Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Filtered-Variate Prior Distributions for Histogram Smoothing

James M. Dickey & Thomas J. Jiang

To cite this article: James M. Dickey & Thomas J. Jiang (1998) Filtered-Variate Prior Distributions for Histogram Smoothing, Journal of the American Statistical Association, 93:442, 651-662

To link to this article: http://dx.doi.org/10.1080/01621459.1998.10473718

Filtered-Variate Prior Distributions for Histogram Smoothing

James M. DICKEY and Thomas J. JIANG

We develop prior distributions for histogram inference favoring smooth population frequencies; that is, probability vectors with small differences for neighboring categories. We give a theory of prior-random probability vectors representable as a linear transform, or "filter," of a standard random probability vector, or equivalently, a random weighted average of nonrandom smooth probability vectors. Promising methods of prior assessment are given based on elicitation of a list of typically smooth probability vectors, the empirical moments of which can then be matched by the mean vector and variance matrix of a constructed continuous-type filtered-variate prior distribution.

KEY WORDS: Bayesian smoothing; Carlson function; Generalized Dirichlet distribution; Generalized hypergeometric function; Multinomial distribution; Multinomial estimation.

1. INTRODUCTION

How can a statistician effectively model a subject-matter expert's prior partial belief in local smoothness of the unknown sampling probabilities of histogram categories? That is, how can a joint prior distribution be chosen to give high prior probability to the event that the sampling probabilities are "smooth," that neighboring categories have probabilities close in value? Equivalently, how can one arrange low prior expected squared differences between neighboring category probabilities? The categories may refer to grouping intervals, for example, and their probabilities may be the integrals of a relatively smooth, but otherwise unknown, density function. This problem is important for a wide range of applications, from uses of one-way histogram data to medical diagnosis, optical image processing, and other uses of multidimensional histograms. The problem has been important for decades (see, e.g., Dickey 1968a; Vardi, Shepp, and Kaufman 1985). In its extreme form, with an infinite or continuous set of categories, it is the problem of Bayesian nonparametric inference, a major embarrassment to Bayesians (L. J. Savage, personal communication, 1970). A review of the literature would be overly lengthy here, but the reader may find interest in the discussions and citations of de Finetti (1935), Diaconis and Freedman (1986), Dickey (1968a), Lenk (1988), Leonard (1978), and Titterington (1985).

In its finite form, the problem can be set out as follows. A vector will be called a *probability k vector* if each of its k coordinates is nonnegative and the coordinates sum to unity. Denote the simplex of probability k vectors by $\Delta^{(k)} \subset R^k$, and let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ be a probability k vector, the coordinates of which comprise the probability masses of the unknown parent distribution of a sampling process with k categories. As called for by Dickey (1968), families of prior distributions for $\boldsymbol{\theta}$ are needed that will have the following four properties, in addition to giving unit probability to $\Delta^{(k)}$:

1. The family must be large enough to allow a choice accurately expressing the real predata expert uncertainty concerning $\boldsymbol{\theta}$. In particular, prior distributions favoring smooth values of $\boldsymbol{\theta}$ should be available.
2. Situationally appropriate assessment methods should facilitate the choice of a meaningful member of the family.
3. Following the arrival of new statistical data, Bayes theorem calculations should be simple to carry out.
4. The resulting posterior distribution of $\boldsymbol{\theta}$ must be tractable, in that one can easily compute inferentially useful summaries of the posterior distribution.

We propose a family of prior distributions on $\Delta^{(k)}$, together with practical methods for their assessment, satisfying these requirements. To appreciate difficulties and establish notation, consider iid sampling from a finite distribution having a probability mass function, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) \in \Delta^{(k)}$, with probability θ_i assigned to the ith category ($x = i$, say), $P(x = i|\boldsymbol{\theta}) = \theta_i$, for $i = 1, \ldots, k$. Under noninformative stopping (Raiffa and Schlaifer 1961, sec. 2.3), the category counts $\mathbf{n} = (n_1, \ldots, n_k)$ suffice for the likelihood function from a sample sequence $\mathbf{x} = (x_1, \ldots, x_N)$ stopped at N,

$$p(N, \mathbf{x}|\boldsymbol{\theta}) \propto \prod_{i=1}^{k} \theta_i^{n_i} \equiv L_n(\boldsymbol{\theta}), \tag{1}$$

where $n_+ = N$ (defining $n_+ \equiv n_1 + \cdots + n_k$). (The proportionality is taken with respect to $\boldsymbol{\theta}$). For example, if the sample size N is prespecified, then $(\mathbf{n}|\boldsymbol{\theta}) \sim$ multinomial$(N, \boldsymbol{\theta})$. The usual conjugate family of prior distributions for likelihoods (1) is the Dirichlet, $\boldsymbol{\theta} \sim D(\mathbf{b}), \mathbf{b} = (b_1, \ldots, b_k)$, each $0 \le b_i \le \infty$, having the density (if each $0 < b_i < \infty$), $p(\boldsymbol{\theta}) = B(\mathbf{b})^{-1} \prod_{i=1}^{k} \theta_i^{b_i - 1}, \boldsymbol{\theta} \in \Delta^{(k)}$, where $B(\mathbf{b}) = [\prod \Gamma(b_i)]/\Gamma(b_+)$. The k coordinates identically sum to unity,

James M. Dickey is Professor, School of Statistics, University of Minnesota, Minneapolis, MN 55455. Thomas J. Jiang is Professor, Department of Mathematical Sciences, National Chengchi University, Wen-Shan, Taipei 11623, Taiwan. The authors are grateful for early discussions with participants at the National Bureau of Economic Research (NBER)-National Science Foundation (NSF) Seminar on Bayesian Inference in Econometrics, Rutgers University, October 26–27, 1984, and the Fifth Workshop on Maximum Entropy Methods, University of Wyoming, August 5–8, 1985. The presentation has benefitted from suggestions by editors and referees. Dickey's research was supported in part by NSF research grants DMS-8614793 and DMS-8911548. Jiang's research was supported in part by NSF research grant DMS-9202161 and Taiwan National Science Council (NSC) research grants NSC-84-2121-M-004-008, NSC-85-2121-M-004-007, and NSC-86-2115-M-004-013.

$\theta_+ \equiv 1$, and the density is the same for *every* choice of $k-1$ coordinate variables; for example, $\theta_1, \ldots, \theta_{k-1}$. The resulting posterior distribution would again be Dirichlet, $\boldsymbol{\theta}|N, \mathbf{x} \sim \boldsymbol{\theta}|\mathbf{n} \sim D(\mathbf{b}+\mathbf{n})$, with the updated parameter vector $\mathbf{b}+\mathbf{n}$.

The general dth mixed moment of $\boldsymbol{\theta} \sim D(\mathbf{b})$, for $\mathbf{d} = (d_1, \ldots, d_k)$, is $E\prod_{i=1}^{k} \theta_i^{d_i} = h(\mathbf{d};\mathbf{b})$, where

$$h(\mathbf{d};\mathbf{b}) = B(\mathbf{b}+\mathbf{d})/B(\mathbf{b}). \tag{2}$$

So the mean vector and variance matrix are $E\boldsymbol{\theta} = \mathbf{w}$, where $\mathbf{w} = \mathbf{b}/b_+$, and $\text{var}(\boldsymbol{\theta}) = (b_+ + 1)^{-1}[\text{diag}(\mathbf{w}) - \mathbf{w}\mathbf{w}^T]$, where $\text{diag}(\mathbf{w})$ is the diagonal matrix with ith diagonal entry w_i, and (taking vectors as column arrays) $\mathbf{w}\mathbf{w}^T$ denotes the $k \times k$ matrix of i,jth entries $w_i w_j$. Distributions on the probability simplex in which the first two moments are related proportionally through such a matrix quadratic function, with the multiplier $(b_+ + 1)^{-1}$ generalized to an arbitrary constant, will be defined as having mean-structured variance (MSV).

Prior expected squared differences can be written as $E[(\theta_i - \theta_j)^2] = [E(\theta_i) - E(\theta_j)]^2 + [\text{var}(\theta_i) + \text{var}(\theta_j)] - 2\,\text{cov}(\theta_i, \theta_j)$. So a high positive prior correlation of adjacent or near-neighbor category probabilities is desirable, to have a small expected squared difference, to express a prior belief in local smoothness. Then the nonsmooth character of Dirichlet distributions is revealed by their moments, every correlation from such a variance matrix being necessarily nonpositive: $\text{corr}(\theta_i, \theta_j) = -\{[w_i/(1-w_i)] \cdot [w_j/(1-w_j)]\}^{1/2}$. Because the posterior distribution from a Dirichlet prior is again Dirichlet, the posterior moments are similar in character to the prior moments. Both prior and posterior correlations are nonpositive between every pair of category probabilities, and this is true of the posterior distribution no matter how smooth the data (or even the posterior mean) may be.

The Dirichlet posterior mean can be written as an estimator that shrinks the usual unbiased maximum likelihood estimate, $\hat{\boldsymbol{\theta}} = \mathbf{n}/N$, toward the prior mean point $\mathbf{w}$, $E(\boldsymbol{\theta}|\mathbf{n}) = (1-u)\,\hat{\boldsymbol{\theta}} + u\mathbf{w}$, where $u = b_+/(b_+ + N)$. If the prior mean $\mathbf{w}$ is smoother than $\hat{\boldsymbol{\theta}}$, then so is the posterior mean, and thus we would have a "smoothed" estimate. However, as recognized by Good and Gaskins (1971, 1980), such smoothing by scale-shrinkage is global rather than local, in the sense that the differences between neighboring probabilities are diminished to no greater extent than are the nonneighboring differences. Indeed, the effect of this global smoothing in a posterior mean difference, $E(\theta_i|\mathbf{n}) - E(\theta_j|\mathbf{n})$, depends only on the corresponding prior mean difference $w_i - w_j$ and not on the distance between categories (e.g., $|i-j|$), because $E(\theta_i|\mathbf{n}) - E(\theta_j|\mathbf{n}) = (1-u)(\hat{\theta}_i - \hat{\theta}_j) + u(w_i - w_j)$, for all i, j, where the weights $1-u$ and u are independent of i, j. If the prior mean is smooth, then $w_i - w_j$ in the second term is close to 0 for short distances $|i-j|$, but $w_i - w_j$ may also be close to 0 for various longer distances. A similar situation holds for the posterior mode.

In *multivariate-normal* sampling, however, with a conjugate multivariate-normal prior, the posterior mean and mode are the same point, and this point is a matrix-convex combination of the prior mean and the usual maximum likelihood estimate. But the family of normal distributions is closed under linear operations on the random vector, and so a normal prior distribution can be assigned an arbitrary prior covariance structure, say $\text{var}(\boldsymbol{\mu}) = \mathbf{V}_\mu$. In obvious notation, $E(\boldsymbol{\mu}|\hat{\boldsymbol{\mu}}) = (\mathbf{I}-\mathbf{U})\hat{\boldsymbol{\mu}} + \mathbf{U}E(\boldsymbol{\mu})$, with $\mathbf{I}-\mathbf{U} = \mathbf{V}_\mu(\mathbf{V}_\mu + \mathbf{V}_{\hat{\mu}|\mu})^{-1}$ and $\mathbf{U} = \mathbf{V}_{\hat{\mu}|\mu}(\mathbf{V}_\mu + \mathbf{V}_{\hat{\mu}|\mu})^{-1}$. As noted by Titterington (1985), the difference of posterior-mean (mode) coordinates can be written as

$$\begin{aligned} &E(\mu_i|\hat{\boldsymbol{\mu}}) - E(\mu_j|\hat{\boldsymbol{\mu}}) \\ &\quad = (D_{i,j}\mathbf{V}_\mu)(\mathbf{V}_\mu + \mathbf{V}_{\hat{\mu}|\mu})^{-1}\hat{\boldsymbol{\mu}} \\ &\qquad + (D_{i,j}\mathbf{V}_{\hat{\mu}|\mu})(\mathbf{V}_\mu + \mathbf{V}_{\hat{\mu}|\mu})^{-1}E(\boldsymbol{\mu}), \end{aligned} \tag{3}$$

where the operator $D_{i,j}$ yields the difference between the ith and jth row vectors, $D_{i,j}\mathbf{V} = (v_{i1} - v_{j1}, \ldots, v_{ik} - v_{jk})$. By (3), we see that normal-theory smoothing can be truly local, in that a gently varying prior variance-covariance would mean a small difference between the ith and jth row vectors of $\mathbf{V}_\mu$ for a short distance $|i-j|$, and thereby a small effect from the local differences in the data $\hat{\boldsymbol{\mu}}$ through the first term of (3).

Lenk (1988), Leonard (1973), and others achieved local smoothing of histogram data by modifying and exploiting the normal conjugate theory. A *nonlinear* (logistic) change of variable on a multivariate normal vector was used to guarantee prior certainty for the event that all the category probabilities are nonnegative and sum to unity. (See Good and Gaskins 1971, 1980 for related methods.) As an alternative theory, we will work directly with a *linear* transform or "filter" of a Dirichlet or other MSV random vector, the support set of which naturally will lie within the probability simplex. Unlike the normal distributions, however, a class of MSV distributions is not closed under linear transformations of the vector variate. The density of a linear transform is complicated, and if such a density were used as the prior density for multinomial sampling, then the posterior density would be even more complicated. But we shall succeed in using the distribution of a linear transform of an MSV vector as the prior distribution of the category probabilities by maintaining the untransformed MSV random vector as the variable of integration of the prior and posterior density functions. Both the prior and the resulting posterior distributions will then be tractable. The posterior mean, and sometimes the posterior mode, will be computable as estimates, together with other inferential summaries and properties of the prior and posterior distributions.

In Section 2 we formalize the concept of linearly filtering a random vector to obtain a "filtered-variate" generalization of its distribution for prior local smoothness. In Section 3, we develop the filtered-variate Dirichlet family of prior distributions and their consequent posterior distributions. The corresponding posterior family will be a filtered-variate form of Dickey's (1983) generalized Dirichlet distribution. In Section 3 we also give the inferences read from the posterior distribution for smoothed estimation and for hypothesis comparison.

A central problem of this research is how to choose or "assess" a specific filtered-variate prior distribution. A

promising line of approach was found by viewing the filtered-variate Dirichlet as a filtered-variate MSV distribution. The crucial property of a filtered-variate MSV is that the low-order moments can be expressed in terms of the corresponding moments of a finite distribution on the set of column vectors of the filter matrix. This will enable the elicitation of expert prior opinion in the form of a list of typically smooth probability vectors, which can then be linearly transformed, or "dilated," to serve as the column vectors in the filtering matrix, thereby yielding a filtered-variate prior distribution having first- and second-order moments identical to the empirical moments of the assessed list of typical vectors. The theory of general MSV distributions and their filtered-variate forms is developed in Section 4, and used for prior assessment in Section 5. In Section 6 we give simple examples of the assessment and use of filtered-variate priors in histogram smoothing problems. We conclude in Section 7 by mentioning generalizations to a continuous sampling variable and/or continuous-type random mixing measure. We provide the relevant proofs in an Appendix.

2. FILTERED-VARIATE DISTRIBUTIONS

We obtain a random probability vector $\boldsymbol{\theta}$ in $\Delta^{(k)}$ by defining $\boldsymbol{\theta}$ as a linear transform of a "standard" random probability vector $\boldsymbol{\alpha}$ in $\Delta^{(m)}$. Let

$$\boldsymbol{\theta} = \mathbf{G}\boldsymbol{\alpha}, \tag{4}$$

where $\mathbf{G}(k \times m)$ is a constant matrix and the probability m vector $\boldsymbol{\alpha}$ (column) has a specified distribution. Then $\boldsymbol{\theta}$ will be said to have a *filtered-variate* form of the distribution of $\boldsymbol{\alpha}$. If $\boldsymbol{\alpha}$ has a Dirichlet distribution, $\boldsymbol{\alpha} \sim D(\mathbf{a}), \mathbf{a} = (a_1, \ldots, a_m)$, then $\boldsymbol{\theta}$ has a *filtered-variate Dirichlet distribution.* Denote the distribution of (4), where $\boldsymbol{\alpha} \sim D(\mathbf{a})$, by $\boldsymbol{\theta} \sim F_G D(\mathbf{a})$. (This distribution is in no way a "mixture of Dirichlet distributions," because there is only one Dirichlet distribution involved.) What are essentially one-dimensional filtered-variate Dirichlet distributions were studied by Bloch and Watson (1967), Cifarelli and Regazzini (1990), Diaconis and Kemperman (1996), Dickey (1983), and Jiang (1984, 1988). Properties in arbitrary dimensions, including results on the density, were given by Dickey, Garthwaite, and Bian (1995). The Dirichlet distribution $D(\mathbf{a})$ itself is the special case $F_I D(\mathbf{a})$.

It is instructive to interpret a filtered-variate distribution (4) in two ways:

1. Each coordinate of $\boldsymbol{\theta}$ is a linear combination (or "filter") of the coordinates of the random vector $\boldsymbol{\alpha}$, $\theta_i = g_{i,1}\alpha_1 + \cdots + g_{i,m}\alpha_m$, each $i = 1, \ldots, k$. Hence our use of the compound adjective "filtered-variate" for the random vector $\boldsymbol{\theta}$. The variate is filtered rather than the distribution, as would be the case with a mixture of prior distributions or a density transformed by an integral operator.

2. The vector $\boldsymbol{\theta}$ is a weighted average of the fixed column vectors of $\mathbf{G}$, with random weights $\boldsymbol{\alpha}$. For an array of column vectors, $\mathbf{G} = (\mathbf{g}_1, \ldots, \mathbf{g}_m)$, write $\boldsymbol{\theta} = \alpha_1\mathbf{g}_1 + \cdots + \alpha_m\mathbf{g}_m$. Because $\boldsymbol{\theta}$ will need to be a probability vector for every realization of $\boldsymbol{\alpha}$, and in particular for each unit-coordinate point, $\alpha_j = 1, \alpha_{j'} = 0$ (all $j' \neq j$), we have the following result.

Lemma 1. If the support of the underlying random probability m vector $\boldsymbol{\alpha}$ includes the extreme points of $\Delta^{(m)}$, then the requirement that the support of $\boldsymbol{\theta}$ (4) be contained in $\Delta^{(k)}$ is equivalent to the requirement that each column vector $\mathbf{g}_j$ of $\mathbf{G}$ be a probability k vector.

Thus we assume that all entries of $\mathbf{G}$ are nonnegative and each column sums to unity. That is, the fixed matrix $\mathbf{G}$ is what is called a (singly) stochastic matrix. If the support set of the random weights $\boldsymbol{\alpha}$ is the full probability simplex $\Delta^{(m)}$ (e.g., if $\boldsymbol{\alpha} \sim D(\mathbf{a})$ and for each parameter coordinate, $0 < a_j < \infty$), then the support set of random $\boldsymbol{\theta}$ is the full convex hull of the (fixed) column vectors of $\mathbf{G}$, CHull($\mathbf{G}$), a convex polytope and subset of $\Delta^{(k)}$. The vertices or extreme points of CHull($\mathbf{G}$) are column vectors of $\mathbf{G}$, but not all columns of $\mathbf{G}$ need be vertices. The polytope CHull($\mathbf{G}$) would be a complicated range to work with if one tried to develop and use a density for $\boldsymbol{\theta}$, with only the case $k = m$ with nonsingular $\mathbf{G}$ being simple. The only densities that we use are densities of the weights vector $\boldsymbol{\alpha}$, whose support is chosen to be the full probability simplex $\Delta^{(m)}$. Note that our focus here is on statistical inference about $\boldsymbol{\theta} = \mathbf{G}\boldsymbol{\alpha}$, and not about $\boldsymbol{\alpha}$ or $\mathbf{G}$, so no problem of identifiability can arise from a choice of lower rank for $\mathbf{G}$. But, as with any prior distribution, the support CHull($\mathbf{G}$) should include, or nearly include, the so-called "true" value of the parameter $\boldsymbol{\theta}$.

3. STATISTICAL INFERENCE

3.1 Bayes Theorem

The posterior density of the weights $p(\boldsymbol{\alpha}|\mathbf{n})$ is proportional to the product of the prior density $p(\boldsymbol{\alpha})$ and the likelihood (1) rewritten as a function of $\boldsymbol{\alpha}, L_n(\boldsymbol{\theta}(\boldsymbol{\alpha}))$. We give details for the case of $\boldsymbol{\alpha}$ prior-distributed Dirichlet.

Theorem 1. The likelihood (1) for iid sampling from a finite distribution with unknown probability vector $\boldsymbol{\theta}$, together with the filtered-variate Dirichlet prior distribution $\boldsymbol{\theta} \sim F_G D(\mathbf{a})$ on $\Delta^{(k)}$, yield the posterior filtered-variate distribution $\boldsymbol{\theta}|\mathbf{n} \sim F_G D(\mathbf{a}, \mathbf{G}^T, \mathbf{n})$, in which again $\boldsymbol{\theta} = \mathbf{G}\boldsymbol{\alpha}$, but $\boldsymbol{\alpha}$ has the posterior generalized Dirichlet distribution $\boldsymbol{\alpha}|\mathbf{n} \sim D(\mathbf{a}, \mathbf{G}^T, \mathbf{n})$, defined by the density on $\Delta^{(m)}$

$$p(\boldsymbol{\alpha}|\mathbf{n}) = \left(B(\mathbf{a})^{-1}\prod_{j=1}^{m}\alpha_j^{a_j-1}\right) \times \prod_{i=1}^{k}\left(\sum_{j=1}^{m}\alpha_j g_{ij}\right)^{n_i} / \Re(\mathbf{a}, \mathbf{G}^T, -\mathbf{n}). \tag{5}$$

The normalizing constant in the density (5) is a special case of Carlson's (1971) symmetrized multiple hypergeometric function, $\Re(\mathbf{a}, \mathbf{G}^T, -\mathbf{n}) = \Re_{n+}(\mathbf{a}, \mathbf{G}^T, -\mathbf{n})$. (Note the difference in sign between the final parameters in our notation for the distribution, $D(\mathbf{a}, \mathbf{G}^T, \mathbf{n})$, and Carlson's function.) This function is the complete integral of the nu-

merator of (5), a Dirichlet expectation of the likelihood, that is, the Bayesian prior-predictive probability.

Corollary 1. If $\boldsymbol{\theta} \sim F_G D(\mathbf{a})$, then the prior predictive probability of a sample sequence $\mathbf{x}$ with frequency counts $\mathbf{n}$, for fixed $N \equiv n_+$, is $p(\mathbf{x}) = EL_n(\boldsymbol{\theta}) = \Re(\mathbf{a}, \mathbf{G}^T, -\mathbf{n})$.

The posterior distribution of the category probabilities $\boldsymbol{\theta}$ is induced by the posterior distribution of the weights vector $\boldsymbol{\alpha}|\mathbf{n} \sim D(\mathbf{a}, \mathbf{G}^T, \mathbf{n})$ with density (5). Such *generalized Dirichlet* distributions $D(\mathbf{a}, \mathbf{B}, \mathbf{c})$ were defined by Dickey (1983) and applied to missing-data problems by Dickey, Jiang, and Kadane (1987). (The Dirichlet $D(\mathbf{a})$ itself is the special case, $\mathbf{B} = \mathbf{0}$ or $\mathbf{c} = \mathbf{0}$.) The posterior distribution of the linearly transformed $\boldsymbol{\theta} = \mathbf{G}\boldsymbol{\alpha}$, $\boldsymbol{\theta}|\mathbf{n} \sim F_G D(\mathbf{a}, \mathbf{G}^T, \mathbf{n})$, is then a "filtered-variate generalized Dirichlet" distribution. A family of such distributions is obviously closed under further sampling from the same sampling distribution: a prior distribution $\boldsymbol{\theta} \sim F_G D(\mathbf{a}, \mathbf{G}^T, \mathbf{c})$ and sample data $\mathbf{n}$ would yield the posterior distribution $\boldsymbol{\theta}|\mathbf{n} \sim F_G D(\mathbf{a}, \mathbf{G}^T, \mathbf{c} + \mathbf{n})$.

But an even greater generality is available without sacrificing tractability. Define the *filtered-variate generalized Dirichlet* distribution, $F_G D(\mathbf{a}, \mathbf{B}, \mathbf{c})$, as the distribution of $\boldsymbol{\theta} = \mathbf{G}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} \sim D(\mathbf{a}, \mathbf{B}, \mathbf{c})$. Then such a prior family is closed under sampling. Writing $\mathbf{B} = (\mathbf{A}, \mathbf{G}^T)$ (without loss of generality; e.g., by writing a concatenated list $\mathbf{c} = (\mathbf{d}, \mathbf{0})$), obtain the posterior distribution from the data $\mathbf{n}$ with likelihood (1), as $\boldsymbol{\theta}|\mathbf{n} \sim F_G D(\mathbf{a}, \mathbf{B}, \mathbf{c} + \mathbf{n}^*)$, where the concatenated list $\mathbf{n}* = (\mathbf{0}, \mathbf{n})$.

We give further details regarding the posterior distribution of the weights $\boldsymbol{\alpha}$ in the case of a Dirichlet prior for $\boldsymbol{\alpha}$. The generalized Dirichlet posterior distribution, (3.1), is tractable in several senses. (For additional properties, see Dickey, Garthwaite, and Bian 1995.)

Corollary 2. From (5), the posterior $\mathbf{c}$th mixed moment of $\boldsymbol{\alpha}$, for $\mathbf{c} = (c_1, \ldots, c_m)$, is proportional to a ratio of Carlson functions,

$$E\left(\prod_1^m \alpha_j^{c_j} | \mathbf{n}\right) = h(\mathbf{c}; \mathbf{a})\Re(\mathbf{a} + \mathbf{c}, \mathbf{G}^T, -\mathbf{n})/\Re(\mathbf{a}, \mathbf{G}^T, -\mathbf{n}). \quad (6)$$

The proportionality factor $h(\mathbf{c}; \mathbf{a})$ is the corresponding prior Dirichlet $\mathbf{c}$th moment (2).

Carlson's functions can be calculated easily with a microcomputer by multinomial expansion for small to medium N, Laplace's asymptotic method for medium to large N, or Monte Carlo in $\boldsymbol{\alpha} \sim D(\mathbf{a})$ on $\Delta^{(m)}$ (for details, see Jiang, Kadane, and Dickey 1992). Typically, in histogram smoothing problems, one would use either Laplace or Monte Carlo methods. We use Monte Carlo for the examples in Section 6.

3.2 Estimates

The posterior mean is an attractive and natural estimate for $\boldsymbol{\theta}$. It minimizes expected squared error and will be smoother than the raw relative frequencies when the prior distribution favors smooth probability vectors. But although the mean tends to be close to true $\boldsymbol{\theta}$, it tends to be smoother than $\boldsymbol{\theta}$. Define the nonsmoothness, or *r-roughness,* of a vector by the average of its squared r-distant differences,

$$Q_r(\boldsymbol{\theta}) = \sum_{i=1}^{k-r} (\theta_i - \theta_{i+r})^2/(k - r), \quad (7)$$

with special interest in adjacent differences, $r = 1$. Then the posterior expected roughness exceeds the roughness of the posterior mean, because $E[Q_r(\boldsymbol{\theta})|\mathbf{n}] - Q_r[E(\boldsymbol{\theta}|\mathbf{n})] = \sum_{i=1}^{k-r} \operatorname{var}(\theta_i - \theta_{i+r}|\mathbf{n})/(k - r) \geq 0$. In our view, achieving an accurate estimation is usually more important than portraying the true smoothness, and so we recommend the posterior mean (or the mode) when the posterior distribution must be summarized in the form of a vector estimate. But there may be occasions when it is reasonable to quote an estimate that exhibits a roughness equal to the posterior expected roughness, an interesting problem for treatment in future work. Of course, a Bayesian posterior distribution contains more information than a single estimate, and Bayesian inference is not restricted to the reporting of an estimate.

The mean of a filtered-variate random vector $\boldsymbol{\theta} = \mathbf{G}\boldsymbol{\alpha}$ is the same linear filter of the mean of the underlying vector $\boldsymbol{\alpha}$, $E(\boldsymbol{\theta}|\mathbf{n}) = \mathbf{G}E(\boldsymbol{\alpha}|\mathbf{n})$. The posterior variance matrix is a linear function of the posterior variances and covariances of $\boldsymbol{\alpha}$, $\operatorname{var}(\boldsymbol{\theta}|\mathbf{n}) = \mathbf{G} \operatorname{var}(\boldsymbol{\alpha}|\mathbf{n})\mathbf{G}^T$. In the prior filtered-variate Dirichlet case we have seen, in (6), that the posterior moments of the underlying weights $\boldsymbol{\alpha}$ are proportional to ratios of computationally feasible Carlson functions. The posterior moments of $\boldsymbol{\theta}$ themselves are similarly expressible.

Corollary 3. The posterior $\mathbf{d}$th moment of $\boldsymbol{\theta}$, for $\mathbf{d} = (d_1, \ldots, d_k)$, is

$$E\left(\prod_1^k \theta_i^{d_i} | \mathbf{n}\right) = \Re(\mathbf{a}, \mathbf{G}^T, -(\mathbf{n} + \mathbf{d}))/\Re(\mathbf{a}, \mathbf{G}^T, -\mathbf{n}). \quad (8)$$

Hence the first two posterior moments are $E(\theta_i|\mathbf{n})$, given by (8) with $\mathbf{d} = \boldsymbol{\delta}_{(i)}$, and $E(\theta_i\theta_{i'}|\mathbf{n})$, given by (8) with $\mathbf{d} = \boldsymbol{\delta}_{(i)} + \boldsymbol{\delta}_{(i')}$, where $\boldsymbol{\delta}_{(i)} = (\delta_{i1}, \ldots, \delta_{ik})$, with $\delta_{ii} = 1$ and $\delta_{ij} = 0$ all $j \neq i$. Equation (8) also gives the posterior predictive probability $p(\mathbf{y}|\mathbf{n})$ of any specific further sample sequence $\mathbf{y}$ with frequency counts $\mathbf{d}$.

The linear relation between the posterior means of $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ does not depend in any way on the rank of the filter matrix $\mathbf{G}$. When $\mathbf{G}$ is nonsingular, the posterior mode also can be conveniently calculated as the linear filter of the posterior mode of $\boldsymbol{\alpha}$, $\operatorname{mode}(\boldsymbol{\theta}|\mathbf{n}) = \mathbf{G} \operatorname{mode}(\boldsymbol{\alpha}|\mathbf{n})$, because in this case the posterior densities of $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ are directly proportional. To appreciate that nonsingularity of $\mathbf{G}$ would be needed for such invariance of the posterior mode, note that, unlike the mean, a mode is not preserved under marginalization (a singular linear transformation). For example, in the case of a Dirichlet distribution, $(\alpha_1, \ldots, \alpha_k) \sim D(a, b, \ldots, b)$, where $a > 1, b > 1$, we have the vector $\operatorname{mode}(\alpha_1, \ldots, \alpha_{k-1}) = (a - 1, b - 1, \ldots, b - 1)/(a + (k - 1)b - k)$, but because $\alpha_1 \sim \operatorname{beta}(a, (k - 1)b)$, the scaler $\operatorname{mode}(\alpha_1) = (a - 1)/(a + (k - 1)b - 2)$, which is not equal to the first coordinate of the joint mode.

3.3 Comparison of Hypotheses

Posterior "scientific reporting" was defined by Dickey (1973) to require the communication (e.g., by table or graphical display) of the dependence of the inference on the prior distribution meaningfully interpreted in real problems. Bayesian comparative judgement of hypotheses is based on the posterior odds for one hypothesis versus another, $P(H_1|\mathbf{x})/P(H_2|\mathbf{x})$, given observed data $\mathbf{x}$. The evidence in the statistical data $\mathbf{x}$ relevant to such a judgment is summarized through the Bayes factor, the ratio of the posterior odds to the prior odds, reportable even without a choice of prior odds. The Bayes factor can be calculated from the data and the conditional prior distributions given each of the two hypotheses, as the ratio of the two conditional predictive probabilities (or densities) of the data, $[P(H_1|\mathbf{x})/P(H_2|\mathbf{x})]/[P(H_1)/P(H_2)] = p(\mathbf{x}|H_1)/p(\mathbf{x}|H_2)$. Such a predictive probability conditional on a hypothesis H_i is the integral function of the conditional prior distribution, $p(\mathbf{x}|H_i) = \int p(\mathbf{x}|\boldsymbol{\theta})\, dP(\boldsymbol{\theta}|H_i)$. In Corollary 1 we obtained the predictive probability of a sample sequence, $\mathbf{x} = (x_1, \ldots, x_N)$, under a filtered-variate Dirichlet prior distribution. An ordinary Dirichlet prior distribution would have the predictive probability $h(\mathbf{n}; \mathbf{b})$ (2). Thus the Bayes factor for comparing our models for histograms is just a simple ratio involving Carlson functions. For example, we obtain the following from Corollary 1.

Corollary 4. The Bayes factor in favor of the point null H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus the composite alternative H_A: $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ under the filtered-variate conditional prior $\boldsymbol{\theta}|H_A \sim F_G D(\mathbf{a})$ is $(\prod_{i=1}^{k} \theta_{0i}^{n_i})/\Re(\mathbf{a}, \mathbf{G}^T, -\mathbf{n})$.

4. MOMENTS OF FINITE AND CONTINUOUS-TYPE DISTRIBUTIONS

In many situations, smooth category probabilities are considered likely, and the specification of a prior distribution favoring smooth values is desired. But how can one choose a particular filtered-variate distribution to express particular prior opinion about smoothness? Here we propose methods involving the specification of typical category-probability vectors and the prior matching of their empirical moments. Toward this end, we first develop a general theory of distributions on the probability simplex having a structured variance matrix depending quadratically on the mean vector. This is followed by a second-order representation theory for filtered variates of such distributions, which can be used to achieve a continuous-type prior distribution with moments matching the empirical moments of the list of typical vectors.

4.1 Distributions with Mean-Structured Variance

Lemma 2. Suppose that a random vector $\mathbf{y} = (y_1, \ldots, y_m)$ has the first two moments,

$$E\mathbf{y} = \boldsymbol{\mu}, \qquad \operatorname{var}(\mathbf{y}) = c[\operatorname{diag}(\boldsymbol{\mu}) - \boldsymbol{\mu}\boldsymbol{\mu}^T], \tag{9}$$

for some fixed vector (vertical array) $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$ and scalar $c \geq 0$. Then for $\delta = 0, 1$, separately, $y_+ = \delta$ with probability 1 iff $\mu_+ = \delta$.

Our interest is in random probability vectors. So, if $\mathbf{y}$ has nonnegative coordinates $y_j \geq 0$, and has moments (9) with $\mu_+ = 1$, we shall say that $\mathbf{y}$ has a distribution with MSV and write

$$\mathbf{y} \sim \text{MSV}(\boldsymbol{\mu}, c). \tag{10}$$

Such $\mathbf{y}$ thus is restricted to $\Delta^{(m)}$. Among others, the normalized multinomial and the Dirichlet are MSV distributions: (a) If $\mathbf{z} = (z_1, \ldots, z_m) \sim$ multinomial$(N, \boldsymbol{\phi})$ and $\mathbf{y} = \mathbf{z}/N$, then $\mathbf{y} \sim \text{MSV}(\boldsymbol{\phi}, N^{-1})$; (b) If $\mathbf{y} \sim D(\mathbf{a})$ on $\Delta^{(m)}$, with $\mathbf{a} = a_+\mathbf{u}$, then $\mathbf{y} \sim \text{MSV}[\mathbf{u}, (a_+ + 1)^{-1}]$. The corresponding Dirichlet-multinomial distribution (by taking $\boldsymbol{\phi} \sim D(\mathbf{a})$) is also in this class, $\mathbf{y} \sim \text{MSV}(\mathbf{u}, c)$, with $c = N^{-1} + (a_+ + 1)^{-1} - [N(a_+ + 1)]^{-1}$. (This is true for *any* MSV mixture over the mean of a conditionally MSV distribution.) MSV distributions are of interest in their own right, and an account of their properties will be published elsewhere. (Related properties in one dimension were studied by Bar-Lev and Enis 1986.) These distributions have a simple limiting case that will be useful in our methods of assessment of uncertainty as a filtered-variate MSV distribution.

Theorem 2. If $\mathbf{z} \sim \text{MSV}(\mathbf{u}, c)$ on $\Delta^{(m)}$, then $0 \leq c \leq 1$. As $c \uparrow 1$, $\mathbf{z}$ takes the limiting distribution $\text{MSV}(\mathbf{u}, 1)$, the (unique to $\mathbf{u}$) finite distribution $\tilde{P}$ supported on the set of vertices of the probability simplex $\Delta^{(m)}$,

$$\tilde{P}\,[\mathbf{z} = \boldsymbol{\delta}_{(j)}] = u_j, \tag{11}$$

$j = 1, \ldots, m$. More generally for $0 \leq c \leq 1$, $\mathbf{z}$ has the first two moments $E(\mathbf{z}) = E(\tilde{\mathbf{z}})$, and $\operatorname{var}(\mathbf{z}) = c \operatorname{var}(\tilde{\mathbf{z}})$, where $\tilde{\mathbf{z}}$ has the distribution $\tilde{\mathbf{P}}$.

Corollary 5. If $\mathbf{z} \sim D(\mathbf{a})$ with $\mathbf{a} = a_+\mathbf{u}$, then the limiting distribution of $\mathbf{z}$, as $a_+ \downarrow 0$ (as $c = (a_+ + 1)^{-1} \uparrow 1$), is the finite distribution $\tilde{P}$ (11).

4.2 Filtered-Variate Mean-Structured-Variance Distributions

We turn again to the idea of filtering a random probability vector, keeping in mind that for a filtered-variate Dirichlet, the underlying Dirichlet distribution is MSV. Consider the random vector $\boldsymbol{\theta} = \mathbf{G}\boldsymbol{\alpha}$ and assume that $\boldsymbol{\alpha} \sim \text{MSV}(\mathbf{u}, c)$ (requiring $0 \leq c \leq 1$). Then we say that the induced distribution of $\boldsymbol{\theta}$ is *filtered-variate mean-structured variance* (FMSV) and write $\boldsymbol{\theta} \sim F_G\text{MSV}(\mathbf{u}, c)$. Again, $\boldsymbol{\theta}$ has a distribution supported on a subset of CHull($\mathbf{G}$), the convex hull of the set of column vectors of $\mathbf{G}$. Our limiting case of an MSV distribution, $\tilde{P}$ (11), implies a finitely supported limiting case $\tilde{P}_G$, for the corresponding FMSV distribution. For $\boldsymbol{\theta} \sim F_G\text{MSV}(\mathbf{u}, c)$, as $c \uparrow 1$, $\boldsymbol{\theta}$ has the limiting distribution $\tilde{P}_G$ finitely supported on the set of column vectors of $\mathbf{G} = (\mathbf{g}_1, \ldots, \mathbf{g}_m)$,

$$\tilde{P}_G\,[\boldsymbol{\theta} = \mathbf{g}_j] = u_j, \qquad j = 1, \ldots, m. \tag{12}$$

The finite support of $\tilde{P}_G$ includes the extreme or vertex points of CHull($\mathbf{G}$). Using Theorem 2 and the relations $E(\boldsymbol{\theta}) = \mathbf{G}E(\boldsymbol{\alpha})$, $\operatorname{var}(\boldsymbol{\theta}) = \mathbf{G} \operatorname{var}(\boldsymbol{\alpha})\mathbf{G}^T$, we express the

low-order moments of an FMSV distribution in terms of its limiting moments.

Corollary 6 (Proportional Moments). If $\boldsymbol{\theta} \sim F_G\text{MSV}(\mathbf{u}, c)$ and $\tilde{\boldsymbol{\theta}}_G$ has the finite distribution $\tilde{P}_G$ (12) supported on the set of column vectors of $\mathbf{G}$, then $E(\boldsymbol{\theta}) = E(\tilde{\boldsymbol{\theta}}_G) = \bar{\mathbf{g}}$ and $\text{var}(\boldsymbol{\theta}) = c\ \text{var}(\tilde{\boldsymbol{\theta}}_G) = c\mathbf{S}_g$, where the u-weighted discrete averages, $\bar{\mathbf{g}} = \mathbf{Gu} = \sum u_j \mathbf{g}_j$ and $\mathbf{S}_g = (\mathbf{G} - \bar{\mathbf{g}}\mathbf{1}_m^T)^T \text{diag}(\mathbf{u})(\mathbf{G} - \bar{\mathbf{g}}\mathbf{1}_m^T) = \sum u_j(\mathbf{g}_j - \bar{\mathbf{g}})(\mathbf{g}_j - \bar{\mathbf{g}})^T$.

5. PRIOR ASSESSMENT

If we define prior smoothness by small prior expectation of a quadratic quantification of roughness, say $Q_r(\boldsymbol{\theta})$, then

$$E[Q_r(\boldsymbol{\theta})] = Q_r[E(\boldsymbol{\theta})] + \sum_{i=1}^{k-r} [\text{var}(\theta_i) + \text{var}(\theta_{i+r}) - 2\ \text{cov}(\theta_i, \theta_{i+r})]/(k-r), \quad (13)$$

interest focuses on the low-order moments of $\boldsymbol{\theta}$, which can be controlled by use of the following theorem. Because a continuous-type distribution $\boldsymbol{\theta} \sim F_G\text{MSV}(\mathbf{u}, c)$ in Corollary 6 would necessitate the strict inequality $0 < c < 1$, such a distribution for $\boldsymbol{\theta}$ cannot have the same moments as the corresponding finite limiting distribution $\tilde{P}_G$, for which $c \uparrow 1$. But the moments of $\boldsymbol{\theta}$ *can* be made to match the moments of a finite distribution $\tilde{P}_H$, of which $\tilde{P}_G$ is a dilation.

Theorem 3 (Second-Order Equivalence to a Finite Distribution). If the (filtering) matrix $\mathbf{G}$ is obtained as a d-dilation of the (elicited) matrix $\mathbf{H}$,

$$\mathbf{G} = d(\mathbf{H} - \bar{\mathbf{h}}\mathbf{1}_m^T) + \bar{\mathbf{h}}\mathbf{1}_m^T, \quad (14)$$

where $\bar{\mathbf{h}} = \sum u_j \mathbf{h}_j$, the $\mathbf{u}$-weighted average of the columns of $\mathbf{H}$, and if such $\mathbf{G}$ is then used to define the filtered-variate distribution $\boldsymbol{\theta} \sim F_G\text{MSV}(\mathbf{u}, c)$, then we have the proportional moments,

$$E(\boldsymbol{\theta}) = E(\tilde{\boldsymbol{\theta}}_H) = \bar{\mathbf{h}}$$

and

$$\text{var}(\boldsymbol{\theta}) = cd^2\text{var}(\tilde{\boldsymbol{\theta}}_H) = cd^2\mathbf{S}_h, \quad (15)$$

where $\tilde{\boldsymbol{\theta}}_H$ has the finite distribution $\tilde{P}_H$ with the probability masses $\mathbf{u}$ at the column vectors of $\mathbf{H}$, with mean $\bar{\mathbf{h}}$ and variance $\mathbf{S}_h = \sum u_j(\mathbf{h}_j - \bar{\mathbf{h}})(\mathbf{h}_j - \bar{\mathbf{h}})^T$. If the columns of $\mathbf{H}$ are all probability vectors and if $d \le [1 - \min_j(h_{i,j}/\bar{h}_i)]^{-1}$ for each $\bar{h}_i > 0$, then the columns of $\mathbf{G}$ (14) are also probability vectors, and further assuming $d \ge 1$, the convex hull CHull($\mathbf{G}$) contains the column vectors of $\mathbf{H}$. For choices $cd^2 = 1$ in (15), the first two moments of the (constructed) distribution of $\boldsymbol{\theta}$ are equal to the moments of the (elicited) finite distribution $\tilde{P}_H$, $E(\boldsymbol{\theta}) = \bar{\mathbf{h}}$ and $\text{var}(\boldsymbol{\theta}) = \mathbf{S}_h$. In this case, the mean roughnesses (13) also match, $E[Q_r(\boldsymbol{\theta})] = E[Q_r(\tilde{\boldsymbol{\theta}}_H)] = \sum_{i=1}^{k-r} \sum_j u_j(h_{i,j} - h_{i+r,j})^2/(k-r)$.

Corollary 7. Both the limiting distribution (12) and the moment representation (15) hold for the continuous-type filtered-variate Dirichlet, $\boldsymbol{\theta} \sim F_G D(\mathbf{a}), \mathbf{a} = a_+\mathbf{u}$, with $c = (a_+ + 1)^{-1}$.

5.1 Constructing the Prior Distribution

There are many variations on the use of Theorem 3 or its Corollary 7 to construct a meaningful distribution for $\boldsymbol{\theta}$ in various situations. We set out the principles and offer specific detailed suggestions, but the approach is too new and the possibilities too rich to advocate, now, a unique "best" assessment procedure.

To construct a filtered-variate form of prior $\boldsymbol{\theta} \sim F_G\text{MSV}(\mathbf{u}, c)$, one first develops a discrete (finite) distribution $\tilde{P}_H$ having the desired mean $\bar{\mathbf{h}}$ and the desired variance $\mathbf{S}_h$, and then fine tunes the remaining proportionality constant d with $c = d^{-2}$, for example, by examining Monte Carlo samples of $\boldsymbol{\theta}$ for overall suitability to depict the expert's prior opinion. We are not advocating use of the tuning parameter d to control the extent to which the prior distribution favors smoothness. This aspect is controlled largely by the smoothness of the column vectors of $\mathbf{H}$, as captured by their low-order empirical moments, matched, for each value of d, by the corresponding moments of the prior distribution.

For simplicity, it is tempting to begin by considering cases where the random vector $\tilde{\boldsymbol{\theta}}_H$ is discrete uniform over its finite support set,

$$P[\tilde{\boldsymbol{\theta}}_H = \mathbf{h}_j] = u_j \equiv 1/m, \qquad j = 1, \ldots, m. \quad (16)$$

We restrict our discussion to this choice, for which $\mathbf{u} = \mathbf{1}_m/m$. Following the specification of "typical-vector" columns $\mathbf{h}_j$ for $\mathbf{H}$, a filtered-variate Dirichlet prior distribution would be constructed as

$$\boldsymbol{\theta} \sim F_G D(a\mathbf{1}_m), \quad (17)$$

where $a = (d^2 - 1)/m$, with $\mathbf{G}$ being the d-dilation (14) of the specified matrix $\mathbf{H}$.

5.2 Using a Small Number of Elicited Typical Vectors With Additional Prior Invariance Assumptions

A practical problem arises immediately. The expert may find it more difficult to devise a whole list of vectors having prior means and variances as the empirical moments of the list than just to state a few vectors that are typically smooth. Enough vectors will be needed to produce a sufficiently rich convex hull. One obvious solution is to use smoothness-preserving transformations to generate a balanced set of vectors from just a few elicited typical vectors. For example, for a one-dimensional histogram with sequentially numbered categories, $i = 1, \ldots, k$, consider a square matrix $\mathbf{H}, k \times k$ $(m = k)$, where each point $\mathbf{h}_j = (h_{1,j}, \ldots, h_{k,j})$ is a cyclically j-shifted version of a single typical probability vector $\mathbf{t} = (t_1, \ldots, t_k)$. That is, for each $i, j = 1, \ldots, k$,

$$h_{i,j} = t_s, \quad (18)$$

where $s = i - j + 1 \pmod k$. This implies an ergodic property for $\tilde{P}_H$, whereby the joint distribution of any subset of coordinates of $\tilde{\boldsymbol{\theta}}_H$ is identical to their serial distribution over the cyclic coordinate shifts of any particular possible outcome vector $\mathbf{h}_j$. Then $E\theta_i = E\tilde{\theta}_{Hi} = \bar{t} = 1/k$, for all i, and similarly, the variance is a cyclic Toeplitz-form matrix,

$\tilde{\sigma}_{i,i+r} \equiv \sum_s t_s t_{s+r}/k - (\bar{t})^2 = f(r)$, for all i, where t_{k+s} is taken to equal t_s. Such cyclical translation invariance in one or more dimensions provides a powerful simplicity, although it may need to be corrected in use for unrealistic edge effects.

Somewhat less simply, one can increase the number of column vectors m to a multiple of k and include different shapes or "frequencies" along with the shifts or "phases"; for example,

$$t_{s,s'} \propto 1 + \sin[(sB_{s'} - C_{s'})2\pi/k]. \tag{19}$$

Such a frequency-and-phase approach is related to that of Lo (1984) in the continuous realm.

We suggest that the expert state a characteristic variety of typical vectors. If he or she avows a symmetry like the shift-invariance (18), or for some other reason would like to express a prior distribution of the form $E\theta_i \equiv 1/k$ and $\operatorname{cov}(\theta_i, \theta_{i+r}) \equiv f(r)$, then moments of this form can be achieved by extending the expert's elicited list of typical vectors to include all their coordinate shifts. Or if a nonconstant mean $E\theta_i$ is desired, the shifts can be performed on the difference vectors, typical vectors minus desired mean vector, as follows. (As mentioned in Sec. 3.2 in connection with equation (7), a mean itself should not be considered a typical vector.) For simplicity, we state results in terms of a single typical vector $\mathbf{h}_0$ and the desired mean vector $\mathbf{p}$.

Theorem 4. Given two probability vectors, $\mathbf{h}_0$ and $\mathbf{p}$, define the difference vector $\mathbf{t}^*$ by

$$\mathbf{h}_0 = \mathbf{p} + \mathbf{t}^*, \tag{20}$$

and denote the shifted difference vectors, constructed by shifting $\mathbf{t}^*$ according to (18), by $\mathbf{h}_j^* = (h_{1j}^*, \ldots, h_{kj}^*), j = 1, \ldots, k$. (Here $\sum t_i^* = 0$, and similarly for each vector $\mathbf{h}_j^*$.) Then if $\min(t_i^*) + \min(p_i) \geq 0$, the new vectors, defined by

$$\mathbf{h}_j = \mathbf{p} + \mathbf{h}_j^*, \tag{21}$$

$j = 1, \ldots, k$, are probability vectors. The new list has the empirical mean, $\bar{\mathbf{h}} = \sum \mathbf{h}_j/k = \mathbf{p}$, and the empirical variance matrix, in Toeplitz form matching the serial covariances from $\mathbf{t}^*, \tilde{\sigma}_{i,i+r} \equiv \sum_s t_s^* t_{s+r}^*/k = f^*(r)$.

In the constant-mean case, the shifts can be applied directly to the typical vectors themselves, because the shifts have no effect on the mean vector. Finally, in this case, if $\boldsymbol{\theta} \sim F_G \mathrm{MSV}(\mathbf{1}_k/k, c)$, with $c = d^{-2}$, where $m = k$ and $\mathbf{G}$ is derived by the dilation (14) from the shifts $\mathbf{H}$ on the single typical vector, $\mathbf{h}_0 = \mathbf{1}_k/k + \mathbf{t}^*$, then the prior expectation of the r-roughness $Q_r(\boldsymbol{\theta})$ (7) will be the same as the empirical serial roughness of the expert's typical vector, $E[Q_r(\boldsymbol{\theta})] = \sum_i (h_{i,0} - h_{i+r,0})^2/(k-r)$, in which $h_{k+s,0}$ is taken to equal $h_{s,0}$. (Each of the $k - r$ terms of $Q_r(\boldsymbol{\theta})$ has the same expected value.)

Table 1. Chondrite Data: Probability Estimates

Silica percentage	Counts	Estimation method A	B	C	D	E
20.00–21.60	1	.0455	.0588	.0645	.0664	.0795
21.61–23.20	3	.1364	.1275	.1237	.0796	.0873
23.21–24.80	0	.0000	.0245	.0349	.0561	.0764
24.81–26.40	0	.0000	.0245	.0349	.0665	.0880
26.41–28.00	6	.2727	.2305	.2124	.1508	.1034
28.01–29.60	2	.0909	.0931	.0941	.1041	.1076
29.61–31.20	1	.0455	.0588	.0645	.0848	.1117
31.21–32.80	1	.0455	.0588	.0645	.0969	.1234
32.81–34.40	7	.3182	.2648	.2420	.1894	.1374
34.41–36.00	1	.0455	.0588	.0645	.1055	.0855

6. EXAMPLES

We illustrate the theory with three examples. In the first example, symmetric-prior methods will be used, and the posterior means of category probabilities will be compared to other estimates in a data problem studied in the literature. In the second example, elicited typical vectors are used to construct the prior distribution, and the posterior means are compared for different values of the tuning parameter d. In our third example, we construct the typical vectors from historical data. In each case we report posterior moments, computed according to (8), which can be viewed as the ratio of prior moments, each in the form of Corollary 1,

$$E[g(\boldsymbol{\theta})|\mathbf{n}] = E[L_n(\boldsymbol{\theta})g(\boldsymbol{\theta})]/E[L_n(\boldsymbol{\theta})], \tag{22}$$

where $\boldsymbol{\theta}$ has the filtered-variate Dirichlet (prior) distribution $\boldsymbol{\theta} \sim F_G D(\mathbf{a})$, say with $\mathbf{a} = a\mathbf{1}_m$. (Note that the expectations on the right side of (22) are taken with the data $\mathbf{n}$ fixed, but not conditioned on.) We use Monte Carlo to compute the numerator and denominator of (22), because a is small and m is not small, in the examples. Because $\boldsymbol{\theta} = \mathbf{G} \cdot \boldsymbol{\alpha}$, where $\boldsymbol{\alpha} \sim D(a\mathbf{1}_m)$, this means pseudorandom sampling of a Dirichlet vector, with coordinates expressible as ratios constructed from independent chi-squared variates, $\alpha_j = X_j/\sum_1^m X_i$, where $X_j \sim \chi_{2a}^2$. The relative errors in the computations can be estimated explicitly and they are small in all of the following examples, with most less than 1%. (Jiang, Kadane, and Dickey 1992 treated the computation of $\Re$ for statistical uses.)

Example 1 (Chondrite Data). The chondrite data of Ahrens (1965) appeared in continuous form, but it has been converted here to counts $\mathbf{n}$ with $k = 10$ categories and total count $n_+ = 22$ (Table 1, column 2). To construct a prior distribution symmetrically, as in (18), we start with a single simple vector, $\mathbf{t} = (1, 2, \ldots, 10)/55$. We omit the dilation step for further simplicity. Then $\tilde{\boldsymbol{\theta}}$ is set up first to have the discrete uniform distribution over the $m = 10$ column vectors $\mathbf{g}_j$ of $\mathbf{G}$, obtained by cyclic shifts of $\mathbf{t}$. The mean (average) of such $\tilde{\boldsymbol{\theta}}$ is $\bar{\mathbf{g}} = \mathbf{1}_{10}/10$, and its variance is the cyclic Toeplitz-form matrix with first row as given (Table 2). If the expert assesses his or her prior variance of $\boldsymbol{\theta}$ to be $c = .5$ times this variance of $\tilde{\boldsymbol{\theta}}$ with the same correlations (so a dilation step would not be needed), then he or she will have the assessed prior distribution, $\boldsymbol{\theta} \sim F_G D(a\mathbf{1}_{10})$, with $a = (c^{-1} - 1)/10 = 1/10$. The corresponding posterior distribution is then provided by Theorem 1 as $\boldsymbol{\theta}|\mathbf{n} \sim F_G D[\mathbf{1}_{10}/10, \mathbf{G}^T, \mathbf{n}]$.

To compare our posterior means to other probability estimates discussed in this example (Titterington 1985, p. 149), we include the following estimates (Table 1, columns 3–7):

Table 2. Chondrite Data: First Row of Cyclic Variance Matrix for $\tilde{\theta}$

13.63	6.20	.41	−3.72	−6.20	−7.02	−6.02	−3.72	.41	6.20

NOTE: All entries are to be multiplied by 10^{-4}.

A, unsmoothed relative frequencies, B, minimum quadratic risk using convex smoothing prescription (Fienberg and Holland 1973), C, cross-validation with convex smoothing prescription and quadratic loss (Stone 1974), D, minimum quadratic risk using our minimum penalized distance prescription (Simonoff 1983), and E, the posterior means using the constructed smooth prior. The plot of these estimates (Fig. 1) shows that all the methods B–E smooth away the zero counts, and method E, leads to more local smoothness than methods A–D. If the vector t were chosen increasingly smooth, then the prior and posterior distributions would both concentrate more closely toward the central vector $\mathbf{1}_{10}/10$.

Example 2 (Women Categorized by Number of Children. In 1960, the U.S. Public Health Service interviewed American women aged 18–79 years and determined a distribution of women by number of children born, $0, \ldots, 9_+$(9 or more) (Table 3, column 2; from Freedman, Pisani, and Purves 1978, p. 38). The example treats this parent distribution as unknown, to be estimated from a sample $\mathbf{n}$. We used the distribution as the $k = 10$ multinomial probabilities to generate $N = 32$ women with the observed sample category counts $\mathbf{n}$ as shown (column 3). Included for comparison is the maximum likelihood estimate, the sample relative frequencies $\mathbf{n}/32$ (column 4).

To construct a prior distribution here, we use the method of elicited typical vectors. We assume that a subject-matter expert has assessed the set of $m = 10$ column vectors of matrix $\mathbf{H}$, as shown (in Table 4), chosen to be typical for their local smoothness and to have an empirical mean vector and variance matrix matching the expert's personal means and variances of vector $\boldsymbol{\theta}$. (The prior mean $\bar{\mathbf{h}}$ is given in

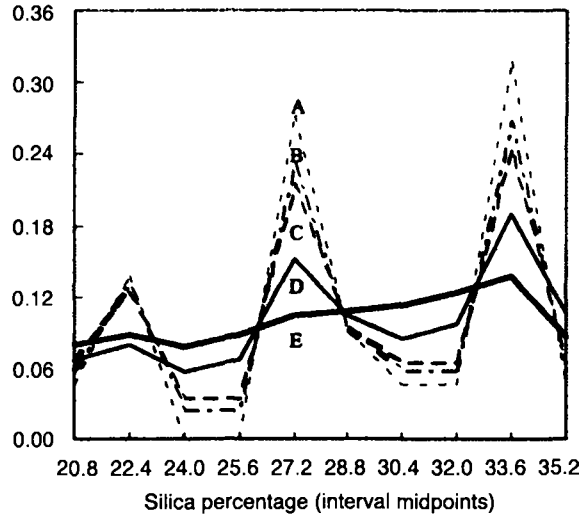

Figure 1. Chondrite Data: Probability Estimates. A, Observed frequencies; B, minimum risk (Fienberg and Holland 1973); C, cross validation; D, minimum risk (Simonoff 1983); E, posterior means.

Table 3. Distribution of Women by Number of Children Born

Number of children	Distribution of women	Sample counts	Sample relative frequency	Prior mean	Posterior mean
0	.22	5	.15625	.083	.1301
1	.17	3	.09375	.096	.1882
2	.21	14	.43750	.104	.2394
3	.16	0	.00000	.108	.1800
4	.10	6	.18750	.109	.1179
5	.05	3	.09375	.109	.0661
6	.03	0	.00000	.108	.0278
7	.02	0	.00000	.104	.0148
8	.02	0	.00000	.096	.0158
9+	.03	1	.03125	.083	.0198

Table 3, column 5.) Denoting the ith coordinate of $\bar{\mathbf{h}}$ by $\bar{h}_i$, the average of entries in the ith row of matrix $\mathbf{H}$, we compute $\min_i\{[1 - \min_j(h_{i,j}/\bar{h}_i)]^{-1}\} = 1.477$. For any value d between 1 and 1.477, the prior distribution would be $\boldsymbol{\theta} \sim F_G D(a\mathbf{1}_{10})$, where $\mathbf{G}$ follows from $\mathbf{H}$ by equation (14) and $a = (d^2 - 1)/10$ (17). Using such a prior distribution, we generated 30 random vector values $\boldsymbol{\theta}$ for each of several tentative choices of d. Assuming that the expert in the example chooses $d = 1.429$ from inspection of our sets of representative outcomes, the prior distribution then has $a = .1041$.

With this prior distribution and the sample data $\mathbf{n}$, our posterior distribution is the filtered-variate generalized Dirichlet, $\boldsymbol{\theta}|\mathbf{n} \sim F_G D(a\mathbf{1}_{10}, \mathbf{G}^T, \mathbf{n})$. The corresponding posterior mean of $\boldsymbol{\theta}$ is as shown (column 6). As depicted by the plot of posterior-mean coordinates and standard deviations (Fig. 2), the posterior mean smooths out the zero counts and is locally smoother than the maximum likelihood estimate. This smoothed estimate is, perhaps surprisingly, close to the parent distribution in the plot. Examining the effect of different choices of d, we found that the posterior mean coordinates change by about .01 as d varies from 1.33 to 1.47, with each coordinate changing monotonically.

Example 3 (Household Size). Consider the distribution of U.S. households by number of persons in the household, $i = 1, \ldots, 6_+$ (6 or more). This distribution is given in Table 5 for each of the seven years at 5-year intervals from 1955 to 1985. We discuss the problem of inferring the distribution for 1985 from a sample of $N = 600$ U.S. households from 1985, with the category counts $\mathbf{n} = (159, 210, 81, 90, 42, 18)$. In the inference problem, this sample and the six previous distributions at 5-year intervals are assumed known, while the 1985 distribution is not known.

Table 4. Distribution of Women: Matrix $\mathbf{H}$ of Typical Vectors

.056	.043	.035	.031	.030	.040	.071	.115	.173	.236
.056	.043	.035	.031	.040	.070	.111	.165	.223	.186
.056	.043	.035	.041	.070	.110	.161	.215	.173	.136
.056	.043	.045	.071	.110	.160	.211	.165	.123	.096
.056	.053	.075	.111	.160	.210	.161	.115	.083	.066
.066	.083	.115	.161	.210	.160	.111	.075	.053	.056
.096	.123	.165	.211	.160	.110	.071	.045	.043	.056
.136	.173	.215	.161	.110	.070	.041	.035	.043	.056
.186	.223	.165	.111	.070	.040	.031	.035	.043	.056
.236	.173	.115	.071	.040	.030	.031	.035	.043	0.56

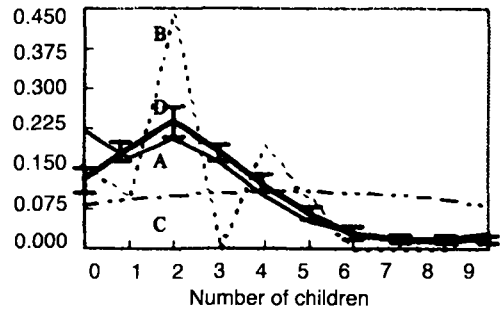

Figure 2. Distribution of Women: Probability Estimates. A, Parent distribution; B, sample relative frequency; C, prior mean; D, posterior mean (and mean ± SD).

In such an inference problem, an expert's prior opinion might be expressible as a list of typical category-probability vectors, either having the needed low-order prior moments as their empirical moments or accompanied by shift invariance or other assumptions further needed to define the prior moments. Our illustration uses the household-size distribution vectors from the previous years to construct typical vectors and prior moments regarding the 1985 vector. The prior information available in this problem has structure beyond the ergodic situation of the category shifts (18), and even beyond the more general setting of the enrichment Theorem 4. So we use a further modified assessment method. Stochastic processes and other sophisticated models using more extensive prior data could provide relatively accurate predictions of the 1985 vector. Also, plots against time of the given prior data show obvious advantages for curvilinear trend analysis and growth-curve analysis for predicting the population frequencies for 1985. But for our illustration, we merely assume simple linear time trends and use the following result about least-squares fitting to construct a list of typical vectors.

Consider as observed data a list of probability vectors in R^k, the list indexed by time. If each category's probability is separately regressed against time, then the vector of predicted values at any common time sums to unity (although the predicted values may fail to fall within [0, 1]). This is because the predicted value for each category is the same linear function of past years' observed sample data, the same linear calculation performed during the least-squares fitting to the respective category's data. Because the sum of each year's data over categories is the constant 1, the sum of predicted values is the predicted value 1. Using this result, we obtain the vector $\mathbf{p}$ of linear-trend values predicted for 1985 based on 1955–1980 (Table 6, column 8). This serves as our prior mean and the empirical mean $\bar{\mathbf{h}}$ for our constructed list of typical vectors $\mathbf{h}_j$. Let $\mathbf{h}_j^*$ be the residual vector back at the jth time under the linear-trends fit and define the typical vectors as the sums, $\mathbf{h}_j = \mathbf{p} + \mathbf{h}_j^*, j = 1, \ldots, 6$. Because the residuals $\mathbf{h}_j^*$ sum to 0 over time j, the empirical mean of the $\mathbf{h}_j$'s is $\bar{\mathbf{h}} = \mathbf{p}$, and we obtain a matrix $\mathbf{H}$ of typical vectors, as displayed in Table 6, columns 2–7.

Proceeding as in Example 2, we have $k = m = 6$ and $1 \leq d \leq 3.156$. But now, to help in choosing d, quantify roughness by $Q_1(\boldsymbol{\theta}) = \sum(\theta_i - \theta_{i+1})^2/5$. Using Monte Carlo samples of size 400, we obtain prior quantiles of $Q_1(\boldsymbol{\theta})^{1/2}$ for various values of d. Boxplots of medians, hinges, and so on (Mendenhall, Reinmuth, Beaver, and Duhan 1986) are given here for a few illustrative d values (Fig. 3). Choosing, for example, $d = 3.125$, we have $a = 1.46$ and $\mathbf{G}$ from $\mathbf{H}$ according to (14). The posterior distribution is again the corresponding filtered-variate Dirichlet. The posterior means and standard deviations are plotted for comparison to the sample relative frequencies, the least-squares linear trend predictions (prior mean vector), and the eventual "true" distribution for 1985 (Fig. 4).

Bayes-Factor Test of Hypothesis. Finally, for illustration of Bayesian comparison of hypotheses, we test in the context of this example the silly model of equal category probabilities, H_0: $\boldsymbol{\theta} = (1/6, \ldots, 1/6)$, nested within the local-smoothness hypothesis H_1 of our assessed prior model. The Bayes factor in favor of H_0 versus H_1 is computed as in Section 3.3, as follows: $P(\mathbf{n}|H_0) = \prod(1/6)^{n_i} = (1/6)^{600} = 1.2896 \times 10^{-467}$; $P(\mathbf{n}|H_1) = \Re(\mathbf{a}, \mathbf{G}^T, -\mathbf{n}) = 2.5935 \times 10^{-411}$, and so the odds for H_0 are diminished by the extreme factor, $[P(H_0|\mathbf{n})/P(H_1|\mathbf{n})]/[P(H_0)/P(H_1)] = 4.972 \times 10^{-57}$.

A deeper treatment and scientific report of a real version of any of these inference problems of course would include an analysis of the sensitivity of the inference to the prior assessment and structure and would report the inferences for a variety of contending expert's prior opinions and extreme bounding opinions, in the spirit of Dickey (1973).

7. CONCLUSION

In our class of inference problems, the filter $\boldsymbol{\theta} = \mathbf{G} \cdot \boldsymbol{\alpha}$

Table 5. Household-Size Historical Distributions; Percentages of U.S. Households by Number of Persons in Household, at 5-Year Intervals

Number of persons	1955[a]	1960[b]	1965[b]	1970[c]	1975[c]	1980[c]	1985[c]
1	10.9	13.1	15.0	17.1	19.6	22.7	23.7
2	28.5	27.8	28.1	28.9	30.6	31.4	31.6
3	20.4	18.9	17.9	17.3	17.4	17.5	17.8
4	18.9	17.6	16.1	15.8	15.6	15.7	15.7
5	11.1	11.5	11.0	10.3	9.0	7.5	7.0
6+	10.2	11.1	11.9	10.6	7.8	5.2	4.2

[a] Data from U.S. Bureau of the Census (1975), p. 42.
[b] Data from Hoffman (1987), p. 777.
[c] Data from U.S. Bureau of the Census (1987), p. 44.

Table 6. Prior-Chosen Typical Household-Size Distributions for 1985; Matrix $\mathbf{H}$ *of Typical Column Vectors*

Number of persons	$\mathbf{h}_1$	$\mathbf{h}_2$	$\mathbf{h}_3$	$\mathbf{h}_4$	$\mathbf{h}_5$	$\mathbf{h}_6$	$\bar{\mathbf{h}} = \mathbf{p}$
1	24.7	24.6	24.2	24.0	24.2	25.0	24.5
2	32.6	31.2	30.8	30.9	32.0	32.1	31.6
3	17.0	16.1	15.7	15.6	16.3	16.9	16.3
4	15.1	14.4	13.6	13.9	14.3	15.1	14.4
5	6.6	7.8	8.0	8.1	7.5	6.7	7.4
6+	4.0	5.9	7.8	7.5	5.7	4.2	5.8

NOTE: The average vector $\bar{\mathbf{h}}$ is the prior mean $\mathbf{p}$ for the 1985 household size percentages.

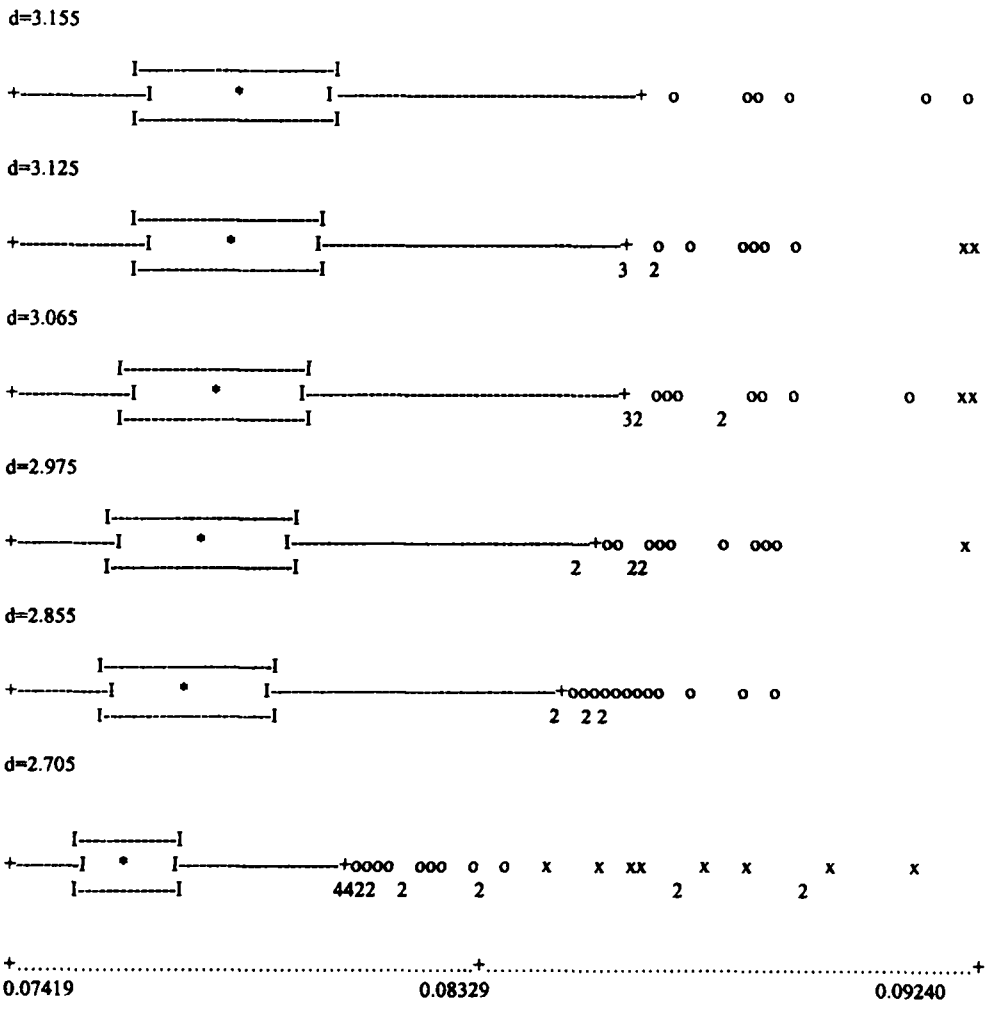

*Figure 3. Household Size: Boxplots of Prior Distributions of Smoothness Measure. *, Median location; I, hinge location; +, adjacent value; o, mild outlier; X, extreme outlier.*

is just a device used in the construction of a tractable hierarchical prior distribution expressing opinion for a locally smooth θ. In another, more general area of inference, $\mathbf{G} \cdot \boldsymbol{\alpha} = \sum \alpha_j \mathbf{g}_j$ might represent a "mixed-distribution" sampling model with component conditional sampling distributions $\mathbf{g}_j$ and unknown mixing proportions parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$. From this viewpoint, at each trial in the sample there would be an independent random selection, according to fixed weights α_j, of a component distribution to further draw from. For the case of known $\mathbf{G}$, interest would focus on estimation of the unknown sampling parameter, $\boldsymbol{\alpha}$. Titterington, Smith, and Markov (1985, p. 107) have proposed Dirichlet prior distributions for such $\boldsymbol{\alpha}$ in finite-mixture distribution problems.

The likelihood function $L_n(\boldsymbol{\theta}(\boldsymbol{\alpha}))$ entering into our posterior density (5) is a product of weighted averages. In the special case where the prior support of the unknown weights $\boldsymbol{\alpha}$ is restricted to the set of vertices of the probability simplex $(c \uparrow 1, \tilde{P}[\boldsymbol{\alpha} = \boldsymbol{\delta}(j)] = u_j, j = 1, \ldots, m)$, the prior belief is that a mixture-distribution sampling model would randomly select a model from a known set of alternate models $\mathbf{g}_j$, once only, and maintain the same such model $\mathbf{g}_j$ for every trial in the sample. The single randomly selected model would be hidden, so one would have in effect a "model-choice" or model-recognition inference problem. In this case the likelihood would degenerate into a product of factors from the coordinates of one vector $\mathbf{g}_j$, where which vector to use is not known. The reader should avoid an unfortunate tendency to confuse this extreme case with our adaptation of the fuller mixture distribution as a sample distribution.

The theory here can be generalized to a continuous sampling variable x (or i) and/or a continuous filtering variable j. A continuous sampling variable alone requires no additional theory. One merely tends to have a different row of $\mathbf{G}$ arise for each observed datum value, with each frequency count n_i equal to 1, and so there would be n_+ distinct factors in the likelihood and in the Carlson functions.

A continuous filtering variable, on the other hand, is more problematic. For example, the mixing distribution might be prior distributed according to a Dirichlet process. A realized outcome distribution of a prior Dirichlet random process is, with probability 1, a discrete distribution (Blackwell 1973; Ferguson 1973). The work here with filtered-variate Dirichlet distributions can be viewed as a finite-dimensional analog of Lo's (1984, 1987) prior process for Bayesian nonparametric inference. Lo used a Dirichlet process to mix over an infinite class of densities that are smooth to various

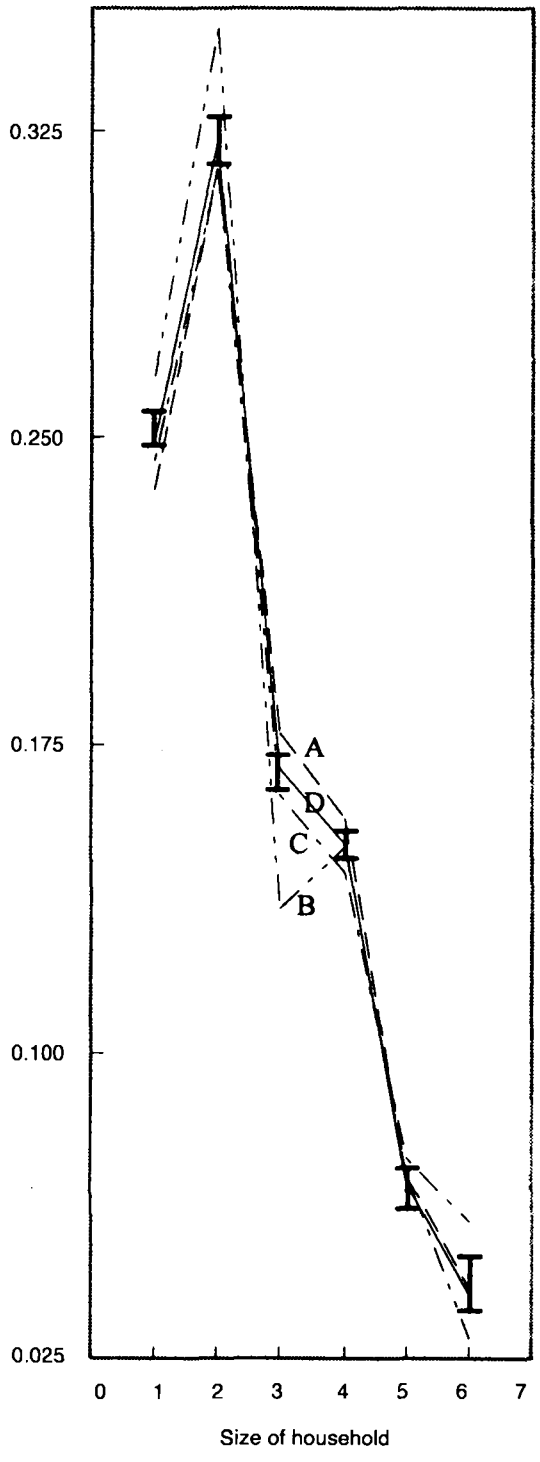

Figure 4. Household Size Distribution. A, Eventual "true" distribution for 1985 (– – –); B, sample relative frequency (– – – – –); C, least squares linear trend prediction (prior mean) (– – – –); D, posterior mean (and mean ± SD) (——).

extents (see also Antoniak 1974, the reply to the discussants in Diaconis and Freedman 1986, Escobar 1988, and Tiwari, Chib, and Jammalamadaka 1987). An entirely different approach that estimates a continuous sampling distribution as a filtered-variate distribution per se was explored by Dickey et al. (1989).

The approach here is motivated by the desire for precise and controlled expression of prior uncertainty about smoothness that can be combined coherently with whatever information on smoothness is inherent in the sample data. The resultant posterior distribution would then express the corresponding coherently updated uncertainty regarding the probability vector and its smoothness. Alternative approaches are less ambitious. One of the authors (JMD) recalls an insightful statement by John Tukey during a discussion in 1971 to the effect that the nonsmooth appearance of a histogram reported as an estimate could be useful to help warn against undue trust in a small sample. For a judgment of relative practicality, more extensive experience is needed with the filtered-variate Dirichlet and other filtered-variate MSV prior distributions and their assessment. The mechanics of their prior assessment, their updating by sample data and their posterior inferences are easy enough with present computing tools, and their potential advantages are strong enough, we hope, to tempt the reader to experiment with their assessment and use.

APPENDIX: PROOFS

Proof of Lemma 2

$$E(y_+) = \mu_+ \quad \text{and} \quad \text{var}(y_+) = c\mu_+(1-\mu_+).$$

So, if $\mu_+ = \delta$, then $\text{var}(y_+) = 0$ and with probability 1 $y_+ = E(y_+) = \mu_+ = \delta$. Conversely, if $y_+ = \delta$ with probability 1, then $E(y_+) = \delta$, so $\mu_+ = \delta$.

Proof of Theorem 2

Lemma. If the (scalar) random quantity z is supported on the unit interval, $0 \le z \le 1$, and $Ez = u$, then the variance of z is maximized for fixed u by the Bernoulli(u) distribution.

Proof. $\text{var}(z) = E(z^2) - u^2 \le E(z) - u^2 = u(1-u)$.

Apply the lemma to each coordinate z_j of $\mathbf{z}$, for which $\text{var}(z_j) = cu_j(1-u_j)$, which is maximized to the Bernoulli variance at $c = 1$. So $c \le 1$. $\tilde{P}$ is the only joint distribution on $\Delta^{(m)}$ having such margins.

Proof of Theorem 3

The proportionality of moments (15) follows from Corollary 6 and the proportionality, $E(\tilde{\boldsymbol{\theta}}_G) = E(\tilde{\boldsymbol{\theta}}_H)$, $\text{var}(\tilde{\boldsymbol{\theta}}_G) = d^2\text{var}(\tilde{\boldsymbol{\theta}}_H)$. To prove that CHull($\mathbf{G}$) contains the columns of $\mathbf{H}$, note that because $\bar{\mathbf{h}} = \bar{\mathbf{g}}$, the columns of $\mathbf{H}$ and $\mathbf{G}$ are related by $\mathbf{h}_j = d^{-1}\mathbf{g}_j + (1-d^{-1})\bar{\mathbf{g}}$, a convex combination when $d \ge 1$.

[Received September 1990. Revised November 1997.]

REFERENCES

Ahrens, L. A. (1965), "Observations on the Fe-Si-Mg Relationship in Chondrite," *Geochimica et Cosmochimica Acta.*, 29, 801–806.

Antoniak, C. E. (1974), "Mixtures of Dirichlet Processes With Applications to Bayesian Nonparametric Problems," *Annals of Statistics*, 2, 1152–1174.

Bar-Lev, S. K., and Enis, P. (1986), "Reproducibility and Natural Exponential Families With Power Variance Functions," *Annals of Statistics*, 14, 1507–1522.

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

Bloch, D. A., and Watson, G. S. (1967), "A Bayesian Study of the Multinomial Distribution," *Annals of Statistics*, 38, 1423–1434.

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

Cifarelli, D. M., and Regazzini, E. (1990), "Distribution Functions of Means of a Dirichlet Process," *Annals of Statistics*, 18, 429–442.

Diaconis, P., and Freedman, D. (1986), "On the Consistency of Bayes Estimates" (with discussion), *Annals of Statistics*, 14, 1–67.

Diaconis, P., and Kemperman, J. (1996), "Some New Tools for Dirichlet Priors," in *Bayesian Statistics 5*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, New York: Oxford University Press, pp. 97–106.

Dickey, J. M. (1968a), "Estimation of Disease Probabilities Conditioned on Symptom Variables," *Mathematical Biosciences*, 3, 249–265.

——— (1968b), "Smoothed Estimates for Multinomial Cell Probabilities," *Annals of Statistics*, 39, 561–566.

——— (1969), "Smoothing by Cheating," *Annals of Statistics*, 40, 1477–1482.

——— (1973), "Scientific Reporting and Personal Probabilities: Student's Hypothesis," *Journal of the Royal Statistical Society*, Ser. B, 35, 285–305.

——— (1978), "Discussion of 'Density Estimation, Stochastic Processes, and Prior Information'," by T. Leonard, *Journal of the Royal Statistical Society*, Ser. B, 40, 113–146.

——— (1983), "Multiple Hypergeometric Functions: Probabilistic Interpretations and Statistical Uses," *Journal of the American Statistical Association*, 78, 628–637.

Dickey, J. M., Garthwaite, P. H., and Bian, G. (1995), "An Elementary Continuous-Type Nonparametric Distribution Estimate," *International Journal of Mathematical and Statistical Sciences*, 4(2), 193–247.

Dickey, J. M., Jiang, J.-M., and Kadane, J. B. (1987), "Bayesian Methods for Censored Categorical Data," *Journal of the American Statistical Association*, 82, 773–781.

Escobar, M. D. (1988), "Estimating the Means of Several Normal Populations by Nonparametric Estimation of the Distribution of the Means," unpublished Ph.D. dissertation, Yale University, Dept. of Statistics.

Ferguson, T. S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *Annals of Statistics*, 1, 209–230.

Fienberg, S. E., and Holland, P. W. (1973), "Simultaneous Estimation of Multinomial Cell Probabilities," *Journal of the American Statistical Association*, 68, 683–689.

Freedman, D., Pisani, R., and Purves, R. (1978), *Statistics*, New York: Norton.

Good, I. J. (1950), *Probability and the Weighing of Evidence*, New York: Hafner.

——— (1965), *The Estimation of Probabilities*, Cambridge, MA: MIT Press.

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," *Journal of the American Statistical Association*, 75, 42–56.

Hill, B. M. (1969), "Foundations of the Theory of Least Squares," *Journal of the Royal Statistical Society*, Ser. B, 31, 89–97.

——— (1990), "A Theory of Bayesian Data Analysis," in *Bayesian and Likelihood Methods in Statistics and Econometrics*, eds. S. Geisser, J. S. Hodges, S. J. Press, and A. Zellner, Amsterdam: North-Holland.

Hoffman, M. S. (1987), *The World Almanac and Book of Facts 1987*, New York: Pharos.

Jaynes, E. T. (1979), "Concentration of Distributions at Maximum Entropy," in *E. T. Jaynes: Papers on Probability, Statistics, and Statistical Physics*, ed. R. D. Rosenkrantz, Boston: Reidel.

Jiang, J.-M. (1984), "Distributional Properties of Linear Forms in a Dirichlet Vector and Applications," Ph.D. dissertation and Research Report 14, State University of New York at Albany, Dept. of Mathematics and Statistics.

——— (1988), "Starlike Functions and Linear Functions of a Dirichlet Distributed Vector," *SIAM Journal of Mathematical Analysis*, 19, 390–397.

Jiang, T. J., Kadane, J. B., and Dickey, J. M. (1992), "Computation of Carlson's Multiple Hypergeometric Function R for Bayesian Applications," *Journal of Computational and Graphical Statistics*, 1, 231–251.

Lenk, P. J. (1988), "The Logistic Normal Distribution for Bayesian Nonparametric Predictive Densities," *Journal of the American Statistical Association*, 83, 509–516.

Leonard, T. (1973), "A Bayesian Method for Histograms," *Biometrika*, 60, 297–308.

——— (1978), "Density Estimation, Stochastic Processes, and Prior Information" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 40, 113–146.

Lo, A. Y. (1984), "On a Class of Bayesian Nonparametric Estimates: I, Density Estimates, *Annals of Statistics*, 12, 351–357.

——— (1987), "Bayes Methods for Mixture Models," Research Report 86-3, State University of New York at Buffalo, Dept. of Statistics.

Mendenhall, W., Reinmuth, J. E., Beaver, R., and Duhan, D. (1986), *Statistics for Management and Economics* (5th ed.), Boston: Duxbury.

Raiffa, H., and Schlaifer, R. (1961), *Applied Statistical Decision Theory*, Cambridge, MA: Division of Research, Harvard Business School.

Simonoff, J. S. (1983), "A Penalty Function Approach to Smoothing Large Sparse Contingency Tables," *Annals of Statistics*, 11, 208–218.

Stone, M. (1974), "Cross-Validation and Multinomial Prediction," *Biometrika*, 61, 509–515.

Tierney, L., and Kadane, J. B. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82–86.

Titterington, D. M. (1985), "Common Structure of Smoothing Techniques in Statistics," *International Statistical Review*, 53, 141–170.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.

Tiwari, R. C., Chib, S., and Jammalamadaka, S. R. (1987), "Nonparametric Bayes Prediction Density Estimation by Random Mixtures of Multivariate Distributions," paper presented at the 34th Meeting of the NBER-NSF Seminar for Bayesian Inference in Econometrics, Duke University, April 24–25, 1987.

U.S. Bureau of the Census (1975), *Historical Statistics in the United States. Part I*, Washington, D.C.: U.S. Government Printing Office.

——— (1987), *Statistical Abstract of the United States 1988* (108th ed.), Washington, D.C.: U.S. Government Printing Office.

Vardi, Y., Shepp, L. A., and Kaufman, L. (1985), "A Statistical Model for Positron Emission Tomography" (with discussion), *Journal of the American Statistical Association*, 80, 8–37.

Whittle, P. (1958), "On the Smoothing of Probability Density Functions," *Journal of the Royal Statistical Society*, Ser. B, 20, 334–343.