalar values ($k=1$),

$x_1, \ldots, x_{20}$, marked by triangles below the horizontal axis (Δ). Two values of β are examined, 0.06 and 0.40, and the positions of the corresponding outwardly extended points, $z_1, \ldots, z_{20}$, from eqs. (4.1), (4.8), are also marked. Note that the points are much more widely spaced for the larger β=0.40. The corresponding density estimates are drawn (solid line and long-dashed line) and they show that the difference in dilation has little effect on the essential tail-length of the two distributions. As is generally the case, the density for the larger β value is smoother and more like a normal distribution. As $\beta \uparrow \infty$, the density approaches the line of short dashes, the normal density with mean and variance matching that of the raw sample data. The sample data are skewed, so the normal distribution does not model the empirical distribution well. In particular, the smallest datum is very far into the left tail of the normal distribution, in contrast to its position relative to the filtered-variate Dirichlet distributions. (The data are drawn from a chi-squared distribution with 6 degrees of freedom.) Effects of n will be addressed in later sections.

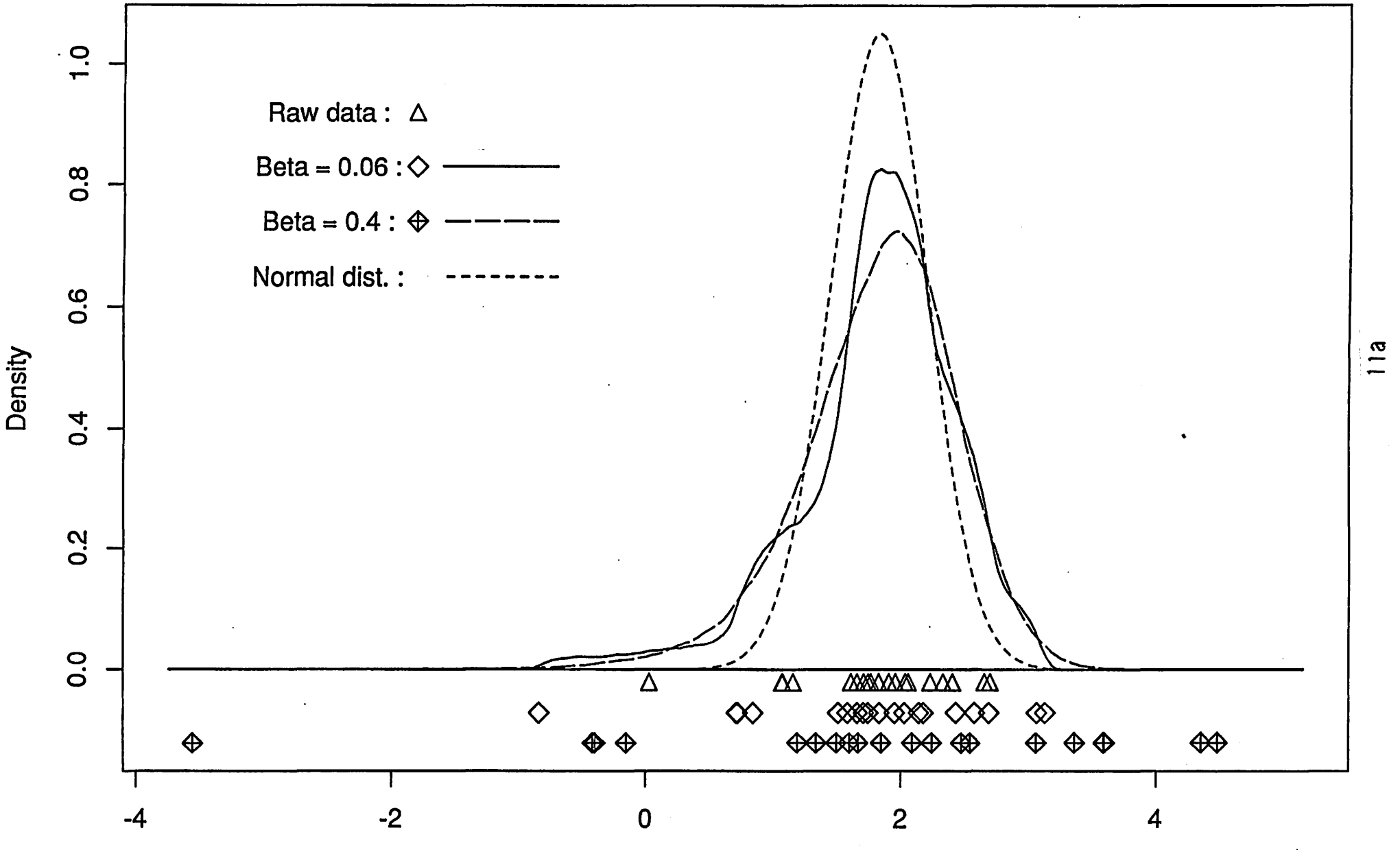
[INSERT FIGURE 1 HERE WITH THE FOLLOWING CAPTION:]

Figure 1. Distribution estimates and limiting normal distribution, as $\beta \uparrow \infty$, with data points and dilated points.

In many problems, a sampling distribution is considered to have its support restricted to a subregion of the space R^k. Often, there is a natural invertible transformation that maps the support onto the whole space, where the proposed method can be applied. Then the distribution estimate obtained for such transformed data can yield, by a change of variable, a distribution estimate on the original subregion. If the data vectors $\boldsymbol{\xi}_j$, $j = 1, \ldots, n$, have all their coordinates positive, then the method can be applied to their logarithms $\mathbf{x}_j = \log(\boldsymbol{\xi}_j)$ (evaluated coordinate-by-coordinate), to produce a random vector $\mathbf{y}$, which then yields a **log filtered-variate Dirichlet distribution** with the random vector $\boldsymbol{\eta} = \exp(\mathbf{y})$. Such a distribution for $\boldsymbol{\eta}$ is nearly as easy to use for simulation as a direct filtered-variate Dirichlet.

A Bayesian posterior predictive distribution is typically more diffuse than the sampling distribution. Consider that $\mathrm{Var}(x_{n+1} \mid x_1, \ldots, x_n) = E[\mathrm{Var}(x_{n+1} \mid \theta) \mid x_1, \ldots, x_n] + \mathrm{Var}[E(x_{n+1} \mid \theta) \mid x_1, \ldots, x_n]$. In the problem of constructing a predictive distribution, unlike the problem of estimating a population distribution, one may wish to assert a more diffuse distribution

Figure 1.


Raw data :
Beta = 0.06 :
Beta = 0.4 :
Normal dist. :
Density
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
11a

than the sample empirical distribution, choosing $c\,d^2 = 1 + \varepsilon$, say, for which Var $\mathbf{y} = (1 + \varepsilon)\,\mathbf{S_x}$.

We end this section by presenting results of simulations in which filtered-variate Dirichlet distributions with various values of β were compared to the population distributions from which the filter points were sampled. Six populations that were used are listed in Table 1. (The "halved-exponential" density there is $p(x) = (1/3)\,e^{x}$, $x < 0$; $(1/3)\,e^{-(x/2)}$, $x > 0$.) A random sample of size 20 was generated from a population distribution and three filtered-variate Dirichlet distributions were determined for the sample, with $\beta = 1/19, 3/20, 8/20$. Each such distribution was then compared to the population distribution and the Kolmogorov-Smirnov statistic (K-S) was computed. This was repeated for 50 samples of size 20 from each population distribution and the observed mean and standard deviation of the K-S statistic were determined, as shown in Table 1.

Table 1. Difference between filtered-variate Dirichlet distribution estimate and population for various populations and values of β. Mean and standard deviation (s.d.) of Kolmogorov-Smirnov statistics for fifty samples of size 20 from each population distribution.

Distribution	β-Value						Std. error of
(deg. fr.)	1/19		3/20		8/20		diff. between
	Mean	S.D.	Mean	S.D.	Mean	S.D.	two means
Chi-sq. (6)	0.115	0.059	0.111	0.061	0.114	0.059	0.0122
Log Chi-s. (6)	0.116	0.055	0.109	0.055	0.106	0.054	0.0054
Stud.-t (4)	0.121	0.053	0.112	0.054	0.114	0.054	0.0031
Normal	0.116	0.058	0.103	0.056	0.097	0.054	0.0021
Halved expon.	0.132	0.061	0.132	0.063	0.135	0.063	0.0024
Uniform	0.130	0.056	0.129	0.055	0.129	0.054	0.0016

For each population distribution a two-way analysis of variance was performed with the different β-values and the different samples as the two factors. The difference in β-values had a statistically significant effect for the Student-t population ($P < .05$) and the normal population ($P < .001$). In each case, the K-S statistic for $\beta = 1/19$ was poorer than for each of the larger β-values, and for the normal population, $\beta = 3/20$ was poorer than $\beta = 8/20$. This ordering for the normal distribution might have been anticipated from Theorem 4.1, according to which the limiting form of the filtered-variate Dirichlet as β increases is normal. The smallest value

that we used, $\beta = 1/19 = 1/(n-1)$, is the smallest possible for a continuous density function, as will be shown in Section 7.

If one were to test whether an empirical distribution for a sample of size 20 came from a particular distribution, the 5% critical value of the K-S statistic would be 0.29. This number was always more than two sample standard deviations above the sample mean of the K-S values for the filtered-variate Dirichlet distributions, suggesting that these distributions could provide a better approximation to the population distribution than is given by the empirical distribution. This does not mean that the filtered-variate Dirichlet distribution will be far from the empirical distribution, as illustrated by Figure 2, where the filtered-variate Dirichlet (solid line) is close to both the true population (dotted line) and the empirical distribution (step-function of dashes). Notice, though, that the solid line in the figure is closer to the population distribution than would be obtained by an arbitrary method of smoothing the empirical distribution.

We have suggested that a transformation could be used if the population distribution has a restricted support set. The effect of a logarithmic transformation for a chi-squared distribution is indicated in Table 1. It suggests that a transformation is more beneficial as β increases. This may be because, with no transformation, the dilation into the region where support is zero is greater for larger β-values.

[INSERT FIGURE 2 HERE WITH THE FOLLOWING CAPTION:]

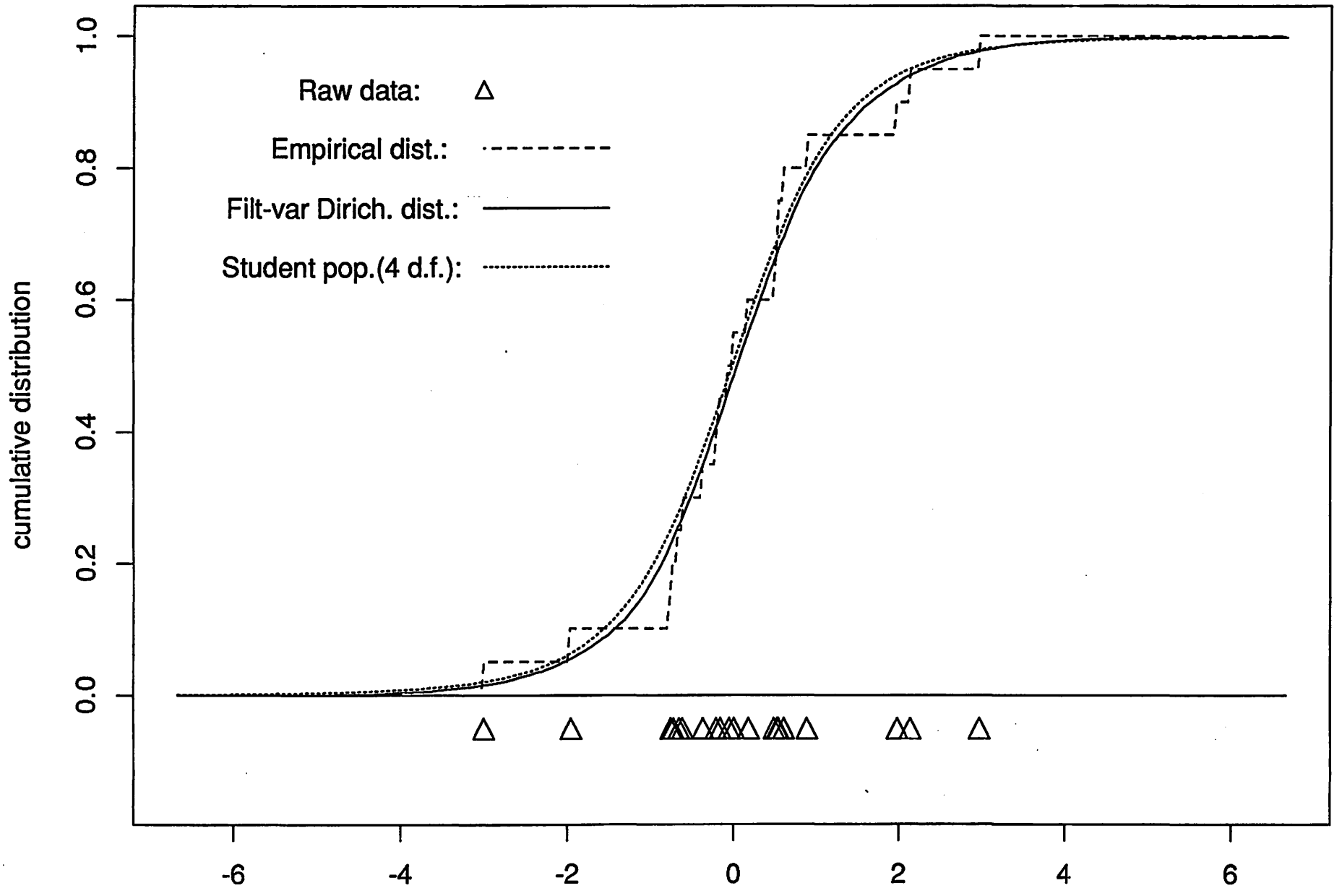
Figure 2. Population distribution, empirical distribution, and filtered-variate Dirichlet with $\beta = 3/20$. The population distribution is a Student-t (4 d.f.), from which a sample of 20 was drawn and used to form the other two distributions.

5. A MARGINALIZATION PROPERTY: B-SPLINES AND THE DISTRIBUTION AS A PROJECTION

The proposed distribution estimate, a filtered-variate Dirichlet distribution with the dilated data as the filter points, is easy to use for simulation, since the Dirichlet random vector has coordinates distributed like the ratios of independent chi-squared or gamma random variables to their sum. We develop mathematics of the distribution and its density in this Section and the next two Sections.

The distribution of a linear function of a Dirichlet vector has been important, at least, since Bloch and Watson (1967) and interest continues in more recent work of Margolin (1977), Jiang (1984), Ignatov and Kaishev

Figure 2.


Raw data:
Empirical dist.:
Filt-var Dirich. dist.:
Student pop.(4 d.f.):
cumulative distribution
0.0
0.2
0.4
0.6
0.8
1.0
-6
-4
-2
0
2
4
6

(1987), and others. [For the extension to the distribution of an integral of a Dirichlet process, see Section 8, to follow, and references cited there.] The density is especially easy in one particular case. This is the full-dimensional or unprojected case of $\mathbf{y} = \mathbf{Z}\,\mathbf{w}$, $\mathbf{w} \sim D(\boldsymbol{\beta})$, $\mathbf{Z}$ $k \times n$, where $k = n-1$ and $\text{rank}(\mathbf{Z} - \mathbf{z}_n \mathbf{1}_n^T) = n-1$. Then the filter transformation is invertible and the density of $\mathbf{y}$ can be obtained from the density of $\mathbf{w}$, just by a simple substitution with constant Jacobian,

$$p_{\mathbf{y}}(\mathbf{y}) = p_{w_1, \ldots, w_{n-1}}[(z_1 - z_n, \ldots, z_{n-1} - z_n)^{-1}\,\mathbf{y}] \Big/ |\det(z_1 - z_n, \ldots, z_{n-1} - z_n)|. \qquad (5.1)$$

Although a value of n as small as $k+1$ is usually not of direct interest, the density (5.1) is useful in the more general case as an initial function in two methods: (i) marginalization, or projective collapse of the dimension k; and (ii) iteration by increase of the sample size n. We discuss the first method (i), and we turn to the second method in the next section.

A marginalization property. A linear transformation of a filtered-variate distribution is again filtered-variate with the same underlying random vector $\mathbf{w}$. If $\mathbf{y} = \mathbf{Z}\,\mathbf{w}$, then $\mathbf{B}\,\mathbf{y} = (\mathbf{B}\,\mathbf{Z})\,\mathbf{w}$. In particular, a coordinate projection preserves the filter structure: if $\mathbf{y} = (\mathbf{y}^{(1)T}, \mathbf{y}^{(2)T})^T$ and $\mathbf{Z} = (\mathbf{Z}^{(1)T}, \mathbf{Z}^{(2)T})^T$ are partitioned conformably, and $\mathbf{y} = \mathbf{Z}\,\mathbf{w}$, then $\mathbf{y}^{(1)} = \mathbf{Z}^{(1)}\,\mathbf{w}$, which is constructed in the same direct way from $\mathbf{w}$ and the data subvector $\mathbf{X}^{(1)}$ of the conformably partitioned $\mathbf{X} = (\mathbf{X}^{(1)T}, \mathbf{X}^{(2)T})^T$. Thus, in a very strong sense, the method is invariant to the dimensionality k of the sample vectors. The distribution estimate obtained directly from a few coordinates of the data vectors is identical to the corresponding marginal distribution of the distribution estimate obtained from the full dimensional data vectors, provided only that the dilation factor does not depend on k. (This is true even if $\mathbf{w}$ is not Dirichlet, or even MSV.) Beyond the logical attraction of this feature, it can be important as a conceptual and computational aid.

Given a set of n points in R^k, $\mathbf{Z} = (z_1, \ldots, z_n)$, such as an outward dilation of sample data by (4.1), we define the **projected-hypervolume distribution** as the distribution of the random k-dimensional vector $\mathbf{y} = \mathbf{Z}\,\mathbf{w}$ where $\mathbf{w}$ has the uniform distribution on $\Delta^{(n)}$. That is, $\mathbf{w}$ has the exchangeable Dirichlet distribution with unit parameter ($\beta = 1$), $\mathbf{w} \sim D(\mathbf{1}_n)$. To motivate this definition, when $n \geq k + 1$, just imagine augmenting the $k \times n$ matrix $\mathbf{Z}$ by appending $(n-1) - k$ further rows, if necessary, to obtain an $(n-1) \times n$ matrix, denoted by $\tilde{\mathbf{Z}} = (\tilde{z}_1, \ldots, \tilde{z}_n)$, for which $\text{rank}(\tilde{\mathbf{Z}} - \tilde{z}_n \mathbf{1}_n^T) = n-1$. Then, by Lemma 4.1, the region $\text{CHull}(\tilde{z}_1, \ldots, \tilde{z}_n)$ is of full dimension in R^{n-1}, and $\tilde{\mathbf{y}} = \tilde{\mathbf{Z}}\,\mathbf{w}$ is

uniformly distributed over this polytope. So $\tilde{\mathbf{y}}$ has an (n-1)-dimensional hypervolume distribution, and its subvector $\mathbf{y}$ has our more general k-dimensional projected-hypervolume distribution. If, as is commonly assumed, the original array of n data $\mathbf{x}_j$'s, and hence the $\mathbf{z}_j$'s, are of full rank, then it is possible, by the following lemma, to construct such augmented $\tilde{\mathbf{Z}}$. For example, in one dimension (k=1), this requires only that the data not have an n-way tie at a single numerical value.

Lemma 5.1 Assume $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$, $k \times n$, $n \geq k+1$, and $\text{rank}(\mathbf{Z} - \mathbf{z}_n \mathbf{1}_n^T) = k$. Without loss of generality, the n vectors can be ordered so that the first k vectors and the nth vector satisfy $\text{rank}(\mathbf{z}_1 - \mathbf{z}_n, \ldots, \mathbf{z}_k - \mathbf{z}_n) = k$. Define the $(n-1) \times n$ matrix $\tilde{\mathbf{Z}} = (\mathbf{Z}^T, \mathbf{J}^T)^T$, where $\mathbf{J} = (\mathbf{0}_{h \times k}, \mathbf{I}_h, \mathbf{0}_{h \times 1})$ with $h = n-1-k$. Then $\text{rank}(\tilde{\mathbf{Z}} - \tilde{\mathbf{z}}_n \mathbf{1}_n^T) = n-1$.

An amazing feature of a projected-hypervolume distribution is that it depends in no way on which (n-1)-dimensional figure has its normalized hypervolume projected to form the probability, provided its vertices $\tilde{\mathbf{z}}_j$ have the specified projections $\mathbf{z}_j$. The distribution of $\mathbf{y} = \mathbf{Z}\mathbf{w}$ on $\text{CHull}(\mathbf{z}_1, \ldots, \mathbf{z}_n)$ in R^k is the marginal distribution of the uniform distribution on $\text{CHull}(\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_n)$ in R^{n-1} for any augmentation $\tilde{\mathbf{Z}}$ of $\mathbf{Z}$ for which $\text{rank}(\tilde{\mathbf{Z}} - \tilde{\mathbf{z}}_n \mathbf{1}_n^T) = n-1$. Figure 1 illustrates a hypervolume projection for one-dimensional data, k=1, consisting of n=4 data points including one tie. The solid body in $n-1 = 3$ dimensions is depicted relative to the first two axes in the plane of the page and the third axis rising perpendicularly from the page. It has $n = 4$ vertices, $\tilde{\mathbf{z}}_1$, $\tilde{\mathbf{z}}_2$, $\tilde{\mathbf{z}}_3$, and $\tilde{\mathbf{z}}_4$, which can be placed arbitrarily, except for their first coordinates , z_1, z_2, z_3, and z_4, which are fixed at 0, 2, 2, and 4, respectively. Equivalent alternative point values are shown for $\tilde{\mathbf{z}}_4$. The projected-volume density plotted here is proportional to the cross-sectional area of the solid figure. This is easily seen to be the piecewise quadratic: $p(y) = (3/16)\, y^2$, if $0 \leq y \leq 2$, and $(3/16)(4-y)^2$, if $2 \leq y \leq 4$.

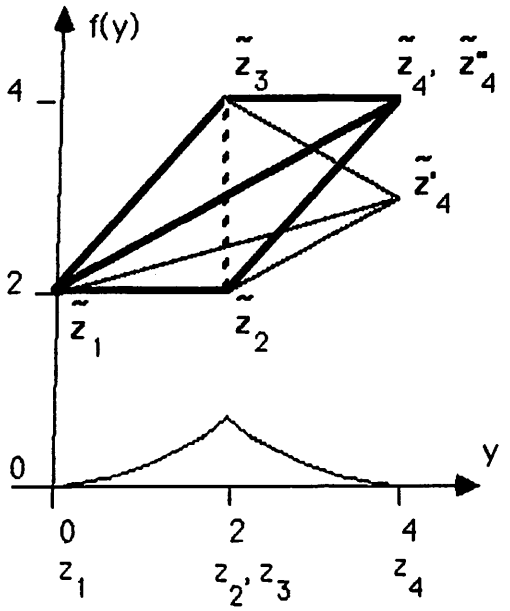


Figure 3. A projected-volume distribution, for k = 1 and n = 4. (The third axis rises perpendicularly from the plane of the page.)

Somewhat more can be said about a projected-hypervolume distribution in only k=1 dimension. Suppose the specified scalar values are ordered and labeled according to $z_1 \le z_2 \le \ldots \le z_n$, and $z_1 < z_n$. Then the corresponding vectors $\tilde{z}_1, \ldots, \tilde{z}_n$, as given by Lemma 5.1, are the vertices of a simplex Smp_{n-1} in n-1 dimensions, and the random variable $y = z_1w_1 + \ldots + z_nw_n$, where $\mathbf{w} \sim D(\mathbf{1}_n)$, has a density proportional to the (n-2)-dimensional hypervolume of the polytope Ptp_{n-2}, the cross-section of Smp_{n-1} at the first coordinate equal to y. Cross sections of simplices are developed in detail in Sommerville (1958), with results that justify the following statements regarding special cases.

Theorem 5.1 Suppose k = 1, $z_1 \le z_2 \le \ldots \le z_n$, and $y = z_1w_1 + \ldots + z_nw_n$, where $\mathbf{w} \sim D(\mathbf{1}_n)$.

(i) If y is just inside the extreme dilated order statistic, $z_1 \le y \le z_2$, then the cross-section Ptp_{n-2} is also a simplex, and y has the density

$$f(y) = (n-1)(y - z_1)^{n-2} / \prod_2^n (z_j - z_1), \qquad (5.2)$$

and analogously for $z_{n-1} \le y \le z_n$.

(ii) If $n = 4$ and $z_2 \le y \le z_3$,

$$f(y) = [3/(z_2-z_1)] \{ (y-z_1)^2/[(z_3-z_1)(z_4-z_1)] - (y-z_2)^2/[(z_3-z_2)(z_4-z_2)] \}. \qquad (5.3)$$

Proof. (i) For the vectors $\tilde{z}_j$ constructed as in Lemma 5.1, Ptp_{n-2} is an orthotope, with volume proportional to the product of its orthogonal measurements. So

$$f(y) = Vol(Ptp_{n-2}) / Vol(Smp_{n-1})$$

$$= \{ [1/(n-2)!] \prod_2^{n-1} [(y - z_1)/(z_j - z_1)] \}$$

$$/ \{ [1/(n-1)!] |det(\tilde{z}_1-\tilde{z}_n, \ldots, \tilde{z}_{n-1}-\tilde{z}_n)| \}. \qquad (5.4)$$

(ii) The derivation is elementary in the case $n = 4$. The plane cross-sectional region is the difference of two right triangles with co-linear legs.

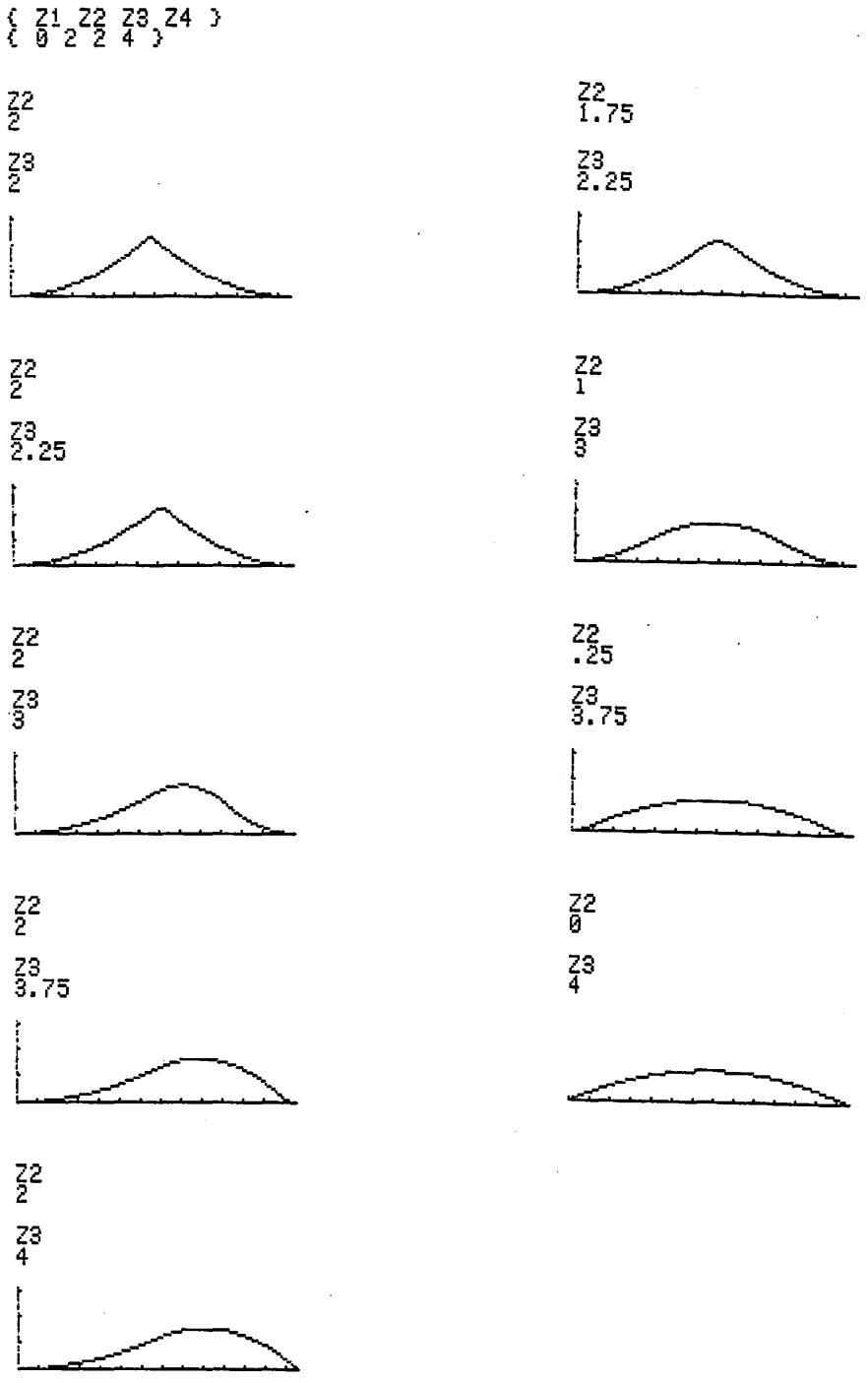
The effect of the placement of the z_j's when $n = 4$ is illustrated in Figure 4. Included as the first plot is the density of Figure 3.

[INSERT FIGURE 4 HERE WITH THE FOLLOWING CAPTION:]

Figure 4. Influence of the filter points on the filtered-variate Dirirchlet with $n = 4$.

Our general filtered-variate Dirichlet distribution, with $\mathbf{w} \sim D(\boldsymbol{\beta})$ for more general $\boldsymbol{\beta}$, is just a straightforward generalization of a projected-hypervolume distribution. Lemma 5.1 and Eq. (5.1) can be used in the same way again to construct an $(n-1)$-dimensional density, which can be marginalized, analytically or numerically. In the example of Figure 3, $y = 2w_2 + 2w_3 + 4w_4$, where $(w_1, w_2, w_3, w_4) \sim D(\beta, \beta, \beta, \beta)$ for $\beta = 1$. If we modify this by taking $\beta = 2$, then y is found to have the marginal density: $p(y) = (63/384)\, y^5[1 - (5/12)y]$ if $0 \le y \le 2$, and $(63/384)\,(4-y)^5 [1 - (5/12)(4-y)]$ if $2 \le y \le 4$.

Figure 4.


{ Z1 Z2 Z3 Z4 }
{ 0 2 2 4 }
Z2
2
Z3
2
Z2
1.75
Z3
2.25
Z2
2
Z3
2.25
Z2
1
Z3
3
Z2
2
Z3
3
Z2
.25
Z3
3.75
Z2
2
Z3
3.75
Z2
0
Z3
4
Z2
2
Z3
4

Univariate B-splines were interpreted geometrically as marginal densities of hypervolume distributions by Curry and Schoenberg (1966), and multivariate B-splines were introduced as such by C. de Boor (1976). Ignatov and Kaishev (1987) have shown that for any dimension k if all (or nearly all) the coordinates of the parameter $\boldsymbol{\beta}$ are integers, then the filtered-variate Dirichlet density is equal to a multivariate B-spline $M(\mathbf{y})$, as follows.

Theorem 5.2 The k-dimensional random vector $\mathbf{y} = \mathbf{Zw}$, $\mathbf{w} \sim D(h_1, \ldots, h_n)$, for integer h_j , $j = 1, \ldots, n$, has probability density equal to the multvariate B-spline,

$$p(\mathbf{y}) = M(\mathbf{y}; z_1, \ldots, z_1, \ldots, z_n, \ldots, z_n), \tag{5.5}$$

where $\mathbf{Z} = (z_1, \ldots, z_n)$ and each vector z_j is repeated h_j times as a parameter of M, for $j = 1, \ldots, n$.

In our application of Theorem 5.2, a common integer-valued parameter coordinate β will imply projection from a constant hypervolume density in $\beta(n-1)$ dimensions. Contrast this with the projection from a more general density in n-1 dimensions, introduced in a paragraph following Theorem 5.1. Ignatov and Kaishev use recurrence formulas for M from the literature and they derive the following divided-difference formulas for computing the density, the c.d.f., and the moments of y, for $k = 1$ and integer-valued Dirichlet parameters. (Similar results hold for **nearly** all of the parameters integer-valued.) Related expressions are given by Margolin (1977).

Define the divided difference,

$$[t_1, \ldots, t_m]_u f(u) = \\ ([t_2, \ldots, t_m]_u f(u) - [t_1, \ldots, t_{m-1}]_u f(u)] / (t_m - t_1), \tag{5.6}$$

with $[t_1]_u f(u) = f(t_1)$. The univariate B-spline is

$$M(t; t_1, \ldots, t_m) = [t_1, \ldots, t_m]_u g(u), \tag{5.7}$$

where $g(u) = (m-1)[(u - t)_+]^{m-2}$, where $(v)_+ = \max[0, v]$. For $k = 1$ the c.d.f. of the distribution in (5.5) can be computed for $y \in \text{CHull}(z_1, \ldots, z_n)$ by

$$P(y) = 1 - [z_1, \ldots, z_1, \ldots, z_n, \ldots, z_n]_u H(u), \tag{5.8}$$

where, for $h_{tot} = \sum_1^k h_i$, $H(u) = [(u - y)_+]^{h_{tot}-1}$, $H(z_1) = 0$. Then the density (5.5) for k = 1 can be computed as the following derivative of (5.8),

$$p(y) = [z_1, \ldots, z_1, \ldots, z_n, \ldots, z_n]_u \, H'(u), \tag{5.9}$$

where $H'(u) = (h_{tot} - 1)\,[(u - y)_+]^{h_{tot}-2}$. Expectations can be computed as divided differences of antiderivatives,

$$E[f(y)^{(h_{tot}-1)}/(h_{tot}-1)!] = [z_1, \ldots, z_1, \ldots, z_n, \ldots, z_n]_u \, f(u), \tag{5.10}$$

for any functions $f \in C^{(h_{tot}-1)}$ on $CHull(z_1, \ldots, z_n)$.

We have seen that the moments of a filtered-variate Dirichlet random vector are Carlson $\mathcal{R}$ functions (Theorem 3.5). By Theorem 5.2, this yields an elegant and apparently new relation between Carlson's functions and multivariate B-splines, expressing $\mathcal{R}$ as a k-dimensional integral; and for k=1, expressing R as a divided difference.

Corollary 5.1 For $h = (h_1, \ldots, h_n)$ with all coordinates h_j integer-valued, $r_{tot} = \sum_1^k r_i$, and M as in (5.5), the **r**th mixed moment of vector **y** takes the form,

$$\mathcal{R}_{r_{tot}}(\mathbf{h}, Z, -\mathbf{r}) = \int_{R^k} (\prod_1^k y_i^{r_i})\, M(\mathbf{y}; z_1, \ldots, z_1, \ldots, z_n, \ldots, z_n)\, d\mathbf{y}; \tag{5.11}$$

and by (5.10), for k = 1, the rth moment of scalar y is

$$R_r(\boldsymbol{\beta}, \mathbf{z}) = [z_1, \ldots, z_1, \ldots, z_n, \ldots, z_n]_u \, u^{r+h_{tot}-1} \,/\, C_r^{r+h_{tot}-1}, \tag{5.12}$$

in terms of the usual combinatorial coefficient.

6. THE DENSITY BY CONVOLUTION

We develop the general filtered-variate Dirichlet density by iteration with respect to increasing sample size, in the following theorem.

Lemma 6.1 (Adapted from Wilks 1962, pp 180-181.) Write, in terms of the scalars w_n and β_n,

$$\mathbf{w}^{(n)} = \begin{pmatrix} (1-w_n)\mathbf{w}^{(n-1)} \\ w_n \end{pmatrix}, \qquad \boldsymbol{\beta}^{(n)} = \begin{pmatrix} \boldsymbol{\beta}^{(n-1)} \\ \beta_n \end{pmatrix}. \tag{6.1}$$

Then $\mathbf{w}^{(n)} \sim D(\boldsymbol{\beta}^{(n)})$ if and only if

$$\mathbf{w}^{(n-1)} \sim D(\boldsymbol{\beta}^{(n-1)}), \qquad w_n \sim \text{beta}(\beta_n, \beta_1+\ldots+\beta_{n-1}), \tag{6.2}$$

and $\mathbf{w}^{(n-1)}$ and w_n are independent.

Theorem 6.2 If $\mathbf{y}^{(n-1)} = (z_1, \ldots, z_{n-1})\mathbf{w}^{(n-1)}$ and $\mathbf{y}^{(n)} = (z_1, \ldots, z_n)\,\mathbf{w}^{(n)}$ where $\mathbf{w}^{(n-1)}$, w_n, and $\mathbf{w}^{(n)}$, satisfy the conditions of Lemma 6.1, then

$$\mathbf{y}^{(n)} = (1 - w_n)\,\mathbf{y}^{(n-1)} + w_n\, z_n, \tag{6.3}$$

where $w_n \sim \text{beta}(\beta_n, \beta_1+\ldots+\beta_{n-1})$ independently of $\mathbf{y}^{(n-1)}$. Thus, the densities of $\mathbf{y}^{(n)}$ and $\mathbf{y}^{(n-1)}$ are related by

$$p_{\mathbf{y}^{(n)}}(\mathbf{y}) = \int_0^1 p_{\mathbf{y}^{(n-1)}}[(1-w)^{-1}(\mathbf{y} - w\, z_n)]\; b(w;\, \beta_n,\, \beta_1+\ldots+\beta_{n-1})\, dw. \tag{6.4}$$

The first factor in the integrand of (6.4) is nonzero only when its argument lies in $\text{CHull}(z_1, \ldots, z_{n-1})$. This restriction can be difficult to handle for a moderate or large dimension k. But for $k = 1$, the range of $\mathbf{y}$ and w is

$$w\, z_n + (1-w)\min\{z_1, \ldots, z_{n-1}\} \le y \le w\, z_n + (1 - w)\max\{z_1, \ldots, z_{n-1}\} \tag{6.5}$$

and $0 \le w \le 1$.

7. POLES IN THE DENSITY

As in the beta distribution, small parameter values in a Dirichlet distribution will result in the presence of poles in the density function, and this is also true of a filtered-variate Dirichlet distribution. This can be an important consideration for a method purporting to estimate a population distribution judged to be smooth and to have a smooth density function. The following theorems establish the conditions under which our distribution estimate has a finite density.

Lemma 7.1 (Poles of a Dirichlet density). If $\mathbf{w} \sim D(\boldsymbol{\beta})$ on $\Delta^{(n)}$, the density of $w_1, \ldots, w_{n-1}$ is finite at $\mathbf{w} = \boldsymbol{\delta}_{(j)}$ for specified j ($j = 1, \ldots, n-1$), if and only if

$$\sum_{j' \neq j} \beta_{j'} \geq n - 1 . \qquad (7.1)$$

That is, all the *other* parameters, $\beta_{j'}$, $j' \in \{1, \ldots, n\} - \{j\}$, must have a large enough sum.

Proof. For simplicity, consider $j = 1$. On a straight line through the point $\boldsymbol{\delta}_{(1)}$, the density takes the following form, for constants $a_j > 0$, $j = 2, \ldots, n$, and $\varepsilon > 0$,

$$p_{w_1, \ldots, w_{n-1}}\left[1 - \left(\sum_{j=2}^{n} a_j\right)\varepsilon, \quad a_2\varepsilon, \ldots, a_{n-1}\varepsilon\right] =$$

$$B(\boldsymbol{\beta})^{-1}\left[1 - \left(\sum_{j=2}^{n} a_j\right)\varepsilon\right]^{\beta_1 - 1} \prod_{j=2}^{n} (a_j\varepsilon)^{\beta_j - 1}. \qquad (7.2)$$

This has a finite limit as $\varepsilon \to 0$ only when $\sum_{j=2}^{n} \beta_j \geq n-1$.

A Dirichlet distribution $\mathbf{w} \sim D(\boldsymbol{\beta})$ on $\Delta^{(n)}$, $n \geq k+1$, has a marginal density in R^k that can be viewed as being the density of a filtered-variate Dirichlet distribution with the n vertices in R^k, $\boldsymbol{\delta}_{(1)}, \ldots, \boldsymbol{\delta}_{(k)}, \mathbf{0}, \ldots, \mathbf{0}$, with $n - k$ ties at zero. Since Dirichlet distributions have a marginalizaton or grouping property (Theorem 3.1 (ii)), this is also the same as the density in R^k of the Dirichlet distribution $D(\boldsymbol{\alpha})$ on $\Delta^{(k+1)}$, where $\alpha_j = \beta_j$ $(j = 1, \ldots, k)$ and $\alpha_{k+1} = \beta_{k+1} + \ldots + \beta_n$. Then according to Lemma 7.1, for this filtered-variate Dirichlet, the condition for a finite density at $\boldsymbol{\delta}_{(1)}$, say, is that $\sum_{j=2}^{k+1} \alpha_j \equiv \sum_{j=2}^{n} \beta_j \geq k$. But this condition applies quite generally to filtered-variate Dirichlet distributions, since the property of having a finite density at a filter point is invariant to local changes in position of other filter points that are distant from the point of interest. Thus, we have the following result for filtered-variate Dirichlet distributions, allowing for ties at filter points.

Theorem 7.1 (Poles of a filtered-variate Dirichlet density.) If $\mathbf{y} = \mathbf{Z}\mathbf{w}$ on R^k, where $\mathbf{w} \sim D(\boldsymbol{\beta})$ on $\Delta^{(n)}$, then for any specified j $(j = 1, \ldots, n)$, the density of $\mathbf{y}$ is finite at $\mathbf{z}_j$ if and only if

$$\sum_{j' \notin T(j)} \beta_{j'} \geq k , \qquad (7.3)$$

where $T(j) = \{j' : \mathbf{z}_{j'} = \mathbf{z}_j\}$.

Corollary 7.1 (Symmetric parameters). If $\mathbf{y} = \mathbf{Zw}$ on R^k, where $\mathbf{w} \sim D(\beta \mathbf{1}_n)$ on $\Delta^{(n)}$, and if there are no ties among the filter points $\mathbf{z}_1, \ldots, \mathbf{z}_n$, then at every such point $\mathbf{y} = \mathbf{z}_j$, the density of $\mathbf{y}$ is finite if and only if

$$\beta \geq k / (n - 1). \tag{7.4}$$

According to this result, a filtered-variate Dirichlet distribution, with a symmetric parameter vector and no ties among its filter points, will have no poles in its density whenever β is large enough to satisfy (7.4). But if $\beta < k / (n - 1)$ there will be poles at all n filter points. This effect is shown in Figures 5 and 6 for $k = 1$ and $n = 5$, for which the threshold becomes $\beta \geq .25$.

[INSERT FIGURES 5 AND 6 HERE WITH THE FOLLOWING CAPTIONS:]

Figure 5. A filtered-variate Dirichlet density with 5 poles, for $\beta = 0.2$. . The filter points were 5 equally spaced points, and the distribution was simulated by generating 100,000 random deviates.

Figure 6. A density similar to Figure 5, but without poles, for $\beta = 0.4$.

In larger dimensions k, whole line segments and other convex closures of subsets of $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ can bear infinite density. Theorem 7.1 generalizes to provide conditions preventing such singularities of the density, as follows.

Theorem 7.2 (Convex sets of singularities). In the setting of Theorem 7.1, if S is a subset of $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ of cardinality $|S| < k + 1$, then the density of $\mathbf{y}$ is finite on the interior of CHull(S) if and only if

$$\sum_{j \notin T(S)} \beta_j \geq k + 1 - |S| , \tag{7.5}$$

where $T(S) = \{j: \mathbf{z}_j \in S\}$.

In the symmetric-parameters case, $\boldsymbol{\beta} = \beta \mathbf{1}_n$, the condition of Theorem 7.2 is that

$$\beta \geq (k + 1 - |S|) / (n - |S|). \tag{7.6}$$

Note that if this is satisfied for single points ($|S| = 1$), it is automatically satisfied for larger subsets ($|S| > 1$).

Figure 5.

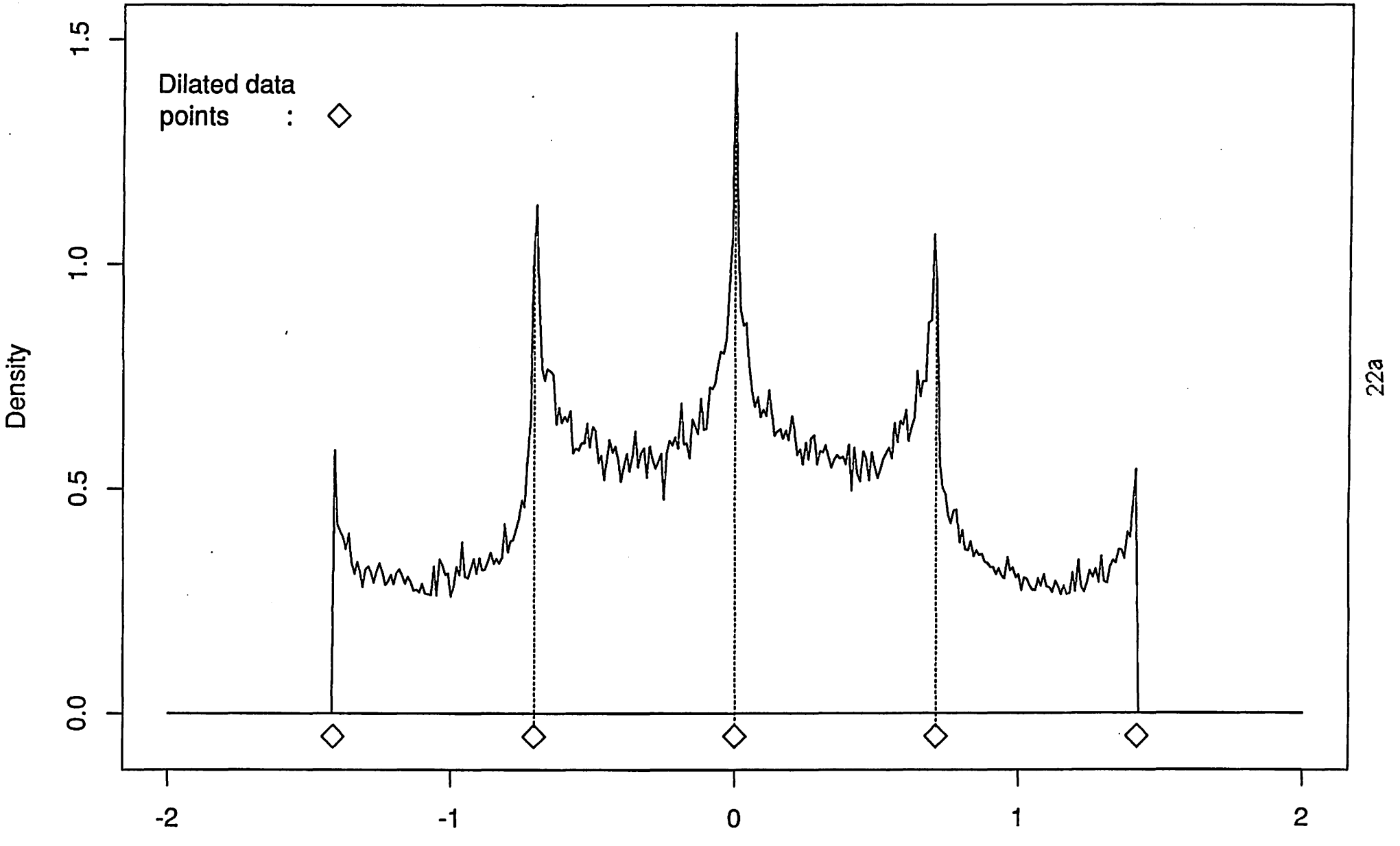
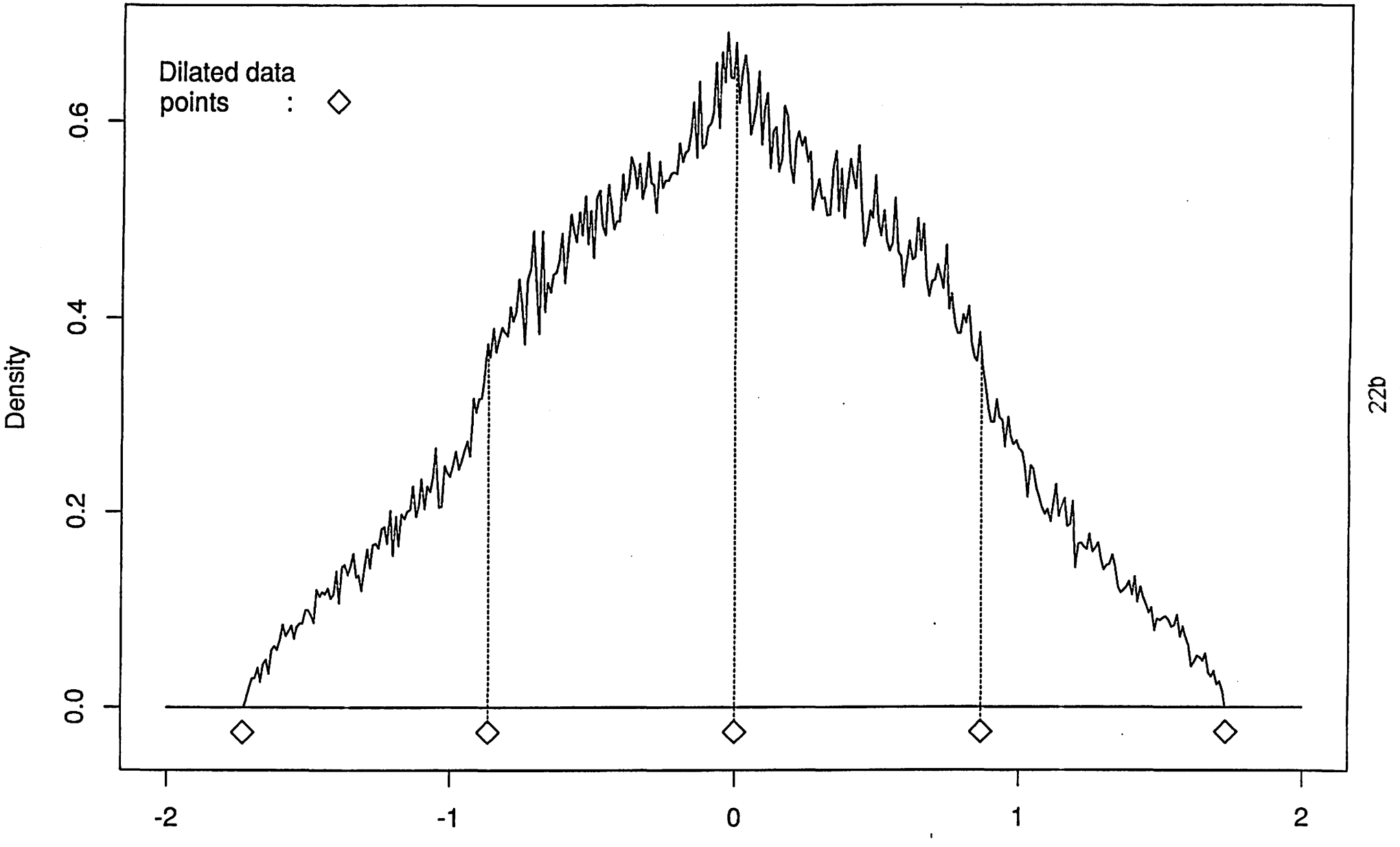


Figure 6.



8. INTERPRETATION AS THE RANDOM MEAN OF A DIRICHLET PROCESS

What happens to our distribution estimate as the sample size n increases? Is it consistent in the traditional sense of convergence to the population distribution as $n \to \infty$? A far more relevant question is, how close will it be to the population distribution for typical sample sizes. "Close" can be defined in many ways, including ways that refer to the smoothness of the distribution estimate relative to the smoothness of the population distribution. A complete answer to this important second question would seem to require extensive experience with the method, simulated or real. We turn in this section to the first question. It will lead quite naturally to consideration of a generalization of our filtered-variate Dirichlet distribution, the distribution of a random mean of a Dirichlet process.

For each sample size n, let

$$\mathbf{y}^{(n)} = \mathbf{Z}^{(n)} \mathbf{w}^{(n)} \tag{8.1}$$

on R^k, where $\mathbf{w}^{(n)} \sim D(\boldsymbol{\beta}^{(n)})$ and where $\mathbf{Z}^{(n)}$ dilates the data $\mathbf{X}^{(n)}$ by (4.1) with $d^{(n)} = (\beta_{tot}^{(n)} + 1)^{1/2}$, where $\beta_{tot}^{(n)} = \sum_{j=1}^{n} \beta_j^{(n)}$. It is apparent from Theorem 4.1 that for the symmetric choice $\boldsymbol{\beta}^{(n)} = \beta^{(n)} \mathbf{1}_n$ we will have consistency in two cases: (i) when $\beta_{tot}^{(n)} = n\beta^{(n)}$ goes to zero fast enough; and (ii) when, in the case of a normal population distribution, $\beta^{(n)}$ goes to infinity fast enough. We shall give such statements a foundation and simultaneously generalize the proposed distribution estimate by viewing $\mathbf{y}^{(n)}$ as the random mean of a Dirichlet random process constructed on the Borel sets $\mathcal{A}$ of the data-variable space R^k.

As defined by Ferguson (1973) (and stated here in the particular space R^k), a random probability distribution Q on $(R^k, \mathcal{A})$ is said to obey a Dirichlet process with finite nonnull parameter measure α on $(R^k, \mathcal{A})$, denoted by $Q \sim \mathcal{D}(\alpha)$, if Q is a random process indexed by elements of $\mathcal{A}$, and Q has the following finite-dimensional marginal distributions. For all positive integers m, and every measurable partition $A_1, \ldots, A_m$ of R^k, the random vector $(Q(A_1), \ldots, Q(A_m))$ has a Dirichlet distribution on $\Delta^{(m)}$ with parameter vector $(\alpha(A_1), \ldots, \alpha(A_m))$. Dirichlet processes appear often as Bayesian prior distributions since they are quite tractable in Bayes' Theorem. But they have been criticized regarding such use because, with probability one, their realizations Q are discrete distributions, and what is

worse, the above-mentioned Dirichlet vectors $(Q(A_1), \ldots, Q(A_m))$ cannot have any positively correlated coordinates. That is, such a prior distribution expresses prior opinion favoring locally highly nonsmooth sampling distributions. (See e.g. Dickey 1968ab.) Even so, Dirichlet processes have been applied with some success in a wide variety of Bayesian problems, as elegantly reviewed in Ferguson, Phadia, and Tiwari (1989).

Random linear functionals of a Dirichlet process have been studied by Hannum, Hollander, and Langberg (1981), Jiang (1984), Lo (1987), Cifarelli and Regazzini (1988), Weng (1989), and others. Sethuraman and Tiwari (1982) have given a powerful theory of weak convergence for Dirichlet processes based on uniform convergence of parameter measures. Our approximating parameter measures, however, will be discrete, and hence not uniformly convergent. Sethuraman and Tiwari show that if a sequence of parameter measures is tight, the corresponding sequence of Dirichlet processes will be tight. Hence, we can simplify the development here by merely assuming a common compact subset Ω of R^k containing the support sets of our parameter measures. Define for a parameter measure α on $(R^k, \mathcal{A})$, for a Dirichlet process $Q \sim \mathcal{D}(\alpha)$, and for a continuous function $f(z)$ on R^k, the Stieltjes integral for outcome distribution Q,

$$\bar{f}_Q = \int_{R^k} f(z)\, dQ(z). \tag{8.2}$$

This is a linear function of Q with an induced distribution. The random mean of a Dirichlet process is the special case of a vector of linear integrands,

$$\bar{z}_Q = \int_{R^k} z\, dQ(z). \tag{8.3}$$

The important insight of this section is to realize that the filtered-variate Dirichlet vector $y^{(n)}$ is the random mean of a simple Dirichlet process. Define for the matrix of n fixed column-vector points in R^k, $Z^{(n)} = (z_1, \ldots, z_n)$, the finitely-supported (atomic) measure $\alpha^{(n)}$ on R^k assigning masses $\beta_j^{(n)} > 0$ to the points z_j,

$$\alpha^{(n)} = \sum_{j=1}^{n} \beta_j^{(n)} \delta_{z_j}, \tag{8.4}$$

where $\delta_{z_j}(S) = 1$ if $z_j \in S$, and 0 otherwise. This is proportional to the empirical distributon $F_z^{(n)}$ of $Z^{(n)}$,

$$\alpha^{(n)} = \beta_{tot}^{(n)} F_z^{(n)}. \tag{8.5}$$

Lemma 8.1 For fixed $\mathbf{z}^{(n)}$, the filtered-variate random vector $\mathbf{y}^{(n)}$ (8.1) is distributed identically to the random mean $\bar{\mathbf{z}}_{Q^{(n)}}$ (8.3) of a Dirichlet process $Q^{(n)} \sim \mathcal{D}(\alpha^{(n)})$ with the finitely-supported (atomic) parameter measure $\alpha^{(n)}$ (8.4), (8.5).

Interesting results follow immediately from this representation. We obtain first a uniform limit and the low-order moments of the random mean of a general Dirichlet process.

Lemma 8.2 A Dirichlet-process random linear fuctional $\bar{f}_Q$, $Q \sim \mathcal{D}(\alpha)$, can be expressed as the **uniform** limit of such functionals having finitely supported (atomic) parameter measures $\alpha^{(n)}$ (8.4), in the sense that, for any given $\varepsilon > 0$ there exist such $\alpha^{(n)}$ and $\bar{f}_{Q^{(n)}}$, where $Q^{(n)} \sim \mathcal{D}(\alpha^{(n)})$, satisfying for **all** outcomes, $|\bar{f}_Q - \bar{f}_{Q^{(n)}}| < \varepsilon$.

Proof. Partition Ω (compact) into a finite number of small enough hyperintervals I_j and choose any points $\mathbf{z}_j \in I_j$. Then $|f(\mathbf{z}) - f(\mathbf{z}_j)| < \varepsilon$ whenever $\mathbf{z} \in I_j$. To finish the construction, let $\alpha^{(n)}$ be defined by (8.4) with $\beta_j^{(n)} = \alpha(I_j)$, and write $\bar{f}_{Q^{(n)}} = \sum_j Q^{(n)}(\{\mathbf{z}_j\})\, f(\mathbf{z}_j)$, defining each $Q^{(n)}(\{\mathbf{z}_j\}) = Q(I_j)$. Write $\bar{f}_Q = \sum_j Q(I_j)\, \bar{f}_{Q|I_j}$, where $\bar{f}_{Q|I_j}$ are conditional Dirichlet random means on I_j, independent of the $Q(I_j)$.

This lemma, showing that a Dirichlet-process random mean is distributed according to the limiting distribution of Dirichlet-weighted averages, can be applied to justify the following distributions on the unit disk, which were obtained by Jiang (1988) as limiting distributions of the filtered-variate Dirichlet distribution, with the the nth roots of unity as the filter points becoming dense on the unit circle for $n \to \infty$.

Corollary 8.1 In the plane (k=2) let $\alpha = \alpha_{tot}\, \tilde{\alpha}$, where $\tilde{\alpha}$ is the uniform probability distribution on the unit circle. Then the Dirichlet-weighted mean $\bar{\mathbf{z}}_Q$ (8.3), where $Q \sim \mathcal{D}(\alpha)$, has the circularly symmetric distribution on the unit disk with squared radius, $r^2 \sim \text{beta}(1, \alpha_{tot})$. In particular, if the parameter measure α is the uniform probability distribution on the unit circle ($\alpha_{tot} = 1$), then $\bar{\mathbf{z}}_Q$ is uniformly distributed on the unit disk.

Theorem 8.1 (Moments of the Random Mean of a Dirichlet Process) If $\bar{z}_Q = \int_{R^k} z\, dQ(z)$, where $Q \sim \mathcal{D}(\alpha)$ on $(R^k, \mathcal{A})$, and if $\alpha = \alpha_{tot}\,\tilde{\alpha}$, where $\alpha_{tot} = \int_{R^k} d\alpha$, then

$$E\,\bar{z}_Q = E(z \mid \tilde{\alpha}) \tag{8.6a}$$

and

$$\mathrm{Var}\,\bar{z}_Q = (\alpha_{tot} + 1)^{-1}\,\mathrm{Var}(z \mid \tilde{\alpha}). \tag{8.6b}$$

If $k = 1$ and $\tilde{\alpha}$ is a symmetric distribution, then $\bar{z}_Q$ has a symmetric distribution, and so has odd moments zero (Hannum, Hollander, and Langberg 1981). If $k = 1$ and $\tilde{\alpha}$ is a continuous-type distribution with zero mean and finite fourth moment,

$$E(\bar{z}_Q^4) = [(\alpha_{tot} + 1)(\alpha_{tot} + 2)(\alpha_{tot} + 3)]^{-1}\,[6E(z^4 \mid \tilde{\alpha}) + 3\alpha_{tot}\,(E(z^2 \mid \tilde{\alpha}))^2]. \tag{8.7}$$

Proof. Apply Lemmas 8.1, 8.2 with $\mathbf{f}(z) = z$, and eqs.(3.4ab). The uniform limit in Lemma 8.2 is stronger than limit in probability which implies weak convergence. In (8.7), the only surviving contributions in the four-way sum as it approaches $E(\bar{z}_Q^4)$ are the constant terms in the Dirichlet moments $E(w_i^4)$ and $E(w_i^2 w_j^2)$ $(i \neq j)$.

We turn back now to our distribution estimate $\mathbf{y}^{(n)}$ (8.1) and we formally assume that the data $\mathbf{X}^{(n)} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is an i.i.d. random sample from a parent population, say $\mathbf{x}_j \sim \Pi_{\mathbf{x}}$ on R^k, $j = 1, \ldots, n$.

Theorem 8.2 With probability one, as sample size n increases, the filtered-variate Dirichlet random vector $\mathbf{y}^{(n)}$ (8.1) converges in probability to the Dirichlet-process random mean, $\mathbf{y}^{(\infty)} = \bar{z}_Q$ (8.3), where $Q \sim \mathcal{D}(\beta_{tot}^{(\infty)}\,\Pi_z)$ where we assume a finite limiting total measure $\beta_{tot}^{(\infty)} = \lim_{n\to\infty} \beta_{tot}^{(n)}$ $(\beta_{tot}^{(n)} = n\beta^{(n)})$, and Π_z is the population distribution of the transformed data variable,

$$z = (\beta_{tot}^{(\infty)} + 1)^{1/2}\,(\mathbf{x} - \bar{\mathbf{x}}_\Pi) + \bar{\mathbf{x}}_\Pi, \tag{8.8}$$

where $\bar{\mathbf{x}}_\Pi = E(\mathbf{x} \mid \Pi_{\mathbf{x}})$.

Proof. Partition Ω into $\{J_i\}_{i=1}^{I}$ such that $\|z - z'\| < \varepsilon$ whenever $z, z' \in J_i$. Choose points $z_i^* \in J_i$, e.g. midpoints. Define the vectors $\mathbf{R}$ and $\boldsymbol{\rho}$ of coordinates $R_i = Q(J_i)$ and $\rho_i = \beta_{tot}^{(\infty)} \Pi_Z(J_i)$, respectively. Then $\mathbf{R} \sim D(\boldsymbol{\rho})$ and

$$| \textstyle\sum_1^I z_i^* R_i - \bar{z}_Q | < \varepsilon, \tag{8.9}$$

for all outcomes. By Lemma 8.1, $y^{(n)} = \bar{z}_{Q^{(n)}}$, where $Q^{(n)} \sim \mathcal{D}(\beta_{tot}^{(n)} F_z^{(n)})$, and $F_z^{(n)}$ is the empirical distribution of $Z^{(n)}$. Define the vectors $\mathbf{R}^{(n)}$ and $\boldsymbol{\rho}^{(n)}$ of coordinates $R_i^{(n)} = Q^{(n)}(J_i)$ and $\rho_i^{(n)} = \beta_{tot}^{(n)} F_z^{(n)}(J_i)$, respectively. Then $\mathbf{R}^{(n)} \sim D(\boldsymbol{\rho}^{(n)})$ and

$$| \textstyle\sum_1^I z_i^* R_i^{(n)} - y^{(n)} | < \varepsilon, \tag{8.10}$$

for all outcomes. Now, as $n \to \infty$, $\boldsymbol{\rho}^{(n)} \to \boldsymbol{\rho}$ with probability one. So $\mathbf{R}^{(n)}$ converges in probability to a random vector distributed like $\mathbf{R}$, and $\sum z_i^* R_i^{(n)}$ converges in probability to a random variable distributed like $\sum z_i^* R_i$. Since ε is arbitrarily small, (8.9) and (8.10) imply that $y^{(n)}$ converges in probability to a random variable distributed like $\bar{z}_Q$.

Corollary 8.2 Applying Theorems 8.1 and 8.2 to a Normal(μ, σ^2) population yields a limiting distribution with its first three moments matching the population, and the ratio of fourth moments,

$$E[(y^{(\infty)} - \mu)^4] / [3\,\sigma^4] = (\beta_{tot}^{(\infty)}+1)(\beta_{tot}^{(\infty)}+6) / [(\beta_{tot}^{(\infty)}+2)(\beta_{tot}^{(\infty)}+3)]. \tag{8.11}$$

This is never smaller than unity, it approaches unity at $\beta_{tot}^{(\infty)} = 0$ and ∞, and it is maximized to 1.20204103 at $\beta_{tot}^{(\infty)} = \sqrt{6}$.

[INSERT FIGURES 7 AND 8 HERE WITH THE FOLLOWING CAPTIONS:]

Figure 7.

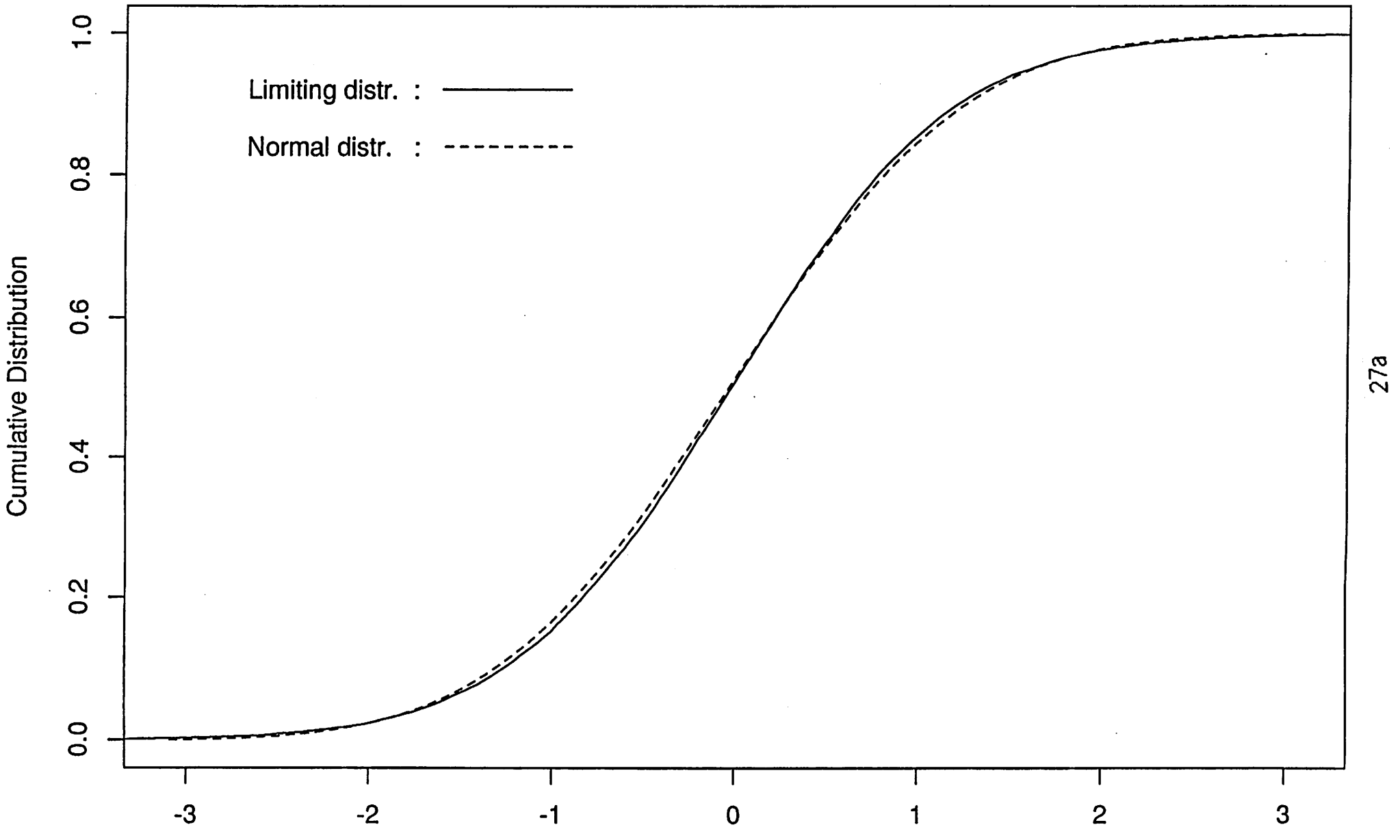

Limiting distr. :
Normal distr. :
Cumulative Distribution
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3

Figure 8.

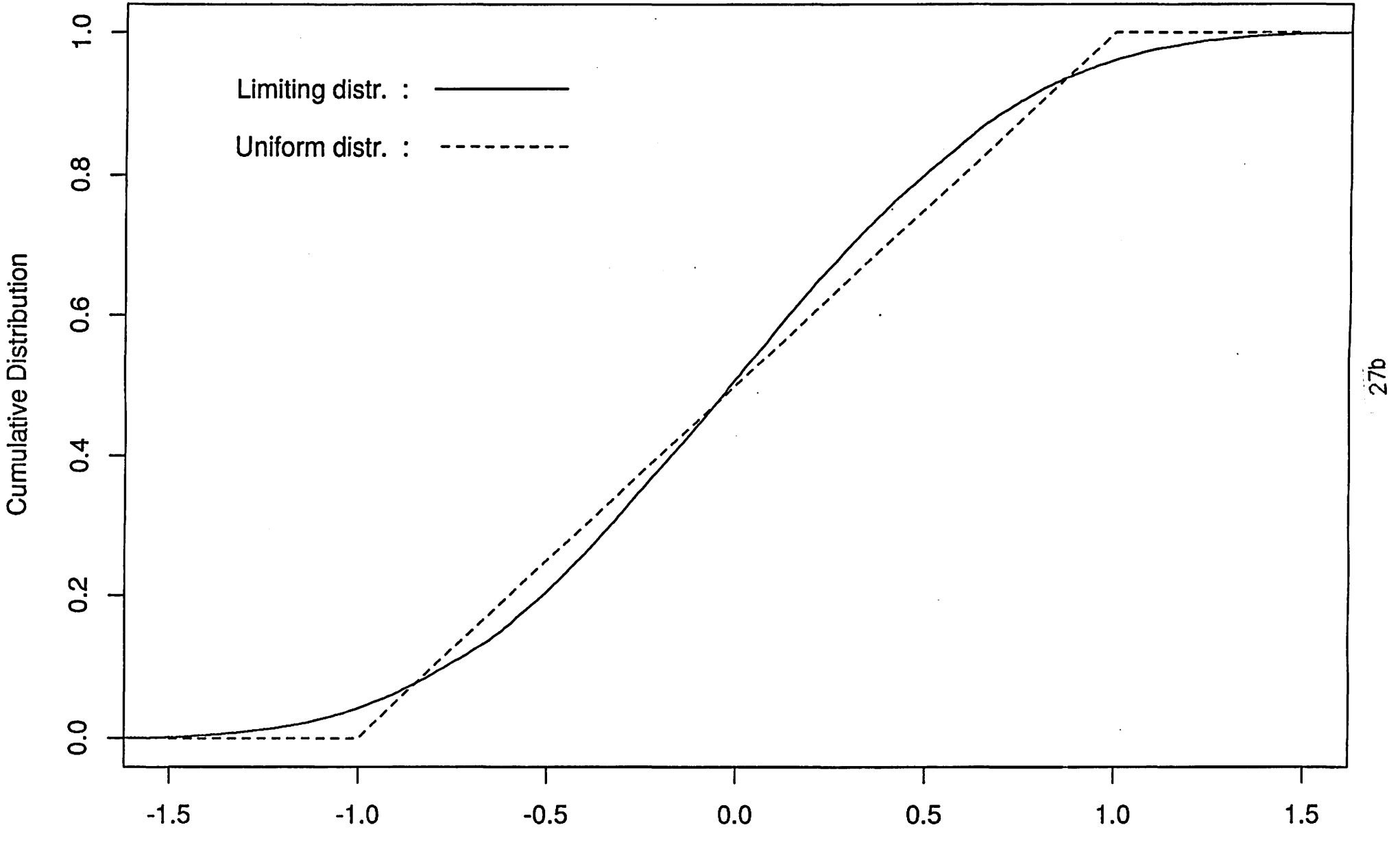

Limiting distr. :
Uniform distr. :
Cumulative Distribution
0.0
0.2
0.4
0.6
0.8
1.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

Figure 7. Cumulative distribution of the limiting filtered-variate Dirichlet distribution estimate with $\beta_{tot}^{(\infty)} = \sqrt{6}$, as sample size $n \to \infty$, for a normal parent population. The normal parent cumulative is also plotted.

Figure 8. The case of a uniform parent population.

The degree of similarity between the (limiting) distribution of $\bar{z}_Q$ and the parent population will depend on $\beta_{tot}^{(\infty)}$ and the parent population (assuming $\beta_{tot}^{(\infty)}$ is nonzero). The distributions will differ little if the parent population is normal, as illustrated in Figure 7. That the distributions can differ more is illustrated in Figure 8, where the parent is a uniform distribution. This difference would be reduced in practice when such an interval range was known, by using an appropriate invertible transformation (e.g. log[x/(1-x)]), as suggested in Section 4.

9. SIMULATING THE RANDOM MEAN OF A DIRICHLET PROCESS

Theorem 8.2 represents the large-sample-size limiting form of the distribution estimate as the distribution of a Dirichlet-process random mean, parameterized by a measure proportional to the dilated population distribution. This can be used to study, by simulation, the limiting form of the distribution estimate as it relates to the population distribution and the choice of limiting value $\beta_{tot}^{(\infty)} = \lim_{n\to\infty} n\beta^{(n)}$. This poses the practical question of how to simulate a Dirichlet process random mean with a continuous-type parameter measure. Two methods are available.

Lemma 8.2 suggests that one should partition the space R^k (or Ω) into many equal small hyperintervals, assign a number of filter points to each hyperinterval proportional to its population probability, and just carry out the same simulation and calculations that one uses to produce the distribution estimate, itself. Of course, the points within a hyperinterval J_i can be combined into a single point z_i^* for which the Dirichlet parameter coordinate would then be proportional to the hyperinterval probability. If the random mean is approximately $\mathbf{y} = \sum_{j=1}^{n} w_j z_j$, where $\mathbf{w} \sim D(\beta \mathbf{1}_n)$, and $n\pi_i$ is the number of z_j's equal to z_i^* in the hyperinterval J_i, $i = 1, \ldots, l$, then we have $\mathbf{y} = \sum_{i=1}^{l} u_i z_i^*$, where each

$u_i = \sum_{j|i} w_j$, and so $\mathbf{u} \sim D(\boldsymbol{\gamma})$, where each $\gamma_i = \sum_{j|i} \beta = (n\beta)\pi_i$. This requires only l, instead of n, filter points and a nonexchangeable Dirichlet distribution on $\Delta^{(l)}$, instead of an exchangeable one on $\Delta^{(n)}$.

A variation on the method is to choose the points z_j or z_i^* within J_i at random, repeatedly throughout the simulation, thus preventing any density poles arising as artifacts of the approximation method, even for very small β. In the next section, we will suggest such a variation in the estimation method, itself.

A second method of simulating $\mathbf{y}^{(\infty)}$ in Theorem 8.2 exploits an elegant representation of the Dirichlet process by Sethuraman and Tiwari (1982). Generate two independent i.i.d. sequences, $z_1, z_2, \ldots$ and $q_1, q_2, \ldots$, as follows: $z_i \sim \Pi_z$; and $q_0 = 0$, and for $i > 0$,

$$(q_i \mid q_0, \ldots, q_{i-1}) \sim (1 - \sum_0^{i-1} q_i) \times \text{beta}(1, \beta_{tot}^{(\infty)}). \tag{9.1}$$

The process has the discrete outcome distribution with probabilities q_i assigned to points z_i, and then the random mean $\mathbf{y}^{(\infty)} = \sum_{i=1}^{\infty} q_i z_i$. The method requires approximating this infinite series by a finite sum.

10. CONTINUOUS PARAMETER MEASURES AND PRIOR INFORMATION

In Section 8., we saw that a filtered-variate Dirichlet distribution is a simple case of the distribution of the random mean of a Dirichlet process, and in Section 9., we have seen that the latter can be simulated for a fully general Dirichlet process. This raises the possibility of using a continuous-type measure, or one having a continuous-type component, as the parameter of a distribution estimate.

The following variation of our method replaces the atomic parameter measure by a continuous measure proportional to a sum of normal probability distributions having variance matrices proportional to the sample variance $\mathbf{S_x}$. Other ellipsoidally symmetric distributions could also serve. One literally replaces the common mass β at each dilated data point z_j in (8.4) by β times the normal distribution $N(z_j, \varepsilon d^2 \mathbf{S_x})$, to obtain the new parameter measure,

$$\alpha = \beta \sum_j \Phi(z_j, \varepsilon d^2 \mathbf{S_x}). \tag{10.1}$$

Then, the low-order moments of the corresponding probability measure $\tilde{\alpha}$ are $\bar{z} = \bar{x}$ and $\mathbf{S_z} + \varepsilon d^2 \mathbf{S_x} = (1+\varepsilon) d^2 \mathbf{S_x}$, and so the moments of $\mathbf{y}$

will again match the empirical sample moments $\bar{\mathbf{x}}$, $\mathbf{S_x}$ for the choice $(1+\varepsilon)d^2(n\beta+1)^{-1} = 1$, or for the choice of dilation parameter,

$$d = (n\beta+1)^{1/2} / (1+\varepsilon). \tag{10.2}$$

Another variation would be to obtain a distribution estimate essentially by adding to $\mathbf{y}$ an extraneous independent random vector. This was done by Lo (1987) for the Bayesian bootstrap. Let $\mathbf{y} = (1-\varepsilon)^{1/2}\mathbf{Zw} + \varepsilon\mathbf{u}$, where independently of $\mathbf{w}$, $\mathbf{u} \sim N(\bar{\mathbf{x}}, \mathbf{S_x})$. Then, again, $E\,\mathbf{y} = \bar{\mathbf{x}}$ and $\mathrm{Var}\,\mathbf{y} = \mathbf{S_x}$. Other variations might mix over parameters, such as β.

The preceding two methods have the advantage of killing any poles in the density of the distribution estimate and extending the support to the whole space R^k, thus avoiding any awkward effects from the arbitrary convex boundary of the points z_j. We end with suggestions for modifying the parameter measure by including a component that may be of continuous type.

The methods that have been given here are not strictly Bayesian, but there may be interest in using prior information, even if not in a formal Bayesian way. One way to use prior information is to choose β with an eye to the likely population distributions and the effect of β on the distribution estimate. That choice will be easier after future work on the detailed effects of β. Another way is to include a pre-data or "prior" component in the parameter measure, as follows.

In the usual Bayesian theory a prior Dirichlet process is assumed for the unknown population distribution, say $\Pi_{\mathbf{x}} \sim \mathcal{D}(n_0 G_{\mathbf{x}})$, where $G_{\mathbf{x}}$ is a probability distribution. A prior Bayesian distribution estimate that has been suggested in the literature is the expected distribution $E\,\Pi_{\mathbf{x}} = G_{\mathbf{x}}$. The Bayesian posterior process would again be Dirichlet with the posterior parameter measure $n_0 G_{\mathbf{x}} + nF_{\mathbf{x}} = (n_0+n)\,G_{\mathbf{x}}'$, where

$$G_{\mathbf{x}}' = (n_0+n)^{-1}(n_0 G_{\mathbf{x}} + nF_{\mathbf{x}}), \tag{10.3}$$

where $F_{\mathbf{x}}$ is the empirical sample distribution. The parameter measures combine in the same way that empirical sample measures combine for a pooled sample. The posterior Bayesian distribution estimate is the posterior expected distribution $G_{\mathbf{x}}'$ (10.3) which, of course, includes atoms from $F_{\mathbf{x}}$. Note that n_0 plays the role of the size of a "prior sample".

Now, the traditional distribution estimate is just $F_{\mathbf{x}}$, which we have suggested replacing in use by the smoother distribution of the random mean of a Dirichlet process with parameter $\beta_{tot}F_{\mathbf{Z}}$ (8.5), where β_{tot} changes only slowly with n. By extension, we suggest replacing in use as an

estimate the posterior Bayesian distribution estimate, $G_{\mathbf{x}}'$ (10.3), by the smoother distribution of the Dirichlet random mean based on the parameter measure $\beta_{tot} G_Z'$, where G_Z' is the distribution induced from $G_{\mathbf{x}}'$ by the transformation,

$$Z = (1+\beta_{tot})^{1/2} [\mathbf{x} - E(\mathbf{x} | G_{\mathbf{x}}')] + E(\mathbf{x} | G_{\mathbf{x}}'). \qquad (10.4)$$

Again as in the Bayesian method, our parameter distributions combine in the same way that empirical distributions combine for pooled samples. However, we control the total parameter measures by our choice of $\beta_{tot} = \beta_{tot}^{(n)}$.

Before knowledge of the data $\mathbf{X}$, our distribution estimate would be the distribution of the Dirichlet random mean based on the parameter measure $\beta_{tot} G_Z$. This pre-data distribution estimate is a distribution having moments

$$E\, \mathbf{y}_0 = E(\mathbf{x} | G_{\mathbf{x}}) \qquad (10.5a)$$

$$\text{Var}\, \mathbf{y}_0 = \text{Var}(\mathbf{x} | G_{\mathbf{x}}). \qquad (10.5b)$$

The ratio of sample sizes n_0 to n, prior versus data, must be specified to calculate $G_{\mathbf{x}}'$ by (10.3); the total parameter measure $\beta_{tot}^{(n_0+n)}$ must be specified as suitable for $n_0 + n$; and then the dilation (10.4) can be performed to obtain the post-data parameter measure $\beta_{tot}^{(n_0+n)} G_Z'$. After the data $\mathbf{X}$, the distribution estimate will have moments

$$E\, \mathbf{y} = E(\mathbf{x} | G_{\mathbf{x}}') \qquad (10.6a)$$

$$\text{Var}\, \mathbf{y} = \text{Var}(\mathbf{x} | G_{\mathbf{x}}'). \qquad (10.6b)$$

Thus, our post-data generalized distribution estimate is a distribution whose low-order moments (10.6) are identical to the moments of the posterior Bayesian distribution estimate. But our distribution estimate is a smoother distribution having no discrete components, unlike the posterior Bayesian distribution estimate.

The final theorem allows simulated sampling from such a distribution estimate based on a composite parameter measure, by simulation of the random means based on the separate components. The proof is straightforward for finite-supported measures and carries over to more general parameter measures by Lemma 8.2.

Theorem 10.1 Suppose that independently $Q \sim \mathcal{D}(\alpha)$, $R \sim \mathcal{D}(\gamma)$, and $u \sim \text{beta}(\alpha_{tot}, \gamma_{tot})$. Then

$$uQ + (1-u)R \sim \mathcal{D}(\alpha + \gamma), \tag{10.7}$$

and so $u\,\bar{z}_Q + (1-u)\,\bar{z}_R$ has the same distribution as $\bar{z}_P$ where $P \sim \mathcal{D}(\alpha + \gamma)$.

The theorem can be applied to simulate the undilated random mean $\tilde{\mathbf{y}}$ with parameter $\beta_{tot}^{(n_0+n)} G_{\mathbf{X}}' = \beta_{tot}^{(n_0+n)}(n_0+n)^{-1} n_0 G_{\mathbf{X}} + \beta_{tot}^{(n_0+n)}(n_0+n)^{-1} n F_{\mathbf{X}} = \alpha + \gamma$, in terms of the simulated random means with the respective parameters, $\alpha = \beta_{tot}^{(n_0+n)}(n_0+n)^{-1} n_0 G_{\mathbf{X}}$ and $\gamma = \beta_{tot}^{(n_0+n)}(n_0+n)^{-1} n F_{\mathbf{X}}$.

Then the post-data distribution estimate has the dilated random vector

$$\mathbf{y} = (\beta_{tot}^{(n_0+n)} + 1)^{1/2} (\tilde{\mathbf{y}} - \bar{\mathbf{y}}) + \bar{\mathbf{y}}, \tag{10.8}$$

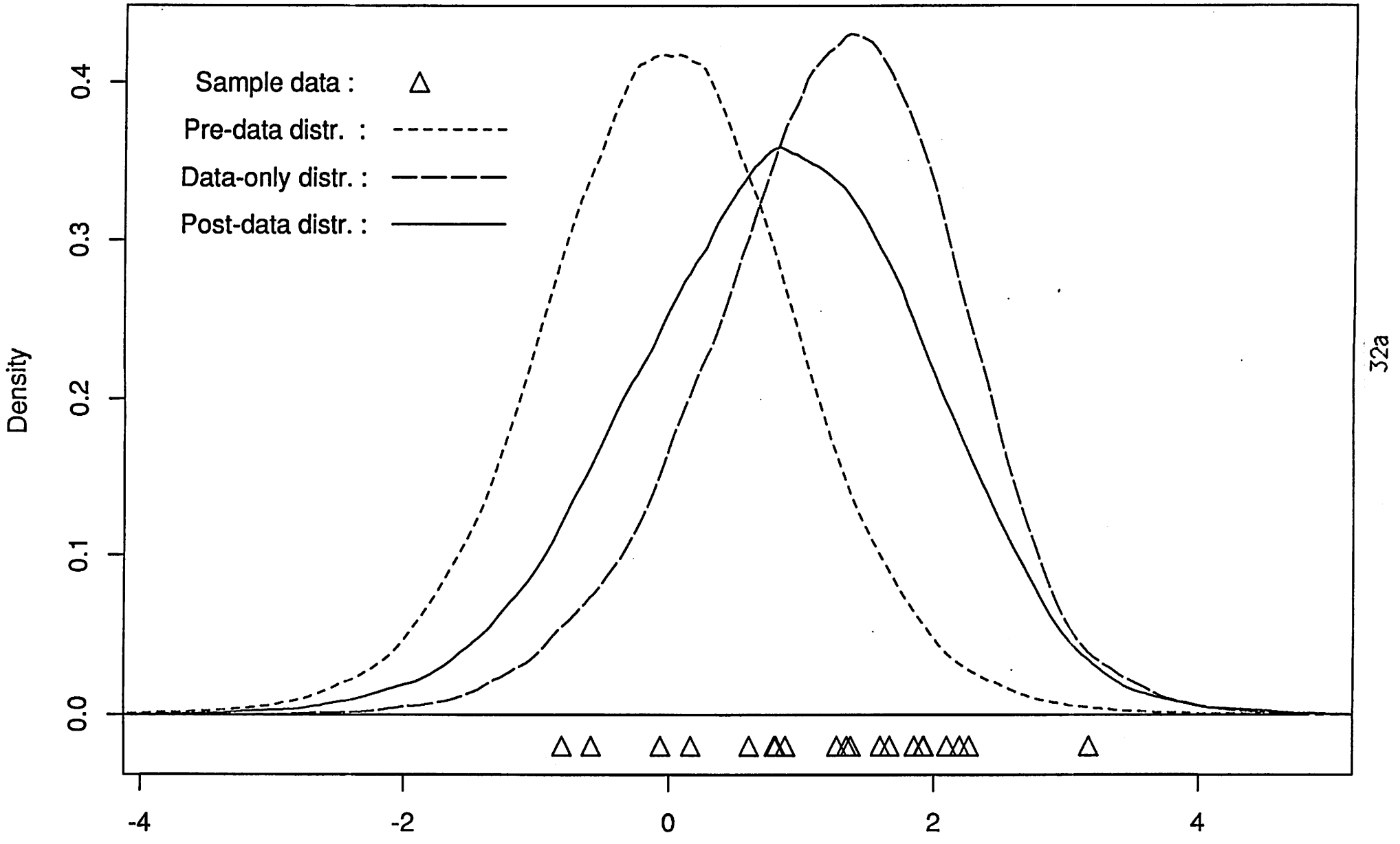
where $\bar{\mathbf{y}} = [\, \alpha_{tot} E(\mathbf{x} | G_{\mathbf{X}}) + \gamma_{tot} \bar{\mathbf{x}} \,] / \beta_{tot}^{(n_0+n)}$.

Figure 9 illustrates, for $k = 1$, the use of a parameter measure having a normal pre-data component $G_X = N(0, 1)$ with relative "prior sample size" $n_0 = 10$, and data X sampled from $N(1, 1)$ with $n = 20$. With the choice $\beta_{tot}^{(30)} = 3\sqrt{6}$, our respective total parameter measures for the simulation use of Theorem 10.1 were $\alpha_{tot} = \sqrt{6}$ and $\gamma_{tot} = 2\sqrt{6}$. (In general, the distribution estimates need not have $\beta_{tot}^{(n_0)}$ and $\beta_{tot}^{(n)}$ sum to $\beta_{tot}^{(n_0+n)}$.)

[INSERT FIGURE 9 HERE WITH THE FOLLOWING CAPTION:]

Figure 9. Combinig a pre-data distribution estimate with sample data. The pre-data distribution estimate is the distribution of the random mean of $\mathcal{D}(\sqrt{6}\, N(0, 1))$. The data are a random sample of size 20 from $N(1, 1)$ and the total parameter measure would be set at $2\sqrt{6}$ for the corresponding filtered-variate Dirichlet distribution estimate. The post-data distribution gives twice as much weight to the data-only parameter as to the pre-data distribution and has total parameter measure $3\sqrt{6}$.

Figure 9.


Sample data :
Pre-data distr. :
Data-only distr. :
Post-data distr. :
Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4

The example in Figure 9 illustrates that the post-data estimate of the parent distribution can be more diffuse than either the pre-data or the data-only distribution. This may seem different than typical results for standard Bayesian analyses, but here it is a distribution being estimated, rather than the parameters of a distribution. In forming the post-data distribution in Figure 9, twice the weight has been given to the data as to pre-data knowledge, so the post-data distributon is consequently closer to the data-only distribution. The figure also shows that the density for the post-data distribution is very small only where the densities are very small for both the pre-data and the data-only distributions. Again, this contrasts with the usual parametric Bayesian methods.

Of course, the generalizations in this Section can be combined. For example, F_z can be replaced by (10.1) in the calculation of G_z' (10.3), thereby avoiding atoms in the postdata parameter measure, too, not merely in the distribution estimate.

11. DISCUSSION

The parameter β in the proposed method is available to tune the distribution estimate to the sample size and likely populations. The effect of this choice is an important area for further work. We have found that values above the threshold for density poles, $1/(n-1)$, can be useful, even as large as $3/n$ or $8/n$. The choice of the total parameter measure $\beta_{tot}^{(n_0+n)}$ when there are "prior" components involves additional considerations.

The representation of Sethuraman and Tiwari for a Dirichlet process offers a powerful tool for simulating general random means. With the availability of increased computing capability, this gives rise to exciting possibilities for mapping out the performance and robustness of the method for useful population distributions and prior weightings over populations. A reemphasis is suggested, to our thinking, on the problem of performance of estimates of aspects of a population from aspects of a nonparametric distribution estimate, and performance of asserted aspects of a predictive distribution. The possibility of using a sample to simulate a distribution estimate raises fundamental questions in statistical inference for, and inference from, applied simulation studies.

We conclude by recalling that the distribution estimate here matches the first two sample moments and approximates further moments. This was achieved by a linear dilation of the filter points. A promising extension is to employ nonlinear dilations that match or approximate collections of moments, generalized moments, and other aspects of the sample.

REFERENCES

Bar-Lev, Shaul K., and Enis, Peter (1986), "Reproducibility and natural exponential families with power variance functions," *Annals of Statistics,* 14, 1507-1522.

Blackwell, D. (1973), "Discreteness of Ferguson selection," *Annals of Statistics,* 1, 356-358.

Bloch, Daniel A., and Watson, Geoffrey S. (1967), "A Bayesian study of the multinomial distribution," *Annals of Mathematical Statistics,* 38, 1423-1434.

Carlson, B. C. (1971), "Appell Functions and Multiple Averages," *SIAM Journal on Mathematical Analysis,* 2, 420-430.

_____ (1977), *Special Functions of Applied Mathematics,* New York: Academic Press.

Cifarelli, Donato Michele, and Regazzini, Eugenio (1988), "Distribution functions of means of a Dirichlet process," Typescript, University of Milan.

Dickey, James M. (1968a), "Estimation of disease probabilities conditioned on symptom variables," *Mathematical Biosciences,* 3, 249-265.

_____ (1968b), "Smoothed estimates for multinomial cell probabilities," *Ann. Math. Statist.*, 39, 561-566.

_____ (1969), "Smoothing by cheating," *Ann. Math. Statist.*, Vol 40, No 4, 1477-82.

_____ (1983). Multiple hypergeometric functions: Probabilistic interpretations and statistical uses. *J. Amer. Statistical Assoc.* Vol. 78, No. 383 (Sept. 1983), 628-637.

_____, Jiang, Jyh-Ming, and Kadane, Joseph B. (1986). Bayesian methods for censored categorical data. *J. Amer. Statist. Assoc.* Vol. 82, No. 399, 773-781.

_____ and Jiang, Jyh-Ming (1988). Filtered-variate prior distributions for local smoothing of histograms. Work in progress.

Ferguson, T. S. (1973), "A Bayesian analysis of some nonparametric problems," *Ann. Statist.*, 1, 209-230.

_____, Eswar G. Phadia, and Ram C. Tiwari (1989), "Bayesian Nonparametric Inference," Manuscript to be published in edited volume of research papers.

Good, I. J., and Gaskins, R. A. (1971), "Nonparametric roughness penalties for probability densities," *Biometrika*, 58, 255-277.

_____ (1980), "Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data," *J. Amer. Statistical Assoc.*, 75, 42-56.

_____ (1980), "Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data," *J. Amer. Statistical Assoc.*, 75, 42-56.

Grunbaum, Branko (1967), *Convex Polytopes*, New York: John Wiley, Interscience.

Hannum, R. C., Hollander, M., and Langberg, N. A. (1981). "Distributional results for random functionals of a Dirichlet process," *Ann. Prob.* 9, 665-670.

Hill, B. M. (1968), "Posterior distribution of percentiles: Bayes' theorem for sampling from a finite population," *J. Amer. Statistical Assoc.* 63, 677-691.

Hill, B. M. (1987a), "De Finetti's theorem, induction, and A_n, or Bayesian nonparametric predictive inference" (With discussion), To appear in *Bayesian Statistics 3*, J. M. Bernardo, M. H. DeGroot, D. V. Lindley, A. F. M. Smith, eds., Oxford: Oxford Univ. Press.

Hill, B. M. (1987b), "Parametric models for A_n: Splitting processes and mixtures," Manuscript, Dept. of Statistics, Univ. of Michigan, Ann Arbor, Mich.

Ignatov, Z. G. and V. L. Kaishev (1987), "Multivariate B-Splines, Analysis of Contingency Tables, and Serial Correlation," P. Bauer et al. (eds.), *Mathematical Statistics and Probability Theory*, Vol. B, 125 -137. D. Reidel Publ.

Jiang, Jyh-Ming (1984), *Distributional properties of linear forms in a Dirichlet vector and applications*, PhD dissertation, Res. Rept. 7/84-#14, Dept. of Mathematics and Statistics, State Univ. of N. Y. at Albany.

_____ (1988), "Starlike functions and linear functions of a Dirichlet distributed vector," *SIAM J. Math. Anal.*, Vol 19, No 2 (March 1988), 390-397.

Lenk, P. J. (1984), *Bayesian nonparametric predictive distribution*, PhD dissertation, Dept. of Statistics, Univ. of Mich., Ann Arbor, Mich.

_____(1988), "The logistic normal distribution for Bayesian nonparametric predictive densities," *J. Amer. Statist. Assoc.* 83.

Leonard, T. (1973), "A Bayesian method for histograms," *Biometrika*. 60, 297-308.

_____ (1978), "Density estimation, stochastic processes, and prior information (with Discussion)," *J. Roy. Statist. Soc., Ser. B*, 40, 113-146.

Lo, A. Y. (1984), "On a class of Bayesian nonparametric estimates: I density estimates," *Ann. Statist.* 12, 351-357.

_____ (1987), "A Large-Sample Study of the Bayesian Bootstrap," *Ann. Statist.*, 15, 360-375.

Margolin, B. H. (1977), "The Distribution of Internally Studentized Statistics Via Laplace Transform Inversion," *Biometrika*, 64, 3, 573-82.

Olkin, I., and Spiegelman, C. H. (1987), "A semiparametric approach to density estimation, *J. Amer. Statist. Assoc.* 82, 858-865.

Rubin, D. B. (1981), "The Bayesian Bootstrap," *Ann. Statist.,* 9, 130-134.

Savage, L. J. (1964), Personal communication.

Sethuraman, Jayaram, and Ram C. Tiwari (1982), "Convergence of Dirichlet Measures and the Interpretation of Their Parameter," In *Statistical Decision Theory and Related Topics III, Vol. 2,* 305-315.

Sommerville, D. M. Y. (1958). *An Introduction To the Geometry of n Dimensions,* New York: Dover.

Titterington, D. Michael (1986), "A general view of statistical smoothing," *A.S.A. Proceedings of Business and Economic Statistics Section,* 79-84.

Weng, C.-S. (1989), "On a Second-Order Asymptotic Property of the Bayesian Bootstrap," *Ann. Statist.,* 17, 705-710.

Whittle, P. (1958), "On the smoothing of probability density functions," *J. Roy. Statist. Soc.* 20, 334-343.

Wilks, S. S. (1962), *Mathematical Statistics,* New York: John Wiley.