Superiority of Bayes Estimators Over the MLE in High Dimensional Models on Compact Riemannian Manifolds and its Implication for Nonparametric Bayes Theory

by

Rachel Oliver

A Dissertation Submitted to the Faculty of the

Department of Mathematics

In Partial Fulfillment of the Requirements
For the Degree of

Doctor of Philosophy

In the Graduate College

The University of Arizona

2 0 2 0

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Rachel Oliver entitled
Superiority of Bayes Estimators Over the MLE in High Dimensional Models on Compact Riemannian Manifolds and its Implication for Nonparametric Bayes Theory
and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

________________________ Date: April 29, 2020
Rabi Bhattacharya

________________________ Date: April 29, 2020
Joseph Watkins

________________________ Date: April 29, 2020
David Glickenstein

________________________ Date: April 29, 2020
Helen Zhang

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

________________________ Date: April 29, 2020
Rabi Bhattacharya

Dedication

For my amazing husband Ben and my son and budding scientist Monty.

Acknowledgments

I am exceptionally grateful to my advisor, Rabi Bhattacharya. I have been humbled to learn from such a giant in this field. His knowledge is only rivaled by his patience and kindness. I also received valuable feedback and support from Lizhen Lin at Notre Dame. Many thanks to David Glickenstein, the head of the graduate program, in addition to being a member of my committee, for geometry advice, and academic and moral support; thanks also for statistics help from Joe Watkins and Helen Zhang. There are many other members of the faculty and staff in the University of Arizona Department of Mathematics who have been invaluable in helping me navigate the PhD program.

Spring 2016 research assistantship funded by the ADAPT study in the Mel & Enid Zuckerman College of Public Health, Health Promotion Sciences Division at the University of Arizona (study funded by the National Heart, Lung, and Blood Insitute). Many thanks to Patricia Haynes for such a wonderful interdisciplinary opportunity. Fall 2017 research assistantship funded by the NSF TRIPODS grant, and Fall 2018, Spring 2019, Fall 2019, and Spring 2020 research assistantships funded by NSF grant DMS 1811317. Part of Fall 2016 and all of Spring 2017 were funded by the Outreach Scholar program in the Mathematics Department. I'm grateful to Patricia O'Rourke and her 8th grade mathematics students at Sierra 2-8 School for teaching me about tenacity. The remaining semesters of my PhD were funded by teaching assistantships. Thanks to my college algebra, pre-calculus, and biostatistics students for helping me see how mathematics and statistics apply to a greater context.

I'm also grateful to the Association for Women in Mathematics and the opportunities for personal and professional growth it has afforded me. The leadership and members of the UA chapter have helped me be a stronger mathematician and woman.

Table of Contents

List of Figures

List of Tables

Abstract

This research focuses on the performance of Bayes estimators, in comparison with the MLE, in multinomial models with a relatively large number of cells. The prior for the Bayes estimator is taken to be the conjugate Dirichlet, i.e., the multivariate Beta, with exchangeable distributions over the coordinates, including the non-informative uniform distribution. The choice of the multinomial is motivated by its many applications in business and industry, but also by its use in providing a simple nonparametric estimator of an unknown distribution on a compact Riemannian manifold. It is striking that the Bayes procedure outperforms the asymptotically efficient MLE over most of the parameter space for even moderately large dimensional parameter spaces and rather large sample sizes.

Chapter 1
Introduction

The present research shows by analytical computations and simulations that Bayes estimators even in moderately high dimensional multinomial models outperform the MLE on most of the parameter space. High dimensional multinomial models are useful for industrial planning. For example, for planning its manufacturing process a big departmental store may try to estimate the proportions of a certain type of clothing, by sizes and/or colors, demanded by its customers (see the example in Chapter 6).

It may be noted that the venerable Bernstein-Von Mises theorem shows that the Bayes estimator under reasonable priors is asymptotically as efficient as the MLE (see, e.g. DasGupta (2008, Chapter 20), Bickel and Doksum (2001, pp. 339-342), and Bhattacharya et al. (2016, pp. 189-190)). These results show that even with the non-informative prior the Bayes estimator outperforms the MLE in high dimensional multinomial models, and that it is the MLE which tries to catch up with the Bayes estimator when the sample size increases, and that too not very successfully.

Indeed, in estimating the probabilities of a moderate to large number k cells of a multinomial, while the uniform prior $\mathrm{Dir}(1, \ldots, 1)$ is optimal in decreasing the size of the expected squared error in proportion to that of the MLE (see Chapter 4 and Figure 4.1), the prior $\mathrm{Dir}\left(\frac{1}{k}, \ldots, \frac{1}{k}\right)$ performs better in terms of the volume of the parameter space on which Bayes outperforms the MLE (Chapter 3, Table 3.1). Both perform extremely well on both counts. For this theoretical comparison exchangeable priors, $\mathrm{Dir}(\alpha_1, \ldots, \alpha_k)$, $\alpha_i = C_{k,n}\ \forall i$, are considered. Designate by $\Omega_{k,R}$ the intersection of simplex E_k with a $(k-1)$-dimensional sphere centered at the middle of the simplex with radius $\sqrt{R}$:

$$\Omega_{k,R} \equiv \left\{(\theta_1, \theta_2, \ldots, \theta_k) \in \mathbb{R}^k : \theta_j \geq 0 \ \forall j, \sum_{j=1}^{k} \theta_j = 1, |\boldsymbol{\theta}|^2 \leq R\right\}.$$

The region of E_k falling outside of $\Omega_{k,R}$ is precisely the region of the parameter space where the MLE has a lower expected squared error than the Bayes for some R depending on $C_{k,n}$, k, and n (see (2.0.6)). The geometry of this excluded region depends on the largest dimensional boundary of the simplex lying outside the sphere (Chapter 3). Since the volume excluded is difficult to compute analytically, an effective conservative upper bound of the volume is provided (Theorem 3.5.1).

Lastly, the Dir $\left(\frac{1}{k}, \ldots, \frac{1}{k}\right)$ prior provides an approximation to Ferguson's non-parametric Bayes estimation, which is briefly described in comparison with frequentist methods such as the histogram method (Chapter 5).

Extensive simulations provide support for the methodology outlined above. Some of the simulations are rather challenging; in addition to the analytical complexity of volume determination as mentioned above, L_1 comparisons between the MLE and the Bayes estimator are computationally intensive (see Chapter 5).

On the data analysis side, Chapter 6 presents a data example to illustrate an industrial application of a high dimensional multinomial. The example clearly demonstrates the significant advantage an industry or a department store may have in ordering or storing needed supplies based on a sample using Bayes estimators proposed here, even with the non-informative uniform prior, as opposed to the usual MLE. In the example, with $k = 60$, a sample size of 100 is used. If 1000 items are to be ordered for the inventory of jeans of 60 different sizes based on just a sample of size 100, then the expected total error (for having too many or too few of a size) of the MLE is about 40% higher than that of the Bayes estimator: 682 versus 486 or 494. A store manager who is highly knowledgeable about the demands for the different sizes may use a different prior, such as the very informative prior used in this paper to estimate the proportions of jeans in different sizes to be stocked, and reduce the total error to 428, a marginal improvement upon the Bayes estimators under the uniform prior. A final Chapter 7 lays down some final remarks.

Chapter 2
The multinomial distribution

Consider the estimation of the parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ in the multinomial distribution, where θ_j is the proportion of the j-th class in a population with $k \geq 2$ classes ($j = 1, 2, \ldots, k$). Based on a simple random sample of size n from the population, let $n_1, n_2, \ldots, n_k$ be the numbers in the sample belonging to each of the k classes. Since $(n_1, n_2, \ldots, n_k)$ is a sufficient statistic for $\boldsymbol{\theta}$, it is enough to consider the distribution of $(n_1, n_2, \ldots, n_k)$ for the estimation of $\boldsymbol{\theta}$. Their joint probability distribution is

$$\begin{aligned} &f(n_1, n_2, \ldots, n_k; \boldsymbol{\theta}) = \frac{n!}{n_1! n_2! \cdots n_k!} \theta_1^{n_1} \theta_2^{n_2} \cdots \theta_k^{n_k}; \\ &\text{for } \boldsymbol{\theta} \in E_k \equiv \left\{ (\theta_1, \theta_2, \ldots, \theta_k) \in \mathbb{R}^k : \theta_j \geq 0 \ \forall j, \sum_{j=1}^{k} \theta_j = 1 \right\}. \end{aligned} \tag{2.0.1}$$

The Maximum Likelihood Estimator (MLE) is $\hat{\theta} \equiv \left(\frac{n_1}{n}, \frac{n_2}{n}, \ldots, \frac{n_k}{n}\right)$. The multivariate Beta, or *Dirichlet*, prior $\mathrm{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_k)$ has density with respect to Lebesgue measure on $\Theta^{\sim}$, where $\Theta^{\sim}$ is given by

$$\Theta^{\sim} \equiv \left\{ (\theta_1, \theta_2, \ldots, \theta_{k-1}) \in \mathbb{R}^{k-1} : \theta_j \geq 0 \ \forall j, \sum_{j=1}^{k-1} \theta_j \leq 1 \right\}.$$

The Dirichlet density is

$$\pi(\theta_1, \theta_1, \ldots, \theta_k) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_k^{\alpha_k - 1}, \quad \text{for } \boldsymbol{\theta} \in \Theta^{\sim}, \tag{2.0.2}$$

where $\theta_k = 1 - \theta_1 - \theta_2 - \cdots - \theta_{k-1}$.

It is well known , and easy to prove, that if the prior is $\text{Dir}(\alpha_1, \ldots, \alpha_k)$, the posterior distribution of $\boldsymbol{\theta}$ is Dirichlet $\text{Dir}(\alpha_1+n_1, \alpha_2+n_2, \ldots, \alpha_k+n_k)$ (see, e.g., Bhattacharya et al. (2016)).

Under squared error loss: $L(\boldsymbol{\theta}, \boldsymbol{\theta}') = |\boldsymbol{\theta} - \boldsymbol{\theta}'|^2 = \sum_{1\le i\le k}(\theta_i - \theta_i')^2$; the risk function of the MLE ($\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$ with $\hat{\theta}_i = n_i/n$) is given by

$$R(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \sum_{1\le i\le k} \frac{\theta_i(1-\theta_i)}{n} = \frac{1 - \sum_{1\le i\le k}\theta_i^2}{n} \tag{2.0.3}$$

We wish to choose an exchangeable prior–invariant under permutation of coordinates. Thus we choose $\alpha_1 = \alpha_2 = \cdots = \alpha_k = C_{k,n}$, where $C_{k,n}$ is some constant which may depend on k and n. The choices of $C_{k,n}$ that lead to better estimators in terms of risk under squared error loss will be investigated.

Under the Dirichlet prior $\text{Dir}(\alpha_1, \ldots, \alpha_k)$ with $\alpha_1 = \alpha_2 = \cdots = \alpha_k = C_{k,n}$ and squared error loss, the Bayes estimator is

$$\boldsymbol{d}_B = (d_{B1}, \ldots, d_{Bk}), \quad \text{with } d_{Bi} = \frac{n_i + C_{k,n}}{n + kC_{k,n}} \quad (i = 1, 2, \ldots, k), \tag{2.0.4}$$

and its risk function is (see, e.g., Bhattacharya et al. (2016))

$$
\begin{aligned}
R(\boldsymbol{d}_B, \boldsymbol{\theta}) &= \sum_{1 \le i \le k} \frac{n\theta_i(1-\theta_i) + (C_{k,n} - k\theta_i C_{k,n})^2}{(n + kC_{k,n})^2} \\
&= \left(\sum_{1 \le i \le k} (\theta_i - \theta_i^2) \right) \frac{n}{(n + kC_{k,n})^2} + \frac{C_{k,n}^2 \left(k - 2k + k^2 \left(\sum_{1 \le i \le k} \theta_i^2\right)\right)}{(n + kC_{k,n})^2} \\
&= \left(1 - \sum_{1 \le i \le k} \theta_i^2 \right) \frac{n}{(n + kC_{k,n})^2} - \frac{kC_{k,n}^2}{(n + kC_{k,n})^2} + \frac{k^2 C_{k,n}^2 \sum_{1 \le i \le k} \theta_i^2}{(n + kC_{k,n})^2} \\
&= \frac{n - kC_{k,n}^2}{(n + kC_{k,n})^2} + \frac{\left(k^2 C_{k,n}^2 - n\right) \sum_{1 \le i \le k} \theta_i^2}{(n + kC_{k,n})^2} \qquad (2.0.5)
\end{aligned}
$$

Hence, $R(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) \le R(\boldsymbol{d}_B, \boldsymbol{\theta})$ (and thus the MLE has lower risk than the Bayes estimator) only on the set

$$
\left\{ \boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k) : \theta_i \ge 0 \ \forall i, \sum_{1 \le i \le k} \theta_i = 1, \sum_{1 \le i \le k} \theta_i^2 \ge \frac{2n + (n+k)C_{k,n}}{2n + (k + kn)C_{k,n}} \right\}. \qquad (2.0.6)
$$

Recall that the parameter space is the simplex E_k

$$
E_k \equiv \{\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k) : \theta_i \ge 0 \ \forall i, \sum_{1 \le i \le k} \theta_i = 1\} \qquad (2.0.7)
$$

We will calculate/simulate the surface area of the region 2.0.6 in E_k for various choices of $C_{k,n}$. This gives the proportion of the parameter space that is better estimated (with regard to risk) by the MLE.

It is seen (see Chapter 3) that even for fairly large sample sizes the Bayes estimator outperforms the MLE after k is large, such as $k \ge 10$.

Chapter 3
The proportion of the parameter space favoring the Bayes estimator

Before calculating the surface area of the region (2.0.6), we will consider, more generally, the simplex E_k defined in (2.0.1), and the region $\Omega_{k,R} \equiv \{\boldsymbol{\theta} \in E_k : |\boldsymbol{\theta}|^2 \leq R\}$, where R is a known constant. Note that the region (2.0.6) is the complement of $\Omega_{k,R}$ (for a specific choice of R). Thus the region $\Omega_{k,R}$, when applied to this problem, represents the region of the parameter space where the Bayes estimator has lower risk than the MLE.

Denote by $\boldsymbol{e_0} \equiv \left(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}\right)$, the centroid of the simplex equidistant from its extreme points, which is also the point in E_k that is closest to the origin, with $|\boldsymbol{e_0}|^2 = \frac{1}{k}$. We can see, then, that if $R < 1/k$ then $\Omega_{k,R} = \emptyset$. Similarly, if $R \geq 1$ then $\Omega_{k,R} = E_k$, since $|\boldsymbol{\theta}|^2 \leq 1 \ \forall \boldsymbol{\theta} \in E_k$.

For $R \geq 1/k$, define $\delta_k(R)$ to be the distance between $\boldsymbol{e_0}$ and the sphere $\{\boldsymbol{\theta} \in E_k : |\boldsymbol{\theta}|^2 = R\}$. Then

$$\delta_k(R) = \sqrt{R - \frac{1}{k}}. \tag{3.0.1}$$

Define ν_j to be the distance between $\boldsymbol{e_0}$ and the $(k-1-j)$-dimensional boundary of E_k. This is the distance between $\boldsymbol{e_0}$ and $\left(0, \ldots, 0, \frac{1}{k-j}, \ldots, \frac{1}{k-j}\right)$, which has the first j coordinates equal to 0, and the remaining $k-j$ coordinates equal to $\frac{1}{k-j}$. We have

$$\nu_j = \sqrt{\frac{j}{k(k-j)}}. \tag{3.0.2}$$

Note that we take $j = 1, \dots, k-1$ and that $\nu_1 < \nu_2 < \cdots < \nu_{k-1}$. Thus, for any $R \in (1/k, 1)$, we can find j such that $\nu_j \le \delta_k(R) < \nu_{j+1}$. The precise shape of $\Omega_{k,R}$, and thus the formula for calculating its surface area, should depend on which j satisfies this condition. Here "surface area" denotes the $(k-1)$-dimensional volume of $\Omega_{k,R}$, since $\Omega_{k,R}$ (and also E_k) is a $(k-1)$-dimensional subspace of $\mathbb{R}^k$.

3.1 The surface area of E_k

Consider in general the simplex $S_k(r)$, defined $S_k(r) \equiv \{\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k) : \ \theta_i \ge 0 \ \forall i, \sum_{1 \le i \le k} \theta_i \le r\}, \ r > 0$.

Let $E_k(r)$ be the boundary of $S_k(r)$, namely, $E_k(r) \equiv \{\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k) : \ \theta_i \ge 0 \ \forall i, \sum_{1 \le i \le k} \theta_i = r\}$, and write $E_k \equiv E_k(1)$.

Lemma 3.1.1. *(i) The volume $V_k(r)$ of $S_k(r)$ is $\frac{r^k}{k!}$, and (ii) the surface area $A_k(r)$ of $E_k(r)$ is $r^{k-1}\frac{\sqrt{k}}{(k-1)!}$.*

In particular, the surface area of the simplex $E_k \equiv E_k(1)$ is

$$A_k \, (\equiv A_k(1)) = \frac{\sqrt{k}}{(k-1)!}. \tag{3.1.1}$$

Proof. (i)

$$
\begin{aligned}
V_k(r) &= \int_{S_k(r)} d\theta_1 d\theta_2 \cdots d\theta_k \\
&= \int_{S_{k-1}(r)} \left(r - \sum_{1 \le i \le k-1} \theta_i \right) d\theta_1 d\theta_2 \cdots d\theta_{k-1} \\
&= \int_{S_{k-2}(r)} \frac{\left(r - \sum_{1 \le i \le k-2} \theta_i \right)^2}{2} d\theta_1 d\theta_2 \cdots d\theta_{k-2} \\
&= \cdots \\
&= \int_{S_1(r)} \frac{(r-\theta_1)^{k-1}}{(k-1)!} d\theta_1 \\
&= \frac{r^k}{k!}.
\end{aligned}
$$

(ii) The difference in volume between $S_k(r)$ and $S_k(r+\Delta r)$ is a slab around $E_k(r)$. Note that

$$
\begin{aligned}
V_k(r+\Delta r) - V_k(r) &= \Delta r \left[\frac{d}{dr} \left(\frac{r^k}{k!} \right) \right] + o(\Delta r) \\
&= \frac{(\Delta r) r^{k-1}}{(k-1)!} \quad \text{as } \Delta r \searrow 0.
\end{aligned}
$$

The unit normal to the surface $E_k(r)$ at every point on it is

$$
(\mathrm{grad} \sum_{1 \le i \le k} \theta_i)/|\mathrm{grad} \sum_{1 \le i \le k} \theta_i| = \left(\frac{1}{\sqrt{k}}, \frac{1}{\sqrt{k}}, \ldots, \frac{1}{\sqrt{k}} \right).
$$

Hence at every point the thickness of the slab $S_k(r+\Delta r) \setminus S_k(r)$ is $\Delta r/\sqrt{k}$. One may also see this by computing the distance between $E_k(r)$ and $E_k(r+\Delta r)$ along the normal through the origin, i.e.

$$\left| \left(\frac{r}{k}, \frac{r}{k}, \ldots, \frac{r}{k} \right) - \left(\frac{r+\Delta r}{k}, \frac{r+\Delta r}{k}, \ldots, \frac{r+\Delta r}{k} \right) \right|.$$

The surface area $A_k(r)$ then is given by

$$\begin{aligned} A_k(r) &= \lim_{\Delta r \to 0} \frac{V_k(r+\Delta r) - V_k(r)}{\Delta r/\sqrt{k}} \\ &= r^{k-1} \frac{\sqrt{k}}{(k-1)!} \ . \end{aligned}$$

□

3.2 The surface area of $\Omega_{2,R}$

Let us calculate the $k = 2$ case (this corresponds to the binomial distribution, a special case of the multinomial distribution with $k = 2$). The simplex $E_2 = \{\boldsymbol{\theta} \in \mathbb{R}^2 : \ \theta_1, \theta_2 \geq 0, \ \theta_1 + \theta_2 = 1\}$ is the line between $(0, 1)$ and $(1, 0)$. The region of interest is $\Omega_{2,R} = \{\boldsymbol{\theta} \in E_2 : \ \theta_1^2 + \theta_2^2 \leq R\}$. See Figure 3.1 for an illustration of this region.

We find the intersection points by solving $\theta_1^2 + (1 - \theta_1)^2 = R$ to obtain the points

$$p_1 = \left(\frac{1}{2} - \frac{1}{2}\sqrt{2R-1}, \frac{1}{2} + \frac{1}{2}\sqrt{2R-1} \right)$$

and

$$p_2 = \left(\frac{1}{2} + \frac{1}{2}\sqrt{2R-1}, \frac{1}{2} - \frac{1}{2}\sqrt{2R-1} \right).$$

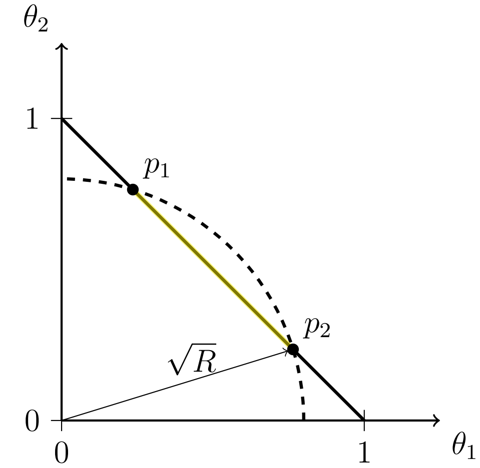

FIGURE 3.1. An illustration of E_2. The region $\Omega_{2,R}$ is the line sement between p_1 and p_2.

The surface area of $\Omega_{2,R}$ is the length of the line segment between these two points, which is $\sqrt{4R-2}$.

We can also find $\delta_2(R) = \sqrt{R - 1/2}$ and $\nu_1 = \sqrt{1/2}$ using equations (3.0.1) and (3.0.2), respectively. These distances are pictured in Figure 3.2. Note that we also have that the 1-dimensional surface area of $\Omega_{2,R}$ is equal to $2\delta_2(R) = 2\sqrt{R-1/2} = \sqrt{4R-1}$.

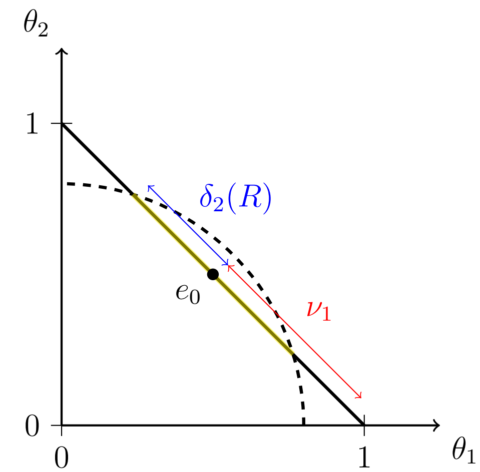

FIGURE 3.2. An illustration of $\Omega_{2,R}$ with the distances $\delta_2(R)$ and ν_1 labeled.

The length of the line E_2 is $\sqrt{2}$, giving that the proportion of the 1-dimensional surface area of E_2 made up by $\Omega_{2,R}$ is

$$\begin{aligned}\frac{\text{Area}(\Omega_{2,R})}{\text{Area}(E_2)} &= \frac{\sqrt{4R-2}}{\sqrt{2}} \\ &= \sqrt{2R-1}\end{aligned}$$

3.3 The surface area of $\boldsymbol{\Omega_{3,R}}$

For $k = 3$, we can also calculate this surface area exactly. The simplex $E_3 = \{\boldsymbol{\theta} \in \mathbb{R}^3 : \theta_1, \theta_2, \theta_3 \geq 0,\ \theta_1 + \theta_2 + \theta_3 = 1\}$ is an equilateral triangle between the points $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$. See Figure 3.3 for an illustration of the space for $R \in \left(\frac{1}{3}, \frac{1}{2}\right)$ and Figure 3.4 for an illustration of the space for $R \in \left(\frac{1}{2}, 1\right)$.

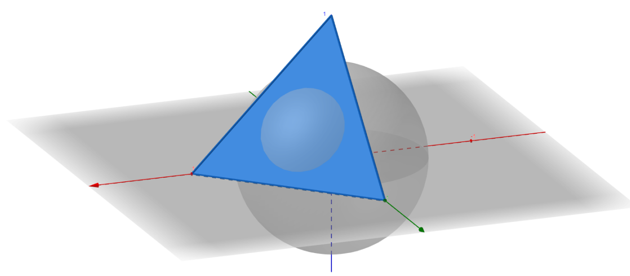

FIGURE 3.3. An illustration of E_3 with the region $\Omega_{3,R}$ in gray for $R \in \left(\frac{1}{3}, \frac{1}{2}\right)$.

We calculate ν_1 and ν_2 using equation (3.0.2) and $\delta_3(R)$ using equation (3.0.1) as follows:

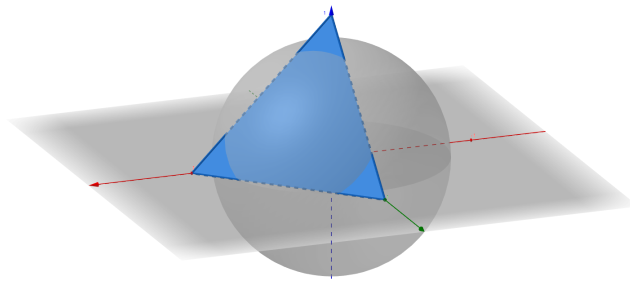

FIGURE 3.4. An illustration of E_3 with the region $\Omega_{3,R}$ in gray for $R \in \left(\frac{1}{2}, 1\right)$.

$$\nu_1 = \sqrt{\left(\frac{1}{3}\right)^2 + 2\left(\frac{1}{3} - \frac{1}{2}\right)^2} = \frac{1}{\sqrt{6}},$$
$$\nu_2 = \sqrt{2\left(\frac{1}{3}\right)^2 + \left(\frac{1}{3} - 1\right)^2} = \sqrt{\frac{2}{3}},$$

and

$$\delta_3(R) = \sqrt{R - \frac{1}{3}}.$$

If $\delta_3(R) \leq \nu_1$, then $\Omega_{3,R}$ is just a circle with radius $\delta_3(R)$. Its surface area is then $\pi\left[\delta_3(R)\right]^2 = \pi\left(R - \frac{1}{3}\right)$. The surface area of E_3 is (using equation (3.1.1)) $A_3 = \sqrt{3}/2$. This gives that the proportion of the 2-dimensional surface area of E_3 made up by $\Omega_{3,R}$ is

$$\frac{\text{Area}(\Omega_{3,R})}{\text{Area}(E_3)} = \frac{2\sqrt{3}}{3}\pi\left(R - \frac{1}{3}\right).$$

If $\nu_1 < \delta_3(R) < \nu_2$ then we can divide up the region as in Figure 3.5 to determine the surface area.

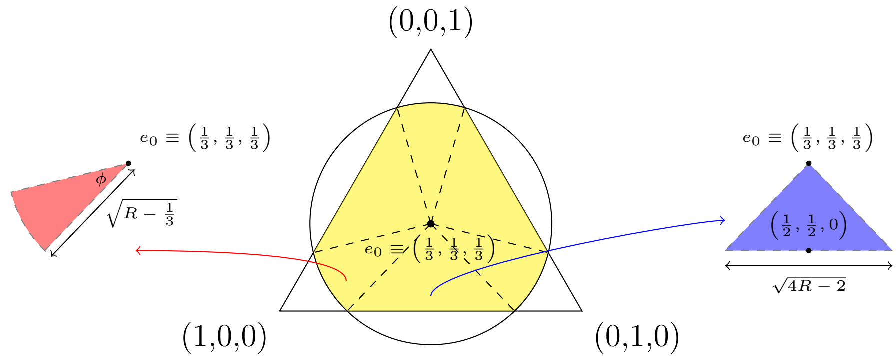

FIGURE 3.5. A diagram of how to divide $\Omega_{3,R}$ to determine its surface area.

The triangles each have area $\frac{1}{2}\frac{1}{\sqrt{6}}\sqrt{4R-2} = \frac{1}{2}\sqrt{\frac{2R-1}{3}}$. The circular sectors have radius $\delta_3(R) = \sqrt{R-1/3}$. The angle ϕ is the angle between the vectors $\left(\frac{1}{6}+\frac{1}{2}\sqrt{2R-1}, -\frac{1}{3}, \frac{1}{6}-\frac{1}{2}\sqrt{2R-1}\right)$ and $\left(\frac{1}{6}+\frac{1}{2}\sqrt{2R-1}, \frac{1}{6}-\frac{1}{2}\sqrt{2R-1}, -\frac{1}{3}\right)$. We have

$$\cos\phi = \frac{\frac{1}{4}(2R-1)+\frac{1}{2}\sqrt{2R-1}-\frac{1}{12}}{R-\frac{1}{3}}.$$

The surface area of $\Omega_{3,R}$ is then

$$\text{Area}(\Omega_{3,R}) = \frac{3}{2}\sqrt{\frac{2R-1}{3}} + \frac{3}{2}\left(R-\frac{1}{3}\right)\arccos\left(\frac{\frac{1}{4}(2R-1)+\frac{1}{2}\sqrt{2R-1}-\frac{1}{12}}{R-\frac{1}{3}}\right).$$

The proportion of the 2-dimensional surface area of E_3 made up by $\Omega_{3,R}$ is then

$$\frac{\text{Area}(\Omega_{3,R})}{\text{Area}(E_3)} = \sqrt{2R-1} + \sqrt{3}\left(R-\frac{1}{3}\right)\arccos\left(\frac{\frac{1}{4}(2R-1)+\frac{1}{2}\sqrt{2R-1}-\frac{1}{12}}{R-\frac{1}{3}}\right).$$

We can bound the surface area of $\Omega_{3,R}$ from below by cutting out triangles in the corners. This is done by drawing a straight line between the intersections of the sphere and the edges with the same value in one of the coordinates. There are six places where the sphere intersects the edges of E_3. If we let $\theta' = \frac{1}{2} + \frac{1}{2}\sqrt{2R-1}$, these solutions are $\boldsymbol{a} = (\theta', 1-\theta', 0)$, $\boldsymbol{b} = (\theta', 0, 1-\theta')$, $\boldsymbol{c} = (1-\theta', 0, \theta')$, $\boldsymbol{d} = (0, 1-\theta', \theta')$, $\boldsymbol{e} = (0, \theta', 1-\theta')$, and $\boldsymbol{f} = (1-\theta', \theta', 0)$. Draw lines between $\boldsymbol{a}$ and $\boldsymbol{b}$, $\boldsymbol{c}$ and $\boldsymbol{d}$, and $\boldsymbol{e}$ and $\boldsymbol{f}$ (see Figure 3.6).

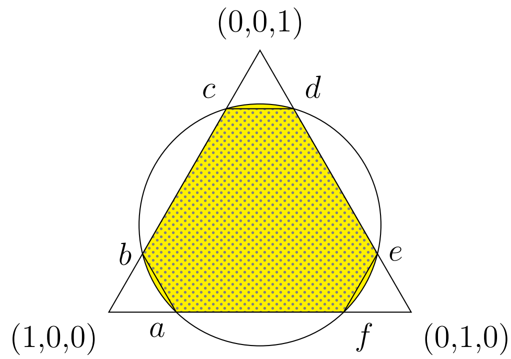

FIGURE 3.6. How to bound the surface area of $\Omega_{3,R}$ from below using similar triangles. $\Omega_{3,R}$ is shaded as before and the lower bound area is crosshatched.

Due to the convexity of the sphere, the lines drawn are entirely inside the sphere. Thus the set $\Omega'_{3,R}$, which is E_k without the three triangles, is entirely contained in $\Omega_{3,R}$. This gives

$$\text{Area}(\Omega_{3,R}) \geq \text{Area}(\Omega'_{3,R}). \tag{3.3.1}$$

The triangles in the corners that are removed are equilateral triangles, with side length $\sqrt{R - \sqrt{2R-1}}$. They are thus similar to E_3, which is equilateral with side length $\sqrt{2}$. The ratio of the areas of the small triangles to E_3 is the ratio of the squared side lengths, which is $\frac{R-\sqrt{2R-1}}{2}$. This gives, finally,

$$\frac{\text{Area}(\Omega_{3,R})}{\text{Area}(E_3)} \geq 1 - \frac{3R - 3\sqrt{2R-1}}{2}. \tag{3.3.2}$$

3.4 The surface area of $\Omega_{k,R}$ for $k \geq 4$

Calculating this surface area of $\Omega_{k,R}$ for $k \geq 4$ explicitly appears to be an open problem. We have not found a formula, but can approximate the area with a fairly sharp lower bound for certain choices of R. We will generalize the "cutting off corners" method in the $k = 3$ case, which is valid for R such that $\nu_{k-2} \leq \delta_k(R) < \nu_{k-1}$, where $\nu_j = \sqrt{j/(k(k-j))}$, $\delta_k(R) = \sqrt{R - 1/k}$.

Lemma 3.4.1 (The Surface Area of $\Omega_{k,R}$ for $R \in [1/2, 1)$)**.** *Let E_k be the standard k-simplex (defined in equation* (2.0.1)*) and $\Omega_{k,R}$ be the region $\{\boldsymbol{\theta} \in E_k : |\boldsymbol{\theta}|^2 \leq R\}$. Assume that $R \in [1/2, 1)$. Then*

$$\frac{\text{Area}(\Omega_{k,R})}{\text{Area}(E_k)} \geq 1 - k\left(\frac{R - \sqrt{2R-1}}{2}\right)^{\frac{k-1}{2}}. \tag{3.4.1}$$

Additionally, the proportion of E_k made up by $\Omega_{k,R}$ approaches 1 as $k \to \infty$.

Proof. Since $1/2 \leq R < 1$, we have

$$\begin{aligned} \frac{1}{2} \leq R < 1 &\Leftrightarrow \frac{k-2}{2k} \leq R - \frac{1}{k} < \frac{k-1}{k} \\ &\Leftrightarrow \nu_{k-2}^2 \leq [\delta_k(R)]^2 < \nu_{k-1}^2 \quad \text{(see equation (3.0.2))} \\ &\Leftrightarrow \nu_{k-2} \leq \delta_k(R) < \nu_{k-1} \end{aligned}$$

In this case, where $\nu_{k-2} \leq \delta_k(R) < \nu_{k-1}$, there are intersections along the 1-dimensional edges of E_k ($\boldsymbol{\theta}$ such that only two of its components are nonzero) with the sphere $\{\boldsymbol{\theta} \in \mathbb{R}^k : |\boldsymbol{\theta}|^2 = R\}$.

Due to the convexity of the sphere, hyperplanes between these points will be contained inside the sphere. As in the $k = 3$ case, we can form k $(k-1)$-dimensional equilateral simplices. The j-th simplex will have as one of its vertices a single vertex from E_k of the form $\theta_j = 1$ and $\theta_i = 0$ for $i \neq j$. Its remaining $k-1$ vertices will be of the form $\theta_j = \theta'$ and $\theta_i = 1-\theta'$, with $i \in \{1, 2, \ldots, j-1, j+1, \ldots, k\}$ (as in the $k = 3$ case, we define $\theta' = \frac{1}{2} + \frac{1}{2}\sqrt{2R-1}$). These have edge length $\sqrt{R - \sqrt{2R-1}}$ and are similar to E_k which has edge length $\sqrt{2}$. The ratio of the areas of the small simplices to E_k is $\left(\frac{R-\sqrt{2R-1}}{2}\right)^{\frac{k-1}{2}}$. This gives the lower bound (3.4.1).

We see that $\frac{\text{Area}(\Omega_{k,R})}{\text{Area}(E_k)}$ approaches 1 as $k \to \infty$ as long as $\frac{R-\sqrt{2R-1}}{2} < 1$. This is, in particular, true when $R \in [1/2, 1)$. □

3.5 Applying the surface area calculations to Bayes estimators

When using the prior $\text{Dir}\left(C_{k,n}, \ldots, C_{k,n}\right)$, we obtained the region (2.0.6) where the MLE has lower risk than the Bayes estimator

$$\left\{\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k) : \theta_i \geq 0 \ \forall i, \sum_{1 \leq i \leq k} \theta_i = 1, \sum_{1 \leq i \leq k} \theta_i^2 \geq R \equiv \frac{2n + (n+k)C_{k,n}}{2n + (k+kn)C_{k,n}}\right\}. \tag{3.5.1}$$

Note that this is the complement of $\Omega_{k,R}$ in E_k. We can then determine which choice of $C_{k,n}$ will yield the exponential convergence in Lemma 3.4.1. This gives the following theorem.

Theorem 3.5.1. *Consider estimating the k-dimensional ($k \geq 3$) parameter $\boldsymbol{\theta}$ in the multinomial distribution based on a simple random sample of size n. Under the prior* $\text{Dir}\left(C_{k,n}, \ldots, C_{k,n}\right)$*, define R as in equation 3.5.1. The proportion of the parameter space where the Bayes estimator has lower risk than the MLE is greater than*

$$1 - k\left(\frac{R - \sqrt{2R-1}}{2}\right)^{\frac{k-1}{2}}, \tag{3.5.2}$$

for $C_{k,n}$ satisfying

$$C_{k,n} < \frac{2n}{n(k-2)-k}. \tag{3.5.3}$$

Proof. As noted above, the region where the Bayes estimator has lower risk than the MLE is $\Omega_{k,R}$ with R defined by equation (3.5.1):

$$R = \frac{2n + (n+k)C_{k,n}}{2n + (k+kn)C_{k,n}}.$$

We have

$$\begin{aligned} R > 1/2 &\Leftrightarrow \frac{2n + (n+k)C_{k,n}}{2n + (k+kn)C_{k,n}} > 1/2 \\ &\Leftrightarrow C_{k,n}(k + 2n - kn) > -2n \\ &\Leftrightarrow C_{k,n} < \frac{2n}{n(k-2)-k}. \end{aligned}$$

We can thus apply Lemma 3.4.1. Since $R > 1/2$, we have

$$\begin{aligned} R > \frac{1}{2} &\Rightarrow R - \sqrt{2R-1} < \frac{1}{2} - \sqrt{\frac{2}{2} - 1} \\ &\Rightarrow k\left(\frac{R - \sqrt{2R-1}}{2}\right)^{\frac{k-1}{2}} < k\left(\frac{1}{4}\right)^{\frac{k-1}{2}} \\ &\Rightarrow 1 - k\left(\frac{R - \sqrt{2R-1}}{2}\right)^{\frac{k-1}{2}} > 1 - k\left(\frac{1}{4}\right)^{\frac{k-1}{2}}. \end{aligned}$$

Since $1 - k\left(\frac{1}{4}\right)^{\frac{k-1}{2}} \to 1$ as $k \to \infty$, we have proved that $\frac{\text{Area}(\Omega_{k,R})}{\text{Area}(E_k)}$ approaches 1 as $n, k \to \infty$. $\square$

One may ask if there is a best choice of $C_{k,n}$. Note that the squared radius R of the region where the Bayes estimator has lower risk ($\Omega_{k,R}$), defined in equation (3.5.1), is a decreasing function of $C_{k,n}$. We have that R approaches 1 as $C_{k,n} \searrow 0$. That is, the proportion of the parameter space where the Bayes estimator has lower risk approaches 1 (the whole space) as $C_{k,n}$ approaches 0. Note that we cannot take $C_{k,n} = 0$ as the Dirichlet prior requires that $\alpha_i > 0$ for all i. Indeed, if we could take $C_{k,n}$ to be identically zero, the Bayes estimator would be equal to the MLE!

One could, however, use the formula for R and the lower bound in (3.5.2) to select $C_{k,n}$ small enough to satisfy a desired level of coverage of the parameter space for a choice of k (and any larger k).

One choice of prior that satisfies (3.5.3) is Dir $\left(\frac{1}{k}, \ldots, \frac{1}{k}\right)$. Using Ferguson's nomenclature for Dirichlet processes, this can be thought of as using a base measure that is a probability measure, since $\sum_{1 \le i \le k} \alpha_i = 1$ (Ferguson, 1973; Ghosal and van der Vaart, 2017; Ghosh and Ramamoorthi, 2003). The region where the Bayes estimator has lower risk than the MLE is $\Omega_{k,R}$ with $R = \frac{2n+1+\frac{n}{k}}{3n+1} > 2/3$.

Thus the proportion of the parameter space where the Bayes estimator has lower risk is greater than $1 - k\left(\frac{\frac{2}{3} - \sqrt{\frac{4}{3} - 1}}{2}\right)^{\frac{k-1}{2}}$. Table 3.1 contains estimates using (3.5.1) and one minus the bound in (3.5.2), giving an *upper bound* of the proportion of the parameter space where the MLE has lower risk for various values of k and n (rather than the lower bound for its complement given in (3.5.2)). It also contains simulated proportions using similar methods as in Section 3.6. The simulation used sample sizes of 10,000,000, and thus the small proportions for $k = 20$ could not be detected.

3.6 Simulation results for other priors

The requirement that $C_{k,n} < \frac{2n}{n(k-2)-k}$ precludes some priors that may be of interest. These correspond to cases with a region of interest $\Omega_{k,R}$ such that $\delta_k(R) < \nu_{k-2}$. We

k	n	Prop (Upper Bound)	Prop (Simulated)
$k = 5$	$n = 10$	2.68×10^{-3}	2.12×10^{-3}
$k = 5$	$n = 25$	2.95×10^{-3}	2.32×10^{-3}
$k = 10$	$n = 20$	1.97×10^{-6}	7.00×10^{-7}
$k = 10$	$n = 100$	2.30×10^{-6}	8.00×10^{-7}
$k = 20$	$n = 40$	6.53×10^{-13}	0
$k = 20$	$n = 400$	7.88×10^{-13}	0

TABLE 3.1. Estimates of the proportion of the parameter space where the MLE has lower risk for various values of k and $n = 2k, k^2$. Note that it is nearly 0 for even the moderate $k = 10$.

have not found a suitable surface area lower bound for such cases. However, we have found simulation examples of a slower convergence in k.

3.6.1 Uniform prior

A common choice of prior is the uniform prior, which is the prior $\text{Dir}(1, \ldots, 1)$. Under our notation, this corresponds to $C_{k,n} = 1$, which clearly does not satisfy $C_{k,n} < \frac{2n}{n(k-2)-k}$ (3.5.3) for $k > 4$, as required in Theorem 3.5.1.

Here the MLE has lower risk than the Bayes estimator only on the set

$$\left\{ \boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k) : \theta_i \geq 0 \ \forall i, \sum_{1 \leq i \leq k} \theta_i = 1, \sum_{1 \leq i \leq k} \theta_i^2 \geq \frac{3n + k}{nk + 2n + k} \right\}. \tag{3.6.1}$$

We used a simulation study to better understand the regions where the MLE has lower risk than the Bayes estimator under this prior. Rearranging the inequality in the region 3.6.1, define the function $g(\boldsymbol{\theta}) = -3n - k + (nk + 2n + k)|\boldsymbol{\theta}|^2$. The MLE has lower risk than the Bayes estimator for $\boldsymbol{\theta} \in E_k$ where $g(\boldsymbol{\theta}) \geq 0$.

To estimate the percent of the surface area of E_k where the MLE has lower risk than the Bayes estimator, we fixed k and took a uniform sample of size 500,000 from E_k using the R package `hitandrun` (van Valkenhoef and Tervonen, 2016). We then calculated $g(\boldsymbol{\theta})$ for $n = k, 2k, 3k, 4k, k^2, 2k^2, 3k^2, 4k^2, k^3, 2k^3, 3k^3, 4k^3, k^4, 2k^4, 3k^4, 4k^4$ and found the percentage of the samples where g is positive for each n. This gives a numeric estimate of the percent of the area of E_k where the MLE has lower risk than the Bayes estimator. The results are summarized in Figure 3.7. Note that eventually we see that the MLE has lower risk in a proportion of the parameter space that is close to zero, but it is a much slower process than the case shown in Table 3.1, only getting close to zero when $k \geq 100$.

3.6.2 Dir $\left(\frac{C}{k}, \ldots, \frac{C}{k}\right)$ prior

Note that the Dirichlet distribution on the space of all probabilities on the Borel sigma-field of a Polish space S is just the Dirichlet, or multivariate Beta, distribution when S is a finite set (see Chapter 5). By studying the limiting behavior in n and k of Bayes estimators in the multinomial distribution, we can hopefully gain insight into the difference between the risks of density estimation via nonparametric Bayes and using the MLE for a parametric model. However, since Ferguson's construction of the Dirichlet process prior relies on a finite base measure $\alpha(S)$, we may want to consider the sum of the prior parameters $\sum_{1 \leq i \leq k} \alpha_i = C$, a constant, rather than $\sum_{1 \leq i \leq k} \alpha_i = k$, which is the case for the uniform prior.

If we choose $C_{k,n} = C/k$, $R(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) \leq R(\boldsymbol{d}_B, \boldsymbol{\theta})$ (and thus the MLE has lower risk than the Bayes estimator) only on the set

$$\left\{ \boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k) : \theta_i \geq 0 \ \forall i, \sum_{1 \leq i \leq k} \theta_i = 1, \sum_{1 \leq i \leq k} \theta_i^2 \geq \frac{2n + C + Cn/k}{2n + C + nC} \right\}. \quad (3.6.2)$$

Note that if $C = 1$, we obtain the example in Section 3.5. If $C > 2$, then $C_{k,n} = C/k$ does not satisfy (3.5.3) for moderately sized k and n.

We again used simulation to study the regions in question. In the simulations, the

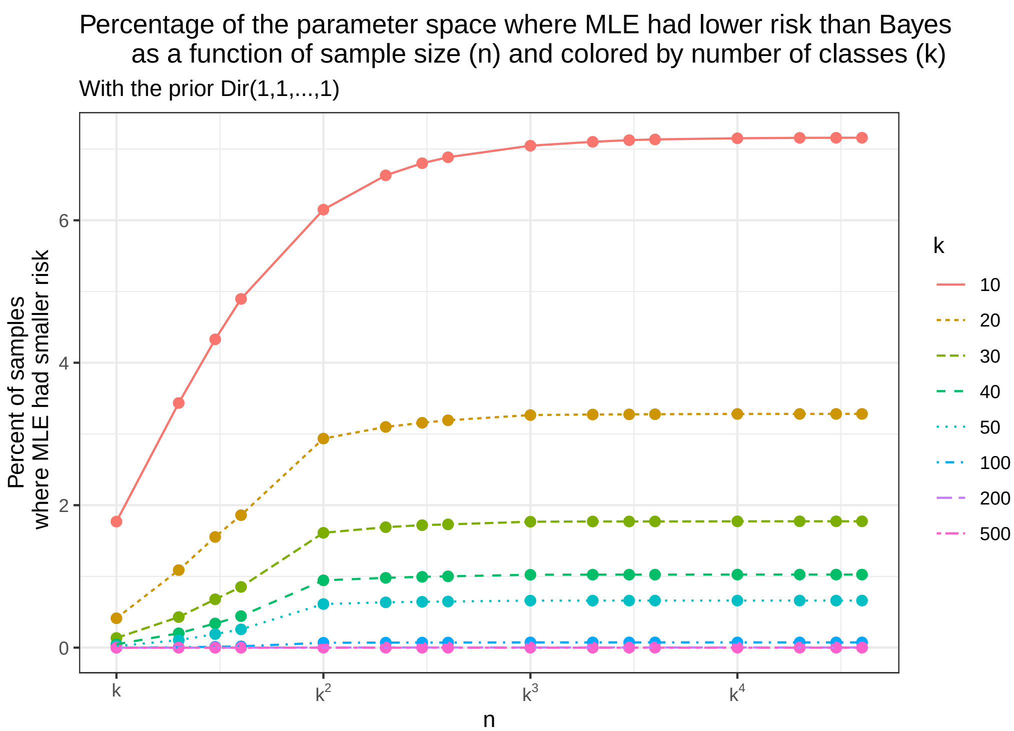

FIGURE 3.7. We see that, although Theorem 3.5.1 can't be applied to the uniform prior, the MLE still has lower risk in a proportion of the parameter that decreases to 0 as k increases. Note that the n-axis is plotted in log-scale and in terms of k to make the samples comparable.

limiting behavior is similar in shape to the uniform prior case, but seems to converge faster in k. For fixed k near C, as n increases, the percent of the surface area of the parameter space where the MLE has lower risk increases to some limiting value. As k increases, this limiting value decreases to zero. For $k >> C$, however, the Bayes estimator had lower risk in 100% of the samples, indicating that the area of the region where the MLE has lower risk is very small.

For example, with $C = 30$ and $k = 10, 20, 30$, the results can be found in Figure 3.8. A relatively large C was chosen so that there would be several k smaller than C to graph.

For $k = 40$, the maximum percentage was 0.12% and for $k = 50$, 0.0076%. For $k = 100, 200, 500$ (the three largest values used), the MLE had lower risk in 0% of the samples.

Again, we see that for large enough k, the Bayes estimator has lower risk than the MLE for almost all of the parameter space. We would need to find a suitable area bound to properly describe this phenomenon.

3.6.3 $\mathrm{Dir}(C, C, \ldots, C)$ prior for $C > 1$

Since we have already considered the uniform prior, which is $\mathrm{Dir}(1, 1, \ldots, 1)$, it is natural to consider priors $\mathrm{Dir}(C, C, \ldots, C)$ with $C > 1$. That is, $C_{k,n}$ is a constant rather than depending on k or n. Priors of this type are unimodal, focusing most of their mass on the center of the parameter space $(1/k, 1/k, \ldots, 1/k)$, with smaller component variances as C increases.

If we choose $C_{k,n} = C$, $R(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) \leq R(\boldsymbol{d_B}, \boldsymbol{\theta})$ (and thus the MLE has lower risk than the Bayes estimator) only on the set

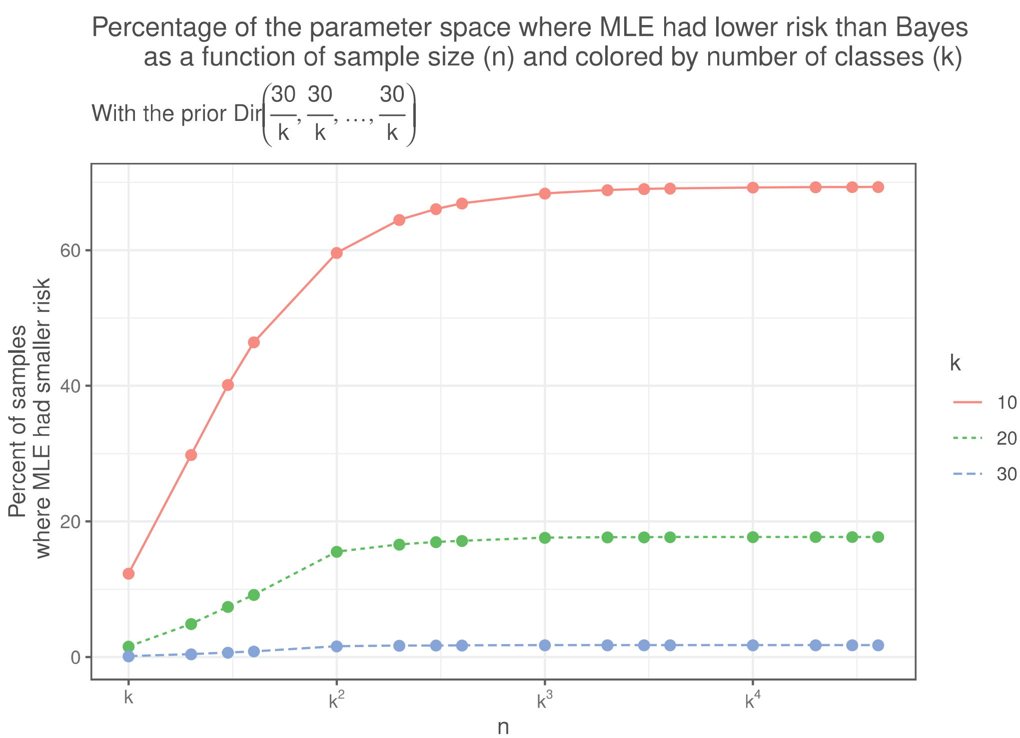

FIGURE 3.8. Simulation results for the $\text{Dir}(C/k, C/k, \ldots, C/k)$ prior, with $C = 30$. For $k = 10, 20, 30$, the maximum percentages were 69.32%, 17.73%, and 1.77%, respectively. Again we see that, although Theorem 3.5.1 does not apply, the MLE has smaller risk for almost none of the space for k large enough.

$$\left\{\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k) : \theta_i \geq 0 \ \forall i, \sum_{1 \leq i \leq k} \theta_i = 1, \sum_{1 \leq i \leq k} \theta_i^2 \geq \frac{2n + C(k+n)}{2n + C(k+kn)}\right\}. \quad (3.6.3)$$

For simulation purposes, define the function $h(\boldsymbol{\theta}) = -2n - c(k+n) + (2n + C(k + kn))|\boldsymbol{\theta}|_2^2$, which is positive when the MLE has lower risk than the Bayes estimator, by rearranging the inequality (3.6.3).

In simulations, a change of behavior is observed at $C = 2$. For $C \in (1, 2)$, the limiting behavior is similar to the uniform prior case, although with slower convergence. However, for $C \geq 2$, the opposite limiting behavior is observed. As k increases, the proportion of the parameter space where the MLE has lower risk *increases* to 1. See Figures 3.9 through 3.11 for illustrative examples with $C = 1.9$, $C = 2$, and $C = 3$.

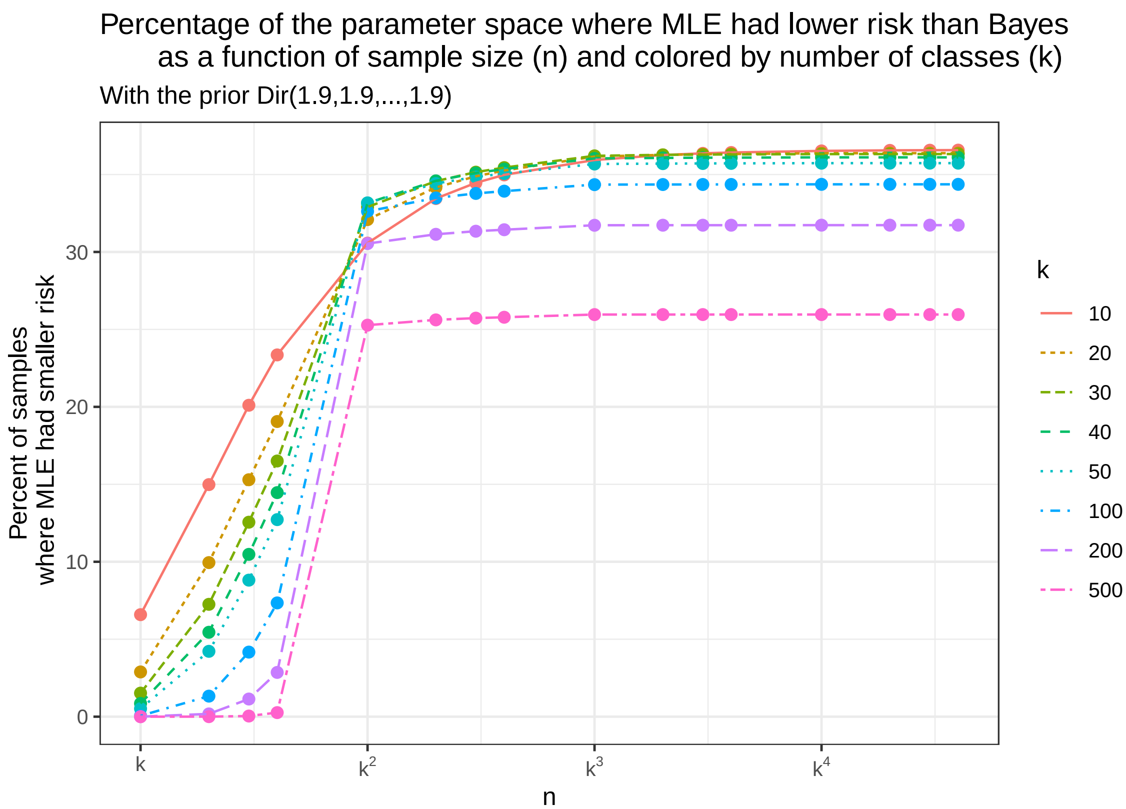

FIGURE 3.9. Simulation results for the prior $\mathrm{Dir}(C, C, \ldots, C)$ with $C = 1.9$. Note the final percent is still decreasing as k increases as in the earlier examples.

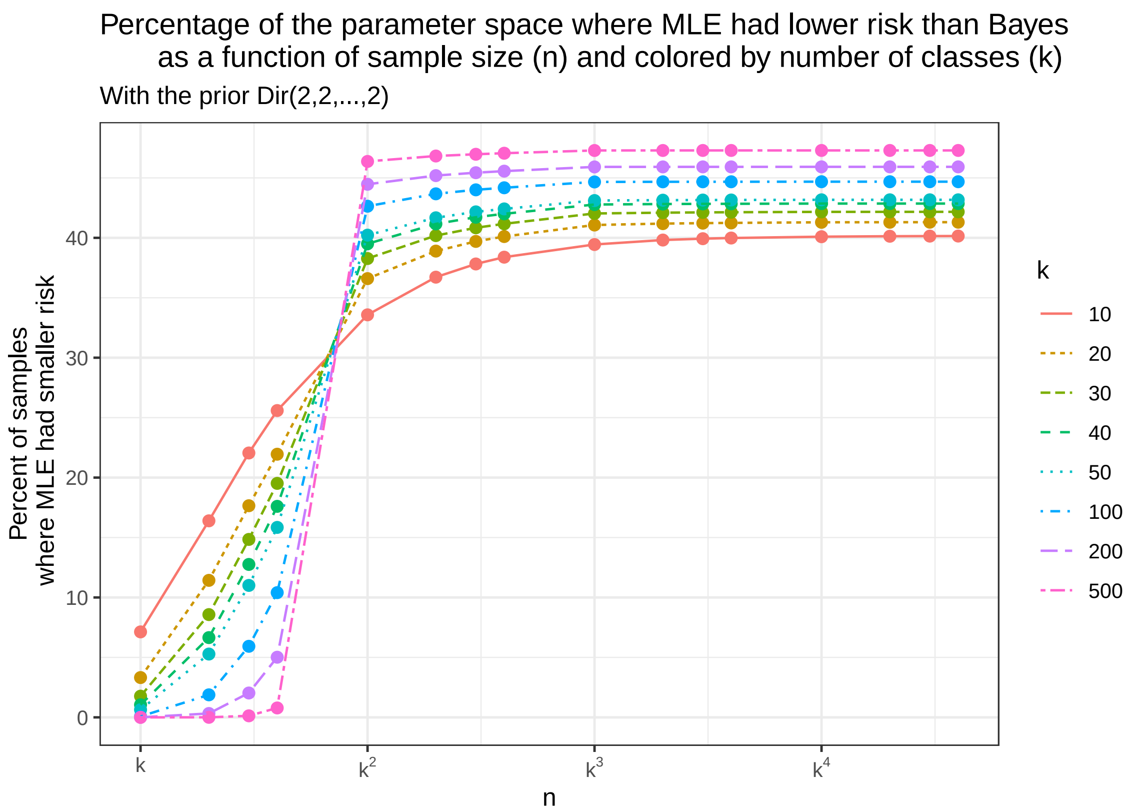

FIGURE 3.10. Simulation results for the prior $\mathrm{Dir}(C, C, \ldots, C)$ with $C = 2$. Note the change in limiting behavior, where the final percent is *increasing* as k increases, rather than decreasing as in the earlier examples.

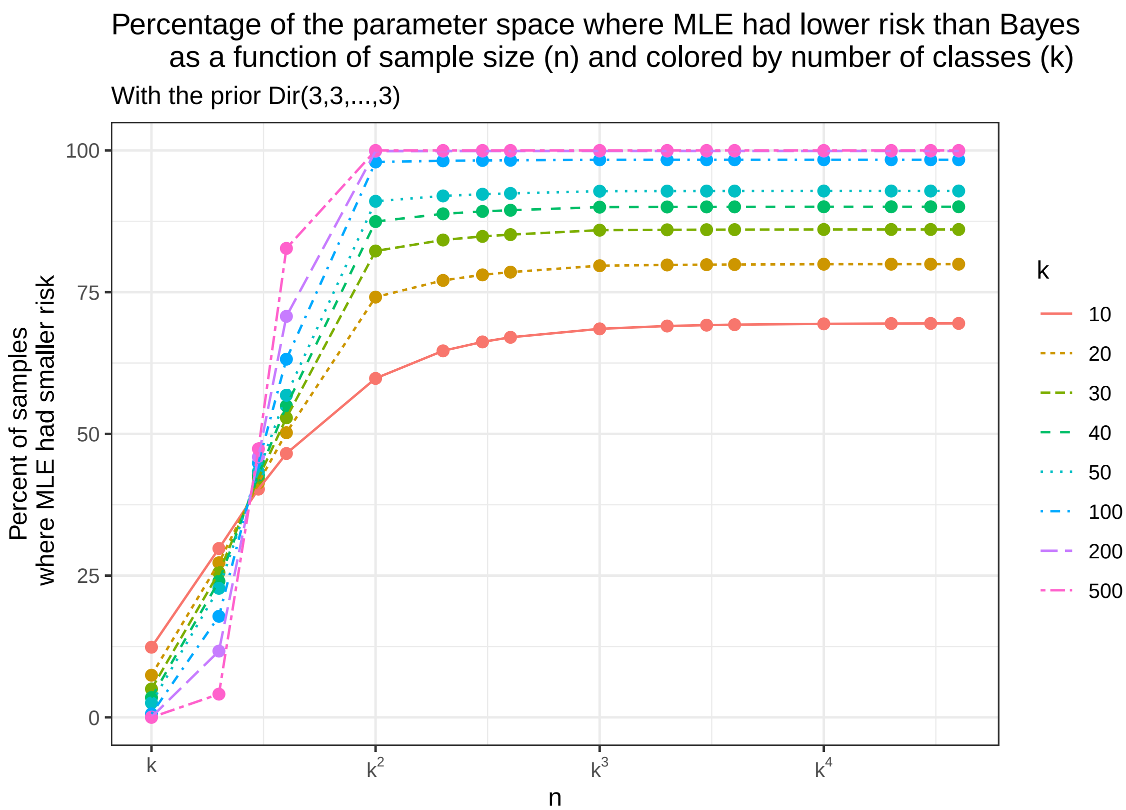

Figure 3.11. Simulation results for the prior $\text{Dir}(C, C, \ldots, C)$ with $C = 3$. Note the change in limiting behavior, where the final percent is *increasing* as k increases, rather than decreasing as in the examples with $C < 2$.

Chapter 4

Average risk across the parameter space

In Chapter 3, we considered *whether* the Bayes estimator had smaller risk than the MLE, and where this occurred within the parameter space. We did not consider the *magnitude* of the decrease in risk. In this section, we look at the average risk (with respect to the volume measure) of the various estimators to understand the magnitude of decrease.

Recall that the risk of the MLE, by (2.0.3), is

$$R(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \frac{1 - \sum_{1 \le i \le k} \theta_i^2}{n}$$

and the risk of the Bayes estimator with Dirichlet prior with $\alpha_i = C_{k,n}, \ \forall i$, is, by (2.0.5),

$$R(\boldsymbol{d}_B, \boldsymbol{\theta}) = \frac{n - kC_{k,n}^2}{\left(n + kC_{k,n}\right)^2} + \frac{\left(k^2 C_{k,n}^2 - n\right) \sum_{1 \le i \le k} \theta_i^2}{\left(n + kC_{k,n}\right)^2}.$$

Note that in each case, for fixed $k, \boldsymbol{\theta}$ and prior (choice of $C_{k,n}$), the risk is decreasing to 0 as $n \to \infty$. Thus any decrease in risk becomes negligible for large enough n.

Integrating the above over E_k, we obtain

$$\int_{E_k} R(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) d\boldsymbol{\theta} = \frac{1}{n} A_k - \frac{1}{n} \int_{E_k} |\boldsymbol{\theta}|^2 d\boldsymbol{\theta}$$

and

$$\int_{E_k} R(\boldsymbol{d}_B, \boldsymbol{\theta})d\boldsymbol{\theta} = \frac{n - kC_{k,n}^2}{\left(n + kC_{k,n}\right)^2} A_k + \frac{\left(k^2C_{k,n}^2 - n\right)}{\left(n + kC_{k,n}\right)^2} \int_{E_k} |\boldsymbol{\theta}|^2 d\boldsymbol{\theta}.$$

By (3.1.1), $A_k = \frac{\sqrt{k}}{(k-1)!}$. It can be shown that $\int_{E_k} |\boldsymbol{\theta}|^2 d\boldsymbol{\theta} = \frac{2k\sqrt{k}}{(k+1)!}$. Thus we obtain that the average risks $\bar{R}_{\hat{\boldsymbol{\theta}}}$ and $\bar{R}_{\boldsymbol{d_B}}$ are, respectively,

$$\begin{aligned} \bar{R}_{\hat{\boldsymbol{\theta}}} = \frac{\int_{E_k} R(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})d\boldsymbol{\theta}}{A_k} &= \left(\frac{\sqrt{k}}{n(k-1)!} - \frac{2k\sqrt{k}}{n(k+1)!}\right) \div \frac{\sqrt{k}}{(k-1)!} \\ &= \frac{k-1}{n(k+1)}, \end{aligned} \tag{4.0.1}$$

and

$$\begin{aligned} \bar{R}_{\boldsymbol{d_B}} = \frac{\int_{E_k} R(\boldsymbol{d}_B, \boldsymbol{\theta})d\boldsymbol{\theta}}{A_k} &= \left[\frac{(n - kC_{k,n}^2)\sqrt{k}}{\left(n + kC_{k,n}\right)^2 (k-1)!} + \frac{2k\left(k^2C_{k,n}^2 - n\right)\sqrt{k}}{\left(n + kC_{k,n}\right)^2 (k+1)!}\right] \div \frac{\sqrt{k}}{(k-1)!} \\ &= \frac{(k-1)(kC_{k,n}^2 + n)}{(kC_{k,n} + n)^2 (k+1)}. \end{aligned} \tag{4.0.2}$$

Then the average decrease in risk for the Bayes estimator *in proportion to the average risk of the MLE* is

$$\begin{aligned} \frac{\bar{R}_{\hat{\boldsymbol{\theta}}} - \bar{R}_{\boldsymbol{d_B}}}{\bar{R}_{\hat{\boldsymbol{\theta}}}} &= \left[\frac{k-1}{n(k+1)} - \frac{(k-1)(kC_{k,n}^2 + n)}{(kC_{k,n} + n)^2 (k+1)}\right] \div \frac{k-1}{n(k+1)} \\ &= 1 - \frac{n\left(kC_{k,n}^2 + n\right)}{\left(kC_{k,n} + n\right)^2}. \end{aligned} \tag{4.0.3}$$

For fixed k and n, this function has a global maximum at $C_{k,n} = 1$, as illustrated in Figure 4.1. When $C_{k,n} = 1$ we obtain, by plugging in to (4.0.3),

$$\frac{\bar{R}_{\hat{\boldsymbol{\theta}}} - \bar{R}_{\boldsymbol{d_B}, \boldsymbol{C_{k,n}=1}}}{\bar{R}_{\hat{\boldsymbol{\theta}}}} = \frac{k}{k+n} = \frac{1}{1 + n/k}. \tag{4.0.4}$$

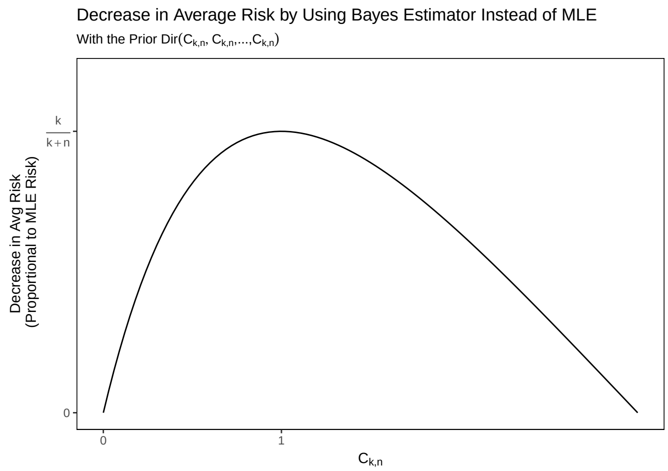

FIGURE 4.1. Graph of the average decrease in risk for the Bayes estimator in proportion to the average risk of the MLE, as a function of $C_{k,n}$ with fixed k and n. The function has a global maximum at $C_{k,n} = 1$. It is positive for the region shown here, but can become negative for large enough $C_{k,n}$. This graph was made using the values $k = 10$ and $n = 30$; however, a similar shape will result from any fixed k and n, with a maximum at $C_{k,n} = 1$ and maximum value $k/(k + n)$.

We see from (4.0.4) that for all k, there is some positive decrease in (proportional) average risk that depends on the relationship between k and n. For example, when $n = k$, the decrease is 50%, when $n = 2k$, the decrease is 33.3%, and when $n = k^2$, the decrease is $100 \cdot \left(\frac{1}{k+1}\right)\%$. The Bayes estimator with the uniform prior is the estimator that has the smallest average risk (with respect to the Lebesgue measure) by definition (see, e.g., Bhattacharya et al. (2016, pp. 22-23)). We see here that this effect, in comparison to the MLE, is most pronounced for n on the order of k.

On the other hand, we can consider $C_{k,n} = 1/k$. This prior gave rise to an estimator that had smaller risk than the MLE for nearly the entire parameter space for even small to moderate k (see example in Section 3.5). Plugging in to (4.0.3),

$$\frac{\bar{R}_{\hat{\boldsymbol{\theta}}} - \bar{R}_{\boldsymbol{d_{B,C_{k,n}=1/k}}}}{\bar{R}_{\hat{\boldsymbol{\theta}}}} = 1 - \frac{kn^2 + n}{k(n+1)^2} = 1 - \frac{n(n+1/k)}{(n+1)^2}. \tag{4.0.5}$$

This decreases to 0 as n become large, but depends less on k for its limiting behavior.

For the nonparametric estimation of a distribution, based on a given sample size n, it is useful to consider the behavior of (4.0.5) as $k \nearrow \infty$. For fixed n, (4.0.5) $\nearrow 1 - \frac{n^2}{(n+1)^2}$.

Comparing the behavior for moderate to large k and n in (4.0.4) and (4.0.5) shows an opposite kind of optimality than in Chapter 3, where the $\mathrm{Dir}(1/k, \ldots, 1/k)$ was favored over the uniform prior. It is a little surprising that, while the proportion of the surface area of the parameter space where the Bayes estimator under the prior $\mathrm{Dir}(1/k, \ldots, 1/k)$ has a smaller risk than the MLE is larger than that under the uniform prior, the latter has a smaller average risk than the former (compare Theorem 3.5.1 and Table 3.1 for $\mathrm{Dir}(1/k, \ldots, 1/k)$ on the one hand, and Figures 3.7 and 4.1 for the uniform prior on the other). Balancing the smaller radius of the region with lower risk for the estimator under the uniform prior with its optimal decrease in average risk indicates that, for moderate to large k and n, this estimator ($\boldsymbol{d_{B,C_{k,n}=1}}$) is preferable over other Bayes estimators. For small k ($k < 10$) or when the true distribution is believed to have an unknown dominating class (and thus requires an exchangeable prior with a large radius for a decrease in risk), the estimator with the $\mathrm{Dir}(1/k, \ldots, 1/k)$ prior may be preferable. For illustration, some comparison values are in Table 4.1. Note the line "$k = 5,\ n = 25$": the proportion of the parameter space where the Bayes estimator under the uniform prior has lower risk that the MLE is only 0.89. This example, with small $k = 5$, is a situation (low k) where the $\mathrm{Dir}(1/k, \ldots, 1/k)$ prior may be preferable since the proportion of the parameter space where the estimator under this prior has lower risk is 0.997, and it still decreases the average risk by almost 7 %.

		MLE	Uniform Prior		1/k Prior	
k	**n**	**Avg. Risk**	**Decrease (%)**	**Vol. Prop.**	**Decrease (%)**	**Vol. Prop.**
$k = 5$	$n = 5$	1.33×10^{-1}	50.00 %	0.9443	27.77 %	0.9977
$k = 5$	$n = 25$	2.67×10^{-2}	16.67 %	0.8926	6.80 %	0.9970
$k = 10$	$n = 10$	8.18×10^{-2}	50.00 %	0.9823	16.53 %	1
$k = 10$	$n = 100$	8.18×10^{-3}	9.09 %	0.9385	1.87 %	1
$k = 50$	$n = 50$	1.92×10^{-2}	50.00 %	0.9998	3.84 %	1
$k = 50$	$n = 2500$	3.84×10^{-4}	1.96 %	0.9939	0.08 %	1
$k = 100$	$n = 100$	9.80×10^{-3}	50.00 %	1	1.96 %	1
$k = 100$	$n = 10000$	9.80×10^{-5}	0.99 %	0.9993	0.02 %	1

TABLE 4.1. Table comparing the average risks for the MLE and the Bayes estimators under the uniform prior and the $\text{Dir}(1/k, \ldots, 1/k)$ priors for $k = 5, 10, 50, 100$ and $n = k, k^2$. Listed as well are estimated proportions of the parameter space where the estimators have lower risk than the MLE. For the uniform prior, these are estimated using the simulation in section 3.6.1. For the $\text{Dir}(1/k, \ldots, 1/k)$ prior, the estimates use (3.5.1) and (3.5.2).

Chapter 5
On simulation of total variation distances between the true distribution and the distributions of (1) MLE-based frequentist estimators and (2) nonparametric Bayes estimators

Let Q be a probability measure on some measurable state space $(S, \mathcal{S})$, which is partitioned into k measurable subsets $A_1, \ldots, A_k$. If one wishes to estimate the probabilities $Q(A_j) = \theta_j > 0$, $j = 1, \ldots, k$, based on the numbers $n_1, \ldots, n_k$ of a random sample of size n from Q falling into these classes, the MLE $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1 = n_1/n, \ldots, \hat{\theta}_k = n_k/n)$ is the time-honored estimate, and $(n_1, \ldots, n_k)$ has the multinomial distribution $M(\boldsymbol{n} : \theta_1, \theta_2, \ldots, \theta_k)$. One may, instead, use the Bayes estimator with the conjugate prior $\mathrm{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_k)$, namely, $\boldsymbol{d_B} = ((n_1 + \alpha_1) / (n + \sum \alpha_j), \ldots, (n_k + \alpha_k) / (n + \sum \alpha_j))$, where $\alpha_i > 0 \, \forall i$. We have seen that under squared error loss, for large or moderately large k, $\boldsymbol{d_B}$ outperforms $\hat{\theta}$ on most of the parameter space in cases (1) $\alpha_i = C > 0 \, \forall i$ (for $C < 2$), and (2) $\alpha_i = 1/k \, \forall i$. For large k these estimators provide approximations to Q. We now consider a way of providing approximations to Q in variation norm for classes of S with a finite volume measure ω and Q absolutely continuous with respect to it.

Consider a closed bounded region S such as a ball or rectangular region in an Euclidean space, or a compact Riemannian manifold such as the sphere S^d, Kendall's planar shape space Σ_2^m (which is the same as the complex projective space $\mathbb{C}P^{m-2}$), etc., each equipped with a volume measure ω. Recall (see, e.g., Bhattacharya et al. (2016, p. 309)) that Σ_2^m or, equivalently, $\mathbb{C}P^{m-2}$ comprises all m-ads or sets of m

($m > 2$) points in the plane from which the effect of translation is removed (by centering), the effect of scaling is removed (by scaling the translated m-ad) and the effect of rotation is removed by identifying all rotations. As above, we consider a partition of S into k subsets $A_{j,k},\ j = 1, \ldots, k$, such that $\omega(A_{j,k}) > 0\,\forall j, k$ and $\omega(A_{j,k}) \to 0$ as $k \to \infty$, and let $Q(A_{j,k}) = \theta_{j,k}$. Let $\hat{\boldsymbol{\theta}} = (\hat{\theta}_{1,k} = n_1/n, \ldots, \hat{\theta}_{k,k} = n_k/n)$ be the MLE of $\boldsymbol{\theta} = (\theta_{1,k}, \ldots, \theta_{k,k})$. Consider the Bayes estimator under the Dirichlet prior with $\alpha_i = 1/k\,\forall i$ and under the *absolute error* loss function, say $\breve{\boldsymbol{\theta}} = (\breve{\theta}_{1,1}, \ldots, \breve{\theta}_{k,k})$, where $\breve{\theta}_{j,k}$ is the median of the posterior distribution of $\theta_{j,k}$, namely, the *median* of Beta$(1/k + n_j, n + 1 - n_j - 1/k),\ j = 1, \ldots, k$.

The risk function of $\breve{\boldsymbol{\theta}}$ is

$$
\begin{aligned}
R\left(\boldsymbol{\theta}, \breve{\boldsymbol{\theta}}\right) &= \sum_{1 \le j \le k} E_{\theta_{j,k}} \left| \breve{\theta}_{j,k} - \theta_{j,k} \right| \\
&= \sum_{j=1}^{k} \sum_{r=0}^{n} C_r^n \theta_{j,k}^r (1 - \theta_{j,k})^{n-r} |F^{-1}_{(r+1/k, n-r+1-1/k)}(1/2) - \theta_{j,k}|, \qquad (5.0.1)
\end{aligned}
$$

where $F_{(\alpha,\beta)}$ is the distribution function of Beta(α, β), $F^{-1}_{(\alpha,\beta)}$ is its inverse, and $F^{-1}_{(\alpha,\beta)}(1/2)$ is the median of Beta(α, β).

The risk function of $\hat{\boldsymbol{\theta}}$ is

$$
\begin{aligned}
R\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}\right) &= \sum_{1 \le j \le k} E_{\theta_{j,k}} |\hat{\theta}_{j,k} - \theta_{j,k}| \\
&= \sum_{j=1}^{k} \sum_{r=0}^{n} C_r^n \theta_{j,k}^r (1 - \theta_{j,k})^{n-r} |r/n - \theta_{j,k}|. \qquad (5.0.2)
\end{aligned}
$$

Suppose Q has a continuous density f (with respect to the volume measure ω). We now consider the problem of estimating the approximate density of Q as f_k, where

$$f_k(x) = \theta_{j,k}/\omega(A_{j,k}) \quad \text{for } x \in A_{j,k} \ (j = 1, \ldots, k). \tag{5.0.3}$$

Assume that $\max\{\text{diam}(A_{j,k}) : j = 1, \ldots, k\} \to 0$ as $k \to \infty$. Here $\text{diam}(A)$ stands for the diameter of the set A. Then $\int |f_k(x) - f(x)|\omega(dx) \to 0$ as $k \to \infty$. We consider now the estimates of f_k given by

$$\hat{f}_k(x) = \hat{\theta}_{j,k}/\omega(A_{j,k}) \quad \text{for } x \in A_{j,k} \ (j = 1, \ldots, k), \tag{5.0.4}$$

$$\breve{f}_k(x) = \breve{\theta}_{j,k}/\omega(A_{j,k}) \quad \text{for } x \in A_{j,k} \ (j = 1, \ldots, k). \tag{5.0.5}$$

The L_1 (and thus the total variation) distances between f_k and its estimates above are given by (5.0.1) and (5.0.2). In particular, the Bayes estimator $\breve{f}_k(x)$ basically provides the nonparametric estimator of f_k under the Dirichlet prior with base measure α_k on S with density $\alpha_k(x) = \frac{1}{k\omega(A_{j,k})}$ for $x \in A_{j,k}$ $(j = 1, \ldots, k)$.

Note that one may consider the above type of approximation for an unbounded state space by requiring that the density decays fast outside a bounded region.

Similarly, the L_2 distance between f_k and its estimators under the squared error loss are related to the corresponding risk functions. If the $A_{j,k}$ are chosen such that $\omega(A_{j,k}) = \omega(S)/k$ $\forall j$ and $\hat{\theta}_{j,k} = n_j/n$ and $d_{j,k} = (n_j + 1/k)/(n + 1)$ (the Bayes estimator under squared error loss with Dirichlet prior $\text{Dir}(1/k, \ldots, 1/k)$), then

$$\|f_k - \hat{f}_k\|_2 = \sqrt{\frac{k}{\omega(S)}}\sqrt{R\left(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}\right)}, \tag{5.0.6}$$

$$\|f_k - \tilde{f}_k\|_2 = \sqrt{\frac{k}{\omega(S)}}\sqrt{R\left(\boldsymbol{d_B}, \boldsymbol{\theta}\right)}, \tag{5.0.7}$$

where $R\left(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}\right)$ and $R\left(\boldsymbol{d_B}, \boldsymbol{\theta}\right)$ are as defined in (2.0.3) and (2.0.5), respectively.

5.1 Simulated L_1 distances

We simulated these L_1 distances using the case where $\omega(S) = 1$ by uniformly sampling 10,000 probability vectors from the standard k-simplex for each of $k = 10, 20, 30, 40, 50, 100, 200, 500$ and calculating the risk using equations (5.0.1) and (5.0.2). Then the 10,000 risk calculations were averaged to obtain an estimate of the average L_1 distance across the parameter space. Note that the number of simulations (10,000) is smaller than that used in the L_2 case (1,000,000) due to increased computational complexity.

Representative results can be found in Tables 5.1 through 5.4. Note that for large n, the estimated average L_1 distance for the Bayes estimator is often slightly larger than that for the MLE. This is not entirely unexpected, since the Bayes estimator in the parametric model with prior $\text{Dir}(1/k, \ldots, 1/k)$ is optimal when the average of the risk function under absolute error (5.0.1) is calculated with respect to $\text{Dir}(1/k, \ldots, 1/k)$, and not with respect to the uniform $\text{Dir}(1, \ldots, 1)$ as is used to compute the L_1 distance $\int |f_k(x) - f(x)|\omega(dx)$.

$\boldsymbol{k = 10}$		
$\boldsymbol{n}$	**MLE**	**Bayes**
$n = 20$	0.4689	0.4541
$n = 30$	0.3827	0.3765
$n = 40$	0.3314	0.3283
$n = 50$	0.2964	0.2947
$n = 100$	0.2095	0.2095
$n = 200$	0.1481	0.1483
$n = 300$	0.1209	0.1211
$n = 400$	0.1047	0.1048
$n = 500$	0.0936	0.0937
$n = 600$	0.0855	0.0856
$n = 700$	0.0791	0.0792
$n = 800$	0.0740	0.0741
$n = 900$	0.0698	0.0698
$n = 1000$	0.0662	0.0662

Table 5.1. Simulated L_1 distances for $k = 10$.

$\boldsymbol{k = 40}$		
$\boldsymbol{n}$	**MLE**	**Bayes**
$n = 20$	0.9611	0.8375
$n = 30$	0.7948	0.7208
$n = 40$	0.6916	0.6436
$n = 50$	0.6200	0.5872
$n = 100$	0.4397	0.4325
$n = 200$	0.3112	0.3111
$n = 300$	0.2541	0.2549
$n = 400$	0.2201	0.2210
$n = 500$	0.1968	0.1977
$n = 600$	0.1797	0.1805
$n = 700$	0.1664	0.1671
$n = 800$	0.1556	0.1562
$n = 900$	0.1467	0.1473
$n = 1000$	0.1392	0.1397

Table 5.2. Simulated L_1 distances for $k = 40$.

$\boldsymbol{k = 100}$		
$\boldsymbol{n}$	**MLE**	**Bayes**
$n = 20$	1.3908	1.1440
$n = 30$	1.2031	1.0075
$n = 40$	1.0681	0.9143
$n = 50$	0.9676	0.8454
$n = 100$	0.6973	0.6499
$n = 200$	0.4958	0.4838
$n = 300$	0.4053	0.4016
$n = 400$	0.3511	0.3503
$n = 500$	0.3141	0.3144
$n = 600$	0.2868	0.2876
$n = 700$	0.2655	0.2665
$n = 800$	0.2484	0.2495
$n = 900$	0.2342	0.2353
$n = 1000$	0.2222	0.2233

TABLE 5.3. Simulated L_1 distances for $k = 100$.

$\boldsymbol{k = 500}$		
$\boldsymbol{n}$	**MLE**	**Bayes**
$n = 20$	1.8490	1.5426
$n = 30$	1.7798	1.4827
$n = 40$	1.7145	1.4246
$n = 50$	1.6527	1.3695
$n = 100$	1.3930	1.1515
$n = 200$	1.0714	0.9189
$n = 300$	0.8932	0.7951
$n = 400$	0.7798	0.7133
$n = 500$	0.7003	0.6532
$n = 600$	0.6407	0.6064
$n = 700$	0.5941	0.5684
$n = 800$	0.5562	0.5367
$n = 900$	0.5248	0.5097
$n = 1000$	0.4981	0.4864

TABLE 5.4. Simulated L_1 distances for $k = 500$.

Chapter 6

Data example: stocking jeans

In recent years, the lack of sizing representation in clothing stores has been decried by many groups. See, for example, the article "Women's Clothing Retailers are Still Ignoring the Reality of Size in the US" from Quartzy (Shendruk, 2018). A large component of this problem is that stores do not tend to stock sizes in proportion to the distribution of clothing sizes reflected in the general population; rather, there is a notion of stocking clothing based on the "typical customer" for the store. This becomes a self-fulfilling prophecy, however, since choosing not to stock for parts of the population not deemed "typical customers" ensures that they cannot ever be customers by definition.

Let us focus on denim jeans, which are widely considered a staple in the American woman's wardrobe. A brick-and-mortar retailer will largely only sell sizes that are currently in stock (while employees may offer to special order sizes not in stock, the majority of patrons will simply leave the store without purchasing if their size is not in stock). Since the purchasing of stock represents a risk by the retailer, it is important to accurately guess which sizes to stock. However, when taking into account both waist size and inseam, as several denim brands do, this can result in a large number of size options for stock. For example, using the Levi's online size chart and their online catalog, we calculated 59 different sizes (Levi's, 2019).

The retailer could use past sales as a guide for how much of each size to stock. However, this has the effect of perpetuating errors in representation, since patrons who desired to purchase jeans but were unable since their sizes were not in stock can not be represented in the sales data. Instead, the retailer could sample the desired sizes of anyone who enters the store, regardless of whether they make a purchase. This would likely reflect the distribution of potential customers more accuately than sales data. The retailer most likely would need to take a small to moderate sample initially since too much time with an inaccurate stock distribution may cause unrepresented segments of the population to stop coming altogether.

To simulate such a sampling scheme, we used the National Health and Nutrition Examination Survey from 2015-16 (CDC, 2018) and the Levi's size chart to estimate the true Levi's jean size distribution of adult women in the United States. After restricting to adult women and excluding those in the sample that were pregnant (as this temporarily skews waist size), there were 2697 adult women surveyed in the NHANES, with sample weighting to properly reflect the uninstitutionalized population of the United States. Using the Levi's website, we calculated 59 different jean sizes, as well as a category for those whose waist size is too high to fit into any of Levi's listed sizes (we estimated that 8.39% of adult women in the United States fit into this category). There was one jean size that was not sampled in the NHANES. We decided that it is unlikely that this jean size does not exist in the entire population of the US, so we gave this size a proportion equal to one half of the minimum nonzero proportion in the other sizes and then renormalized.

We then simulated random samples of size 100 from the multinomial distribution with 60 categories using the calculated size distribution for the US adult women population. This simulates the following scenario: the retailer hopes to estimate the distribution of jean sizes his clientele desire by recording the desired jean size of a sample of 100 potential customers, and his potential custormers reflect the size distribution of the US adult female population as a whole, rather than a (potentially smaller-waisted) subpopulation. We included the 60-th category of "no size" since, under this scenario in which the potential customers reflect the true distribution, it is possible that customers may arrive at the store hoping to buy jeans before learning that the sizes are not large enough.

We then calculated the MLE and used a uniform prior to calculate two different Bayes estimators: the estimator under squared-error (L_2) loss–the posterior mean–and the estimator under absolute-error (L_1) loss–the posterior median. Estimating under a uniform prior tries to balance between two ideas of "fairness": representing all sizes (a uniform prior) and representing the size distribution (the posterior mean or median given the sample).

We repeated this simulation 1000 times. Each time, at least 18 of the size categories were unrepresented in the sample of 100; the MLE estimated zero probabilities for almost one third of the sizes. On the other hand, the Bayes estimators are never zero and thus lean toward being more inclusive of sizes while still taking into account the sample data. We also calculated the L_1, L_2, and infinity (maximum) distance between the estimators and the calculated size distribution. The results are in Table 6.1. The Bayes estimators tended to be closer than the MLE to the true size distribution by all three distance measures despite being a biased estimator. The L_2 Bayes estimator was closer in L_1 in 85.8% of the simulations, closer in L_2 in 94.9% of the simulations, and had a smaller maximum distance (i.e. the largest absolute difference among all 60 categories) in 67.8% of the simulations. The L_1 Bayes estimator was closer in L_1 in 95.6% of the simulations, closer in L_2 in 93.5% of the simulations, and had a smaller maximum distance in 62.5% of the simulations.

	L_1	L_2	Infinity (maximum)
Bayes estimator (L_2 Loss)	0.4599	0.0816	0.0377
Bayes estimator (L_1 Loss)	0.4368	0.0830	0.0395
MLE	0.5081	0.0969	0.0447

TABLE 6.1. The mean L_1, L_2, and infinity (maximum) distances between the estimators and the true size distribution in 1000 simulations.

Table 6.2 shows the estimated (true) size distribution based on the NHANES, the numbers of a stock of 1000 jeans in a Levi's store this would represent, and stock based on the MLE and Bayes estimators from a sample of 100 customers. Three sizes have such low probabilities that the true size distribution recommends not stocking them. In this particular sample, 35 of the sizes were unrepresented; thus the MLE recommended stocking less than half of the sizes. Due to rounding, the stocks are not exactly 1000. The MLE-based stock has 993, the Bayes (L_2) has 1006, and the Bayes (L_1) has 985. To remedy this, for the ones that were under, one pair was added to sizes randomly chosen from those with the fewest, and for those that were over, pairs were subtracted from the sizes with the most. This led to 28 sizes being unrepresented in the MLE-based stock, and 7 with one pair each. The L_2 distances between the true size distribution and the MLE, Bayes (L_2), and Bayes (L_1) estimators were, respectively, 0.122, 0.077, and 0.079. The absolute errors in the stock (the sum of all absolute differences between stock numbers in each size) for the MLE, Bayes (L_2) and Bayes (L_1) were, respectively, 682, 486, and 494.

A Bayesian would most likely prefer to use an informative prior in such a situation. For comparison, we repeated the simulations using a prior that reflects plausible beliefs about the size distribution. In Figure 6.1 is the layout of the sizes over the waist and height measurements (cm) they should fit. The outer boundaries are estimated, but the inner boundaries come from the Levi's size chart as mentioned earlier. The sizes have been divided into four groups, where group 4 is believed to have the highest proportion, while group 1 the lowest. These are based on the belief that sizes toward the center of the chart are more likely, with some consideration for the area of the size block. We assigned a prior of the form $\text{Dir}(\alpha_1, \ldots, \alpha_i, \ldots, \alpha_k)$ where $\alpha_i = 0.7$, 0.9, 1.6, 1.9 for i in groups 1 through 4, respectively.

We tested this prior on the same 1000 samples as were used in the 1000 simulations previously, only computing the new Bayes estimator under L_2 loss. The estimator using the informative prior was closer to the truth than in MLE in maximum distance

Size (Waist.Inseam)	True p	Stock (true)	Stock (MLE)	Stock (Bayes, L_2 Loss)	Stock (Bayes, L_1 Loss)
24.28	1.386e-04	0	1	7	5
24.30	8.657e-04	1	0	7	6
24.32	2.020e-04	0	0	7	6
25.28	6.930e-05	0	0	7	6
25.30	1.949e-03	2	0	7	6
25.32	1.000e-03	1	0	7	6
26.28	6.218e-04	1	0	7	5
26.30	5.900e-03	6	11	14	13
26.32	4.262e-03	5	0	7	6
27.28	6.986e-04	1	0	7	5
27.30	1.268e-02	14	0	7	6
27.32	8.139e-03	9	11	14	13
27.34	1.214e-03	1	0	7	5
28.28	1.356e-03	1	0	7	5
28.30	5.628e-03	6	1	7	6
28.32	1.305e-02	14	0	7	5
28.34	2.675e-03	3	0	7	6
29.28	1.847e-03	2	0	7	5
29.30	1.181e-02	13	46	34	37
29.32	1.471e-02	16	1	7	5
29.34	3.006e-03	3	0	7	6
30.28	2.008e-03	2	0	7	6
30.30	1.203e-02	13	23	21	21
30.32	1.350e-02	15	0	7	5
30.34	4.348e-03	5	0	7	5
31.28	4.102e-03	4	1	7	5
31.30	2.713e-02	30	11	14	13
31.32	3.483e-02	38	0	7	5
31.34	9.686e-03	11	0	7	5
32.28	3.122e-03	3	0	7	6
32.30	3.314e-02	36	46	34	37
32.32	4.211e-02	46	69	48	52
32.34	1.271e-02	14	34	27	29
33.28	3.128e-03	3	1	7	5
33.30	2.157e-02	24	11	14	13
33.32	3.553e-02	39	23	21	21
33.34	3.100e-03	3	11	14	13
34.28	9.037e-03	10	0	7	5
34.30	4.477e-02	49	46	34	37
34.32	7.274e-02	79	103	65	76
34.34	1.153e-02	13	0	7	6
16W.S	8.729e-03	10	0	7	5
16W.M	9.149e-03	10	0	7	5
16W.L	1.929e-03	2	0	7	5
18W.S	4.694e-02	51	80	55	60
18W.M	5.803e-02	63	92	62	68
18W.L	8.997e-03	10	11	14	13
20W.S	4.353e-02	48	1	7	5
20W.M	4.290e-02	47	46	34	37
20W.L	8.351e-03	9	0	7	5
22W.S	3.617e-02	39	69	48	52
22W.M	3.962e-02	43	57	41	45
22W.L	8.016e-03	9	34	27	29
24W.S	2.187e-02	24	1	7	6
24W.M	3.911e-02	43	103	65	76
24W.L	7.272e-03	8	11	14	13
26W.S	1.851e-02	20	34	27	29
26W.M	2.248e-02	25	11	14	13
26W.L	2.546e-03	3	0	7	6
No size	8.393e-02	0	0	0	0

TABLE 6.2. The true size distribution based on NHANES 15-16 as well as the stock of 1000 based on the truth and estimators from a sample of size 100. The total absolute errors of the three estimated stocks are 682, 486, and 494, respectively.

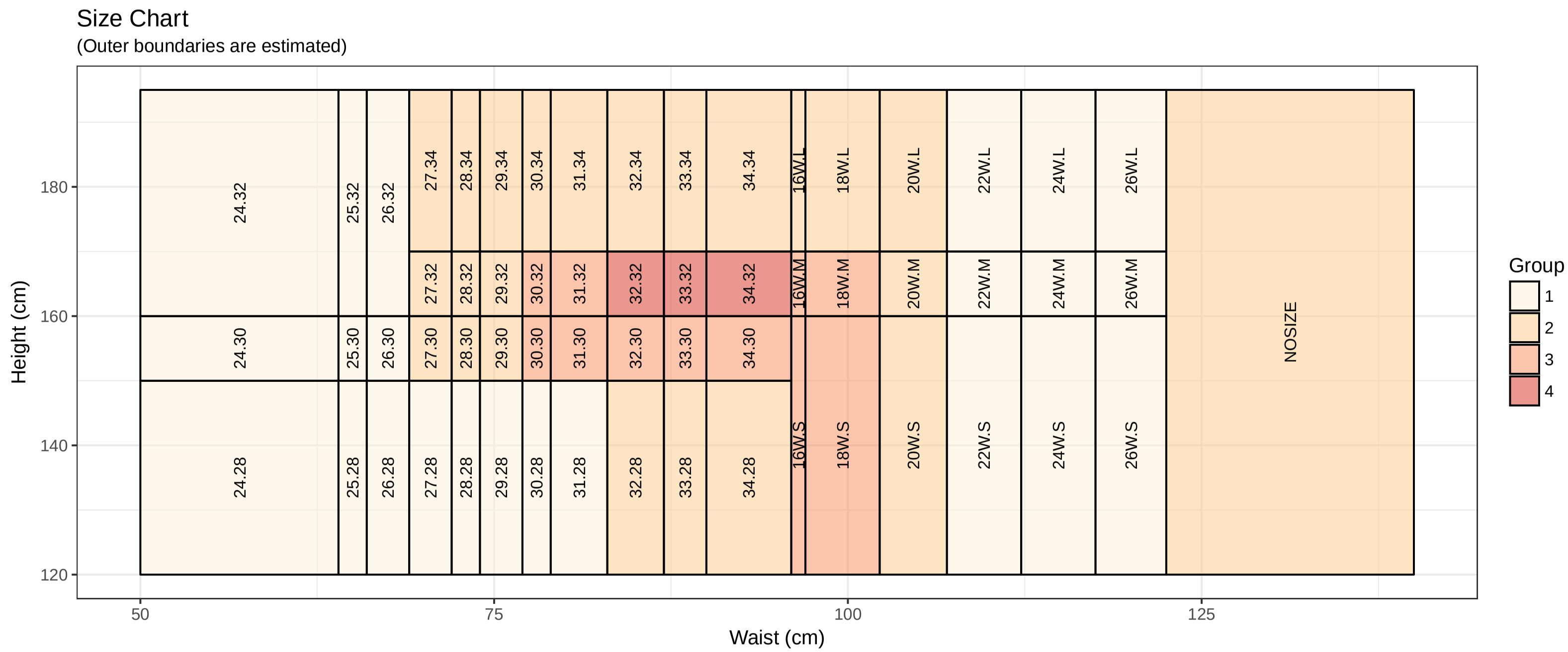

FIGURE 6.1. Figure showing the waist and height measurements (cm) encompassed by each pant size. The outer boundaries on the chart are estimated. The sizes are colored based on the author's prior assumptions about the size distribution, with group 4 being the most likely and group 1 being the least.

76.3% of the time, closer in L_1 98% of the time, and closer in L_2 98.7% of the time. The mean L_1, L_2, and infinity (maximum) distances were 0.4198, 0.0772, and 0.0362, respectively. Comparing to Table 6.1, we see that this is only a modest improvement over the Bayes estimators under the uniform prior. It is actually striking that the uniform prior gets as close as it does to the highly informative prior.

We also computed the recommended stock with this new estimator for the sample in Table 6.2. The total absolute error for the new stock is 428, which is again a modest improvement over the estimators with the uniform prior.

Chapter 7
Final remarks

In this research it is shown that in a multinomial model with a moderately large number of k cells and even a reasonably large sample size, the Bayes estimator with a multivariate Beta (or Dirichlet) $\mathrm{Dir}(C_{k,n}, \ldots, C_{k,n})$ prior has a smaller risk under squared error loss than that of the MLE on most of the parameter space, for the cases $C_{k,n} = 1$, and $C_{k,n} = 1/k$. Simulation studies also show the surprising fact that for $C_{k,n} \geq 2$ the performance of the Bayes estimator rapidly declines. In future research, we would like to investigate absolute error loss in a similar way. Unfortunately, it proved too intensive, both analytically and computationally, for this current work.

It has been pointed out that the prior $\mathrm{Dir}(1, \ldots, 1)$ shrinks "the MLE toward the centroid of the simplex $e_0 = (1/k, \ldots, 1/k)$", somewhat similar to the James-Stein estimator. It may be pointed out, however, that using integer value of $C_{k,n} \geq 2$ shrinks the MLE more, but does not perform so well (see Section 3.6.3).

The choice $C_{k,n} = 1/k$ is motivated by the fact that it provides a simple approximation of the nonparametric estimation of an unknown distribution with a Dirichlet process prior à la Ferguson (1973). In this context, there have been some simulation studies where nonparametric Bayes procedures have been found to outperform frequentist ones. A dramatic example may be found in Bhattacharya and Dunson (2010). Here a random sample is drawn from a parametric distribution on Kendall's planar shape space with density $f_0 = f(\cdot, \theta_0)$ with a given parameter value $\theta = \theta_0$, and three estimates of f_0 are compared: $\hat{f} = f(\cdot, \hat{\theta})$ with $\hat{\theta}$ as the MLE of the parameter θ, the standard nonparametric kernel density estimator g of f, and a nonparametric Bayes estimator h of f. One would expect that the asymptotically efficient MLE of a correctly specified parametric model would perform better than its nonparametric competitors. But, surprisingly, a set of 20 simulations, each with a fresh random sample of size 200, show the following average L_1-distances d: (1) $d(h, f_0) = 0.44$, (2) $d(g, f_0) = 1.03$, (3) $d(\hat{f}, f_0) = 0.75$. Although the present study points to the superiority of the Bayes procedure compared to frequentist ones such as the histogram method, the differences do not appear to be that dramatic. Perhaps the method of representing an unknown density as a mixture of an appropriate parametric family and estimating the mixture by Ferguson's Dirichlet process, as used by Bhattacharya and Dunson (2010), should be preferred (also see Ghosh and Ramamoorthi (2003) and Ghosal and van der Vaart (2017)). Still the present research provides a simple and widely applicable Bayes estimation of a nonparametric distribution, which perhaps may be sharpened to be more effective.

Appendix
R Code

We gratefully acknowledge the creators of the packages used here: van Valkenhoef and Tervonen (2016), Wickham (2007), Wickham (2016), Wickham and Seidel (2019), Microsoft Corporation and Weston (2019), Dahl et al. (2019), Harrell Jr (2020), and Warnes (2020).

```
### R version info:

### R version 3.6.3 (2020-02-29)
### Platform: x86_64-w64-mingw32/x64 (64-bit)
### Running under: Windows >= 8 x64 (build 9200)


### Packages required ###
library(hitandrun)
library(reshape2)
library(ggplot2)
library(scales)
library(doParallel)
library(xtable)
library(Hmisc)
library(SASxport)

########################################################
## Section Applying the volume calculations to Bayes estimators


### C=1/k for k=5,10,20 only

set.seed(2019)

klist2<-c(5,10,20)
area1k<-vector("list", length = length(klist2))

for (k in 1:length(klist2)){
  samples1k<-simplex.sample(n=klist2[k],N=10000000)$samples
  nlist2 <- c(2*klist2[k],klist2[k]^2)

  area1k[[k]]<-sapply(nlist2, function(n){sum(riskdiff2(samples1k,n,c=1)>=0)/10000000})
}


areaDF1k<-as.data.frame(do.call(cbind, area1k))

names(areaDF1k)<-paste0("k",klist2)

areaDF1k[["N.ord"]]<-c(2*klist2[1],klist2[1]^2)
```

```
perc1k<-melt(areaDF1k, id.vars = "N.ord", variable.name = "k", value.name = "Percent")

rm(samples1k)

##Estimating areas for 1/k prior

mlearea<-function(k,n){
  r<- (2*n+1+n/k)/(3*n+1)

  k*((r-sqrt(2*r-1))/2)^((k-1)/2)
}

mlearea(5,10)
mlearea(5,25)
mlearea(10,20)
mlearea(10,100)
mlearea(20,40)
mlearea(20,400)

#######################################################
## Section Simulation results for other priors


riskdiff<-function(mat,n){
  k<-ncol(mat)

  return(
    -3*n - k + (n*k+2*n+k)*apply(mat, 1, function(x) sum(x^2))


  )
}

### Fix k
### Let n=k, 2k, 3k, 4k, k^2, 2k^2, 3k^2, 4k^2, k^3, k^4 and find % Bayes is better

set.seed(2018)

klist<-c(10*(1:5),100,200,500)
area<-vector("list", length = length(klist))

for (k in 1:length(klist)){
  samples<-simplex.sample(n=klist[k],N=500000)$samples
  nlist <- c(klist[k]*(1:4), klist[k]^2*(1:4), klist[k]^3*(1:4), klist[k]^4*(1:4))

  area[[k]]<-sapply(nlist, function(n){sum(riskdiff(samples,n)>=0)/5000})
}

areaDF<-as.data.frame(do.call(cbind, area))

names(areaDF)<-paste0("k",klist)

areaDF[["N.ord"]]<-c(klist[1]*(1:4), klist[1]^2*(1:4), klist[1]^3*(1:4), klist[1]^4*(1:4))

perc<-melt(areaDF, id.vars = "N.ord", variable.name = "k", value.name = "Percent")

labs<-paste0("k",c("","^2","^3","^4"))
lab.exp<-parse(text=labs)

pdf("comp1.pdf", width = 7, height = 5)
ggplot(perc, aes(N.ord, Percent)) +
```

```
  geom_point(aes(color=k),size =2 ) +
  geom_line(aes(color=k, group=k, linetype=k)) +
  scale_x_log10(breaks = c(10,100,1000,10000),
                labels = lab.exp,
                name = "n") +
  ylab("Percent of samples\n where MLE had smaller risk") +
  scale_color_discrete(labels = klist) +
  scale_linetype_discrete(labels = klist) +
  labs(title = "Percentage of the parameter space where MLE had lower risk than Bayes
       as a function of sample size (n) and colored by number of classes (k)",
       subtitle = "With the prior Dir(1,1,...,1)") +
  theme_bw() +
  guides(colour = guide_legend(override.aes = list(shape = NA)))
dev.off()

rm(samples)

## Redo to get k=5

set.seed(2019)

samples<-simplex.sample(n=5,N=500000)$samples
nlist <- c(5,25)

area5<-sapply(nlist, function(n){sum(riskdiff(samples,n)>=0)/5000})

## Redo computations in case of alpha=c/k

riskdiff2<-function(mat,n,c){
  k<-ncol(mat)

  return(
    -2*n - c - c*n/k + (n*c+2*n+c)*apply(mat, 1, function(x) sum(x^2))


  )
}

### Fix k
### Let n=k, 2k, 3k, 4k, k^2, 2k^2, 3k^2, 4k^2, k^3, k^4 and find % Bayes is better

set.seed(2018)

klist<-c(10*(1:5),100,200,500)
area2<-vector("list", length = length(klist))

for (k in 1:length(klist)){
  samples2<-simplex.sample(n=klist[k],N=500000)$samples
  nlist <- c(klist[k]*(1:4), klist[k]^2*(1:4), klist[k]^3*(1:4), klist[k]^4*(1:4))

  area2[[k]]<-sapply(nlist, function(n){sum(riskdiff2(samples2,n,c=30)>=0)/5000})
}
# If I run with c=1, 0% for all k,n combos

areaDF2<-as.data.frame(do.call(cbind, area2))

names(areaDF2)<-paste0("k",klist)

areaDF2[["N.ord"]]<-c(klist[1]*(1:4), klist[1]^2*(1:4), klist[1]^3*(1:4), klist[1]^4*(1:4))

perc2<-melt(areaDF2, id.vars = "N.ord", variable.name = "k", value.name = "Percent")
```

```
labs<-paste0("k",c("","^2","^3","^4"))
lab.exp<-parse(text=labs)

pdf("compC30-k.pdf", width = 7, height = 5)
ggplot(perc2[1:48,], aes(N.ord, Percent)) +
  geom_point(aes(color=k),size =2 ) +
  geom_line(aes(color=k, group=k, linetype=k)) +
  scale_x_log10(breaks = c(10,100,1000,10000),
                labels = lab.exp,
                name = "n") +
  ylab("Percent of samples\n where MLE had smaller risk") +
  scale_color_discrete(labels = klist[1:3]) +
  scale_linetype_discrete(labels = klist[1:3]) +
  labs(title = "Percentage of the parameter space where MLE had lower risk than Bayes
       as a function of sample size (n) and colored by number of classes (k)",
       subtitle = expression(
         'With the prior Dir'*bgroup("(",list(frac(30,k),frac(30,k),...,frac(30,k)),")"))) +
  theme_bw() +
  guides(colour = guide_legend(override.aes = list(shape = NA)))
dev.off()

rm(samples2)

## Redo computations in case of alpha = C

riskdiff3<-function(mat,n,c){
  k<-ncol(mat)

  return(
    -2*n - c*k - c*n + (n*c*k+2*n+c*k)*apply(mat, 1, function(x) sum(x^2))


  )
}

### Fix k
### Let n=k, 2k, 3k, 4k, k^2, 2k^2, 3k^2, 4k^2, k^3, k^4 and find % Bayes is better

## C=5
set.seed(711)

klist<-c(10*(1:5),100,200,500)
area3<-vector("list", length = length(klist))

for (k in 1:length(klist)){
  samples3<-simplex.sample(n=klist[k],N=500000)$samples
  nlist <- c(klist[k]*(1:4), klist[k]^2*(1:4), klist[k]^3*(1:4), klist[k]^4*(1:4))

  area3[[k]]<-sapply(nlist, function(n){sum(riskdiff3(samples3,n,c=5)>=0)/5000})
}


areaDF3<-as.data.frame(do.call(cbind, area3))

names(areaDF3)<-paste0("k",klist)

areaDF3[["N.ord"]]<-c(klist[1]*(1:4), klist[1]^2*(1:4), klist[1]^3*(1:4), klist[1]^4*(1:4))


perc3<-melt(areaDF3, id.vars = "N.ord", variable.name = "k", value.name = "Percent")

labs<-paste0("k",c("","^2","^3","^4"))
```

```
lab.exp<-parse(text=labs)

pdf("compC5.pdf", width = 7, height = 5)
ggplot(perc3, aes(N.ord, Percent)) +
  geom_point(aes(color=k),size =2 ) +
  geom_line(aes(color=k, group=k)) +
  scale_x_log10(breaks = c(10,100,1000,10000),
                labels = lab.exp,
                name = "n") +
  ylab("Percent of samples\n where MLE had smaller risk") +
  scale_color_discrete(labels = klist) +
  labs(title = "Percentage of the parameter space where MLE had lower risk than Bayes
       as a function of sample size (n) and colored by number of classes (k)",
       subtitle = "With the prior Dir(5,5,...,5)") +
  theme_bw()
dev.off()

rm(samples3)

## C=2
set.seed(711)

klist<-c(10*(1:5),100,200,500)
area6<-vector("list", length = length(klist))

for (k in 1:length(klist)){
  samples6<-simplex.sample(n=klist[k],N=500000)$samples
  nlist <- c(klist[k]*(1:4), klist[k]^2*(1:4), klist[k]^3*(1:4), klist[k]^4*(1:4))

  area6[[k]]<-sapply(nlist, function(n){sum(riskdiff3(samples6,n,c=2)>=0)/5000})
}


areaDF6<-as.data.frame(do.call(cbind, area6))

names(areaDF6)<-paste0("k",klist)

areaDF6[["N.ord"]]<-c(klist[1]*(1:4), klist[1]^2*(1:4), klist[1]^3*(1:4), klist[1]^4*(1:4))


perc6<-melt(areaDF6, id.vars = "N.ord", variable.name = "k", value.name = "Percent")


labs<-paste0("k",c("","^2","^3","^4"))
lab.exp<-parse(text=labs)

pdf("compC2.pdf", width = 7, height = 5)
ggplot(perc6, aes(N.ord, Percent)) +
  geom_point(aes(color=k),size =2 ) +
  geom_line(aes(color=k, group=k, linetype = k)) +
  scale_x_log10(breaks = c(10,100,1000,10000),
                labels = lab.exp,
                name = "n") +
  ylab("Percent of samples\n where MLE had smaller risk") +
  scale_color_discrete(labels = klist) +
  scale_linetype_discrete(labels = klist) +
  labs(title = "Percentage of the parameter space where MLE had lower risk than Bayes
       as a function of sample size (n) and colored by number of classes (k)",
       subtitle = "With the prior Dir(2,2,...,2)") +
  theme_bw() +
  guides(colour = guide_legend(override.aes = list(shape = NA)))
dev.off()
```

```
rm(samples6)

## C=3
set.seed(616)

klist<-c(10*(1:5),100,200,500)
area4<-vector("list", length = length(klist))

for (k in 1:length(klist)){
  samples4<-simplex.sample(n=klist[k],N=500000)$samples
  nlist <- c(klist[k]*(1:4), klist[k]^2*(1:4), klist[k]^3*(1:4), klist[k]^4*(1:4))

  area4[[k]]<-sapply(nlist, function(n){sum(riskdiff3(samples4,n,c=3)>=0)/5000})
}


areaDF4<-as.data.frame(do.call(cbind, area4))

names(areaDF4)<-paste0("k",klist)

areaDF4[["N.ord"]]<-c(klist[1]*(1:4), klist[1]^2*(1:4), klist[1]^3*(1:4), klist[1]^4*(1:4))

library(reshape2)

perc4<-melt(areaDF4, id.vars = "N.ord", variable.name = "k", value.name = "Percent")

labs<-paste0("k",c("","^2","^3","^4"))
lab.exp<-parse(text=labs)

pdf("compC3.pdf", width = 7, height = 5)
ggplot(perc4, aes(N.ord, Percent)) +
  geom_point(aes(color=k),size =2 ) +
  geom_line(aes(color=k, group=k, linetype = k)) +
  scale_x_log10(breaks = c(10,100,1000,10000),
                labels = lab.exp,
                name = "n") +
  ylab("Percent of samples\n where MLE had smaller risk") +
  scale_color_discrete(labels = klist) +
  scale_linetype_discrete(labels = klist) +
  labs(title = "Percentage of the parameter space where MLE had lower risk than Bayes
       as a function of sample size (n) and colored by number of classes (k)",
       subtitle = "With the prior Dir(3,3,...,3)")+
  theme_bw()+
  guides(colour = guide_legend(override.aes = list(shape = NA)))
dev.off()

rm(samples4)

## C=1.9
set.seed(411)

klist<-c(10*(1:5),100,200,500)
area5<-vector("list", length = length(klist))

for (k in 1:length(klist)){
  samples5<-simplex.sample(n=klist[k],N=500000)$samples
  nlist <- c(klist[k]*(1:4), klist[k]^2*(1:4), klist[k]^3*(1:4), klist[k]^4*(1:4))

  area5[[k]]<-sapply(nlist, function(n){sum(riskdiff3(samples5,n,c=1.9)>=0)/5000})
}
```

```
areaDF5<-as.data.frame(do.call(cbind, area5))

names(areaDF5)<-paste0("k",klist)

areaDF5[["N.ord"]]<-c(klist[1]*(1:4), klist[1]^2*(1:4), klist[1]^3*(1:4), klist[1]^4*(1:4))

perc5<-melt(areaDF5, id.vars = "N.ord", variable.name = "k", value.name = "Percent")

labs<-paste0("k",c("","^2","^3","^4"))
lab.exp<-parse(text=labs)

pdf("compC1-9.pdf", width = 7, height = 5)
ggplot(perc5, aes(N.ord, Percent)) +
  geom_point(aes(color=k),size =2 ) +
  geom_line(aes(color=k, group=k, linetype=k)) +
  scale_x_log10(breaks = c(10,100,1000,10000),
                labels = lab.exp,
                name = "n") +
  ylab("Percent of samples\n where MLE had smaller risk") +
  scale_color_discrete(labels = klist) +
  scale_linetype_discrete(labels = klist) +
  labs(title = "Percentage of the parameter space where MLE had lower risk than Bayes
       as a function of sample size (n) and colored by number of classes (k)",
       subtitle = "With the prior Dir(1.9,1.9,...,1.9)") +
  theme_bw()+
  guides(colour = guide_legend(override.aes = list(shape = NA)))
dev.off()

rm(samples5)

#######################################################
## Chapter Average risk across the parameter space

## Plotting average risk improvement as a function of ckn

avgriskdiff<-function(k,n,ckn){
  1 - (n*(k*ckn^2+n))/(n+k*ckn)^2
}

pdf("avgrisk.pdf", width = 7, height = 5)
ggplot(data.frame(ckn = c(0,3)), aes(x = ckn)) +
  stat_function(fun = avgriskdiff, args = list(k=10,n=30), col = "black") +
  scale_x_continuous(breaks = c(0,1),
                     name = expression(C["k,n"])) +
  scale_y_continuous(limits = c(0,.3),
                     breaks = c(0,.25),
                     labels = parse(text = c("0","frac(k,k+n)")),
                     name = "Decrease in Avg Risk\n (Proportional to MLE Risk)") +
  labs(title = "Decrease in Average Risk by Using Bayes Estimator Instead of MLE",
       subtitle = expression('With the Prior Dir'(C["k,n"],C["k,n"]*',...,'*C["k,n"]))) +
  theme(panel.background = element_rect(fill = "white",
                                        color = "black",
                                        size = 0.5,
                                        linetype = "solid"),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())
dev.off()

avgriskdiff(10,40,2)
avgriskdiff(100,400,2)
```

```
unif.ard<-unlist(lapply(c(5,10,50,100), function(k){c(avgriskdiff(k,k,1),avgriskdiff(k,k^2,1))}))
p1k.ard<-unlist(lapply(c(5,10,50,100), function(k){c(avgriskdiff(k,k,1/k),avgriskdiff(k,k^2,1/k))}))

mleavg<-function(k,n){
  (k-1)/(n*(k+1))
}

mle.ar<-unlist(lapply(c(5,10,50,100), function(k){c(mleavg(k,k),mleavg(k,k^2))}))

kvals<-c(5,10,50,100)
avgtable<-data.frame(k=rep(kvals,each = 2),n=c(rbind(kvals,kvals^2)),
                     mle.ar,
                     unif.ard=100*unif.ard,
                     uniformprop,
                     p1k.ard=100*p1k.ard,
                     prop1k)

##Get proportions where bayes is better for table
##Requires areaDF from uniform computation and area5 just below it, as well as mlearea above

uniformprop<-1-c(area5,
                 areaDF$k10[c(1,5)],
                 areaDF$k50[c(1,5)],
                 areaDF$k100[c(1,5)])/100

prop1k<-1-unlist(lapply(c(5,10,50,100), function(k){c(mlearea(k,k),mlearea(k,k^2))}))

#######################################################
## Section Simulated L1 distances

##### Simulation of TV Distances between distributions #####

ins.mle <- function(j,r,theta,n){
  choose(n,r)*theta[j]^r*(1-theta[j])^(n-r)*abs(r/n-theta[j])
}
ins.b <- function(j,r,theta,n,k){
  choose(n,r)*theta[j]^r*(1-theta[j])^(n-r)*abs(qbeta(.5,r+(1/k),n-r+1-(1/k))-theta[j])
}

rf.mle2 <- function(x,n) {
  k <- length(x)
  sum( sapply(1:k, function(j){
    sum(sapply(0:n, function(r){ins.mle(j=j,r=r,theta=x,n=n)} )) }) )
}
rf.b2 <-   function(x,n) {
  k <- length(x)
  sum( sapply(1:k, function(j){
    sum(sapply(0:n, function(r){ins.b(j=j,r=r,theta=x,n=n,k=k)} )) }) )
}

klist3<-c(10*(1:5),100,200,500)
nlist <- c(10*(2:5),100*(1:10))


numCores<-detectCores()

cl<-makeCluster(11)
registerDoParallel(cl)

longcomp<-function(i){

  avg.risk1K3<-vector("list", length = length(klist3))
```

```
  avg.riskMLE3<-vector("list", length = length(klist3))

  for (k in 1:length(klist3)){

    samples<-simplex.sample(n=klist3[k],N=100)$samples


    avg.risk1K3[[k]] <- sapply(nlist, function(n){
      sum(apply(samples,1,function(x) {rf.b2(x=x,n=n)}))/100
    })
    avg.riskMLE3[[k]] <- sapply(nlist, function(n){
      sum(apply(samples,1,function(x) {rf.mle2(x=x,n=n)}))/100
    })

  }
  return(list(avg.risk1K3,avg.riskMLE3))
}

trials<-1000

set.seed(616)
system.time(
  results<-foreach(k=klist3) %:% foreach(i=1:100,
                                         .packages="hitandrun") %dopar% {


                                           samples<-simplex.sample(n=k,N=100)$samples


                                           avg.risk1K3<- sapply(nlist, function(n){
                                             sum(apply(samples,1,function(x) {rf.b2(x=x,n=n)}))/100
                                           })
                                           avg.riskMLE3 <- sapply(nlist, function(n){
                                             sum(apply(samples,1,function(x) {rf.mle2(x=x,n=n)}))/100
                                           })

                                           data.frame(ar.1k=avg.risk1K3,ar.mle=avg.riskMLE3)
                                         }
)
stopCluster(cl)

meanresults<-data.frame(k=rep(klist3, each = 14),
                        n = rep(nlist, times = 8),
                        ar.1k = rep(NA, 8*14),
                        ar.mle = rep(NA, 8*14))

for (k in 1:8){
  for (n in 1:14){
    meanresults[(k-1)*14+n,3] <-
      mean(sapply(results[[k]], function(lst){lst[n,1]}))
    meanresults[(k-1)*14+n,4] <-
      mean(sapply(results[[k]], function(lst){lst[n,2]}))
  }
}

tables3<-vector("list", length = length(klist3))

for (k in 1:length(klist3)){
  tables3[[k]]<-data.frame(n = paste0("$n=",nlist,"$"),
                           MLE=meanresults[meanresults$k==klist3[k],]$ar.mle,
                           Bayes=meanresults[meanresults$k==klist3[k],]$ar.1k
  )
```

```
}

for (k in 1:length(klist3)){
  tab<-xtable(tables3[[k]],auto = TRUE, display = c(rep("s",2),rep("f",2)),
              digits = 4)
  colnames(tab)<-c("$\\boldsymbol n$","{\\bf MLE}","{\\bf Bayes}")

  print(tab, math.style.exponents = TRUE, include.rownames = FALSE,
        sanitize.text.function=function(str) gsub("%","\\%",str, fixed = TRUE))
}

#######################################################
## Chapter Data example

lookup.xport("Data/BMX_I.XPT")
lookup.xport("Data/DEMO_I.XPT")

## Read in data
demo<-read.xport("Data/DEMO_I.XPT")
bmx<-read.xport("Data/BMX_I.XPT")


#Remove pregnant (RIDEXPRG=1)
#Keep females (RIAGENDR=2)
#Keep age 18+ (RIAGEYR>=18)

target<-subset(demo, RIAGENDR==2 & RIDAGEYR>=18 & !(RIDEXPRG %in% 1),
               select = c(SEQN,WTMEC2YR))


table(demo$RIDEXPRG,useNA = "ifany")
#Note that only 70 were reported as pregnant; most are missing

# summary(bmx$BMXWAIST)
# summary(bmx$BMXHT)

waistheight<-bmx[complete.cases(bmx[,c("BMXWAIST","BMXHT")]),]
#1245 were missing one of these measurements

pt<-intersect(target$SEQN,waistheight$SEQN)
#returns the 2697 from "target" that are in "waistheight"


# length(intersect(target$SEQN,bmx$SEQN))
#There were 2906 in target and bmx
#209 did not have waist and height measurements: this may affect weighting a little


adultF<-merge(target,waistheight,by="SEQN")

waistbreaks<-c(-Inf,64,66,69,72,74,77,79,83,87,90,96,97,102.25,107,112.25,117.5,122.5,Inf)
waistlabs<-c(24:34,paste0(2*(8:13),"W"),"NOSIZE")
inseambreaks.reg<-c(-Inf,150,160,170,Inf)
inseamlabs.reg<-c(2*(14:17))
## plus starts for waist > 96
inseambreaks.plus<-c(-Inf,160,170,Inf)
inseamlabs.plus<-c("S","M","L")
plus<-paste0(2*(8:13),"W")

adultF["waistsize"]<-cut(adultF$BMXWAIST, breaks = waistbreaks, labels = waistlabs)
```

```
adultF["inseam"]<-rep(NA,nrow(adultF))
adultF$inseam[adultF$waistsize %in% plus]<-as.character(cut(adultF$BMXHT[adultF$waistsize %in% plus],
                                                    breaks = inseambreaks.plus,
                                                    labels = inseamlabs.plus))


adultF$inseam[adultF$waistsize %in% "NOSIZE"]<-"NOSIZE"
adultF$inseam[!(adultF$waistsize %in% c(plus,"NOSIZE"))]<-
  as.character(cut(adultF$BMXHT[!(adultF$waistsize %in% c(plus,"NOSIZE"))], breaks = inseambreaks.reg,
                   labels = inseamlabs.reg))

## 34 doesn't start until waist size 27. Change these to "32"
adultF$inseam<-ifelse(adultF$waistsize %in% c("24","25","26") & adultF$inseam=="34","32",
                      adultF$inseam)

# table(adultF$inseam)

waists<-c(as.character(24:34),plus,"NOSIZE")
adultF$waistsize<-ordered(adultF$waistsize,
                          levels=waists)
inseams<-c(as.character(2*14:17),"S","M","L")
adultF$inseam<-ordered(adultF$inseam,
                       levels=inseams)

# table(adultF$waistsize,adultF$inseam)

adultF["pantsize"]<-paste0(adultF$waistsize,".",adultF$inseam)

# table(adultF$pantsize)

sizes.ord1<-paste(rep(waists[1:3],each=3),inseams[1:3],sep = ".")
sizes.ord2<-paste(rep(waists[4:11],each=4),inseams[1:4],sep = ".")
sizes.ord3<-paste(rep(waists[12:17],each=3),inseams[5:7],sep = ".")

sizes<-c(sizes.ord1,sizes.ord2,sizes.ord3,"NOSIZE")

adultF$pantsize<-ifelse(adultF$pantsize=="NOSIZE.NA","NOSIZE",adultF$pantsize)

adultF$pantsize<-ordered(adultF$pantsize,
                         levels=sizes)


sizeweights<-aggregate(WTMEC2YR ~ pantsize, adultF, sum)
sizeweights["P"]<-sizeweights$WTMEC2YR/sum(sizeweights$WTMEC2YR)



weights0<-data.frame(pantsize=factor(setdiff(sizes,sizeweights$pantsize),ordered=TRUE,
                                     levels = sizes),
                     WTMEC2YR=0,
                     P=0
)

sizeweights<-rbind(sizeweights,weights0)

sizeweights<-sizeweights[order(sizeweights$pantsize),]

## Consider sizeweights$P to be the true distribution of pant sizes among adult women in
## US. This is slightly problematic since one of the sizes has 0 probability.

probs1<-sizeweights$P #Original probabilities with one zero
```

```
## Create new probabilities where the zero one is now equal to half of the smallest nonzero
## probability. This seems reasonable since people of this size probably do exist but were
## missed in the survey.

probs2<-sizeweights$P

newmin<-min(probs2[probs2!=0])/2

probs2[probs2==0]<-newmin
probs2<-probs2/sum(probs2)

#### Testing

size.sim<-function(times,sampsize,prob,dir=1){
  samps<-rmultinom(times, size = sampsize, prob = prob)

  if (length(dir)==1){
    dirichlet<-apply(samps,2,function(x){(x+dir)/(sampsize+length(prob)*dir)})
  }
  else {
    dirichlet<-apply(samps,2,function(x){(x+dir)/(sampsize+sum(dir))})
  }
  mle<-apply(samps,2,function(x){x/sampsize})

  if (length(dir)==1){
    dirabs<-apply(samps,2,function(x){qbeta(.5,x+dir,sampsize+length(prob)*dir-x-dir)})
  }
  else {
    dirabs<-apply(samps,2,function(x){qbeta(.5,x+dir,sampsize+sum(dir)-x-dir)})
  }

  l1.dirichlet<-apply(dirichlet,2,function(x){sum(abs(x-prob))})
  l2.dirichlet<-apply(dirichlet,2,function(x){sqrt(sum((x-prob)^2))})
  max.dirichlet<-apply(dirichlet,2,function(x){max(abs(x-prob))})

  l1.mle<-apply(mle,2,function(x){sum(abs(x-prob))})
  l2.mle<-apply(mle,2,function(x){sqrt(sum((x-prob)^2))})
  max.mle<-apply(mle,2,function(x){max(abs(x-prob))})

  l1.dirabs<-apply(dirabs,2,function(x){sum(abs(x-prob))})
  l2.dirabs<-apply(dirabs,2,function(x){sqrt(sum((x-prob)^2))})
  max.dirabs<-apply(dirabs,2,function(x){max(abs(x-prob))})

  return(list(samps=samps,dir.est=dirichlet, mle=mle,dir.abs=dirabs,
              l1.dir=l1.dirichlet,l2.dir=l2.dirichlet,max.dir=max.dirichlet,
              l1.mle=l1.mle,l2.mle=l2.mle,max.mle=max.mle,
              l1.dirabs=l1.dirabs, l2.dirabs=l2.dirabs,max.dirabs=max.dirabs))

}

set.seed(616)

est<-size.sim(times=1000,sampsize=100,prob=probs2)
sum(est$max.dir<est$max.mle)
## 676
sum(est$l1.dir<est$l1.mle)
## 858
sum(est$l2.dir<est$l2.mle)
## 949

which.max(est$l1.mle-est$l1.dir)
## 419
```

```
sum(est$max.dirabs<est$max.mle)
## 625
sum(est$l1.dirabs<est$l1.mle)
## 956
sum(est$l2.dirabs<est$l2.mle)
## 935


zeroes<-apply(est$samps,2,function(x) sum(x==0))


summary(zeroes)

## make tables


sizetable<-data.frame(sizes=c(sizes[-60],"No size"),
                      p=probs2,
                      stock1=
                        c(round((probs2[-60]/sum(probs2[-60]))*1000,digits = 0),0),
                      stockMLE=
                        c(round((est$mle[-60,419]/sum(est$mle[-60,419]))*1000,
                                digits = 0),0),
                      stockBayes=
                        c(round((est$dir.est[-60,419]/sum(est$dir.est[-60,419]))*1000,
                                digits = 0),0),
                      stockB.abs=c(round((est$dir.abs[-60,419]/sum(est$dir.abs[-60,419]))*1000,
                                         digits = 0),0)

)
sum(sizetable$stock1)
## 1000
sum(sizetable$stockMLE)
## 993
sum(sizetable$stockBayes)
## 1006
sum(sizetable$stockB.abs)
## 985

### Make stock exactly 1000
set.seed(111)

zeroesMLE<-which(sizetable$stockMLE==0)

## purchase 1 pair in 7 of the "zero" sizes to make stockMLE go from 993 to 1000
sizetable$stockMLE[sample(zeroesMLE,7)]<-1

## remove 3 pairs from each of the two largest to make stockBayes go from 1006 to 1000
sizetable$stockBayes[which(sizetable$stockBayes==max(sizetable$stockBayes))]<-
  max(sizetable$stockBayes) - 3

minB.abs<-which(sizetable$stockB.abs==min(sizetable$stockB.abs[-60]))

## add 1 additional pair to 15 of the smallest number sizes to make stockB.abs go from 985 to 1000

sizetable$stockB.abs[sample(minB.abs,15)]<-min(sizetable$stockB.abs[-60]) + 1

######

sizetab<-xtable(sizetable,auto=TRUE,display = c(rep("s",2),"e",rep("d",4)),
```

```
                digits = c(0,0,3,0,0,0,0), caption = c(
                  "The true size distribution based on NHANES 15-16 as well as the stock of 1000
                    based on the true distribution and estimators from a sample of size 100.",
                  "Size distribution (NHANES) and stock"),
                label = "tab:jeans")
colnames(sizetab)<-c("Size (Waist.Inseam)",
                     "True $p$",
                     "Stock (true)",
                     "Stock (MLE)",
                     "Stock (Bayes, $L_2$ Loss)",
                     "Stock (Bayes, $L_1$ Loss)")
print(sizetab, sanitize.text.function = identity, include.rownames = FALSE,
      latex.environments = "center",
      size = 'footnotesize')

#L2 distances for this sample
round(est$l2.mle[419],3)
## 0.122
round(est$l2.dir[419],3)
## 0.077
round(est$l2.dirabs[419],3)
## 0.079

round(est$l1.mle[419],3)
## 0.643
round(est$l1.dir[419],3)
## 0.442
round(est$l1.dirabs[419],3)
## 0.419

## stock errors for the sizetable sample
sum(abs(sizetable$stock1-sizetable$stockMLE)) #682
sum(abs(sizetable$stock1-sizetable$stockBayes)) #486
sum(abs(sizetable$stock1-sizetable$stockB.abs)) #494

disttab<-data.frame(l1=c(mean(est$l1.dir),mean(est$l1.dirabs),mean(est$l1.mle)),
                    l2=c(mean(est$l2.dir),mean(est$l2.dirabs),mean(est$l2.mle)),
                    maxn=c(mean(est$max.dir),mean(est$max.dirabs),mean(est$max.mle)))
distx<-xtable(disttab, auto = TRUE, digits = 4, caption = c(
  "The mean $L_1$, $L_2$, and infinity (maximum) distances between the estimators and the true
    size distribution in 1000 simulations.",
  "Distance between estimators and true probability"),
  label = "tab:dist")
colnames(distx)<-c("$L_1$","$L_2$",
                   "Infinity (maximum)")
rownames(distx)<-c("Bayes estimator ($L_2$ Loss)","Bayes estimator ($L_1$ Loss)","MLE")
print(distx, sanitize.text.function = identity, latex.environments = "center")


## informative prior

sizes.tab<-data.frame(
  waist = c(rep(waists[1:3],each=3),rep(waists[4:11],each=4),rep(waists[12:17],each=3),"NOSIZE"),
  inseam = c(rep(inseams[1:3],times = 3),rep(inseams[1:4],times = 8),
                      rep(inseams[5:7],times = 6),"NOSIZE")
)

waistrange<-data.frame(
  waist = waists,
  min.waist = waistbreaks[-19],
  max.waist = c(waistbreaks[-1])
)
```

```
inseamrange<-data.frame(
  inseam = c(inseams,"NOSIZE"),
  min.inseam = c(inseambreaks.reg[-5],inseambreaks.plus[-4],-Inf),
  max.inseam = c(inseambreaks.reg[-1],inseambreaks.plus[-1],Inf)
)

sizes.tab<-merge(sizes.tab,waistrange,by = "waist", sort = FALSE)

sizes.tab<-merge(sizes.tab,inseamrange, by = "inseam", sort = FALSE)

sizes.tab$max.inseam[sizes.tab$max.waist <= 69 & sizes.tab$min.inseam >= 160]<-Inf

sizes.tab$waist<-ordered(sizes.tab$waist, levels = waists)
sizes.tab$inseam<-ordered(sizes.tab$inseam, levels = inseams)

sizes.tab<-sizes.tab[order(sizes.tab$waist,sizes.tab$inseam),c(2,1,3:6)]

sizes.tab2<-sizes.tab

sizes.tab2$min.inseam[sizes.tab2$min.inseam<0]<-120
sizes.tab2$min.waist[sizes.tab2$min.waist<0]<-50

sizes.tab2$max.inseam[sizes.tab2$max.inseam>200]<-195
sizes.tab2$max.waist[sizes.tab2$max.waist>130]<-140

sizes.tab2["label"]<-paste(sizes.tab2$waist,sizes.tab2$inseam, sep = ".")
sizes.tab2$label[60]<-"NOSIZE"

sizes.tab2["plus"]<-ifelse(sizes.tab2$min.waist < 96, "Regular",
                           ifelse(sizes.tab2$max.waist <= 122.5, "Plus", "No size"))

sizes.tab2["groups"]<-rep(NA,60)
sizes.tab2[sizes.tab2$inseam %in% c("28"),]$groups<-c(rep(1,8),2,2,2)
sizes.tab2[sizes.tab2$inseam %in% c("30"),]$groups<-c(rep(1,3),rep(2,3),rep(3,5))
sizes.tab2[sizes.tab2$inseam %in% c("32"),]$groups<-c(rep(1,3),rep(2,3),rep(3,2), rep(4,3))
sizes.tab2[sizes.tab2$inseam %in% c("34"),]$groups<-2
sizes.tab2[sizes.tab2$inseam %in% c("S"),]$groups<-c(rep(3,2),rep(2,1),rep(1,3))
sizes.tab2[sizes.tab2$inseam %in% c("M"),]$groups<-c(rep(3,2),rep(2,1),rep(1,3))
sizes.tab2[sizes.tab2$inseam %in% c("L"),]$groups<-c(rep(2,3),rep(1,3))
sizes.tab2[sizes.tab2$waist == "NOSIZE",]$groups<-2

sizes.tab2["p"]<-ifelse(sizes.tab2$groups==4,1.9,
                        ifelse(sizes.tab2$groups==3,1.6,
                               ifelse(sizes.tab2$groups==2,.9,.7)))

my_cols<-c("#FEF0D9","#FDCC8A","#FC8D59","#D7301F")
pdf("C:/Users/Rachel Oliver/Documents/Research/Multinomial Article/Submission/sizes.pdf",
    width = 12, height = 5)
ggplot() +
  scale_x_continuous(name = "Waist (cm)") +
  scale_y_continuous(name = "Height (cm)") +
  geom_rect(data = sizes.tab2, mapping = aes(xmin = min.waist, xmax = max.waist,
                                             ymin = min.inseam, ymax = max.inseam,
                                             fill = factor(groups)),
            color = "black", alpha = .5) +
  geom_text(data = sizes.tab2, aes(x = min.waist + (max.waist - min.waist)/2,
                                   y = min.inseam + (max.inseam - min.inseam)/2,
                                   label = label), size = 3, angle = 90) +
  scale_fill_manual(values = my_cols) +
  labs(title = "Size Chart",
       subtitle = "(Outer boundaries are estimated)",
```

```
       fill = "Group")+
  theme_bw() +
  guides(colour = guide_legend(override.aes = list(shape = NA)))
dev.off()

## Testing

set.seed(616)

est2<-size.sim(times=1000,sampsize=100,prob=probs2,dir = sizes.tab2$p)

sum(est$samps!=est2$samps)
# 0

sum(est2$max.dir<est2$max.mle)
## 763
sum(est2$l1.dir<est2$l1.mle)
## 980
sum(est2$l2.dir<est2$l2.mle)
## 987

mean(est2$l1.dir)
## 0.4198473
mean(est2$l2.dir)
## 0.07723804
mean(est2$max.dir)
## 0.03624882

sum(est$max.dir>est2$max.dir)
## 476
sum(est$l1.dir>est2$l1.dir)
## 999
sum(est$l2.dir>est2$l2.dir)
## 977


stockInf<-c(round((est2$dir.est[-60,419]/sum(est2$dir.est[-60,419]))*1000,
                  digits = 0),0)
sum(abs(sizetable$stock1-stockInf))
## 428
```

References

A. Bhattacharya and D. Dunson. Nonparametric Bayesian density estimation on manifolds with applications to planar shapes. *Biometrika*, 97:851–865, 2010.

R. Bhattacharya and R. Oliver. Superiority of Bayes estimators over the MLE in high dimensional multinomial models and its implication for nonparametric bayes theory. *Computational Statistics & Data Analysis*, 150, 2020. doi: 10.1016/j.csda.2020.107011.

R. N. Bhattacharya, L. Lin, and V. Patrangenaru. *A Course in Mathematical Statistics and Large Sample Theory*. Springer Texts in Statistics, Springer, 2016.

P. Bickel and K. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*. Number v. 1 in Holden-Day series in probability and statistics. Prentice Hall, 2001.

CDC. National health and nutrition examination survey 2015-16 [data files], 2018. URL `https://wwwn.cdc.gov/nchs/nhanes/`.

D. B. Dahl, D. Scott, C. Roosen, A. Magnusson, and J. Swinton. *xtable: Export Tables to LaTeX or HTML*, 2019. URL `https://CRAN.R-project.org/package=xtable`. R package version 1.8-4.

A. DasGupta. *Asymptotic theory of statistics and probability*. Springer Texts in Statistics. Springer-Verlag, 2008.

T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

S. Ghosal and A. van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2017.

J. K. Ghosh and R. V. Ramamoorthi. *Bayesian nonparametrics*. Springer, 2003.

F. E. Harrell Jr. *Hmisc: Harrell Miscellaneous*, 2020. URL `https://CRAN.R-project.org/package=Hmisc`. R package version 4.4-0.

Levi's. Levi's size chart and guide, 2019. URL `https://www.levi.com/US/en_US/cms/sizeguide`.

Microsoft Corporation and S. Weston. *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*, 2019. URL `https://CRAN.R-project.org/package=doParallel`. R package version 1.0.15.

A. Shendruk. Women's clothing retailers are still ignoring the reality of size in the US. *Quartzy*, Jun 2018. URL `https://qz.com/quartzy/1309509/we-analyzed-750-pairs-of-jeans-and-found-definitive-skinny-bias-by-us-retailers/`.

G. van Valkenhoef and T. Tervonen. *hitandrun: "Hit and Run" and "Shake and Bake" for Sampling Uniformly from Convex Shapes*, 2016. URL `https://CRAN.R-project.org/package=hitandrun`. R package version 0.5-3.

G. R. Warnes. *SASxport: Read and Write 'SAS' 'XPORT' Files*, 2020. URL `https://CRAN.R-project.org/package=SASxport`. R package version 1.7.0.

H. Wickham. Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12):1–20, 2007. URL `http://www.jstatsoft.org/v21/i12/`.

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `https://ggplot2.tidyverse.org`.

H. Wickham and D. Seidel. *scales: Scale Functions for Visualization*, 2019. URL `https://CRAN.R-project.org/package=scales`. R package version 1.1.0.