Bayesian Analysis (2013) **8**, Number 1, pp. 89–110

Asymptotics for Constrained Dirichlet Distributions

Charles Geyer* and Glen Meeden†

Abstract. We derive the asymptotic approximation for the posterior distribution when the data are multinomial and the prior is Dirichlet conditioned on satisfying a finite set of linear equality and inequality constraints so the posterior is also Dirichlet conditioned on satisfying these same constraints. When only equality constraints are imposed, the asymptotic approximation is normal. Otherwise it is normal conditioned on satisfying the inequality constraints. In both cases the posterior is a root-n-consistent estimator of the parameter vector of the multinomial distribution. As an application we consider the constrained Polya posterior which is a non-informative stepwise Bayes posterior for finite population sampling which incorporates prior information involving auxiliary variables. The constrained Polya posterior is a root-n-consistent estimator of the population distribution, hence yields a root-n-consistent estimator of the population mean or any other differentiable function of the vector of population probabilities.

Keywords: Dirichlet distribution, sample survey, constraints, Polya posterior, consistency, Bayesian inference

1 Introduction

Given the fact that prior information about quantities of interest can often involve constraints there is surprisingly little literature on the topic. Here we will consider the asymptotic behavior of constrained Dirichlet distributions with applications to finite population sampling. In the Bayesian approach to survey sampling, given a sample, inferences are based on a posterior distribution of the unobserved units in the population given the observed units in the sample. Today, this usually involves simulating complete copies of the population from one's posterior distribution. Simulating many complete copies of the population makes point and interval estimation of population parameters of interest straightforward. To ensure that these estimators are sensible and possess good frequentist properties, the posterior distribution should incorporate the kinds of prior information usually available. The Polya posterior is an objective posterior distribution, which is appropriate when one believes that the observed units are roughly exchangeable with the unobserved units. This assumption is often made when little is known a priori about the population and simple random sampling is the design. The constrained Polya posterior was introduced in Lazar, Meeden, and Nelson (2008) and is a generalization of the Polya posterior which incorporates prior information about population means and quantiles of auxiliary variables. The resulting posterior is a constrained Dirichlet distribution which must satisfy certain linear equality and inequality constraints. Here

*School of Statistics, University of Minnesota, Minneapolis, MN, charlie@stat.umn.edu
†School of Statistics, University of Minnesota, Minneapolis, MN, glen@stat.umn.edu

 DOI:10.1214/13-BA804

we will find the asymptotic form of this posterior and prove that it produces consistent estimators of population parameters.

Even if there are only equality constraints, the constrained Dirichlet posterior has no closed-form expressions — an equality-constrained Dirichlet is not Dirichlet — and cannot be sampled by ordinary Monte Carlo (it can, of course, be sampled by Markov chain Monte Carlo). In contrast, an equality-constrained normal distribution is another (degenerate) normal distribution (Cramér 1951, Section 24.3, Anderson 2003, Definition 2.4.1), and this makes tractable the asymptotic approximation for an equality-constrained Dirichlet. When inequality constraints are added, the asymptotic approximation is no longer normal but is easily sampled by ordinary Monte Carlo (simulate the equality-constrained normal distribution and reject simulations that don't satisfy the inequality constraints). These simulations can be used for calculations about the exact posterior via importance sampling (Section 5). In addition to these computational considerations, asymptotic approximation serves its usual role in providing theoretical understanding, for example, our root-n-consistency results (Corollaries 4.4 and 4.8).

2 The Polya Posterior

We begin by briefly reviewing the Polya posterior. Let s be the set of labels of a sample of size n from a population of size N. For convenience we assume the members of s are $1, 2, \ldots, n$ and we also suppose that n/N is very small. Let $y = (y_1, y_2, \ldots, y_N)$ be the characteristic of interest and y_s be the observed sample values.

The Polya posterior is based upon Polya sampling from an urn. It works as follows: suppose that the values from n observed or seen units are marked on n balls and placed in urn 1. The remaining unseen $N-n$ units of the population are represented by $N-n$ unmarked balls placed in urn 2. One ball from each urn is drawn with equal probability, and the ball from urn 2 is assigned the value of the ball from urn 1. Both balls are then returned to urn 1. Thus at the second stage of Polya sampling, urn 1 has $n+1$ balls and urn 2 has $N-n-1$ balls. This procedure is repeated until urn 2 is empty, at which point the N balls in urn 1 constitute one complete simulated copy of the population. Any finite population quantity — means, totals, regression coefficients — may now be calculated from the complete copy. By creating K complete copies in the same manner, the Polya posterior for the desired population quantity is generated. The mean of these simulated values is the point estimate and a 95% Bayesian credible interval is calculated from the 2.5% and 97.5% quantiles of the posterior distribution.

For the sample unit i let p_i denote the proportion of units in a full, simulated copy of the population which have the value y_i. One can show that under the Polya posterior $E(p_i) = 1/n$, and from this it follows that under the Polya posterior the posterior expectation of the population mean is the sample mean and the posterior variance is $(n-1)/(n+1)$ times the usual design-based variance of the sample mean under simple random sampling without replacement. The Polya posterior has a decision-theoretic justification based on its stepwise Bayes nature. Using this fact many standard estimators can be shown to be admissible. Details can be found in Ghosh and Meeden

(1997). The Polya posterior is the Bayesian bootstrap of Rubin (1981) applied to finite population sampling. Lo (1988) also discusses the Bayesian bootstrap in finite population sampling. Some early related work can be found in Hartley and Rao (1968) and Binder (1982).

It is of interest to compare the Polya posterior to the usual bootstrap methods in finite population sampling. Both approaches are based on an assumption of exchangeability. Gross (1980) introduced the basic idea for the bootstrap. Assume simple random sampling without replacement and suppose it is the case that $N/n = m$ is an integer. We then create a good guess for the population by combining m replicates of the sample. By taking repeated random samples of size n from this created population we can study the behavior of an estimator of interest. Booth, Butler, and Hall (1994) studied the asymptotic properties of such estimators. This is in contrast to the Polya posterior which considers the sample fixed and repeatedly generates complete versions of the population. This in turn generates a distribution for the population parameter of interest. Inferences for the population parameter are made using this predictive distribution.

3 The Constrained Polya Posterior

In many problems, in addition to the variable of interest, y, the sampler has in hand auxiliary variables for which prior information is available. A very common case is when the population mean of an auxiliary variable is known. In this situation either the ratio or the regression estimator is often used when estimating the population mean. Let w be an auxiliary variable and μ_w be its known population mean. It is possible to combine such information with the Polya posterior as follows. Given a sample, say $(y_1, w_1), \ldots, (y_n, w_n)$, and a simulated complete copy of the population generated from the Polya posterior one just checks to see if the simulated copy satisfies the mean constraint, i. e., the mean of the w values in the simulated population equals μ_w. That is we will just consider simulated completed copies of the population which satisfy the known mean constraint. In other words our posterior is just the Polya posterior restricted to this set. We call this restricted distribution the constrained Polya posterior.

In theory one could use rejection sampling to simulate from this distribution, but this is not practical. To get around this difficulty we proceed as follows. As before let $p = (p_1, \ldots, p_n)$ but now p_i is the proportion of units which are assigned the value (y_i, w_i) in a simulated complete copy of the population under the Polya posterior. When n/N is small and N is large it is well known that p has approximately a Dirichlet distribution with a parameter vector of all ones, i. e., it is uniform on the $(n-1)$-dimensional simplex, where $\sum_{j=1}^{n} p_j = 1$. Any linear constraint on the population value of an auxiliary variable translates in an obvious way to a linear constraint on the vector p involving the observed values of the auxiliary variable. For example, when the population mean of w is known then for the simulated population this translates to the constraint

$$\sum_{i=1}^{n} p_i w_i = \mu_w. \tag{1}$$

Lazar et al. (2008) discussed the constrained Polya posterior more generally and showed how it can incorporate various types of prior information involving auxiliary variables.

Let $\mathcal{P}$ denote the subset of the n-dimensional simplex that satisfies (1). Note that the lefthand side of this equation depends on the sampled values and so it is possible that even when μ_w is the true population value there is a positive probability under the sampling design that $\mathcal{P}$ will be empty. When $\mathcal{P}$ is non-empty, it is a non-full-dimensional polytope. In this case the approximate version of the Polya posterior which includes the prior information about the mean of w will just be the uniform distribution over $\mathcal{P}$. We call this distribution the constrained Polya posterior (CPP). It is not possible to generate independent observations from the CPP, but using Markov chain Monte Carlo methods one can generate dependent samples which will allow one to compute approximately point and interval estimates. This is easy to do in R (R Development Core Team 2011) by using the R package `polyapost` (Meeden and Lazar 2011).

Our goal in this paper is to study the asymptotic behavior of constrained Dirichlet distributions. In order to do asymptotics we assume that the population is classified into finitely many categories, which remain fixed as the sample size goes to infinity. Collapsing categories of a Dirichlet distribution gives another Dirichlet distribution. When each individual is distinguished, the Dirichlet approximation to the CPP has parameter vector having all components equal to one. When individuals are not distinguished, the Dirichlet approximation to the CPP has parameter vector having components equal to the numbers in the sample falling in each category. Thus we study Dirichlet(α_n) distributions where α_n goes to infinity in the sense described by (6a), (6b), and (6c) below or in the sense described by (15) below.

We also generalize the constraint (1) to allow population means of finitely many auxiliary variables to be known, either exactly or imprecisely, which gives rise to finitely many linear equality or inequality constraints on the Dirichlet distributed parameter vector. Thus we allow the constraint set to be an arbitrary non-full-dimensional convex polytope.

Making use of constraints on auxiliary variables is not the only way to exploit the available information in survey sampling. Calibration is another that has been widely discussed in the design approach since being introduced by Deville and Särndal (1992).

Since by definition a finite population has only finitely many elements, asymptotic results in survey sampling normally require additional machinery. Typically one assumes the existence of an infinite sequence of values leading to an infinite sequence of finite populations. For more details see Särndal, Swensson, and Wretman (1992) or Fuller (2009). However, as was shown in Hájek (1960), one gets the same asymptotic distribution assuming the sampling design is random sampling with replacement as one does under random sampling without replacement as long as one makes suitable assumptions about how the sample size and population size simultaneously go to infinity. Thus we assume that α_n goes to infinity having the same asymptotics (15) as if the data were multinomial. This does not assume sampling with replacement, only that one is doing asymptotics, like Hájek (1960), so as to get the same asymptotics as under sampling with replacement. Under this setup we will find the asymptotic form of the

CPP and show that estimators derived from it are consistent.

4 Asymptotics for the Dirichlet Distribution

4.1 Unconstrained Asymptotics

We begin our proof of consistency by reproving a well-known result (originally proved by Bienaymé in 1838, according to Gupta and Richards 2001), asymptotic normality of the Dirichlet distribution. We do this because we need the method of proof used here for constrained Dirichlet distributions.

Let Q_d denote the unit simplex in $\mathbb{R}^d$

$$Q_d = \{\, x \in \mathbb{R}^d : (\forall i)(x_i \geqslant 0) \text{ and } x_1 + \cdots + x_d = 1 \,\},$$

and for any set S let I_S denote its indicator function. The Dirichlet distribution of dimension d with parameter vector α has joint density

$$f_\alpha(x_1, \ldots, x_{d-1}) = \Gamma(\alpha_1 + \cdots + \alpha_d) \prod_{i=1}^{d} \frac{x_i^{\alpha_i - 1}}{\Gamma(\alpha_i)} I_{Q_d}(x), \tag{2}$$

where

$$x_d = 1 - x_1 - \cdots - x_{d-1}, \tag{3}$$

an abbreviation that will be used throughout this section.

The log unnormalized densities have the form

$$l_\alpha(x_1, \ldots, x_{d-1}) = \sum_{i=1}^{d} (\alpha_i - 1) \log(x_i), \tag{4}$$

where we adopt the convention that $\log(s) = -\infty$ whenever $s \leqslant 0$. This makes log an extended-real-valued strictly concave function (Rockafellar and Wets 2004, p. 1 and Section 2A), and makes (4) well-defined for all values of the variables. Since the composition of a strictly concave function and an affine function is strictly concave, and since a positive combination of strictly concave functions is strictly concave (Rockafellar and Wets 2004, Exercises 2.18 and 2.20), we conclude that (4) defines a strictly concave function whenever $\alpha_i > 1$ for all i. From this we conclude that, if there exists a point where the gradient of (4) is equal to zero, then this point is the unique mode of the distribution with density (2). It is easily checked that the point $\hat{x}(\alpha)$ having coordinates

$$\hat{x}(\alpha)_i = \frac{\alpha_i - 1}{\alpha_1 + \cdots + \alpha_d - d} \tag{5}$$

is such a point when $\alpha_i > 1$ for all i.

We now consider a sequence of parameter vectors α_n having components $\alpha_{n,i}$ satisfying

$$\alpha_{n,1} + \cdots + \alpha_{n,d} \to \infty \tag{6a}$$

and

$$\frac{\alpha_{n,i}}{\alpha_{n,1} + \cdots + \alpha_{n,d}} \to \lambda_i, \qquad 1 \leqslant i \leqslant d, \tag{6b}$$

where

$$\lambda_i > 0, \qquad i \leqslant 1 \leqslant d. \tag{6c}$$

For notational convenience we define

$$\nu_n = \alpha_{n,1} + \cdots + \alpha_{n,d}. \tag{7}$$

Then (6a) can be written more simply as $\nu_n \to \infty$, and (6b) can be written more simply as $\hat{x}(\alpha_n) \to \lambda$, where λ is the vector having components λ_i.

With an eye toward the eventual asymptotic result, we now define the variable transformation

$$z = \sqrt{\nu_n}\big(x - \hat{x}(\alpha_n)\big)$$

having inverse transformation

$$x = \hat{x}(\alpha_n) + \nu_n^{-1/2} z$$

and look at the log unnormalized density of the Dirichlet distribution written in terms of the new variables

$$\begin{aligned} r_n(z) &= l_{\alpha_n}\big(\hat{x}(\alpha_n) + \nu_n^{-1/2} z\big) - l_{\alpha_n}\big(\hat{x}(\alpha_n)\big) \\ &= \sum_{i=1}^{d} (\alpha_{n,i} - 1)\Big[\log\big(\hat{x}(\alpha_n)_i + \nu_n^{-1/2} z_i\big) - \log\big(\hat{x}(\alpha_n)_i\big)\Big] \end{aligned} \tag{8}$$

where now we have the abbreviation

$$z_d = -z_1 - \cdots - z_{d-1} \tag{9}$$

which operates the same way as the abbreviation (3).

Lemma 4.1. *With r_n defined by (8) and the sequence α_n satisfying the conditions (6a), (6b), and (6c)*

$$r_n(z) \to -\sum_{k=1}^{d} \frac{z_k^2}{2\lambda_k}, \qquad z \in \mathbb{R}^{d-1}, \tag{10}$$

where λ_k is defined in (6b). Moreover, this convergence is uniform on compact subsets of $\mathbb{R}^{d-1}$.

Proof. First derivatives of r_n are given by

$$\frac{\partial r_n(z)}{\partial z_k} = \nu_n^{-1/2} \left[\frac{\alpha_{n,k} - 1}{\hat{x}(\alpha_n)_k + \nu_n^{-1/2} z_k} - \frac{\alpha_{n,d} - 1}{\hat{x}(\alpha_n)_d + \nu_n^{-1/2} z_d} \right].$$

Second derivatives are given by

$$\frac{\partial^2 r_n(z)}{\partial z_k^2} = -\nu_n^{-1}\left[\frac{\alpha_{n,k}-1}{\big(\hat{x}(\alpha_n)_k + \nu_n^{-1/2} z_k\big)^2} + \frac{\alpha_{n,d}-1}{\big(\hat{x}(\alpha_n)_d + \nu_n^{-1/2} z_d\big)^2}\right]$$
$$\frac{\partial^2 r_n(z)}{\partial z_j \partial z_k} = -\nu_n^{-1}\frac{\alpha_{n,d}-1}{\big(\hat{x}(\alpha_n)_d + \nu_n^{-1/2} z_d\big)^2}, \qquad j \neq k.$$

We have $r_n(0) = 0$ and $\nabla r_n(0) = 0$. The integral form of the remainder gives the Maclaurin series

$$r_n(z) = \int_0^1 z^T \nabla^2 r_n(sz) z (1-s)\, ds,$$

where

$$\begin{aligned} z^T \nabla^2 r_n(sz) z &= -\sum_{k=1}^{d-1} \frac{\nu_n^{-1}(\alpha_{n,k}-1) z_k^2}{\big(\hat{x}(\alpha_n)_k + s\nu_n^{-1/2} z_k\big)^2} \\ &\quad - \frac{\nu_n^{-1}(\alpha_{n,d}-1)}{\big(\hat{x}(\alpha_n)_d + s\nu_n^{-1/2} z_d\big)^2}\left(\sum_{j=1}^{d-1} z_j\right)^2 \\ &= -\sum_{k=1}^{d} \frac{\nu_n^{-1}(\alpha_{n,k}-1) z_k^2}{\big(\hat{x}(\alpha_n)_k + s\nu_n^{-1/2} z_k\big)^2}. \end{aligned}$$

Hence

$$\begin{aligned} r_n(z) &= -\sum_{k=1}^{d} \int_0^1 \frac{\nu_n^{-1}(\alpha_{n,k}-1) z_k^2}{\big(\hat{x}(\alpha_n)_k + s\nu_n^{-1/2} z_k\big)^2}(1-s)\, ds \\ &\to -\sum_{k=1}^{d} \frac{z_k^2}{\lambda_k}\int_0^1 (1-s)\, ds \\ &= -\sum_{k=1}^{d} \frac{z_k^2}{2\lambda_k}, \end{aligned}$$

the limit here being dominated convergence, the integrand being strictly positive and dominated by

$$\frac{\nu_n^{-1}(\alpha_{n,k}-1) z_k^2}{\hat{x}(\alpha_n)_k^2}(1-s),$$

the fraction here being a sequence converging to z_k^2/λ_k and hence being dominated by $z_k^2/\lambda_k + \varepsilon$ for any $\varepsilon > 0$ and sufficiently large n.

That pointwise convergence of concave functions implies uniform convergence on compact sets is Theorem 7.17 in Rockafellar and Wets (2004). □

Theorem 4.2. *Let X_n be a d-dimensional random vector having the Dirichlet distribution with parameter vector α_n. Suppose the sequence α_n satisfies the conditions* (6a),

(6b), *and* (6c). *Then the densities of the random vectors* Z_n *having components*

$$Z_{n,i} = \sqrt{\nu_n}\left(X_{n,i} - \frac{\alpha_{n,i} - 1}{\nu_n - d}\right) \tag{11}$$

where $X_{n,i}$ *are the components of* X_n *and* ν_n *is given by* (7) *converge to a multivariate normal density*

$$\frac{e^{r_n(z)}}{\int e^{r_n(z)}\,dz} \to c \exp\left(-\sum_{k=1}^{d} \frac{z_k^2}{2\lambda_k}\right), \tag{12}$$

where r_n *is given by* (8), c *is chosen to make the right-hand side integrate to one, and finiteness of the integral on the left-hand side is part of the assertion. Moreover, the distribution of* Z_n *converges in total variation to this normal distribution.*

Although the right-hand side of (12) looks like a d-dimensional distribution with independent components, it is not. It is a $(d-1)$-dimensional distribution with correlated components because of (9).

Proof. Let B denote the boundary of the unit ball in $\mathbb{R}^{d-1}$. This is a compact set; hence there exists an $N \in \mathbb{N}$ such that

$$r_n(z) \leqslant -\sum_{k=1}^{d} \frac{z_k^2}{4\lambda_k}, \qquad n \geqslant N \text{ and } z \in B.$$

Define

$$\lambda_{\max} = \max_{1 \leqslant k \leqslant d-1} \lambda_k.$$

Then

$$r_n(z) \leqslant -\frac{1}{4\lambda_{\max}}, \qquad n \geqslant N \text{ and } z \in B.$$

By the concavity inequality, for $s > 1$ we have

$$r_n(z) \geqslant \left(1 - \frac{1}{s}\right) r_n(0) + \frac{1}{s} \cdot r_n(sz),$$

or, since $r_n(0) = 0$,

$$r_n(sz) \leqslant s r_n(z) \leqslant -\frac{s}{4\lambda_{\max}}, \qquad s > 1 \text{ and } n \geqslant N \text{ and } z \in B. \tag{13}$$

Let S_n and V_n denote the surface area and volume of the unit sphere in $\mathbb{R}^n$, let D denote the exterior of the unit ball, and define the function $h : \mathbb{R}^{d-1} \to \mathbb{R}$ by

$$h(z) = -\frac{I_D(z)\|z\|_2}{4\lambda_{\max}},$$

where $\|\cdot\|_2$ denotes the Euclidean norm. Then $e^{r_n} \leqslant e^h$ for all n by (13), and

$$\int e^{h(z)}\,dz = V_{d-1} + S_{d-1} \int_1^\infty \exp\left(-\frac{t}{4\lambda_{\max}}\right) t^{d-2}\,dt, \tag{14}$$

and this integral, being an incomplete gamma integral, is clearly finite. This proves the integrability assertion.

Let $q(z)$ denote the right-hand side of (10) so Lemma 4.1 asserts $e^{r_n} \to e^q$ pointwise. Then

$$\int e^{r_n(z)}\, dz \to \int e^{q(z)}\, dz$$

by dominated convergence, e^h being a dominating function, and this shows that the right-hand side of (12) must integrate to one.

That pointwise convergence of densities implies convergence in total variation of distributions is Scheffé's lemma (Scheffé 1947). □

And what is this asymptotic normal distribution?

Theorem 4.3. *The normal distribution having density on the right-hand side of* (12) *is, considered as a d-dimensional distribution,* $\text{Normal}(0, \Lambda - \lambda\lambda^T)$*, where* Λ *is the diagonal matrix having* λ *as its vector of diagonal elements.*

Proof. First note that the normal distribution described by the theorem is degenerate (Cramér 1951, Section 24.3, Anderson 2003, Definition 2.4.1). If u is the vector having all components equal to one, then $(\Lambda - \lambda\lambda^T)u = 0$. Hence if Z is a random vector having this distribution $u^T Z = 0$ with probability one.

Thus the distribution described by both Theorems 4.2 and 4.3 is actually $(d-1)$-dimensional, and we use the abbreviation (9) to make it so. The exponent in (12) is $-\frac{1}{2} z^T A z$, where A is the matrix having components

$$\begin{aligned} a_{ii} &= \frac{1}{\lambda_i} + \frac{1}{\lambda_d} \\ a_{ij} &= \frac{1}{\lambda_d}, \qquad i \neq j. \end{aligned}$$

The variance matrix in Theorem (4.3) is B having components

$$\begin{aligned} b_{ii} &= \lambda_i - \lambda_i^2 \\ b_{ij} &= -\lambda_i \lambda_j, \qquad i \neq j. \end{aligned}$$

To check that both theorems describe the same distribution, we must check that A and B are inverse matrices. Letting δ_{ij} be the Kronecker delta (the components of the

identity matrix),

$$\begin{aligned}
(AB)_{ik} &= \sum_{j=1}^{d-1} a_{ij} b_{jk} \\
&= \sum_{j=1}^{d-1} \left(\frac{\delta_{ij}}{\lambda_i} + \frac{1}{\lambda_d} \right) (\lambda_j \delta_{jk} - \lambda_j \lambda_k) \\
&= \sum_{j=1}^{d-1} \left(\frac{\delta_{ij} \lambda_j \delta_{jk}}{\lambda_i} + \frac{\lambda_j \delta_{jk}}{\lambda_d} - \frac{\delta_{ij} \lambda_j \lambda_k}{\lambda_i} - \frac{\lambda_j \lambda_k}{\lambda_d} \right) \\
&= \delta_{ik} + \frac{\lambda_k}{\lambda_d} - \lambda_k - \frac{(1 - \lambda_d) \lambda_k}{\lambda_d} \\
&= \delta_{ik},
\end{aligned}$$

so it does check. □

4.2 Random Sampling

The result in the preceding section describes the asymptotic behavior of the posterior but is not enough by itself to discuss consistency. Theorems 4.2 and 4.3 describe the spread of the posterior distribution around its mode, but we also need to consider how far that mode is from the true unknown parameter value. We shall see that both of these are $O_p(n^{-1/2})$ so the sum is also $O_p(n^{-1/2})$.

In applications of interest to us α_n is random and the Dirichlet distribution is the posterior distribution. The Dirichlet distribution is conjugate to the multinomial distribution, so this occurs when the data are multinomial and a conjugate prior is used. If a Dirichlet(ξ) prior distribution is adopted and the data distribution is Multinomial(n, λ), then the posterior distribution is Dirichlet(α_n), where $\alpha_n = \xi + y_n$, where y_n is the multinomial data vector. The central limit theorem (CLT) says

$$\sqrt{n} \left(\frac{Y_n}{n} - \lambda \right) \xrightarrow{\mathcal{D}} \text{Normal}(0, \Lambda - \lambda \lambda^T),$$

where Λ is the diagonal matrix having λ as its vector of diagonal elements. From this

$$\sqrt{n} \left(\frac{\alpha_n}{n} - \lambda \right) \xrightarrow{\mathcal{D}} \text{Normal}(0, \Lambda - \lambda \lambda^T) \tag{15}$$

follows by Slutsky's theorem. The hyperparameter ξ of the prior distribution plays no role in the asymptotics. We can even use improper priors determined by hyperparameters ξ having nonpositive components, although then we only have proper posteriors when $y_{n,i} > \max(0, -\xi_i)$ for all i, which happens with probability converging to one as n goes to infinity but fails to happen with positive probability for all n.

Since we have $\nu_n / n \to 1$ almost surely, where ν_n is given by (7), we may, as in (15), use n where we had ν_n in preceding sections.

As discussed at the end of Section 3 above, (15) does not actually assume sampling with replacement or independent and identically distributed data, only that one gets the same asymptotics for α_n as under these assumptions, as does Hájek (1960).

Convergence-in-distribution asymptotics do not (in general) handle conditional distributions well, but we want to consider the constrained Dirichlet distribution given α_n, which is now considered random. For this it helps to consider a Skorohod representation. By the Skorohod theorem (Billingsley 1999, Theorem 6.7), there exist random vectors α_n^* and Z^* defined on the same probability space such that α_n^* has the same distribution as α_n and Z^* has the distribution on the right-hand-side of (15) and

$$\sqrt{n}\left(\frac{\alpha_n^*}{n} - \lambda\right) \to Z^*, \qquad \text{almost surely.} \tag{16}$$

This device allows us to state the more interesting asymptotics of the deviation of X_n from the true unknown parameter value λ. The combination of Theorems 4.2 and 4.3 gives

$$\sqrt{n}\left(X_n - \frac{\alpha_n^* - 1}{n - d}\right) \xrightarrow{\mathcal{D}} \text{Normal}(0, \Lambda - \lambda\lambda^T),$$

and this combined with Slutsky's theorem and (16) gives

$$\sqrt{n}(X_n - \lambda) \xrightarrow{\mathcal{D}} \text{Normal}(Z^*, \Lambda - \lambda\lambda^T). \tag{17}$$

We may consider this either a conditional or unconditional result. It is true conditionally, taking the left-hand side to refer to the conditional distribution of X_n given α_n^* and the distribution on the right-hand side to refer to the conditional distribution given Z^*. It is also true unconditionally, taking α_n^* and Z^* to be random vectors satisfying (16) (Appendix), and this implies the following.

Corollary 4.4. *The unconstrained posterior distribution of the parameter vector, which is* Dirichlet(α_n)*, is root-n-consistent, that is,*

$$X_n = \lambda + O_p(n^{-1/2}). \tag{18}$$

4.3 Constrained Asymptotics

Linear Equality Constraints

Now we consider the case where we know $B\lambda = a$ for some known matrix B and some known vector a and impose these constraints as prior information that is reflected in the posterior. So now X_n has the Dirichlet(α_n^*) distribution conditioned on the event $BX_n = a$. Since this is an event of measure zero, we mean it in the non-measure-theoretic sense of conditional distributions: we find a conditional density using restriction of the density to the constraint set and renormalization.

Since we can also write the constraint as $B(X_n - \lambda) = 0$, the constraint constrains the asymptotic normal distribution to lie in the vector subspace

$$V = \{\, w \in \mathbb{R}^d : Bw = 0 \,\}. \tag{19}$$

For convenience, we take the matrix B to incorporate the constraint $u^T z = 0$ that we have even in the "unconstrained" case, that is, we insist that some linear combination of rows of B is equal to u^T.

Theorem 4.5. *Let X_n have the* Dirichlet(α_n^*) *distribution conditioned on $BX_n = a$, and assume (16) holds. Then*

$$\sqrt{n}(X_n - \lambda) \xrightarrow{\mathcal{D}} W, \tag{20}$$

where W is a random vector having density

$$f(w) = c \exp\left(-\frac{1}{2}(w - z^*)^T \Lambda^{-1} (w - z^*) \right) \tag{21}$$

with respect to Lebesgue measure on the vector subspace (19).

It is not completely clear what we mean by Lebesgue measure on V. Of course, V is linearly isomorphic to $\mathbb{R}^k$ for some k, but there is a Jacobian term in the change-of-measure formula mapping Lebesgue measure from $\mathbb{R}^k$ to V. The Jacobian of a linear mapping is a constant function, however, so this defines Lebesgue measure on V up to an arbitrary constant. Since (21) contains a constant c that must be determined by normalization, changing the constant in Lebesgue measure just changes the constant c in (21) without changing the fact that (21) integrates to one.

Proof. In the notation of Theorem 4.2 the random vector

$$U_n = \sqrt{n} \left(X_n - \frac{\alpha_n^* - 1}{\nu_n^* - d} \right)$$

has log unnormalized density r_n, where ν_n^* is the sum of the components of α_n^*, and this converges uniformly on compact sets to the function

$$q(u) = -\frac{1}{2} u^T \Lambda^{-1} u$$

on V (and in fact on a larger subspace). That

$$\int_V e^{r_n(u)} \, du \to \int_V e^{q(u)} \, du$$

is proved similarly to the proof in Theorem 2. This implies the density of U_n converges uniformly on compact sets to the multivariate normal density proportional to e^q.

The random vector

$$Y_n = U_n + \sqrt{n} \left(\frac{\alpha_n^* - 1}{\nu_n^* - d} - \lambda \right),$$

which is the left-hand side of (20), has density

$$f_{U_n} \left[y - \sqrt{n} \left(\frac{\alpha_n^* - 1}{\nu_n^* - d} - \lambda \right) \right],$$

where $f_{U_n} = e^{r_n} / \int_V e^{r_n}$, and this converges uniformly on compact sets to $y \mapsto ce^{q(y-z^*)}$, where c is chosen to make this integrate to one over V. □

Remark 4.6. In a real application, n does not go to infinity. We have a Dirichlet(α_n) random vector X_n conditioned on $BX_n = a$. We do not know λ (it is the unknown quantity we are estimating with our constrained Dirichlet posterior), hence we do not know Λ.

We approximate the distribution of X_n by the normal distribution having density

$$f(x) = c \exp\left(-\frac{n}{2}(x - \hat{\lambda}_n)^T \widehat{\Lambda}_n^{-1}(x - \hat{\lambda}_n)\right), \tag{22}$$

with respect to Lebesgue measure on the affine subspace

$$C = \{\, x : Bx = a \,\}, \tag{23}$$

where

$$\hat{\lambda}_n = \alpha_n / \nu_n, \tag{24}$$

where ν_n is given by (7), and $\widehat{\Lambda}_n$ is the diagonal matrix having $\hat{\lambda}_n$ as its vector of diagonal elements.

The point of introducing z^* and (16) is an artifact of the conventional way of discussing asymptotics. The practical point is that $\hat{\lambda}_n$ is not λ and does not necessarily satisfy the constraints, that is, $\hat{\lambda}_n \in C$ may be false. Hence, despite appearances, $\hat{\lambda}_n$ may not be the mean of the normal distribution having density (22).

In order to use this normal approximation in practice, we need to know more about it. To do that, we change our characterization of the constraint (23). Let M be a matrix whose columns are a basis for V, so every $x \in C$ has the form $x = \lambda_0 + M\beta$ for some $\beta \in \mathbb{R}^k$, where λ_0 is any point in C and k is the dimension of V. The mean of the asymptotic normal distribution is the same as the mode, which is the minimizer of

$$\beta \mapsto (\lambda_0 + M\beta - \hat{\lambda}_n)^T \widehat{\Lambda}_n^{-1}(\lambda_0 + M\beta - \hat{\lambda}_n)$$

which is recognizable as a weighted least squares problem having solution

$$\beta_n^* = \left(M^T \widehat{\Lambda}_n^{-1} M\right)^{-1} M^T \widehat{\Lambda}_n^{-1}(\hat{\lambda}_n - \lambda_0) \tag{25}$$

(Weisberg 2005, Section 5.1). Thus we can rewrite the asymptotic normal distribution as

$$f(\beta) = c \exp\left(-\frac{n}{2}(\beta - \beta_n^*)^T M^T \widehat{\Lambda}_n^{-1} M(\beta - \beta_n^*)\right),$$

from which we see that β is multivariate normal of dimension k having mean vector β_n^* and variance matrix $n^{-1}(M^T \widehat{\Lambda}_n^{-1} M)^{-1}$. Thus X_n is approximately (degenerate, Cramér 1951, Section 24.3, Anderson 2003, Definition 2.4.1) multivariate normal of dimension d with mean vector $\lambda_0 + M\beta_n^*$ and (singular) variance matrix $n^{-1} M(M^T \widehat{\Lambda}_n^{-1} M)^{-1} M^T$.

Polyhedral Convex Sets and Tangent Cones

Now suppose C is a polyhedral convex set in $\mathbb{R}^d$, that is, the solution set of a finite family of linear equality and inequality constraints (Rockafellar and Wets 2004, Example 2.10). Then it can be written

$$C = \{\, x \in \mathbb{R}^d : \langle b_i, x\rangle \leqslant a_i,\ i \in J \text{ and } \langle b_i, x\rangle = a_i,\ i \in E \,\},$$

where E and J are disjoint finite sets, each b_i is a nonzero vector, each a_i is a real number, and $\langle\, \cdot\,, \cdot\,\rangle$ denotes the usual inner product. It may be that some of what are nominally inequality constraints actually hold with equality. Define

$$E^* = \{\, i \in J \cup E : \langle b_i, x\rangle = a_i,\ \forall x \in C \,\}$$

and $J^* = J \backslash E^*$. Then we can also write

$$C = \{\, x \in \mathbb{R}^d : \langle b_i, x\rangle \leqslant a_i,\ i \in J^* \text{ and } \langle b_i, x\rangle = a_i,\ i \in E^* \,\}, \tag{26}$$

knowing that every constraint with index $i \in J^*$ is an actual inequality constraint.

The *tangent cone* of (26) at a point $x \in C$ is given by

$$T_C(x) = \{\, y \in \mathbb{R}^d : \langle b_i, y\rangle \leqslant 0,\ i \in A(x) \text{ and } \langle b_i, y\rangle = 0,\ i \in E^* \,\}, \tag{27}$$

where

$$A(x) = \{\, i \in J^* : \langle b_i, x\rangle = a_i \,\}$$

is called the *active set* (Rockafellar and Wets 2004, Theorem 6.46). The vector subspace

$$V = \{\, x \in \mathbb{R}^d : \langle b_i, x\rangle = 0,\ i \in E^* \,\} \tag{28}$$

plays the same role in inequality constrained problems as (19) did in equality constrained problems; V is the affine hull of $T_C(x)$.

Linear Equality and Inequality Constraints

Now we have a theorem very similar to Theorem 4.5 except we add inequality constraints and the asymptotic constraint set turns out to be the tangent cone. For convenience, we assume the constraint $u^T X_n = 1$ is included among the equality constraints determining C, that is, $\langle u, x\rangle = 1$ for all $x \in C$.

Theorem 4.7. *Let X_n have the* Dirichlet(α_n^*) *distribution conditioned on $X_n \in C$, where C is given by (26), and assume (16) holds with $\lambda \in C$. Then (20) holds, where W is a random vector having density (21) with respect to Lebesgue measure on the tangent cone $T_C(\lambda)$.*

In the comments following Theorem 4.5 we explained what we mean by Lebesgue measure on a subspace, and our representation (27) shows that we can always determine a subspace V such that $T_C(\lambda)$ has nonempty interior relative to V, hence positive Lebesgue measure relative to V (every nonempty convex set has nonempty interior relative to its affine hull, Rockafellar and Wets 2004, Theorem 2.40). Thus if we know how to condition on V, then we also know how to condition on $T_C(\lambda)$.

Proof. The proof is almost the same as the proof of Theorem 4.5 (if there were no inequality constraints, it would be the same). Let Y_n denote the left-hand side of (20) conditioned on V given by (28). Then Theorem 4.5 says that the density f_{Y_n} converges uniformly on compact sets to the density (21), both restricted to V. We only need to show the effect of the inequality constraints.

Since $\lambda \in C$, we have $x \in C$ if and only if $y = \sqrt{n}(x - \lambda)$ lies in V given by (28) and satisfies

$$\langle b_i, \lambda + n^{-1/2} y \rangle \leqslant a_i, \qquad i \in J^*. \tag{29}$$

For $i \in J^* \backslash A(\lambda)$, we have $\langle b_i, \lambda \rangle < a_i$, and hence for such i, the inequality in (29) becomes

$$\langle b_i, y \rangle \leqslant n^{1/2} [a_i - \langle b_i, \lambda \rangle],$$

and such constraints have no effect asymptotically because the right-hand side goes to infinity as $n \to \infty$. Thus we are left with the constraints

$$\langle b_i, \lambda + n^{-1/2} y \rangle \leqslant a_i, \qquad i \in A(\lambda); \tag{30}$$

the constraint that y lies in V and satisfies (30) is the same as constraining $y \in T_C(\lambda)$. This shows that if we define

$$D_n = \{\, \sqrt{n}(x - \lambda) : x \in C \,\}$$

then $I_{D_n} \to I_{T_C(\lambda)}$ pointwise. Hence $f_{Y_n} I_{D_n} \to f I_{T_C(\lambda)}$ pointwise, where f is given by (21). A now familiar argument (like those in Theorems 4.2 and 4.5) says that these unnormalized densities also converge pointwise when normalized. Hence we have convergence in total variation by Scheffé's lemma. □

Corollary 4.8. *In the setup of Theorem 4.7, X_n is a root-n-consistent estimator of λ, that is, (18) holds.*

Remark 4.9. In a real application, n does not go to infinity. We have a Dirichlet(α_n) random vector X_n conditioned on the event $X_n \in C$. We do not know λ (it is the unknown quantity we are estimating with our constrained Dirichlet posterior), hence we do not know Λ, nor do we know the tangent cone $T_C(\lambda)$.

We do have the estimate $\hat{\lambda}_n$ given by (24) and the corresponding diagonal matrix $\widehat{\Lambda}_n$ having $\hat{\lambda}_n$ as its vector of diagonal elements.

We approximate the distribution of X_n by taking the normal distribution having mean vector $\lambda_0 + M\beta_n^*$ and variance matrix $n^{-1} M (M^T \widehat{\Lambda}_n^{-1} M)^{-1} M^T$, where λ_0 is any point in

$$\operatorname{aff} C = \{\, x \in \mathbb{R}^d : \langle b_i, x \rangle = a_i,\ i \in E^* \,\},$$

where M is a matrix whose columns are a basis for V given by (28), and where β_n^* is given by (25), and further conditioning this normal distribution to lie in C.

Since the normal distribution just described is concentrated on aff C, we only need to apply the inequality constraints to condition it to lie in C. We need to apply all the

constraints, since we do not know λ and hence have no notion of which constraints are active at λ. While we are at it, we might as well apply the original equality constraints, constraining all components of X_n to be nonnegative. This assures that our asymptotic approximation makes sense in practice.

Remark 4.10. In applications, calculating the vector subspace V given by (19) or (28) and the set of non-redundant inequality constraints, those involved in the representation (27), can be very difficult to do by hand, but functions in the R package `rcdd` (Geyer, Meeden, and Fukuda 2011) do this easily. The `redundant` function determines a minimal set of equality constraints determining aff C and a minimal set of inequality constraints that need to be added to these to determine C. The `scdd` function, given the minimal set of equality constraints determining aff C, produces a point $\lambda_0 \in \operatorname{aff} C$ and a set of basis vectors for V, that can be used as the columns of the matrix M used in Remarks 4.6 and 4.9. All of this can be done using infinite-precision rational arithmetic so the results are exact.

Remark 4.11. A frequentist analysis of the same multinomial data as used by the Bayesian agrees asymptotically with the Bayesian analysis as long as there are only linear equality constraints. Under the "usual" regularity conditions for Bayesian and frequentist asymptotics, Bayesians and frequentists disagree about whether the parameter p or the maximum likelihood estimate $\hat{p}_n$ is random, but they agree that $\hat{p}_n - p$ is asymptotically normal with mean vector zero and variance matrix equal to the inverse of the Fisher information matrix (and our results in this paper agree). Consequently asymptotic Bayesian credible regions and frequentist confidence regions will also agree.

When there are inequality constraints, the Bayesian and frequentist inferences become radically different, and the Bayesian procedure is much simpler. The Bayesian simply produces a highest posterior density region using the intersection of the constraint set with an elliptical contour of the density of the asymptotic normal distribution for equality constraints (described in Remark 4.6). The contour that gives the desired posterior probability cannot be determined by a chi-square critical value when there are inequality constraints but can easily be determined by simulation. The frequentist wants to use the asymptotic distribution of the maximum likelihood estimate (MLE), which is known (LeCam 1970; Self and Liang 1987; Geyer 1994) but has the problem that that asymptotic distribution depends on the tangent cone $T_C(\lambda)$ and λ is unknown. Simply plugging in the MLE $\tilde{\lambda}_n$, that is, using $T_C(\tilde{\lambda}_n)$ does not work because the tangent cone depends on the point discontinuously. Thus constructing valid frequentist confidence regions is, to our knowledge, an open research question.

5 Example

A technical report (Geyer and Meeden 2012) gives full details of a worked example. It is produced by the R command `Sweave` so all calculations are actually done by the code shown therein and are reproducible by anyone who has R. Rather than simulate many variations of a problem, we have written the code so changing two statements defining the dimension d and sample size n of the problem is all that is needed to do a different

example. Interested readers can do their own experiments.

In the example of the technical report the dimension is $d = 10$ and the sample size is $n = 1000$, that is, we have (simulated) data on n individuals who are classified in d categories. We simulated a d-dimensional multinomial data vector y. The posterior distribution of the vector p of category probabilities is Dirichlet with hyperparameter vector y because we used an improper Dirichlet(ξ) prior with hyperparameter vector $\xi = 0$ in the notation of Section 4.2.

To formulate constraints, suppose there is a random variable X taking values $1, \ldots, d$ whose distribution conditional on the random vector p having the posterior distribution is

$$\Pr(X = i) = p_i, \qquad i = 1, \ldots, d.$$

We put constraints on the mean, median, and variance of X, and these correspond to linear constraints on the vector p. First, we assume the mean of X is the midpoint of the range, that is,

$$E(X \mid p) = \sum_{i=1}^{d} i p_i = \mu, \tag{31a}$$

where $\mu = (d + 1)/2$. Second, we assume the median of X is between $\mu - 2$ and $\mu + 2$, that is

$$E(X < \mu - 2 \mid p) = \sum_{i=1}^{\lceil \mu - 3 \rceil} p_i \leqslant \frac{1}{2} \tag{31b}$$

$$E(X > \mu + 2 \mid p) = \sum_{i=\lfloor \mu + 3 \rfloor}^{d} p_i \leqslant \frac{1}{2}. \tag{31c}$$

Third, we assume the variance of X is between some numbers a and b, that is

$$\operatorname{var}(X \mid p) = \sum_{i=1}^{d} (i - \mu)^2 p_i \geqslant a \tag{31d}$$

$$\operatorname{var}(X \mid p) = \sum_{i=1}^{d} (i - \mu)^2 p_i \leqslant b \tag{31e}$$

but do not know how to choose a and b sensibly for this example so we find the set of values of $\operatorname{var}(X \mid p)$ as p ranges over the set of all possible probability vectors that satisfy (31a), (31b), and (31c) and then take a to be the 25th percentile and b to the the 75th percentile of this set of values. In the example worked out in the technical report, with $d = 10$, this procedure gives $a = 19/2$ and $b = 23/2$.

In addition to the equality constraint (31a) we also have the equality constraint that the components of p sum to one. Thus the constrained posterior distribution for p actually has dimension $d - 2$. In addition to the inequality constraints (31b), (31c), (31d), and (31e), we also have the d inequality constraints that each component of p is nonnegative. Thus there are 2 equality constraints and $d + 4$ inequality constraints.

This results in a fairly complex constraint set. In the $d = 10$ case, it is a convex polytope with 46 vertices and 13 facets.

We then chose a "simulation truth" value of p that satisfies (31c) and (31e) with equality (in addition to the equality constraints) and simulated a multinomial data vector having this "simulation truth" probability vector p as its parameter vector and n as its sample size.

All of the work of the example up to this point does not mimic real data analysis. In real life, the data y would be obtained from a survey, and the constraints on p would be obtained from prior information (perhaps other survey or census data).

The technical report shows how to simulate the asymptotic constrained normal approximation to the constrained Dirichlet posterior distribution, using the R package `rcdd` to find the matrix M and the vector λ_0 described in Remark 4.6, using the command `mvrnorm` in the R package `MASS` to simulate multivariate normal random vectors having the mean vector and variance matrix described in Remark 4.6, and using rejection sampling to impose the inequality constraints. In the $d = 10$ and $n = 1000$ case presented in the technical report 53.9% of the unconstrained normal simulations satisfied the inequality constraints and were accepted in the Monte Carlo sample of the constrained posterior.

Since there is no good methodology known to us for comparing high-dimensional multivariate distributions, in this case the asymptotic constrained normal approximation to the posterior versus the exact posterior, we hit upon the idea of using the former as an importance sampling distribution for the latter and using the size of the normalized importance weights as a criterion. The normalized importance weights are not highly variable in the case presented in the technical report, the maximum being 43 times the average. This shows the normal approximation is not perfect (which would be all importance weights equal to the average), but it is quite good enough for importance sampling to work well. Thus we realize that if we are actually going to do a Monte Carlo calculation based on the constrained normal approximation to the posterior, then we might as well also calculate the normalized importance weights and do the same calculation based on the exact constrained Dirichlet distribution via importance sampling. (Calculating the normalized importance weights is a minor fraction of the work.)

For very small n, this scheme will not work; the importance weights will be too variable and importance sampling will be very bad. For moderate n, this scheme will work; the constrained normal approximation may not be perfect but it will be a good enough importance sampling distribution. For very large n, this scheme will still work, although the importance sampling will be unnecessary because the constrained normal approximation will be nearly equal to the exact posterior.

References

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Hoboken: Wiley, third edition. 90, 97, 101

Billingsley, P. (1999). *Convergence of Probability Measures*. New York: Wiley, second edition. 99, 108

Binder, D. (1982). "Non-parametric Bayesian models for samples from a finite population." *Journal of the Royal Statistical Society, Series B*, 44: 388–393. 91

Booth, J. G., Butler, R. W., and Hall, P. (1994). "Bootstrap methods for finite population sampling." *Journal of the American Statistical Association*, 89: 1282–1289. 91

Cramér, H. (1951). *Mathematical Methods of Statistics*. Princeton: Princeton University Press. 90, 97, 101

Deville, J. and Särndal, C. (1992). "Calibration estimators in survey sampling." *Journal of the American Statistical Association*, 87: 376–382. 92

Fuller, W. (2009). *Sampling Statistics*. New York: John Wiley and Sons. 92

Geyer, C. J. (1994). "On the Asymptotics of Constrained M-Estimation." *Annals of Statistics*, 22: 1993–2010. 104

Geyer, C. J. and Meeden, G. D. (2012). "Supplementary Material for the paper "Asymptotics for Constrained Dirichlet Distributions"." Technical Report 691, School of Statistics, University of Minnesota.
URL http://purl.umn.edu/126224 104

Geyer, C. J., Meeden, G. D., and Fukuda, K. (2011). *rcdd: Computational Geometry*. R package version 1.1-4.
URL http://CRAN.R-project.org/package=rcdd 104

Ghosh, M. and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. London: Chapman and Hall. 90

Gross, S. (1980). "Median estimation in survey sampling." In *Proceedings of the Section on Survey Research Methods*, 181–184. American Statistical Association. 91

Gupta, R. D. and Richards, D. S. P. (2001). "The history of the Dirichlet and Liouville distributions." *International Statistical Review*, 69: 433–446. 93

Hájek, J. (1960). "Limiting Distributions in Simple Random Sampling from a Finite Population." *Publications of the Mathematical Institute of the Hungarian Academy of Sciences, Series A*, 5: 361–374. 92, 99

Hartley, H. O. and Rao, J. N. K. (1968). "A new estimation theory for sample surveys." *Biometrika*, 55: 159–167. 91

Lazar, R., Meeden, G., and Nelson, D. (2008). "A noninformative Bayesian approach to finite population sampling using auxiliary variables." *Survey Methodology*, 34: 51–64. 89, 92

LeCam, L. (1970). "On the Assumptions Used to Prove Asymptotic Normality of Maximum Likelihood Estimates." *Annals of Mathematical Statistics*, 41: 802–828. 104

Lo, A. (1988). "A Bayesian Bootstrap for a finite population." *Annals of Statistics*, 16: 1684–1695. 91

Meeden, G. and Lazar, R. (2011). *polyapost: Simulating from the Polya posterior*. R package version 1.1-1.
URL http://CRAN.R-project.org/package=polyapost 92

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
URL http://www.R-project.org/ 92

Rockafellar, R. T. and Wets, R. J.-B. (2004). *Variational Analysis*. Berlin: Springer-Verlag. Corrected 2nd printing. 93, 95, 102

Rubin, D. (1981). "The Bayesian bootstrap." *Annals of Statistics*, 9: 130–134. 91

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer. 92

Scheffé, H. (1947). "A Useful Convergence Theorem for Probability Distributions." *Annals of Mathematical Statistics*, 18: 434–438. 97, 103

Self, S. G. and Liang, K.-Y. (1987). "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions." *Journal of the American Statistical Association*, 82: 605–610. 104

Weisberg, S. (2005). *Applied Linear Regression*. New York: Wiley, third edition. 101

Appendix: Conditional Convergence in Distribution

This appendix justifies the "also true unconditionally" comment just before Corollary 4.4.

Let Y denote the random variable on the right-hand side of (17). What we want to show is that, assuming (17) holds conditionally, then it also holds jointly, that is, for any bounded uniformly continuous function g we have

$$Eg\big(\sqrt{n}(X_n - \lambda), \sqrt{n}(n^{-1}\alpha_n^* - \lambda)\big) \to Eg(Y, Z^*)$$

(Billingsley 1999, Theorem 2.1). We are to derive this from the conditional theorem in the paper, which says

$$E\big\{g\big(\sqrt{n}(X_n - \lambda), z\big) \,\big|\, \alpha_n^*\big\} \to E\{g(Y, z) \mid Z^*\}, \qquad \text{for all } z \in \mathbb{R}^d. \tag{32}$$

Fix $\varepsilon > 0$. Because g is uniformly continuous, there exists a $\delta > 0$ such that

$$|g(u_1, v_1) - g(u_2, v_2)| < \varepsilon, \qquad \text{whenever } \|u_1 - u_2\| + \|v_1 - v_2\| < \delta,$$

where $\|\cdot\|$ is any norm that generates the usual topology for $\mathbb{R}^d$.

Now we have

$$\begin{aligned}\big|Eg\big(\sqrt{n}(X_n - \lambda), \sqrt{n}(n^{-1}\alpha_n^* - \lambda)\big) - Eg(Y, Z^*)\big| \\ \leqslant \big|Eg\big(\sqrt{n}(X_n - \lambda), \sqrt{n}(n^{-1}\alpha_n^* - \lambda)\big) - Eg\big(\sqrt{n}(X_n - \lambda), Z^*\big)\big| \\ + \big|Eg\big(\sqrt{n}(X_n - \lambda), Z^*\big) - Eg(Y, Z^*)\big|\end{aligned}$$

by the triangle inequality. Because

$$g\big(\sqrt{n}(X_n - \lambda), \sqrt{n}(n^{-1}\alpha_n^* - \lambda)\big) - g\big(\sqrt{n}(X_n - \lambda), Z^*\big) \to 0, \qquad \text{almost surely}$$

by uniform continuity of g — it is less than $M\varepsilon$, where M is a bound for g, whenever $\|\sqrt{n}(n^{-1}\alpha_n^* - \lambda) - Z^*\| < \delta$) — the first term on the right-hand side converges to zero by dominated convergence. The second term on the right-hand side converges to zero by (32), the iterated expectation theorem, and dominated convergence.