Bayesian Nonparametric Models for Ranked Set Sampling

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the
Graduate School of The Ohio State University

By

Nader M. Gemayel, B.A.
Graduate Program in Statistics

The Ohio State University

2010

Dissertation Committee:

Douglas A. Wolfe, Co-Adviser
Elizabeth A. Stasny, Co-Adviser
Steven N. MacEachern

Abstract

Ranked Set Sampling (RSS) is a data collection technique that combines measurement with judgment ranking for statistical inference. After a brief review of the basics of RSS, this dissertation lays out a formal and natural Bayesian framework for RSS that is analogous to its frequentist justification, and that does not require the assumption of perfect ranking or use of any imperfect ranking models. Prior beliefs about the judgment order statistic distributions and their interdependence are embodied by a nonparametric prior distribution. Posterior inference is carried out by means of Markov Chain Monte Carlo (MCMC) techniques, and yields estimators of the judgment order statistic distributions (and of functionals of those distributions). Because of non-conjugacy, different MCMC algorithms are used for continuous and discrete data. Judgment post-stratification is introduced to answer questions about handling information from multiple rankers, the quality of judgment ranking, and the role of set size. Finally, a more specific model is proposed for RSS with judgment ranking via a concomitant variable.

Dedication

To my parents, Nina and Michel, for their love and support

Acknowledgments

I am extremely indebted to my co-advisers, Professors Elizabeth Stasny and Doug Wolfe, for their patience, encouragement, and mentorship, ever since we embarked on this journey which began as a random walk through an infinite-dimensional research space, and culminated in the results presented in this dissertation. I am also grateful to Professor Steve MacEachern, who guided me through some critical junctures in the research process and introduced me to nonparametric Bayesian methods.

Vita

2004	B.A. Statistics, American University of Beirut
2004 - 2006	Graduate Teaching Assistant, Department of Statistics, The Ohio State University
2006 - Present	Graduate Research Assistant, Department of Statistics, The Ohio State University

Publications

Optimal Ranked Set Sampling Estimation Based on Medians from Multiple Set Sizes, with E. A. Stasny and D. A. Wolfe, Journal of Nonparametric Statistics, in press.

Fields of Study

Major Field: Statistics

Contents

Page

Chapters:

List of Figures

Figure **Page**

List of Tables

Chapter 1

An Unorthodox Introduction to Ranked Set Sampling

1.1 Motivation: Tallying Termites in Trenches is Tedious.

Reticulitermes flavipes, better known as the Eastern Subterranean Termite, is a common economic pest in North America. In the early 1980's, the Southern Forest Experiment Station, a branch of the U.S. Forest Service, initiated research on this and other *Reticulitermes* species in southern Mississippi. An important step in developing a fuller understanding of termite ecology was estimating the mean number of termites in a mature *R. flavipes* colony. Howard et al. [1982] describe in great detail the painstaking procedure of estimating the number of termites in a single colony:

> A trench (30 cm wide by 122 cm deep) was dug with a tractor-mounted backhoe at a radius of ca. 20 m completely around the colony site to constrain both emigration of the colony's foraging population and immigration of termites from other colonies. The log and associated stump were assumed to constitute the epicenter of the colony. Any logs or other large cellulose debris without termites were placed outside the trench. As the trench was dug, the log portion of the colony was sawed into ca. 25-cm lengths, and each was examined for termites. Segments containing termites were placed in sequence into garbage cans and taken to the laboratory. Segments without termites were removed from the site. The stump and taproot associated with the log then were extracted from the soil by severing lateral roots with an axe and digging out the stump/taproot with the tractor-mounted backhoe. This wood was sawed into ca. 25-cm segments and processed as above.

> To sample termites remaining in the soil, a double layer of corrugated cardboard, 60 cm wide by 2 cm deep, was laid on the ground covering the previous site of the log and stump. The cardboard then was moistened and covered with black polyethylene plastic and held in place with small amounts of soil at the edges. One to 3 weeks later, the cardboard was examined for termites, and penetration sites were noted. The cardboard from each penetration site was taken to the laboratory. Uninfested cardboard was discarded. Fresh sections of cardboard, 60 cm wide by 5 cm thick, were placed over only the active infestation sites, covered with black plastic, and completely buried with soil. This cardboard was examined weekly and replaced with new cardboard until termite infestation either ceased or declined to negligible levels (500 or fewer termites). The logs and stump/taproots were sampled from 23 March to 12 April 1981; the cardboard was sampled from 16 April to 12 June.
>
> Termites were exhaustively extracted by splitting the wood along growth rings or by pulling layers of cardboard apart. Extraneous debris was removed and the termites from each garbage can were weighed. Representative subsamples (n = 3 to 16) from each colony then were reweighed and counted. The resulting average numbers of termites per gram were used to estimate the numbers of termites in each colony.

If estimating the number of termites in a *single* colony is so tedious, costly, and time-consuming, what hope is there of estimating the *average* number of termites in a mature *R. flavipes* colony? Obviously, measuring a large random sample of colonies is not feasible. Since the standard error of the sample mean for a Simple Random Sample (SRS) varies in inverse proportion to the square root of the sample size, smaller samples will produce estimators marred by a higher margin of error. However, even a small sample can be made less "random" and more "representative" by allocating its units across the spectrum of the population.

In this example, the experimenters used a method first proposed by McIntyre [1952] (in an agricultural context involving pasture yields) that had lain dormant until it was put to use again by Halls and Dell [1966] for "estimating weights of browse and herbage in a pine-hardwood forest of east Texas." Howard et al. identified 18 mature *R. flavipes* colonies in the study site, and allocated them at random into 6 sets, each containing 3 colonies. The members of each set were ranked visually by estimated size as "small," "medium," or "large." "The ranking of each of the three colonies within a set was done on the same day, to facilitate visual comparisons of termite numbers." In two sets, those two colonies labeled "small" were designated as sampling units and chosen for measurement. In another two sets, the two colonies labeled "medium" were chosen for measurement. In the remaining two sets, the two colonies deemed "large" were chosen for measurement. The 6 selected colonies were measured as described above. The remaining 12 colonies were not measured. Howard et al. provide the data in Table 1 of their paper. This method is called Ranked Set Sampling (RSS), for entirely conspicuous reasons.

Positing a population of mature *R. flavipes* colonies, the researchers were interested in estimating the population mean number of termites in such colonies, μ. Their estimate was the RSS sample mean, about 244,445 termites per colony, merely the average of the 6 sample measurements. It will be shown in Section 1.3 of this chapter that the RSS mean is an unbiased estimator of the population mean. The standard error of the RSS mean, that is, the square root of its variance, is unknown, but an unbiased estimate of the variance of the sample mean may be calculated, and its square root is 53,901.69. (Howard et al. incorrectly give a standard error for the RSS mean of 53,156, which is the standard error of the SRS mean for a SRS of size 6 consisting of these data.) Needless to say, the population variance σ^2 is also unknown, but an unbiased estimate (see Section 1.5) is 17,033,026,187 (square root is 130,510.6 termites).

1.2 Judgment Ranking

The sampling scheme described in the previous section is known as *balanced RSS*, since an equal number of measurement units is allocated to the judgment ranks "small," "medium," and "large." The terms *judgment ranks* and *judgment order statistics* are used to emphasize that these are not necessarily perfect rankings, and to distinguish them from order statistics. In RSS, judgment ranking is usually done visually (by a field expert, say), or via a concomitant variable. Typical set sizes used in RSS are in the range 2 - 5, since small sets are fairly easy to rank effectively, but larger set sizes can also be considered if they do not hinder judgment ranking.

If judgment ranking is the least bit accurate, then measurements for units assigned to different judgment ranks are not identically distributed. That is, if the judgment ranking mechanism can distinguish effectively between population units, then the hypothetical population of measurements for units assigned judgment rank "small," say, will differ from that of those units deemed "medium" or "large," and the same may be said, *mutatis mutandis*, for the other judgment ranks.

To further develop these insights into the statistical properties of judgment order statistics, suppose that the sets under consideration are all of size K, and that the underlying population has CDF F. Then there are K judgment order statistic distributions, $F_{[1]}$, ..., $F_{[K]}$. For each r, $F_{[r]}$ may be interpreted as the conditional distribution of the measurement Y from a hypothetical population unit given that it was assigned judgment rank r in a set of size K. Intuitively, it is easy to see that these judgment order statistic distributions may run the gamut from being the distributions of the K order statistics (i.e., $F_{[r]} = F_{(r)}$, for each r) in the ideal case when judgment ranking is always perfect, to being indistinguishable from the general population (i.e.,

$F_{[r]} = F$, for each r) when judgment ranking is no better than random. In practice, they will usually lie somewhere in between the two extremes.

Consider a generic SRS Y_1, ..., Y_K from the CDF F. These are i.i.d. draws from F. Suppose these K units are judgment ranked to obtain the ranked set $Y_{[1]}$, ..., $Y_{[K]}$. The members of the ranked set are obviously no longer independent. Nor are they identically distributed (unless judgment ranking is no better than random), for the specialized information implies that $Y_{[r]}$ is distributed according to $F_{[r]}$, $r = 1$, ..., K.

The following argument, inspired by Stokes [1980a], relates the judgment order statistic distribution functions to the parent distribution. Let t be a real number, and define $W_r = \mathbf{1}\left(Y_r \leq t\right)$ and $W_{[r]} = \mathbf{1}\left(Y_{[r]} \leq t\right)$, $r = 1$, ..., K, where $\mathbf{1}\left(\cdot\right)$ is the indicator function. It is easy to see that $\sum_{r=1}^{K} W_r = \sum_{r=1}^{K} W_{[r]}$, since both sides of the equation count the number of Y's no greater than t. Taking expectations of both sides and dividing throughout by K gives

$$F\left(t\right) = \frac{1}{K}\sum_{r=1}^{K} F_{[r]}\left(t\right). \tag{1.1}$$

This relation, known to hold for order statistics [David and Nagaraja, 2003, p. 38], turns out to hold in general for judgment ranking as well, thus providing a constraint on how "different" judgment order statistic distributions can be, both from one another and from the population as a whole.

When the population distribution function F is absolutely continuous, a similar result for densities can be derived by differentiating both sides of 1.1 with respect to t, yielding

$$f\left(t\right) = \frac{1}{K}\sum_{r=1}^{K} f_{[r]}\left(t\right), \tag{1.2}$$

where f is the density associated with F and $f_{[r]}$ is the density associated with $F_{[r]}$, $r = 1, ..., K$. The relationship 1.2 also holds in the discrete case for probability mass functions. This may be seen by differencing both sides of equation 1.1.

When the first moment of the distribution F exists, it also follows from 1.1 that

$$\mu = \frac{1}{K} \sum_{r=1}^{K} \mu_{[r]}, \tag{1.3}$$

where $\mu = \int_{-\infty}^{\infty} x \, dF(x)$ is the population mean and $\mu_{[r]} = \int_{-\infty}^{\infty} x \, dF_{[r]}(x)$ is the mean of the r^{th} judgment order statistic, $r = 1, ..., K$.

When F has a finite second moment as well, the population variance σ^2 may be broken down into two quantities, as follows:

$$\begin{aligned} \sigma^2 &= \int_{-\infty}^{\infty} (x - \mu)^2 \, dF(x) \\ &= \frac{1}{K} \sum_{r=1}^{K} \int_{-\infty}^{\infty} (x - \mu)^2 \, dF_{[r]}(x) \\ &= \frac{1}{K} \sum_{r=1}^{K} E\left[\left(Y_{[r]} - \mu\right)^2\right] \\ &= \frac{1}{K} \sum_{r=1}^{K} E\left[\left(Y_{[r]} - \mu_{[r]}\right)^2\right] + \frac{1}{K} \sum_{r=1}^{K} \left(\mu_{[r]} - \mu\right)^2. \end{aligned}$$

That is,

$$\sigma^2 = \frac{1}{K} \sum_{r=1}^{K} \sigma_{[r]}^2 + \frac{1}{K} \sum_{r=1}^{K} \left(\mu_{[r]} - \mu\right)^2, \tag{1.4}$$

where $\sigma_{[r]}^2$ is the variance of the r^{th} judgment order statistic, $r = 1, ..., K$. Equation 1.4 is best understood as a decomposition of the population variance into two components, one accounting for within-rank variability, and the other measuring between-rank variability. It will become apparent in the next section that this is, in fact, a recurring theme in RSS.

Little can be said in general about judgment order statistic distributions beyond the results presented in this section. Unlike order statistics, which have well-known formulas for CDFs, joint, marginal, and conditional densities (in the absolutely continuous case), as well as moments [David and Nagaraja, 2003], no such information is available about judgment order statistic distributions. It is therefore a common assumption in the RSS literature that judgment ranked data *are* order statistics from their respective sets. (See, for example, Takahasi and Wakimoto, 1968, and Wolfe, 2004.) This is regarded as a simplifying assumption, and many RSS procedures in the literature depend very heavily on it. (For a relevant example which will be discussed in greater detail in Chapter 2, see Kvam and Samaniego, 1994.) Naturally, one is skeptical of the assumption that judgment ranking is perfect and does, in fact, identify the exact order statistic in question from every set. In the example of Section 1.1, the field experts had no way of knowing (short of outright measurement) whether the termite colonies they labeled "small," "medium," and "large" in a given ranked set were indeed the smallest, the median, and the largest, respectively. The assumption of perfect ranking is dubious at best, and one only made to reap the benefits of using the many explicit results known about the distributions of order statistics.

Those whose perception of RSS revolves around order statistics, but who are understandably suspicious of the assumption of perfect ranking, have introduced imperfect ranking models, in an attempt to account for ranking error and include it in statistical inference. An early such model was the "measurement error" model proposed by Dell and Clutter [1972], and analyzed in the appendix of that paper by David and Levine. Bohn and Wolfe [1994] endeavor to understand how different order statistics are assigned to given judgment ranks by postulating fixed transition probabilities. Yet another imperfect ranking model is given by Frey [2007]. In lieu of a review of these models, a different imperfect ranking model, under which judgment ranking is accomplished by means of a concomitant variable, will be examined in depth and its difficulties illustrated in Chapter 5.

It is the aim of the present chapter to provide a formulation of RSS and an understanding of its advantages over SRS that do not involve order statistics. Given their pervasive ubiquity throughout the RSS literature, it is the circumvention of order statistics that makes this introduction to RSS so unorthodox. Instead, this chapter argues for another understanding of the benefits of RSS that even a crude judgment ranking procedure can procure. This alternative line of reasoning will become abundantly clear by the end of the next section.

1.3 Balanced RSS and Properties of the Sample Mean

As described in Sections 1.1 and 1.2, a balanced RSS uses a fixed set size and allocates the same number of measurement units to every judgment rank. The process of drawing a balanced RSS can be visualized in the following manner (which has the added benefit of being computationally efficient for programming purposes). A single cycle is obtained by drawing K^2 units from the population and placing them in a $K \times K$ table. Thinking of each row (or each column) as a set, the statistician proceeds to judgment rank each row (or column), sorts the row (or column) according to these rankings, and measures the K diagonal elements. The off-diagonal units are discarded. This procedure is repeated m times to obtain a balanced RSS of size $n = m \times K$.

The measurements obtained from the i^{th} cycle ($i = 1, ..., m$), $Y_{[1]i}, ..., Y_{[K]i}$, are independent (since they are obtained from independent sets) but not identically distributed. In fact, their distributions are $F_{[1]}, ..., F_{[K]}$, respectively. Looking across cycles, it is easy to see that all measurements from units assigned the same judgment rank are identically distributed. That is, $Y_{[r]1}, ..., Y_{[r]m}$ is a SRS from $F_{[r]}$, $r = 1, ..., K$. Thus, this balanced RSS consists of K independent SRS, each of size m, from $F_{[1]}, ..., F_{[K]}$, respectively.

Let $Y_{[r]i}$, $r = 1, ..., K$, $i = 1, ..., m$, be a balanced RSS from a population with distribution function F, mean μ, and finite variance. Suppose the statistician wishes to estimate the population mean μ. The first intuitive estimator that comes to mind is the RSS sample mean

$$\bar{Y}_{RSS} = \frac{1}{m\,K}\sum_{i=1}^{m}\sum_{r=1}^{K} Y_{[r]i} = \frac{1}{K}\sum_{r=1}^{K}\bar{Y}_{[r]}, \tag{1.5}$$

where $\bar{Y}_{[r]} = \frac{1}{m}\sum_{i=1}^{m} Y_{[r]i}$, $r = 1, ..., K$. Standard calculations show that, for each r, $E\,\bar{Y}_{[r]} = \mu_{[r]}$ and $Var\,\bar{Y}_{[r]} = \frac{\sigma^2_{[r]}}{m}$. It follows from these results and 1.3 that

$$E\,\bar{Y}_{RSS} = \mu \tag{1.6}$$

and

$$Var\,\bar{Y}_{RSS} = \frac{1}{m\,K^2}\sum_{r=1}^{K}\sigma^2_{[r]}. \tag{1.7}$$

Equation 1.6 states that the RSS sample mean is an unbiased estimator of the population mean. Moreover, it follows from 1.3, 1.5, and the Strong Law of Large Numbers that $\bar{Y}_{RSS}$ is a strongly consistent estimator of μ as $m \to \infty$, even if the population variance is infinite.

The variance of $\bar{Y}_{RSS}$ may be estimated by using unbiased estimators of the unknown quantities in 1.7:

$$V\hat{a}r\,\bar{Y}_{RSS} = \frac{1}{m\,K^2}\sum_{r=1}^{K} S^2_{[r]}, \tag{1.8}$$

where $S^2_{[r]} = \frac{1}{m-1}\sum_{i=1}^{m}\left(Y_{[r]i} - \bar{Y}_{[r]}\right)^2$ is the within-rank sample variance of the r^{th} judgment rank, $r = 1, ..., K$.

To compare RSS to SRS, let $\bar{Y}_{SRS}$ denote the mean of a (generic) SRS of size $n = m \times K$ from F. It is well known that $\bar{Y}_{SRS}$ is also an unbiased estimator of the population mean, and that its variance is given by $\frac{\sigma^2}{n}$. It follows from this and 1.4 that

$$Var\,\bar{Y}_{SRS} = \frac{1}{m\,K^2}\sum_{r=1}^{K}\sigma^2_{[r]} + \frac{1}{m\,K^2}\sum_{r=1}^{K}\left(\mu_{[r]} - \mu\right)^2, \tag{1.9}$$

which, by comparison with 1.7, immediately implies that

$$Var\,\bar{Y}_{SRS} \geq Var\,\bar{Y}_{RSS}, \tag{1.10}$$

for the same sample size n.

This result yields the classical argument in favor of RSS relative to SRS: that RSS produces unbiased estimators which have variances no larger than their SRS counterparts. Judgment ranking need not be perfect for this improvement to hold, but the improvement may be minimal if judgment ranking is too far off the mark. Dell and Clutter derive properties of the relative precision of RSS to SRS under imperfect ranking, and describe situations in which RSS does much better, as well as distributions for which RSS offers no advantage over SRS even with perfect ranking.

There are two common interpretations of the variance reduction in 1.10. The first is heuristic, observing that a balanced RSS of size n consisting of m cycles and with set size K uses information about far more units than a SRS of the same sample size.

The RSS includes $n = m \times K$ measurements and uses judgment ranking information about $m \times K \times (K-1)$ unmeasured units, whereas the SRS includes measurements only and has no mechanism for making use of judgment ranking information.

The second argument, both more rigorous and more instructive, is that RSS creates an ordered covariate, namely the judgment ranks, which induces an ANOVA-style variance decomposition of the form 1.9, and explains away between-rank variation, leaving only within-rank variability unaccounted for.

It is noteworthy that for a fixed sample size, both RSS and SRS have the same cost of measurement, yet RSS produces a more efficient estimator. Thus, if the cost of judgment ranking is relatively low, the gains in RSS may be purchased at a bargain. This reduction in variance may be translated into a reduction in cost: as the example in Section 1.4 amply illustrates, a SRS may have to be much larger (thereby incurring a significantly higher cost of measurement) than a RSS for their sample means, say, to have the same standard error.

It is straightforward to show that $\bar{Y}_{RSS}$ has a limiting normal distribution when the set size K is held fixed, but the number of cycles m tends to ∞. This may be proved, for example, by applying the usual Central Limit Theorem to the component means $\bar{Y}_{[r]}$ in 1.5. It follows that $\bar{Y}_{RSS} \pm z_{\alpha/2} \sqrt{V\hat{a}r\, \bar{Y}_{RSS}}$ is an approximate $(1-\alpha)$ 100% confidence interval for μ, where $z_{\alpha/2}$ is the $\alpha/2$ upper percentile of the standard normal distribution.

As a curious tangent, note that, unlike its SRS counterpart, the RSS mean is not, in general, the best linear unbiased estimator of the population mean [Barnett and Moore, 1997].

1.4 An Application to Auditing

The data set used in this section was provided by Professor James A. Tackett of the Department of Accounting at Youngstown State University. It consists of a hypothetical population of 5000 different inventory items derived from the financial records of a retail clothing store. Each of these 5000 items represents the inventory value of a particular retail item. In the presence of inventory fraud, the true value of each item remains unknown until it is audited. In creating this data set, a typical inventory fraud scenario was simulated by randomly overstating the book values of 750 (15%) inventory items. The amount of overstatement for each fraudulent account was calculated using random numbers, with an average overstatement factor of twice the true value. Thus, whereas the total audited value of the inventory is $1,877,837, its total value "on the books" is actually $2,140,057, namely an overstatement of almost 14%. This data set represents a modest inventory overstatement, consistent with a company fraudulently looking to boost its stated earnings by a material amount, yet without attracting inordinate attention.

Initially, the auditor only has access to the book values. Suppose the auditor wishes to estimate the mean audited value (or total audited value) of the inventory items. Since auditing inventory items (i.e., measurement) is costly and time-consuming, but items may be ranked with a fairly high degree of accuracy by their book values, this seems like an ideal application for RSS. Here, the book values serve as the concomitant variable for judgment ranking. The purpose of this example is two-fold: first, to compare balanced RSS to SRS, and second, to examine what happens in RSS when the set size varies with the sample size held fixed.

To allow for a wide range of set sizes, the auditor uses samples of size $n = 400$ audited accounts and compares the SRS mean based on this sample size with seven different balanced RSS means corresponding to set sizes $K = 5$, 8, 10, 20, 25, 40, and 50. The associated numbers of cycles to maintain the common sample size $n = 400$ for the RSS means are $m = 80$, 50, 40, 20, 16, 10, and 8, respectively.

The SRS mean is known to be an unbiased estimator of the population mean. In this example, the population mean audited value is \$375.57. The population standard deviation for the audited values is $\sigma = \$112.69$, so that the standard error of the RSS mean based on a sample size of $n = 400$ is $\frac{\sigma}{\sqrt{n}} = \$5.63$. The means and standard errors of the RSS means were obtained via Monte Carlo simulation. (The sampling distribution of each RSS mean was approximated by 10,000 RSS means of sample size $n = 400$ and the same set size.)

Figure 1.1 illustrates the overall gain obtained from using RSS estimators over the SRS estimator through a comparison of the approximate (i.e., large sample) sampling distributions of the various RSS estimators and the SRS estimator. The three solid curves represent the densities of the approximate sampling distributions of the RSS means for set sizes 50, 25, and 10 (in order from highest to lowest peak), and the dotted curve is the density of the approximate sampling distribution of the SRS mean. The vertical line passes through the population mean.

Notice that while the approximate sampling distributions of all four estimators are centered about the population mean, those for the RSS means are considerably tighter than that of the SRS mean. Moreover, the precision increases (i.e., the curves become narrower and more peaked) as the set size increases. Another visual that

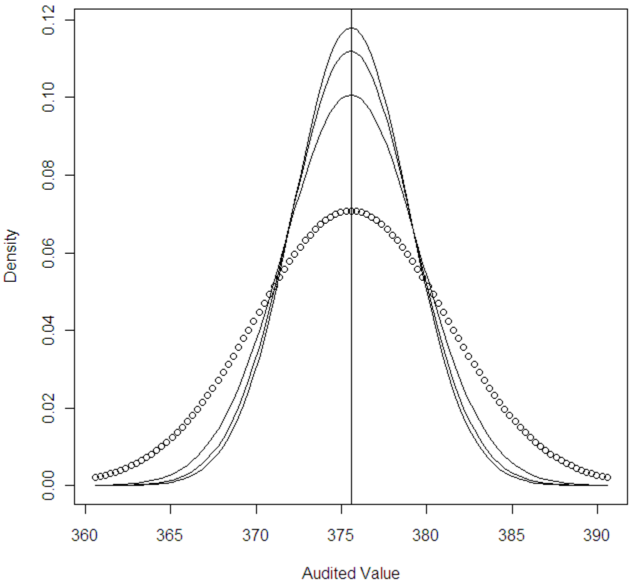

Figure 1.1: Densities of the approximate sampling distributions of the RSS means for $n = 400$ and set sizes 50, 25, and 10 (solid curves, in order of highest peak) and the SRS mean (dotted curve) for $n = 400$. The vertical line is drawn at the population mean \$375.57.

conveys the same point is Figure 1.2, a plot of the standard errors of the estimators against set size. (The SRS mean corresponds to set size 1.)

As Figure 1.2 shows, the standard errors of all the RSS estimators under consideration are significantly smaller than the standard error of the RSS mean, for the same sample size. Moreover, the standard error decreases as set size increases. Thus, increased precision may be obtained by using a larger set size. However, this improvement tapers off and once the set size exceeds a certain threshold, there is little to be gained from using an even larger set size. This phenomenon may be referred to as *the diminishing marginal returns of increasing set size.*

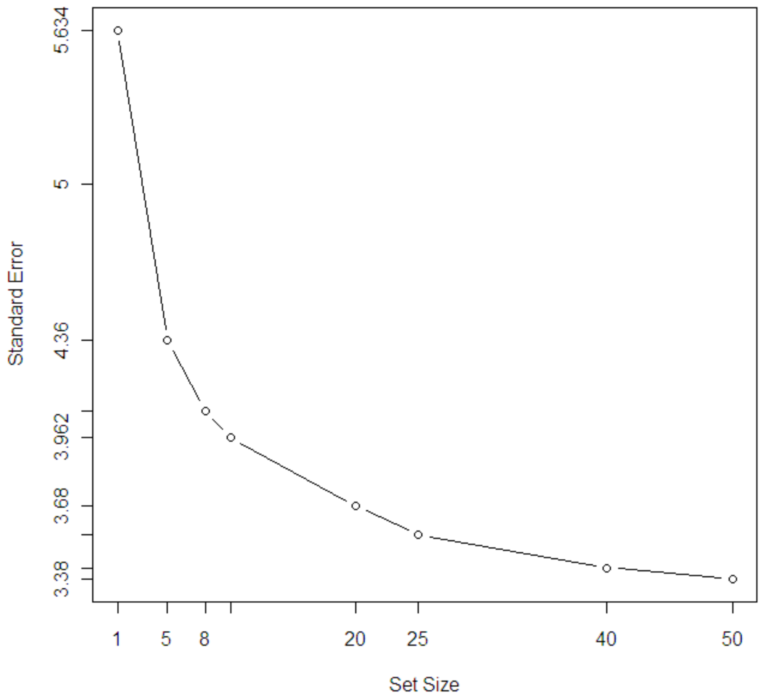

Figure 1.2: Plot of the standard errors of the SRS mean (set size = 1) and the RSS means for set sizes 5, 8, 10, 20, 25, 40, and 50.

Another way to understand the advantages of RSS over SRS is to calculate the relative sample sizes needed to attain the same level of precision. Table 1.1 gives the standard errors for the RSS means (rounded to 3 decimal digits) for the set sizes under consideration, and provides the minimum sample size necessary for the SRS mean to have a standard error at least as small. For example, a SRS would need a sample size of at least 1001 for its mean to have a standard error at least as small as that of the RSS mean with sample size $n = 400$ and set size $K = 25$.

These results demonstrate that RSS produces a much more precise estimator of the population mean than a SRS of the same sample size or, equivalently, that a SRS

Set Size	5	8	10	20	25	40	50
Standard Error of RSS Mean	4.360	4.072	3.962	3.680	3.562	3.428	3.380
Minimum SRS n	668	766	810	938	1001	1081	1112

Table 1.1: Standard errors of the RSS means for set sizes 5, 8, 10, 20, 25, 40, and 50, and the smallest sample sizes necessary for the SRS mean to be at least as precise as the RSS means with $n = 400$ and these set sizes.

would have to be much larger than a RSS to equal the precision of its sample mean. Given the high cost of auditing (measurement) and the negligible effort required for judgment ranking (using the accessible concomitant, book value), the advantages of RSS over SRS in assessing fraud in this scenario are striking.

1.5 Estimating the Population Variance

Section 1.3 provides the elementary theory for estimation of the population mean using a balanced RSS. These results were illustrated in Section 1.4. In that toy example, the population variance was known since the audited values for the entire population were available. Needless to say, this is not usually the case, and so estimators of the population variance are important in their own right. Stokes [1980a] proposes a RSS estimator of the variance that is analogous to the usual SRS estimator, i.e., the sample variance. Stokes' estimator is biased, but unbiased in the limit. MacEachern et al. [2002] observe that

> Stokes's estimator overestimates the population variance; a within-judgment-class estimator of variation will underestimate σ^2. Intuitively we want an estimator that lies somewhere between these two extremes and has little or no bias. The approach that we take is to combine within-judgment-class and between-judgment-class estimators in such a way that the resulting estimator is unbiased for the variance of the underlying population.

This estimator may be written, after algebraic simplification, as the weighted linear combination

$$\hat{\sigma^2} = \frac{1}{K}\, S_W^2 + \left(1 - \frac{1}{K}\right)\, S_B^2, \tag{1.11}$$

where $S_W^2 = \left(1 - \frac{K-1}{m\,K}\right) \sum_{r=1}^{K} S_{[r]}^2$ represents within-judgment-rank variability and $S_B^2 = \frac{1}{K-1} \sum_{r=1}^{K} \left(\bar{Y}_{[r]} - \bar{Y}_{RSS}\right)^2$ measures between-judgment-rank variability. $\hat{\sigma^2}$ is an unbiased estimator of the population variance σ^2, and MacEachern et al. show that it is always more efficient than both Stokes' RSS variance estimator, and the sample variance of a SRS.

1.6 Beyond Balanced RSS

Many procedures involving judgment ranking fall under the banner of RSS. In addition to balanced RSS, one may devise unbalanced RSS procedures, where the set size K is held fixed, but unequal numbers of measurement units are allocated to the different judgment ranks. Wolfe [2004] gives the simple example of estimating the population median using an odd set size and measuring only the ranked medians from each set. Such a design is unbalanced to the extreme. Moreover, cost considerations can sometimes make it worthwhile to measure more than one member of each ranked set, in spite of the increase in the variance of the mean caused by dependence among sample measurements [Wang et al., 2004]. Beyond that, RSS procedures can be envisaged that involve multiple set sizes in a single application [Gemayel et al., in press].

1.7 Description of This Work

This chapter develops the basic elements of RSS necessary for an understanding of the core results in this dissertation. Chapter 2 provides a Bayesian view of RSS and proposes a framework for Bayesian modeling of RSS data. It also includes a model for continuous data. In Chapter 3, the framework of Chapter 2 is applied to discrete data, and the resulting computational problems are resolved. Chapter 4 introduces judgment post-stratification and explores the role of the set size K from the Bayesian perspective. It also proposes a method for combining the judgments of multiple rankers. Chapter 5 considers the special case when judgment ranking is induced by a concomitant variable. Finally, Chapter 6 lists conclusions and suggestions for future research.

Chapter 2

Ranked Set Sampling as an ANOVA-type Procedure: A Bayesian Nonparametric Perspective

2.1 On the Likelihood Principle and Survey Sampling

In his classic paper on the role of the sufficiency and likelihood principles in sample survey theory, Basu [1969] reaches the provocative (but entirely justified) conclusion that the statistician at the analysis stage need not pay any attention to the nature of the sampling design that produced the data. In fact, the statistician does not even need to know the specifics of the sampling design, beyond understanding it well enough to obtain the likelihood function. The role of the data is merely to change the prior scale of preference (or prior probability distribution) about unknown parameters to the posterior scale. This change is represented by the likelihood function, evaluated at the measured data.

Ericson [1969] adds that "when reasonable prior distributions are introduced, their revision by sample data can lead to meaningful and useful inferences." Ericson's paper is also an early example of the use of the multinomial distribution in modeling "extreme prior vagueness" in the finite population setting, which re-interprets some common frequentist procedures as Bayesian procedures based on a noninformative prior distribution. Meeden and Vardeman [1991] note that

> It is somewhat paradoxical that in sampling, the one area of statistical practice where prior information is routinely used, formal Bayesian methods of inference are seldom called upon. One reason for this is the difficulty of specifying sensible prior distributions over a large dimensional parameter space.

Early work on noninformative Bayesian procedures spawned the Bayesian bootstrap [Rubin, 1981], and the finite population Bayesian bootstrap [Lo, 1988], which revolves around the Polya posterior [Nelson and Meeden, 1998]. In turn, the Polya posterior can be seen as an approximation to the posterior under a "flat" Dirichlet process prior. The Dirichlet process is a fundamental building block of Bayesian nonparametrics and will be discussed at length in Section 2.4.

The implications of likelihood principle-based thinking for RSS are striking. Much of the current RSS literature is dominated by stringent attachment to design assumptions, from balanced RSS to unbalanced RSS and beyond. (See, for example, the annotated bibliography by Kaur et al. [1995].) In light of the likelihood principle, the role of RSS as a design is altered: its task is solely to deliver the data (and the likelihood). It is no longer called upon to provide "average" or "long-run" performance characteristics of statistical procedures, such as bias, variance, or efficiency, since these measures average over "all possible samples," whereas the likelihood principle exhorts the statistician to base inference only on the sample at hand. To put it more bluntly, any statistician committed to the likelihood principle (and by extension, any dyed-in-the-wool Bayesian) will not be swayed by the frequentist argument in favor of RSS made in the previous chapter.

A difficult series of questions ensue: how can a Bayesian be convinced of the value of RSS? Does RSS have any benefits at all from a Bayesian or likelihood principle perspective? How can these benefits (if, in fact, there are any) be assessed and understood? It is the aim of this chapter to construct a framework in which these questions may be made more rigorous, and subsequently answered.

2.2 Some Common Assumptions in the RSS Literature

The design structure of RSS is sometimes not sufficient, in and of itself, for devising statistical procedures and evaluating them, and typically must be supplemented with additional assumptions. Many authors make no qualms about their mistrust of these assumptions, the most salient being that of perfect ranking, already discussed in Section 1.2. It is well known that even a crude judgment ranking mechanism for RSS can provide significant improvement over SRS. In fact, the mere use of judgment order statistic distributions implicitly presupposes that the measurements from each judgment rank are homogeneous relative to measurements from other judgment ranks. When ranking is haphazard, units assigned the same judgment rank may differ little from the general population at large. Moreover, it is conceivable in some applications that units in the same ranked set may resemble each other more than they do units assigned the same judgment rank in other ranked sets. This may be the case, for example, when sets are taken to be clusters of adjacent population units. That is, an unranked set is not necessarily a SRS from the population, yet many RSS procedures are derived with this assumption built in from the start.

Another frequent presupposition, coupled to perfect ranking, is absolute continuity of the underlying population. The reason for this assumption is that if the population has a density (with respect to Lebesgue measure) and ranking is perfect, then the possibility of ties in ranking is excluded and, more importantly, judgment order statistic distributions have explicit densities (namely, those of the usual order statistics). Many RSS procedures shun discrete data for these reasons. (Discrete data will be considered in Chapter 3.)

Finally, when concomitant variables are used for judgment ranking, they are often assumed to be linearly related to the variable of interest. This is clearly an unnecessarily stringent assumption, since a concomitant variable need only be associated in some manner with the quantity of interest for it to be of potential assistance in statistical inference. The auditing application in Section 1.4 is a simple example of a concomitant variable that contains a great deal of information about the variable of interest, yet is not linearly related to it.

Assuming a linear relationship between the concomitant and the variable of interest, however, negates the need for a deeper understanding of the relationship between the two variables. Instead, that relationship is entirely summarized by the correlation coefficient. It is only a short step from there to assume that the joint distribution of the concomitant and the variable of interest is bivariate normal. (See, for example, Stokes, 1980b.) In Chapter 5, an approach to handling concomitants will be proposed that does not require a linear relationship. Instead, it tries to "learn" the relationship between the concomitant and the variable of interest from the data and embed that relationship into the analysis.

While there is skepticism of the usual RSS assumptions among researchers and practitioners of RSS, many argue that such strong postulates are necessary for deriving properties of the design and evaluating procedures based upon it. By adopting the likelihood principle, RSS is liberated from this burden, and the statistician is in a position to do away with some unnecessary assumptions or restrictions (although they may be supplanted by different assumptions). Providing an alternative that is entirely likelihood principle-driven (from model construction to evaluation) in a fully nonparametric framework is tricky. The next section deals with some of the thorny aspects of this construction.

2.3 A Bayesian View of RSS

It is by no means an easy or direct leap to incorporate Bayesian ideas into RSS. For one thing, RSS has emerged in the nonparametrics literature, which seems diametrically opposed to the strong parametric assumptions made by Bayesians. (As the previous section amply demonstrates, however, RSS practitioners are not immune to making strong assumptions of their own.) Little work has been done on the Bayesian aspect of RSS, and most of that work employs order statistics and/or emphasizes scalar or other low-dimensional parameters, making RSS seem rather incidental to the problem (e.g., Lavine, 1999). This section seeks to place RSS in a deeper, and less contrived, Bayesian framework.

Perhaps the most "nonparametric" of all objectives in statistics is to estimate a population distribution function. As a result, that objective is very well suited to illustrating how skepticism of the assumptions discussed in Section 2.2 can lead a statistician to view RSS through a Bayesian lens. The simplest approach to estimating a CDF is to estimate it point-wise as a population proportion. This is, indeed, the approach taken by Stokes and Sager [1988]. In the usual notation, let $Y_{[r]i}$, $r = 1, ..., K$, $i = 1, ..., m$, denote a balanced RSS from a distribution function F. Stokes and Sager's proposed estimator of F is just the sample empirical distribution function

$$F^* (t) = \frac{1}{m\,K} \sum_{i=1}^{m} \sum_{r=1}^{K} \mathbf{1} \left(Y_{[r]i} \leq t \right) .$$

It follows from Equations 1.1 and 1.6 that $F^*(t)$ is an unbiased estimator of $F(t)$, and the fact that it has a variance no larger than its SRS counterpart follows from Equation 1.10. Moreover, the estimator $F^*(t)$ is strongly consistent and has a simple normal limiting distribution for each t.

While establishing the superiority of RSS to SRS, the empirical CDF proposed by Stokes and Sager makes hardly any use of the additional information provided by judgment ranking, since it treats all measurements across judgment ranks in the same way, as though they were i.i.d. This was observed by Kvam and Samaniego [1994], who rectified the problem at the cost of some rigor. Their solution was to assume perfect ranking, which allowed them to write the joint p.d.f. of the sample as a product of densities of order statistics. Such a product may be re-written in terms of the population distribution function and density evaluated at the measured data. Upon closer scrutiny, this likelihood function bears a strong resemblance to a product of binomial probability mass functions with ordered success probabilities, which suggests a simple parametrization of the likelihood function. Kvam and Samaniego's proposed estimator is the maximizer of this likelihood function, the NonParametric MLE, or NPMLE [Kiefer and Wolfowitz, 1956]. Huang [1997] obtained some asymptotic results for Kvam and Samaniego's NPMLE, and Ozturk (in press) extended Kvam and Samaniego's work to incorporate the Bohn-Wolfe model for imperfect ranking [Bohn and Wolfe, 1994].

For sake of simplicity, let us illustrate Kvam and Samaniego's procedure and highlight our objections to it in the case when the data consist of a balanced RSS with set size $K = 2$ and $m = 1$ cycle. In this simple setting, the statistician considers a ranked set of size 2 from F, say $Y_{(1)1} < Y_{(2)1}$, but measures only $Y_{(1)1}$. (Note the use of parentheses instead of brackets in the subscript, in line with the authors' assumption that the judgment ranking procedure orders the set perfectly.) The statistician then considers another ranked set (independent of the first), $Y_{(1)2} < Y_{(2)2}$, and measures only $Y_{(2)2}$. The units corresponding to $Y_{(2)1}$ and $Y_{(1)2}$ are discarded: their sole purpose was to provide information for judgment ranking.

The joint density function of $Y_{(1)1}$ and $Y_{(2)2}$ is given by

$$4\left(1 - F\left(y_{(1)1}\right)\right) F\left(y_{(2)2}\right) dF\left(y_{(1)1}\right) dF\left(y_{(2)2}\right). \tag{2.1}$$

Let $p_1 = dF\left(y_{(1)1}\right)$ and $p_2 = dF\left(y_{(2)2}\right)$ be the point masses assigned to the two measurements. Suppose it turns out that $y_{(1)1} < y_{(2)2}$. Then the likelihood function 2.1 may be written as $\mathcal{L} = 4\,(1 - p_1)\,(p_1 + p_2)\,p_1\,p_2$. Since it must be the case that $p_1 + p_2 = 1$, the likelihood function reduces to $\mathcal{L} = 4\,p_1\,(1 - p_1)^2$, which is maximized at $(p_1,\, p_2) = \left(\frac{1}{3},\, \frac{2}{3}\right)$. The maximum value of the likelihood function is $\mathcal{L}\left(\frac{1}{3},\, \frac{2}{3}\right) = \frac{16}{27}$, whereas its value at the empirical CDF is $\mathcal{L}\left(\frac{1}{2},\, \frac{1}{2}\right) = \frac{1}{2}$.

If, however, the statistician observes $y_{(2)2} < y_{(1)1}$, it can no longer be said that $p_1 + p_2 = 1$, since the population distribution function F is known to assign positive mass to the interval $\left(y_{(1)1},\, \infty\right)$. The unwanted side effect of this is that the NPMLE is a step function that never reaches one on the right (and is therefore not a CDF). This occurs, regardless of set size, whenever the sample maximum is drawn from any rank other than the largest rank. At any rate, the likelihood function in this case can be parametrized as $\mathcal{L} = 4\,(1 - p_1 - p_2)\,p_2^2\,p_1$, which is maximized at $(p_1,\, p_2) = \left(\frac{1}{4},\, \frac{1}{2}\right)$. The maximum value attained by the likelihood function is $\mathcal{L}\left(\frac{1}{4},\, \frac{1}{2}\right) = \frac{1}{16}$, whereas its value at the empirical CDF is 0.

While the assumption of perfect ranking provides a framework for a tractable likelihood function combining densities of order statistics, it is too strong in the sense that it betrays the essence of judgment ranking. Nowhere is this more evident than in the EM algorithm proposed by Kvam and Samaniego to maximize their likelihood function. To use the EM algorithm, the authors frame the problem in terms of censored data. For instance, the censored observation $Y_{(2)1}$ is known to necessarily lie to the right of the measurement $Y_{(1)1}$. This need not be the case when judgment ranking is not perfect. In other words, Kvam and Samaniego's assumption of perfect ranking translates into the assumption that the variable of interest is censored by its own measured values. Judgment ranking, however, is by its very nature fuzzy and error-prone. It is one thing to seek to incorporate such information into data analysis; it is an entirely different matter to treat it as infallible, especially when this assumption can produce a deficient estimator.

To elucidate this point further, consider using an accessible concomitant variable, X, for judgment ranking. In this instance, all the information about the judgment ranking mechanism is contained in the joint distribution of (X, Y). In the simple example above, the statistician draws a pair of units from the population, (X_{11}, Y_{11}) and (X_{21}, Y_{21}), characterized by a known concomitant variable X and a not-yet-known variable of interest Y. The X's are ranked in increasing order. (Assume for the moment that there are no ties in the concomitant.) The ranked set is now $\left(X_{(1)1}, Y_{[1]1}\right)$ and $\left(X_{(2)1}, Y_{[2]1}\right)$. The statistician measures the judgment ranked minimum, $Y_{[1]1}$, which may or may not be the actual minimum. Similarly, the statistician obtains a second ranked set (independent of the first) $\left(X_{(1)2}, Y_{[1]2}\right)$ and $\left(X_{(2)2}, Y_{[2]2}\right)$ and measures the judgment ranked maximum, $Y_{[2]2}$. In this instance, the variable of interest is being censored not by its own values, but by the values of another variable. This means that one cannot tell with any certainty how many of the non-measured units are below or above the measured Y values. As a result, Kvam and Samaniego's approach is not replicable when ranking by an imperfect concomitant.

To show why strictly likelihood-based inference is futile in this setting, start by writing the likelihood function for this simple example with self-evident notation. Following Yang [1977], the distribution of $Y_{[r]}$ given $X_{(r)}$ is merely the population conditional distribution of Y given X. Hence the likelihood function in this setting may be written as

$$f\left(y_{[1]1} \middle| x_{(1)1}\right) f_X\left(x_{(1)1}\right) f_X\left(x_{(2)1}\right) f\left(y_{[2]2} \middle| x_{(2)2}\right) f_X\left(x_{(1)2}\right) f_X\left(x_{(2)2}\right). \tag{2.2}$$

As a moment's thought should make clear, inference based on this likelihood function alone is problematic, since it estimates the joint distribution of (X, Y) by placing masses of $1/4$ at the points $\left(x_{(1)1}, y_{[1]1}\right)$ and $\left(x_{(2)2}, y_{[2]2}\right)$ and on the vertical lines $x = x_{(2)1}$ and $x = x_{(1)2}$. Marginalizing to estimate the distribution of Y alone merely recovers the empirical distribution function. That is, purely likelihood-based inference does not make full use of the structure of RSS.

There are many joint distributions that fit the framework of the estimate above. Obviously, not all are equally credible to the statistician. That is, the statistician has a preconceived idea about the nature of the relationship between the concomitant and the variable of interest. (A concomitant that is not relevant to the variable of interest would not have been used in the first place.) For example, the statistician may believe that the variable of interest "tends to increase" with the concomitant. Such prior knowledge implies an assumption of some vague stochastic order on the joint distribution. This may be taken to mean the usual (weak) stochastic order, or monotone likelihood ratio (MLR), or that the population regression function $E[Y \mid X = x]$

is an increasing function of x, or any other way of putting in concrete terms such fuzzy prior information.

As this example illustrates, an essential feature of inference for RSS is the ability to learn about the (conditional) distribution of Y across the range of the concomitant (if there is one) or, more generally, over the various judgment ranks. A Bayesian framework for RSS is, then, a tool that weaves these vague prior beliefs about the judgment ranking mechanism into inference in a formal and natural fashion. That is, the statistician's parameter space is really a set of collections of distribution functions on the data space that are related in some manner. It will come as no surprise that specifying workable prior distributions on such a parameter space is no easy feat.

Between the Bayesian penchant for over-generous use of parametric assumptions for the likelihood function and prior on one hand, and the austerity of nonparametrics on the other, an amalgam of the two approaches seems rather like a contradiction in terms. Nonetheless, it turns out that Bayesian nonparametric methods are more than just a grudging concession on the part of Bayesian modeling to the minimalist style of nonparametrics (which, after all, is the home of RSS). In fact, they provide powerful tools for modeling in parameter spaces as abstract as the one discussed above. The next section deals with Bayesian nonparametric modeling, starting with a review of the basic properties of the Dirichlet process, and on through dependent Dirichlet processes and mixture of Dirichlet process models, in a *crescendo* leading up to a general Bayesian nonparametric model for RSS data.

2.4 A General Model for RSS Data

Ferguson [1973] lists two desirable properties of a nonparametric prior:

1. It should have “large support” on the space of probability distributions on the data space, and

2. “Posterior distributions given a sample of observations from the true probability distribution should be manageable analytically.”

Both these properties are satisfied by the Dirichlet process, the nonparametric prior presented in Ferguson’s paper. Ferguson shows that a random probability distribution can be characterized by the joint distributions of the probabilities it assigns to finite partitions of the data space.

More specifically, let $\mathcal{X}$ denote the data space, let $\mathcal{A}$ denote a σ-field of subsets of $\mathcal{X}$, and let α denote a non-null, finite measure on the measurable space $(\mathcal{X},\, \mathcal{A})$. To say that a random probability P on $(\mathcal{X},\, \mathcal{A})$ follows a *Dirichlet process with parameter* α (abbreviated $P \sim DP\,(\alpha)$) means that for any positive integer m, and for any finite measurable partition $(A_1,\, ...,\, A_m)$ of $\mathcal{X}$, $(P\,(A_1)\,,\, ...,\, P\,(A_m))$ follows a Dirichlet distribution with parameters $\alpha\,(A_1)\,,\, ...,\, \alpha\,(A_m)$. (When $\alpha\,(A) = 0$, this means that $P\,(A) = 0$ with probability 1.) P is called a random probability measure since $P\,(\mathcal{X})$ is degenerate at 1. Ferguson shows that P is discrete with probability one.

The positive number $M = \alpha(\mathcal{X})$, called the *precision parameter* (or *mass parameter*) of the Dirichlet process, and the probability measure $\alpha_0 = \frac{\alpha}{M}$, together characterize the Dirichlet process. Ferguson shows that $E\,P\,(A) = \alpha_0\,(A)$, for any $A \in \mathcal{A}$, and that a single draw X from P is just a draw from α_0.

Moreover, if $P \sim DP\,(\alpha)$ and $X_1,\, ...,\, X_n$ is a random sample of size n from P, then the posterior distribution of P given $X_1,\, ...,\, X_n$ is also a Dirichlet process, with parameter $\alpha + \sum_{i=1}^{n} \delta_{X_i}$, where δ_x denotes a unit point mass at x. The precision parameter of this posterior is $M + n$, so that

$$E\left[P(A) \mid X_1, \ldots, X_n\right] = \frac{\left(\alpha + \sum_{i=1}^{n} \delta_{X_i}\right)(A)}{M+n} = \frac{M}{M+n}\alpha_0(A) + \frac{n}{M+n}F_n(A), \tag{2.3}$$

where $F_n(\cdot) = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}(\cdot)$. Equation 2.3 shows that the posterior mean of P given the sample is a convex linear combination of the base measure α_0 and the empirical distribution measure of the sample, F_n. This equation also highlights the role of the precision parameter M.

Under squared error loss, the Bayes estimator of the distribution P is the posterior mean, given by 2.3. When M is large, the posterior mean is close to the base measure of the prior, whereas when M approaches 0, the posterior mean is well approximated by the empirical distribution of the data. Thus, inference based on the posterior mean when $M \searrow 0$ is similar to strictly data-based inference from the empirical distribution function, and a sample from the posterior mean can be approximated closely by bootstrapping the data [Lo, 1986, 1988].

Blackwell and MacQueen [1973] relate the Dirichlet process to the *Polya urn scheme*, which is tantamount to saying that $X_1, \ldots, X_n$ are exchangeable and that

$$p(x_n|x_1, \ldots, x_{n-1}) = \begin{cases} \delta_{x_i} & \text{with probability } \frac{1}{M+n-1},\ i = 1, \ldots, n-1 \\ \alpha_0 & \text{with probability } \frac{M}{M+n-1} \end{cases}.$$

Sethuraman [1994] offers another representation of the Dirichlet process as a *stick-breaking process*. Specifically, if $P \sim DP(\alpha)$, then P may be written in stick-breaking form as

$$P(\cdot) = \sum_{h=1}^{\infty} p_h \delta_{\theta_h}(\cdot), \tag{2.4}$$

where the weights $p_1, p_2, \ldots$ are derived from the i.i.d. Beta $(1,\, M)$ random variables $V_1, V_2, \ldots$ via $p_1 = V_1$ and, for $h \geq 2$, $p_h = V_h \prod_{l<h}(1 - V_l)$, and the locations

θ_1, θ_2, ... are i.i.d. draws from α_0, independent of the V's. The stick-breaking representation 2.4 of the Dirichlet process provides a new way of looking at the parameter α of the process: the discrete random distribution P is broken down into a set of random locations and random weights assigned to those locations (and independent of them). The locations θ_h are determined solely by the base measure α_0, while the precision parameter M alone influences the weights p_h. Both the Polya urn scheme and the stick-breaking representations of the Dirichlet process yield a simple way of thinking about the sample X_1, ..., X_n from P when α is absolutely continuous with respect to Lebesgue measure: the distinct X values are i.i.d. draws from α_0, but some of the sample values may be tied. By contrast, in a fully parametric setup with absolutely continuous α_0 and no intervening Dirichlet process, the probability of ties among X_1, ..., X_n is zero. The importance of this clustering caused by the Dirichlet process will become apparent in the modeling stage (Section 2.5).

MacEachern [1999, 2000] exploits the stick-breaking representation of the Dirichlet process to develop an elegant and strikingly simple formulation of Dependent Dirichlet Processes (DDP). Suppose the statistician seeks a nonparametric prior on d-tuples of dependent random distributions (P_1, ..., P_d). Such a need arises, for example, when modeling the distribution of a response over d values of a covariate. It is immediately obvious how this can be done using the breakdown 2.4. By representing each component of (P_1, ..., P_d) as $P_x(\cdot) = \sum_{h=1}^{\infty} p_{xh}\delta_{\theta_{xh}}(\cdot)$, $x = 1$, ..., d, and maintaining dependence across x between the weights p_{xh} and between the locations θ_{xh}, MacEachern's DDP provides the statistician with a nonparametric prior over the space of d-dimensional distributions, with each component P_x having a marginal Dirichlet process distribution, and with an intuitive visualization of the structure of the dependence across components. A major simplifying assumption for the DDP, and one that significantly reduces the computational burden of posterior calculations, is to keep the weights p_h constant over the range of x, letting only the locations θ_{xh} vary. This is called the "single-p" DDP. This assumption does not diminish the broad scope of the DDP; in fact, it forces the component distributions to be dependent. For, if a d-tuple of locations (θ_{1h}, ..., θ_{dh}) carries a large clump of mass p_h, then the distributions (P_1, ..., P_d) are restricted in how much they can vary from one another, and thus they cannot be independent. (See MacEachern, 2000, for details.)

On first thought, the discreteness of a random draw from a Dirichlet process may seem an obstacle to modeling continuous data. The actual use of Dirichlet processes in Bayesian modeling, however, is to embed the Dirichlet process prior in a hierarchical model and to smooth away its discreteness by convolving it with a continuous density. Even when the data are discrete, the Dirichlet process may be convolved with an appropriate likelihood function, such as the multinomial for count data. These models are known as Mixture of Dirichlet Process (MDP) models.

As with many Bayesian innovations, posterior inference for MDP models is carried out via Markov Chain Monte Carlo (MCMC) methods [MacEachern, 1998]. As MacEachern [2000] shows, a single-p DDP can be treated as a single Dirichlet process with a multivariate (d-dimensional) base measure specifying the locations. It follows that any MCMC algorithm for MDP models can also be used for posterior inference in a single-p DDP model.

As argued in Section 2.3, a natural Bayesian model for RSS should account for the variability in judgment order statistic distributions that is caused by the judgment ranking mechanism. In fact, the model presented in this section re-casts RSS as a Bayesian nonparametric ANOVA [De Iorio et al., 2004], which suggests a direct analogy with the ANOVA-style frequentist argument for RSS presented in Section 1.3. In light of Equation 1.1, the judgment order statistic distributions $F_{[1]}$, ..., $F_{[K]}$ are clearly dependent. Intuitively speaking, since units assigned different judgment ranks are drawn from the same population, they must all contain information about all the judgment order statistic distributions. The DDP provides a simple way of borrowing this information across judgment ranks.

Suppose the data are a balanced or unbalanced RSS with fixed set size K. It is assumed that only one measurement is taken from each ranked set. Let $Y_{[r]i}$, $i =$ 1, ..., m_r, $r = 1$, ..., K, denote the sample, and suppose $f(y|\theta, \phi)$ is an appropriate density for the quantity Y, as discussed above. Here, θ denotes a parameter that may vary between observations, while ϕ is an optional parameter common to all observations. (For example, if θ is a location parameter, then ϕ may be taken to be a scale parameter.)

The individual θ's for the r^{th} judgment rank are modeled as i.i.d. draws from a random distribution P_r, $r = 1$, ..., K, and the K-tuple (P_1, ..., P_K) is itself a random draw from a single-p DDP, specified by a mass parameter M and a base measure. The base measure of the DDP is taken to be a multivariate (K-dimensional) normal distribution, since the normal distribution allows for easy specification of the mean and the covariance structures. The precision parameter M and the covariance matrix $\mathbf{\Sigma}$ of this normal distribution regulate how information is borrowed across judgment ranks.

For added flexibility, the mean vector $\boldsymbol{\mu}$ of the base measure is itself assigned a normal prior distribution with known mean $\boldsymbol{\mu}_0$ and covariance matrix $\mathbf{\Sigma}_0$. Finally, the parameter ϕ is given its own (independent) prior distribution. All parameters not explicitly assigned a prior are taken to be known and fixed. In particular, the mass parameter M is kept fixed because it would be very difficult to elicit a meaningful prior distribution for it.

The model may be summarized as

$$\begin{aligned} Y_{[r]i}|\theta_{[r]i},\, \phi &\sim f\left(y|\theta_{[r]i}, \phi\right),\, i = 1,\, ...,\, m_r,\, r = 1,\, ...,\, K \\ \theta_{[r]1},\, ...,\, \theta_{[r]m_r}|P_r &\sim P_r,\, r = 1,\, ...,\, K \\ \left(P_1,\, ...P_K\right)|\boldsymbol{\mu} &\sim DDP\left(M,\, N\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\right) \\ \boldsymbol{\mu} &\sim N\left(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) \\ \phi &\sim [\phi] \end{aligned} \tag{2.5}$$

with the usual convention of independence holding among the terms at each level of the hierarchy.

To rephrase the model in terms of a single Dirichlet process, write the data as $(r_1,\, Y_1),\, ...,\, (r_n,\, Y_n)$, where n is the sample size. Let d_i denote a $1 \times K$ design vector consisting of 1 in the r_i^{th} position and 0 elsewhere, $i = 1,\, ...,\, n$. Then the model becomes

$$\begin{aligned} Y_i|\boldsymbol{\theta}_i,\, \phi &\sim f\left(y|d_i\boldsymbol{\theta}_i, \phi\right),\, i = 1,\, ...,\, n \\ \boldsymbol{\theta}_1,\, ...,\, \boldsymbol{\theta}_n|P &\sim P \\ P|\boldsymbol{\mu} &\sim DP\left(M,\, N\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\right) \\ \boldsymbol{\mu} &\sim N\left(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) \\ \phi &\sim [\phi] \end{aligned} \tag{2.6}$$

MacEachern [1998] provides the basic MCMC algorithm for sampling from the posterior distribution of MDP models such as model 2.6, along with some add-ons for increased efficiency. The basic algorithm, a Gibbs sampler, relies to a great extent on conjugacy of the likelihood and hyperprior. (Here, conjugacy is used in the informal sense, to mean that the posterior resulting from the likelihood/prior pair is easy to work with.) In the next section, a simple model for continuous data is proposed and its associated Gibbs sampler is described in detail. In Chapter 3, the model is extended to discrete data and the resulting non-conjugacy problem is explored.

2.5 A Model for Continuous Data

When the variable of interest Y is continuous, the density $f(y|\theta, \phi)$ in models 2.5 and 2.6 is merely an expedient kernel that smooths away the discreteness of the Dirichlet process. Since the hyperprior α_0 is taken to be multivariate normal (for reasons explained in the previous section), a natural, conjugate choice is to take f to be the normal density. This is not to say, however, that the statistician believes for a second that the data themselves may be normally distributed. Rather, the stick-breaking representation 2.4 of the Dirichlet process implies that the judgment order statistic distributions are being modeled as infinite mixtures of normals, while maintaining dependence across judgment ranks among the means of each mixture component. The parameter ϕ is taken to be the data variance σ_Y^2, and is assigned an Inverse Gamma $IG\ (a, b)$ prior, with density proportional to $(\sigma_Y^2)^{-(a+1)} \exp\left(-\frac{1}{b\sigma_Y^2}\right)$.

Model 2.6 for continuous data becomes

$$\begin{aligned} Y_i|\boldsymbol{\theta}_i,\, \sigma_Y^2 &\sim N\left(d_i\boldsymbol{\theta}_i, \sigma_Y^2\right),\, i=1,\, ...,\, n \\ \boldsymbol{\theta}_1,\, ...,\, \boldsymbol{\theta}_n|P &\sim P \\ P|\boldsymbol{\mu} &\sim DP\left(M,\, N\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\right) \\ \boldsymbol{\mu} &\sim N\left(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) \\ \sigma_Y^2 &\sim IG\ (a, b) \end{aligned} \tag{2.7}$$

Note that the first line of model 2.7 may be rewritten as $Y_i = d_i\boldsymbol{\theta}_i + \epsilon_i$, where ϵ_1, ..., ϵ_n are i.i.d. $N\left(0, \sigma_Y^2\right)$, lending itself to a natural interpretation as a random-effects ANOVA model. The Gibbs sampler for sampling from the posterior under model 2.7 makes use of the clustering of the $\boldsymbol{\theta}$'s at $k \leq n$ distinct values, $\boldsymbol{\theta}_1^*$, ..., $\boldsymbol{\theta}_k^*$, owing to the fact that the distribution P is discrete. In any given scan of the Gibbs sampler, the individual $\boldsymbol{\theta}$'s (and their associated observations) are allowed to leave their current clusters to join another existing cluster, or to start a brand new cluster, in accordance with the Polya urn scheme representation of the Dirichlet process.

To formalize this idea, define the random variables s_1, ..., s_n by $s_i = j$ iff $\boldsymbol{\theta}_i = \boldsymbol{\theta}_j^*$, $j = 1$, ..., k, $i = 1$, ..., n, to identify observations with clusters. Thus, every cluster consists of all those observations whose means (at the current stage of the algorithm) are components of the same $\boldsymbol{\theta}$ vector. It is easy to see that these means may be different if observations come from different judgment ranks. In the sequel, $\varphi(\cdot|m,\, v)$ and $\Phi(\cdot|m,\, v)$ denote, respectively, the (univariate or multivariate) normal density and CDF with mean m and variance v. To save space, the conditional posterior distribution of every parameter is denoted by $\cdot|rest$, where *rest* means all the model components not to the left of the conditioning bar.

The Gibbs sampler for model 2.7 consists of the following steps:

- Step 0: Set the values of the fixed parameters M, $\mathbf{\Sigma}$, $\boldsymbol{\mu}_0$, $\mathbf{\Sigma}_0$, a, and b. Initialize all the other parameters to "plausible" values.

- Step 1: Update $(s_i,\, \boldsymbol{\theta}_i)\,|rest$. Excluding $\boldsymbol{\theta}_i$, there are k^- clusters left, containing n_1^-, ..., $n_{k^-}^-$ members. Thinking of $\boldsymbol{\theta}_i$ as a fresh draw from a Polya urn containing the remaining clusters, it is clear that it can join one of the existing k^- clusters, or it can start a new cluster. In fact, it joins cluster j, i.e. $(s_i,\, \boldsymbol{\theta}_i) = \left(j,\, \boldsymbol{\theta}_j^*\right)$, with probability

$$q_j \propto \frac{n_j^-}{M+n-1}\varphi\left(y_i|d_i\boldsymbol{\theta}_j^*, \sigma_Y^2\right),$$

for $j = 1$, ..., k^-, or it starts a new cluster, i.e. $(s_i,\, \boldsymbol{\theta}_i) = (k^- + 1,\, \boldsymbol{\theta}_{new}^*)$, with probability

$$\begin{aligned} q_0 &\propto \frac{M}{M+n-1}\int_{\mathbb{R}^K}\varphi\left(y_i|d_i\boldsymbol{\theta}, \sigma_Y^2\right)\,\varphi\left(\boldsymbol{\theta}|\boldsymbol{\mu}, \mathbf{\Sigma}\right)\,d\boldsymbol{\theta} \\ &= \frac{M}{M+n-1}\varphi\left(y_i|d_i\boldsymbol{\mu}, \sigma_Y^2 + \mathbf{\Sigma}_{r_ir_i}\right), \end{aligned} \tag{2.8}$$

where $\mathbf{\Sigma}_{r_ir_i}$ is the r_i^{th} diagonal element of $\mathbf{\Sigma}$. In the latter case, $\boldsymbol{\theta}_{new}^*$ is drawn from the distribution with density proportional to $\varphi\left(y_i|d_i\boldsymbol{\theta}_{new}^*, \sigma_Y^2\right)\cdot\varphi\left(\boldsymbol{\theta}_{new}^*|\boldsymbol{\mu}, \mathbf{\Sigma}\right)$, which turns out to be the normal distribution with mean

$$\left(\frac{1}{\sigma_Y^2}d_i'd_i + \mathbf{\Sigma}^{-1}\right)^{-1}\left(\frac{y_i}{\sigma_Y^2}d_i' + \mathbf{\Sigma}^{-1}\boldsymbol{\mu}\right)$$

and covariance matrix $\left(\frac{1}{\sigma_Y^2}d_i'd_i + \mathbf{\Sigma}^{-1}\right)^{-1}$. Repeat this step for $i = 1$, ..., n.

- Step 2: Update $\boldsymbol{\theta}_1^*$, ..., $\boldsymbol{\theta}_k^*|rest$. The $\boldsymbol{\theta}_j^*$'s are independent, and

$$[\boldsymbol{\theta}_j^*|rest] \propto \prod_{i=1,\, s_i=j}^{n}\varphi\left(y_i|d_i\boldsymbol{\theta}_j^*, \sigma_Y^2\right)\cdot\varphi\left(\boldsymbol{\theta}_j^*|\boldsymbol{\mu}, \mathbf{\Sigma}\right),$$

which turns out to be the normal density with mean

$$\left(\frac{1}{\sigma_Y^2}\boldsymbol{D}'\boldsymbol{B}\boldsymbol{D}+\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\frac{1}{\sigma_Y^2}\boldsymbol{D}'\boldsymbol{B}\boldsymbol{y}+\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \tag{2.9}$$

and covariance matrix $\left(\frac{1}{\sigma_Y^2}\boldsymbol{D}'\boldsymbol{B}\boldsymbol{D}+\boldsymbol{\Sigma}^{-1}\right)^{-1}$, where $\boldsymbol{D}$ is a $n \times K$ matrix with rows $d_1, ..., d_n$, $\boldsymbol{B} = \text{diag}\left(\mathbf{1}\left(s_1 = j\right), ..., \mathbf{1}\left(s_n = j\right)\right)$, and $\boldsymbol{y} = \left(y_1, ..., y_n\right)'$. Generating $\boldsymbol{\theta}_j^*$ from this distribution is akin to setting $\boldsymbol{\theta}_j^*$ equal to its posterior mean (given by 2.9) and contaminating it with random error. Gelfand and Smith [1990] suggest Rao-Blackwellization as a straightforward improvement of the accuracy of posterior inference. The Rao-Blackwellized update of $\boldsymbol{\theta}_j^*$ merely sets it equal to its posterior mean, given by 2.9. Repeat this step for $j = 1, ..., k$.

- Step 3: Update $\boldsymbol{\mu}|rest$. A familiar calculation shows that

$$[\boldsymbol{\mu}|rest] \propto \prod_{j=1}^{k}\varphi\left(\boldsymbol{\theta}_j^*|\boldsymbol{\mu},\boldsymbol{\Sigma}\right)\cdot\varphi\left(\boldsymbol{\mu}|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0\right),$$

which is proportional to the density of the normal distribution with mean

$$\left(k\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}_0^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}^{-1}\sum_{j=1}^{k}\boldsymbol{\theta}_j^*+\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right) \tag{2.10}$$

and variance $\left(k\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}_0^{-1}\right)^{-1}$. Once again, a Rao-Blackwellized update for $\boldsymbol{\mu}$ is given by 2.10.

- Step 4: Update $\sigma_Y^2|rest$. From

$$[\sigma_Y^2|rest] \propto \prod_{i=1}^{n}\varphi\left(y_i|d_i\boldsymbol{\theta}_i,\sigma_Y^2\right)\cdot\left(\sigma_Y^2\right)^{-(a+1)}\exp\left(-\frac{1}{b\sigma_Y^2}\right),$$

it is easy to see that the posterior distribution of σ_Y^2 is IG (a', b'), where $a' = a + \frac{n}{2}$ and $\frac{1}{b'} = \frac{1}{b} + \frac{1}{2}\sum_{i=1}^{n}\left(y_i - d_i\boldsymbol{\theta}_i\right)^2$.

- Step 5: Repeat steps 1 - 4 until convergence, and from then on, until the desired number of posterior draws is obtained.

Convergence and mixing for this Gibbs sampler are assessed visually by means of time series plots and scatter plots of parameters. The Gibbs sampler is run for an initial *burn-in period*, allowing it to converge to the stationary distribution, then it is run until T posterior draws are accumulated for inference. These recorded draws are separated by a suitable *lag* to diminish the impact of dependence.

Once the Gibbs sampler has completed its run, it is straightforward (though computationally intensive) to obtain posterior estimates of the judgment order statistic distributions. Following De Iorio et al. [2004], the posterior mean of a judgment order statistic distribution may be expressed as a predictive distribution for a future measurement from that judgment rank. That is, $E\left[f_{[r]}|data\right] = p\left(y_{n+1}|r_{n+1} = r, data\right)$, for a future observation y_{n+1} having judgment rank r_{n+1} in a set of size K. For brevity, let $\boldsymbol{\psi}$ denote a vector consisting of all the unknown parameters in the model. Then

$$\begin{aligned} p\left(y_{n+1}|r_{n+1} = r, data\right) &= E\left[p\left(y_{n+1}|r_{n+1} = r, data, \boldsymbol{\psi}\right)|data\right] \\ &\approx \frac{1}{T}\sum_{t=1}^{T} p\left(y_{n+1}|r_{n+1} = r, data, \boldsymbol{\psi}^{(t)}\right) \\ &= \frac{1}{T}\sum_{t=1}^{T} p\left(y_{n+1}|r_{n+1} = r, \boldsymbol{\psi}^{(t)}\right) \end{aligned}$$

which immediately suggests as an estimator of the judgment order statistic density $f_{[r]}$ its estimated posterior mean

$$\hat{f}_{[r]}\left(y\right) = \frac{1}{T}\sum_{t=1}^{T} p\left(y|r, \boldsymbol{\psi}^{(t)}\right), \tag{2.11}$$

where

$$p\left(y|r,\boldsymbol{\psi}^{(t)}\right)=\sum_{j=1}^{k^{(t)}}\frac{n_j^{(t)}}{M+n}\varphi\left(y|d\boldsymbol{\theta}_j^{*(t)},\sigma_Y^{2\,(t)}\right)+\frac{M}{M+n}\varphi\left(y|d\boldsymbol{\mu}^{(t)},\sigma_Y^{2\,(t)}+\boldsymbol{\Sigma}_{rr}\right), \tag{2.12}$$

and d is the $1 \times K$ vector with a one in the r^{th} position and zero elsewhere. Note that 2.12 approximates a draw from the posterior of $f_{[r]}$.

Similarly, the posterior mean of the judgment order statistic CDF $F_{[r]}$ may be estimated by

$$\hat{F}_{[r]}\left(y\right)=\frac{1}{T}\sum_{t=1}^{T}P\left(y|r,\boldsymbol{\psi}^{(t)}\right), \tag{2.13}$$

where $P\left(y|r,\boldsymbol{\psi}^{(t)}\right)$ has the same form as 2.12 with φ replaced by Φ.

Finally, equations 1.1 and 1.2 suggest estimating the population density f and CDF F by

$$\hat{f}\left(y\right)=\frac{1}{K}\sum_{r=1}^{K}\hat{f}_{[r]}\left(y\right) \tag{2.14}$$

and

$$\hat{F}\left(y\right)=\frac{1}{K}\sum_{r=1}^{K}\hat{F}_{[r]}\left(y\right), \tag{2.15}$$

respectively.

2.6 Applications

The downside to using the model proposed in the previous section is its reliance on MCMC methods and computationally intensive posterior calculations. Once these hurdles are overcome, the natural question is whether the model is any good. In this section, the model is first tested on a sample drawn from the normal distribution with perfect ranking (so that judgment order statistic distributions are fully known). This toy example is meant to investigate whether the model can recover judgment order statistic distributions with minimal prior knowledge about how exactly the data were generated. Finally, the model is applied to a real-life data set.

2.6.1 The Normal Distribution with Perfect Ranking

The "data" in this example are a balanced RSS of size $n = 30$ with set size $K = 3$ generated from the standard normal distribution with perfect ranking. That is, the data set consists of three independent random samples, each of size $m = 10$, from the distributions of the order statistics $Z_{1:3}, Z_{2:3}$, and $Z_{3:3}$, respectively, where $Z \sim N(0, 1)$. However, this information is not shared with the statistician fitting the model, who is only given the data. The purpose of this example is to test the model's ability to estimate the "correct" judgment order statistic distributions and the population distribution. The statistician chooses the hyperparameters of the model by a cursory inspection of the data.

A plot of the data versus the judgment ranks shows an increasing trend, with the data from the three judgment ranks centered roughly about -1, 0, and 1, respectively, so the statistician takes $\boldsymbol{\mu}_0 = (-1, 0, 1)'$. Without giving the matter much thought, the statistician takes $\boldsymbol{\Sigma}_0 = \text{diag}(16, 16, 16)$. To allow for some variability in the data, the Inverse Gamma parameters are taken to be $a = 2$ and $b = 0.3$, giving σ_Y^2 a prior mean of 3⅓. Since the data from the tails seem more spread out than the judgment ranked medians, the diagonal of $\boldsymbol{\Sigma}$ is set to $(1.5, 1, 1.5)'$. For the sake of simplicity, the covariance matrix $\boldsymbol{\Sigma}$ is given a constant correlation coefficient $\rho = 0.5$, so that

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1.5 & 0.6124 & 0.75 \\ 0.6124 & 1 & 0.6124 \\ 0.75 & 0.6124 & 1.5 \end{bmatrix}.$$

Finally, not being too confident in the prior specification, the statistician sets the precision parameter to $M = 0.1$ to reduce reliance on the prior.

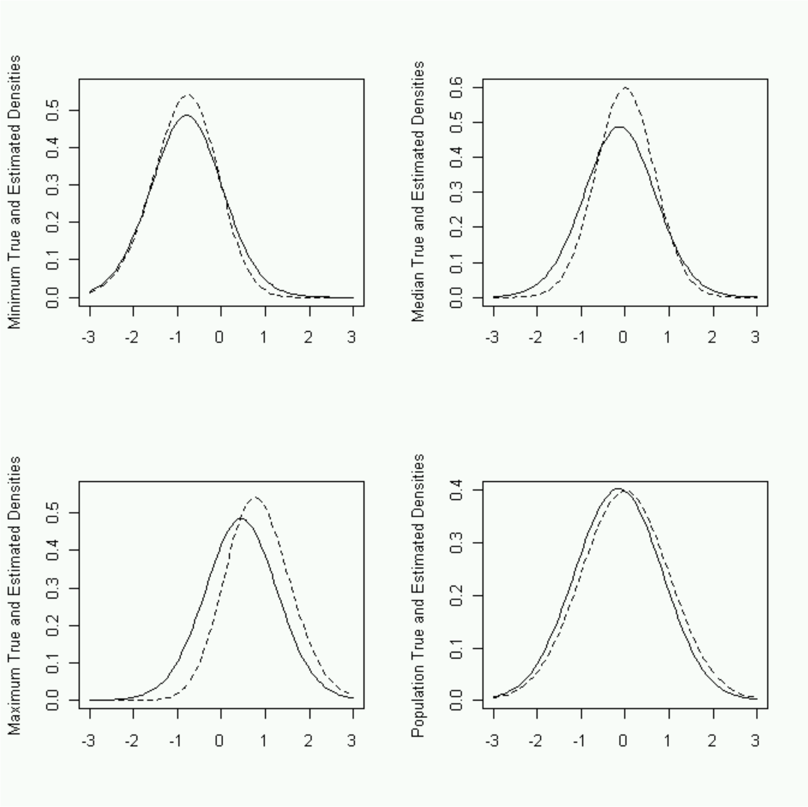

Figure 2.1: Estimated posterior means of the judgment order statistic densities (solid curves) for judgment ranks 1 (top left), 2 (top right), and 3 (bottom left) along with the true order statistic densities (dashed curves), and estimated posterior mean of the population density (solid curve, bottom right) along with the standard normal density (dashed curve).

The Gibbs sampler is run for a burn-in period of 1000 iterations. Subsequently, $T = 5000$ iterations are saved (with a lag of 20 between recorded draws). Posterior inference is based on these 5000 posterior draws. Since M is small, the Gibbs sampler rarely starts new clusters, and the mean number of clusters is 1.329. The judgment order statistic densities $f_{1:3}(t) = 3\,(1 - \Phi(t))^2\,\varphi(t)$, $f_{2:3}(t) = 6\Phi(t)\,(1 - \Phi(t))\,\varphi(t)$, and $f_{3:3}(t) = 3\Phi(t)^2\varphi(t)$ are compared to their estimated posterior means $\hat{f}_{[1]}$, $\hat{f}_{[2]}$, and $\hat{f}_{[3]}$, given by Equation 2.11, and the population density φ is estimated by Equation

2.14. These four density estimates are compared to the corresponding true densities in Figure 2.1. The estimates are evaluated on a grid of points from -3 to 3 that are 0.05 apart. Observe that even though the prior parameters were chosen without much care, the density estimates are reasonably close to their targets. This author is not aware of any other RSS procedures that can produce estimates of the judgment order statistic distributions (either density or CDF) which do not assume perfect ranking or a specific imperfect ranking model and can borrow information across judgment ranks.

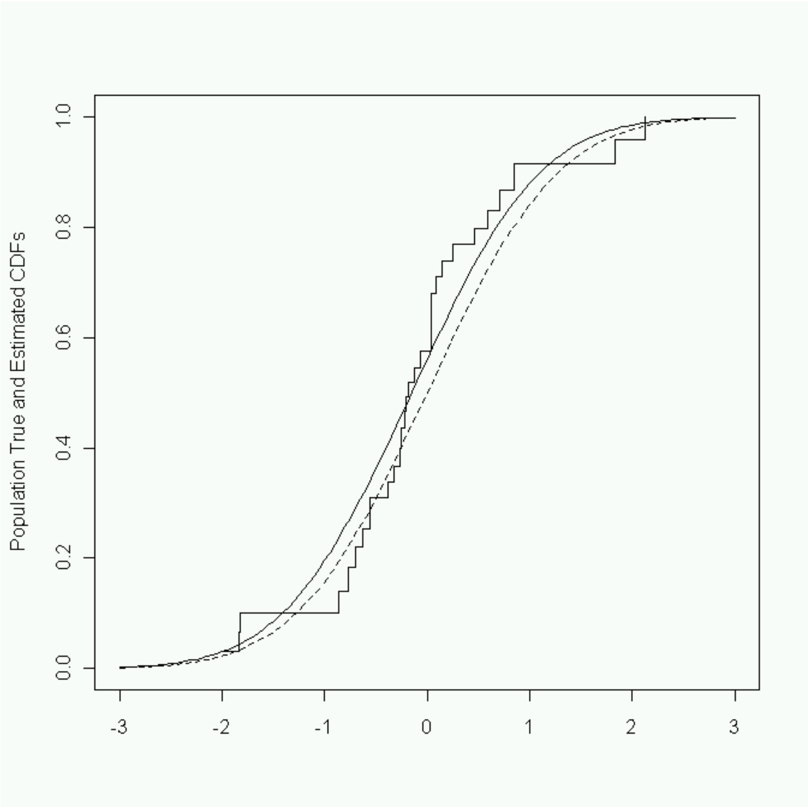

Figure 2.2: Estimated posterior mean of the population CDF (solid curve), standard normal CDF (dashed curve), and NPMLE (step function).

Likewise, posterior estimates can be calculated for the the judgment order statistic distribution functions, and these in turn can be averaged to produce an estimate of the population distribution function, using Equation 2.15. This CDF estimate is compared in Figure 2.2 to the standard normal distribution and Kvam and Samaniego's NPMLE. (The EM algorithm for the NPMLE converged in 16 iterations using tolerance level 0.0001.) The Kolmogorov-Smirnov statistic of the CDF estimate, $\sup_{y\in\mathbb{R}}\left|\hat{F}\left(y\right)-\Phi\left(y\right)\right|$, is approximately 0.061.

2.6.2 Median Household Income by State

The data set used in this example illustrates how fuzzy and amorphous judgment ranking can be. It consists of the two-year-average median household incomes of all 50 states and the District of Columbia for the years 2007-2008 (U.S. Census Bureau [2009]). Of the 51 entries, three were discarded at random, and the remaining 48 were allocated into 12 sets of size $K = 4$. A U.S. resident was shown the names of the states in each set and asked to rank them to the best of their ability in order of (perceived) increasing median household income. The 12 ranked sets were divided into $m = 3$ cycles to obtain a balanced RSS of size $n = 12$. The sample is given in Table 2.1.

The continuous data model 2.7 was fit on the log-scale. The prior parameters were set to $M = 1$, $a = b = 20$, $\boldsymbol{\mu}_0 = \left(\bar{Y}_{[1]}, \bar{Y}_{[2]}, \bar{Y}_{[3]}, \bar{Y}_{[4]}\right)'$, and $\boldsymbol{\Sigma}_0 = 0.1 \cdot \boldsymbol{I}_4$, and the matrix $\boldsymbol{\Sigma}$ was designed so that its diagonal terms reflect the higher variance of the middle two judgment order statistics relative to the extremes. Intuitively, the correlation should be highest between adjacent judgment ranks, and should decay as the judgment ranks grow further apart. Moreover, the correlation should be greater between adjacent judgment ranks when the set size is larger. Thus, a reasonable choice for the correlation in $\mathbf{\Sigma}$ between judgment ranks r_1 and r_2 is $\left(\frac{K}{K+1}\right)^{|r_1-r_2|}$.

State	Median Household Income ($)	Judgment Rank
Tennessee	41240	1
Ohio	48960	2
Utah	59062	3
California	57445	4
Kentucky	41058	1
Wisconsin	52224	2
Oregon	51947	3
Colorado	62217	4
Mississippi	37579	1
Louisiana	41232	2
New Mexico	44081	3
Delaware	53695	4

Table 2.1: RSS of size $n = 12$ used in Subsection 2.6.2.

The Gibbs sampler was run for a burn-in period of 1000 iterations, and subsequently 5000 draws (20 iterations apart) were saved for posterior inference. Time series plots and scatter plots of the parameters indicated convergence. The estimated posterior mean of the population CDF F is compared to the true CDF in Figure 2.3.

The judgment order statistic mean $\mu_{[r]}$ can be estimated using the judgment order statistic density estimate $\hat{f}_{[r]}$ via

$$\begin{aligned}\hat{\mu}_{[r]} &= \int_{-\infty}^{\infty} y\hat{f}_{[r]}\left(y\right)dy \\ &= \frac{1}{T\left(M+n\right)}\sum_{t=1}^{T}\left[\sum_{j=1}^{k^{(t)}} n_j^{(t)} d\boldsymbol{\theta}_j^{*(t)} + Md\boldsymbol{\mu}^{(t)}\right],\end{aligned}$$

and the corresponding estimator of the population mean μ is $\hat{\mu} = \frac{1}{K}\sum_{r=1}^{K}\hat{\mu}_{[r]}$. The Bayesian and frequentist estimates of $\mu_{[r]}$ are given in Table 2.2. Note that the $\bar{Y}_{[r]}$

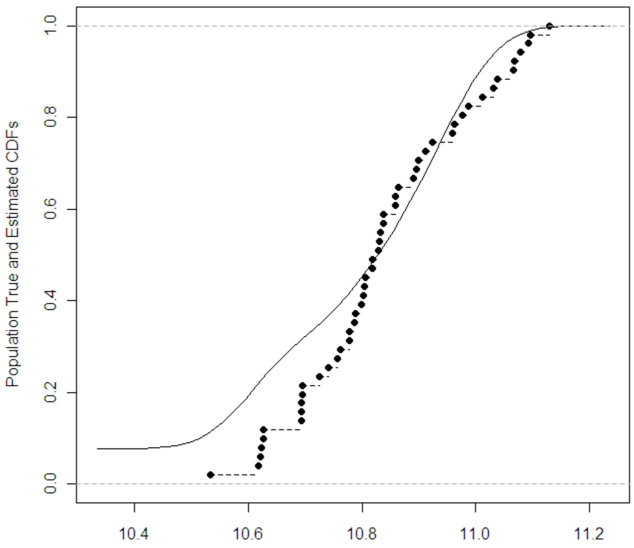

Figure 2.3: Estimated posterior mean of the population CDF (smooth curve) and true population CDF (step function) for the log median household income data.

are consistently smaller than the $\hat{\mu}_{[r]}$, and that $\hat{\mu} = 10.81917$ is closer to the true population mean 10.83998 than the balanced RSS sample mean $\bar{Y}_{RSS} = 10.79161$.

One of the features of judgment ranking highlighted by this data set is the fact that within-set ranking errors depend to a large extent on the units allocated to that set. For instance, one set under consideration consisted of Arkansas, West Virginia, Mississippi, and the District of Columbia. While the ranker felt confident that the District of Columbia had the highest median household income in the set, and less confident that Mississippi had the lowest, the remaining two states were effectively a toss-up for second and third place. The ranker did not have the option to assign them a tied rank and had them in the reverse order. The lesson here is that within-set

r	$\hat{\mu}_{[r]}$	$\bar{Y}_{[r]}$
1	10.60172	10.5947
2	10.79515	10.76301
3	10.91157	10.84604
4	10.96825	10.96268

Table 2.2: Comparison of the Bayesian estimators $\hat{\mu}_{[r]}$ and frequentist estimators $\bar{Y}_{[r]}$ of the judgment order statistic means $\mu_{[r]}$, $r = 1, 2, 3, 4$.

ranking errors vary with the units allocated to the set, but are more likely when the members of the set resemble each other (in the ranker's opinion, at least). Moreover, if the design strategy dictates in advance which ranked unit is to be measured from a given set, then units that are ranked with high accuracy (extremes in particular) may be passed over for units assigned judgment ranks that may not reflect their true rank.

Chapter 3

Bayesian Nonparametric Models for Discrete Data

3.1 Non-conjugacy: The Bane of Bayesian Statistics

When eliciting prior distributions, the statistician must always keep in mind how difficult posterior inference can be. More complicated prior structures may constitute a far more accurate representation of prior beliefs, but they may also be much harder to work with. The standard approach in low-dimensional parameter spaces is to use conjugate priors, if possible. A likelihood/prior pair is usually said to be conjugate if the resulting posterior belongs to the same family of distributions as the prior, or if the prior and the likelihood function have matching kernels. Naturally, specifying conjugate priors in high-dimensional parameter spaces can be difficult. While the Dirichlet process itself is a conjugate prior (Section 2.4), MDP models such as Model 2.6 are not. However, the MCMC algorithm for posterior simulation for the continuous data Model 2.7 was simplified a great deal by the conjugacy of the normal hyperprior, the Inverse Gamma prior on σ_Y^2, and the likelihood function.

What happens when the data density $f(y|\theta,\phi)$ is taken to be non-normal? For example, when modeling count data, the usual choices for $f(y|\theta,\phi)$ are the Bernoulli, binomial, or Poisson probability mass functions. More generally, one can assume that

$f\left(y|\theta,\phi\right)$ has the exponential family form given by McCullagh and Nelder [1999]:

$$f\left(y|\theta,\phi\right)=\exp\left[\frac{y\theta-b\left(\theta\right)}{a\left(\phi\right)}-c\left(y,\phi\right)\right], \tag{3.1}$$

where θ is the (real-valued) canonical parameter. For example, when $f\left(y|\theta,\phi\right)$ is the Bernoulli density, $p^y\left(1-p\right)^{1-y}$, for $y=0,1$, it is easy to see that $\theta=\log\left(\frac{p}{1-p}\right)$ and $b\left(\theta\right)=\ln\left(1+e^{\theta}\right)$. Similarly, when $f\left(y|\theta,\phi\right)$ is the Poisson density, $\frac{\lambda^y e^{-\lambda}}{y!}$, for $y=0$, 1, 2, ..., then $\theta=\ln\lambda$ and $b\left(\theta\right)=\exp\theta$.

Suppose one wishes to perform posterior inference for Model 2.6 with $f\left(y|\theta,\phi\right)$ as in 3.1. Naturally, one may start off trying to mimic the algorithm proposed in Section 2.5. Every iteration of the algorithm would then consist of the following updates:

- Update the clustering structure s_1, ..., s_n.
- Update the cluster locations $\boldsymbol{\theta}_1^*$, ..., $\boldsymbol{\theta}_k^*$.
- Update $\boldsymbol{\mu}$.
- Update ϕ, if it is included in the model.

(These variables were defined in Sections 2.4 and 2.5.) Whereas the third update above (of $\boldsymbol{\mu}$) requires a simple normal-normal posterior calculation, the first two updates are significantly more complicated as a result of the non-conjugacy of the normal hyperprior and $f\left(y|\theta,\phi\right)$.

3.1.1 Updating the Clustering Structure

In Step 1 of the Gibbs sampler of Section 2.5, each $\left(s_i,\boldsymbol{\theta}_i\right)$, $i=1$, ..., n, was removed from the clustering structure in turn and, following the Polya urn scheme, it was allowed to rejoin an existing cluster or start a new cluster on its own. The latter event occurs with probability q_0, which involves an integral of the form

$$\int_{\mathbb{R}^K} f\left(y_i|d_i\boldsymbol{\theta},\phi\right)\varphi\left(\boldsymbol{\theta}|\boldsymbol{\mu},\boldsymbol{\Sigma}\right)d\boldsymbol{\theta}. \tag{3.2}$$

In Equation 2.8, the integral of the form 3.2 reduced to a single evaluation of a normal density. In most other cases, however, this integral can be very hard to compute. West et al. [1994] suggest using either numerical integration techniques or Monte Carlo integration to evaluate the integral. Unfortunately, these techniques are both computationally costly (especially since they need to be performed at every stage of the sampler) and not accurate enough. MacEachern and Muller [1998] observe that the resulting Markov chain may have a stationary distribution that differs substantially from the posterior, and that the quality of the approximation is difficult to evaluate because it impacts the transition probabilities.

Rather than evaluate the integral 3.2, MacEachern and Muller propose their "No Gaps" algorithm for non-conjugate models which circumvents the integration entirely. Reasoning that the number of clusters cannot exceed the sample size n, they augment the k occupied clusters with $n-k$ unoccupied, or empty, clusters. Thus, rather than having to start a new cluster, the algorithm may simply borrow one of the empty clusters on stand-by and place it with the occupied clusters. Calculating the probability of this event (up to a constant of proportionality) requires only likelihood evaluations.

Neal [2000] notes that "there is a puzzling inefficiency in the algorithm's mechanism (...) for assigning an observation to a newly created mixture component." Neal, therefore, also proposes several algorithms, some of which are Gibbs samplers and others Metropolis-Hastings algorithms. In the algorithm of the next section, the clustering structure will be updated by combining a Metropolis step and a partial Gibbs step (Neal's Algorithm 7). The justification for this procedure is that while Metropolis steps are sufficient to obtain an ergodic chain, “such a chain would often sample inefficiently, however, since it can move an observation from one existing component to another only by passing though (sic) a possibly unlikely state in which that observation is a singleton. Such changes can be made more likely by combining these Metropolis-Hastings updates with partial Gibbs sampling updates, which are applied only to those observations that are not singletons,” and which may only move such an observation to a component associated with some other observation [Neal, 2000].

In addition, more advanced algorithms for non-conjugate models in the literature include the split-merge algorithm [Jain and Neal, 2007] and a Metropolis-Hastings algorithm based on the Laplace approximation to exponential-family likelihood functions [Guha, 2008].

3.1.2 Updating the Cluster Locations

The natural update for $\boldsymbol{\theta}_j^*$, $j = 1, ..., k$, is a random draw generated from its posterior distribution given all y_i such that $s_i = j$ (and the other model components). In Section 2.5, conjugacy led to a normal posterior, and a Rao-Blackwellized update (the posterior mean of $\boldsymbol{\theta}_j^*$) was deemed more accurate than generating a random draw that amounted to contaminating the posterior mean with random error. In the non-conjugate case, the posterior is not in general a well-known distribution, so generating a random draw from it requires a non-standard sampling sub-routine. Note that this problem arises even in fully parametric Bayesian settings. Moreover, since it is not obvious whether the mean represents a "typical" value for this posterior distribution (or even how to approximate the mean accurately), Rao-Blackwellization may not be a good idea in this setting.

In the next section, an algorithm is proposed for fitting non-conjugate models.

3.2 MCMC Algorithm for Non-conjugate Models

The algorithm proposed here uses Neal's Algorithm 7 for updating the cluster identifiers s_1, ..., s_n (Steps 1 and 2), a Random Walk Metropolis-Hastings algorithm to update the cluster locations $\boldsymbol{\theta}_1^*$, ..., $\boldsymbol{\theta}_k^*$ (Step 3), and a familiar conjugate normal-normal update for $\boldsymbol{\mu}$. The update step for ϕ is omitted, but it is quite obvious how to insert it if it is needed.

- Step 0: Set the values of the fixed parameters M, $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}_0$, and $\boldsymbol{\Sigma}_0$. Initialize all the other parameters to "plausible" values.

- Step 1: Update $(s_i,\, \boldsymbol{\theta}_i)\,|rest$ via a Metropolis step, as follows.

  - If $n_{s_i} > 1$, set $(s_i, \boldsymbol{\theta}_i) = (k+1, \boldsymbol{\theta}_{new})$, where $\boldsymbol{\theta}_{new}$ is drawn from $N\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$, with probability
  $$\min\left[1, \frac{M}{n-1}\frac{f\left(y_i|d_i\boldsymbol{\theta}_{new}, \phi\right)}{f\left(y_i|d_i\boldsymbol{\theta}_i, \phi\right)}\right]. \tag{3.3}$$

  - Otherwise, when $n_{s_i} = 1$, draw j from 1, ..., k^- with probability $\frac{n_j^-}{n-1}$, and set $(s_i, \boldsymbol{\theta}_i) = \left(j, \boldsymbol{\theta}_j^*\right)$ with probability
  $$\min\left[1, \frac{n-1}{M}\frac{f\left(y_i|d_i\boldsymbol{\theta}_j^*, \phi\right)}{f\left(y_i|d_i\boldsymbol{\theta}_i, \phi\right)}\right].$$

  If $(s_i, \boldsymbol{\theta}_i)$ is not changed, it remains at its present value. Repeat this step for $i = 1$, ..., n.

- Step 2: Update $(s_i,\, \boldsymbol{\theta}_i)\,|rest$ via a partial Gibbs step, as follows.

  - If $n_{s_i} = 1$, do nothing.
  - Otherwise, set $(s_i, \boldsymbol{\theta}_i) = \left(j, \boldsymbol{\theta}_j^*\right)$ with probability proportional to $\frac{n_j^-}{n-1} f\left(y_i | d_i \boldsymbol{\theta}_j^*, \phi\right)$, $j = 1,\, ...,\, k$. If $(s_i, \boldsymbol{\theta}_i)$ is not changed, it remains at its present value. Repeat this step for $i = 1,\, ...,\, n$.

- Step 3: Update $\boldsymbol{\theta}_1^*,\, ...,\, \boldsymbol{\theta}_k^*|rest$. The $\boldsymbol{\theta}_j^*$'s are independent, and

  $$\begin{aligned}[\boldsymbol{\theta}_j^*|rest] &\propto \prod_{i=1,\, s_i=j}^{n} f\left(y_i|d_i\boldsymbol{\theta}_j^*, \phi\right) \cdot \varphi\left(\boldsymbol{\theta}_j^*|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \\ &\propto \exp\left[-\frac{1}{a\left(\phi\right)}\mathbf{1}_n'\boldsymbol{BDb}\left(\boldsymbol{\theta}_j^*\right)\right] \cdot \varphi\left(\boldsymbol{\theta}_j^*|\boldsymbol{m}, \boldsymbol{\Sigma}\right),\end{aligned}$$

  where $\mathbf{1}_n'$ is the row vector of n ones, $\boldsymbol{B}$, $\boldsymbol{D}$, and $\boldsymbol{y}$ are as in Section 2.5, and $\boldsymbol{m} = \frac{1}{a(\phi)}\boldsymbol{\Sigma D}'\boldsymbol{By} + \boldsymbol{\mu}$. This posterior can be sampled using a Random Walk Metropolis-Hastings algorithm [Robert and Casella, 2005, page 288].

  Repeat this step for $j = 1,\, ...,\, k$.

- Step 4: Update $\boldsymbol{\mu}|rest$. This is identical to Step 3 of the Gibbs sampler in Section 2.5, and a Rao-Blackwellized update for $\boldsymbol{\mu}$ is given by the quantity 2.10.

- Step 5: Repeat Steps 1 - 4 until convergence and, subsequently, until the desired number of posterior draws is accumulated.

Note that this formulation of the discrete data problem can be expanded in obvious ways. For example, the parameter ϕ can be included to account for overdispersion. One may opt for a non-canonical link function (such as the probit link for binary data), or even a nonparametric link function [Mallick and Gelfand, 1994]. Finally, one can also introduce latent variables to model categorical or ordinal data [De Iorio et al., 2004].

Posterior inference can be carried out in much the same way as in Section 2.5, except that evaluating an expression similar to Equation 2.12 becomes problematic, since the second term on the right-hand side requires numerous evaluations of an intractable integral of the form 3.2 [Mukhopadhyay and Gelfand, 1997]. Monte Carlo approximations are deemed adequate in this setting since they are carried out in parallel, rather than in series. That is, when calculating features of the posterior distribution from the MCMC output (as opposed to running the chain), replacing one of these integrals by an approximate value will not affect all subsequent calculations. Finally, the intervening link function necessitates care when moving back and forth between the scales of the canonical and actual parameter. This is especially the case when specifying the parameters of the hyperprior.

3.3 On Questions of Allocation

Chen et al. [2005] discuss the use of balanced RSS for estimating a population proportion, and propose logistic regression as a way of pooling information from multiple concomitants for performing judgment ranking. Chen et al. [2006b] show that an unbalanced RSS scheme using Neyman allocation is optimal (with especially noticeable improvements over balanced RSS when the proportion is close to 0 or 1) "in the sense that it leads to minimum variance within the class of RSS estimators that are simple averages of the means of the order statistics."

Kohlschmidt [2009] analyzes RSS estimation of a population proportion under various missing data models. Both Kohlschmidt and Terpstra and Nelson [2005] consider optimal allocations for the naive (sample proportion) estimator and the MLE of the population proportion. In the former case, the optimal allocation corresponds to Neyman allocation, which draws measurements from all the judgment ranks (in varying proportions), whereas in the latter, the optimal allocation is extremely unbalanced and assigns all the measurements to a single judgment rank.

This dissertation's main focus is Bayesian modeling of RSS data, namely, how to analyze data once it has been collected. It does not propose any optimal rules for designing the sample or allocating measurement units to the judgment ranks. However, it is still worthwhile to think about the role that allocation plays from a Bayesian point of view. Early work indicates that good Bayesian sampling schemes are usually sequential [Basu, 1969, Zacks, 1969, 1970, Solomon and Zacks, 1970]. Ericson [1965] generalizes Neyman allocation to include prior information about the unknown stratum means in stratified sampling. (This prior information is represented by a multivariate normal distribution.) Draper and Guttman [1968] obtain optimal allocation results for the second phase of a two-phase stratified sampling procedure, using information obtained from the first stage.

Answering questions about allocation for hierarchical Bayesian nonparametric models such as Model 2.5 is difficult because there is no closed-form representation of the posterior distribution or useful summaries of the posterior. Our modeling strategy incorporates dependence between the judgment order statistic distributions $F_{[1]}$, ..., $F_{[K]}$ via the DDP, i.e., by maintaining dependence within the (symbolic) K-tuples $\boldsymbol{\theta}_j^*$ of locations about which the $F_{[r]}$'s concentrate their mass. From an intuitive standpoint, allocating measurement units to judgment rank r is a way of "pinning down" the curves $\boldsymbol{\theta}_j^*$ at judgment rank r. (This is evident, for example, in the update 2.9, in which the term $\boldsymbol{D}'\boldsymbol{By}$ is a $K \times 1$ vector whose r^{th} term is the sum of only those y_i's in cluster j which come from units assigned judgment rank r in their respective ranked sets.)

A more formal decision-theoretic approach invokes arguments from the Bayesian design of experiments. Chaloner [1984] and Lohr [1995] consider optimal Bayesian designs for linear models and one-way random effects models, respectively. Chaloner and Verdinelli [1995] present an extensive review of Bayesian experimental design. Following Lindley [1972, page 19-20], decision-making is done in two parts: first, the statistician must *choose a design* (from a set of possible designs) and collect data according to the chosen design. The data are used for inference about unknown model parameters, and finally the statistician *makes a terminal decision* about the problem at hand. The statistician's preferences and the goals of the experiment can be described by a *utility function*, $U(\cdot,\cdot,\cdot,\cdot)$, which is a function of the design, the data, the model parameters, and the terminal decision. Since the design needs to be selected before the data are collected or the terminal decision is made, the search for an optimal design requires *preposterior analysis*, i.e., averaging the utility function over the data and unknown model parameters (under a utility-maximizing terminal decision).

When utility is understood as a measure of the information gained from an experiment, a common choice for U is the expected gain in Shannon information or, equivalently, the Kullback-Leibler divergence between the posterior and prior distributions [Lindley, 1956]. This choice gives rise to the various Bayesian "alphabet" optimality criteria [Chaloner and Verdinelli, 1995]. For nonlinear design problems (i.e., when the model is non-linear or the experimenter is interested in a nonlinear function of the parameters in a linear model), the expected utility is a very complicated integral and it is often replaced by an approximation (usually based on a normal approximation to the posterior distribution). For binary response data, a common choice of design is one that is most efficient for a “best guess” of the parameters. Such a design, however, may be inefficient even for parameter values in a close neighborhood of that best guess. Numerical optimization techniques can be used to derive optimal designs that take into account uncertainty about prior guesses of the parameter values [Chaloner and Larntz, 1989]. Muller [1999] and Amzal et al. [2006] propose MCMC-based optimization techniques for Bayesian design. Finally, Hamada et al. [2001] use a genetic algorithm to find near-optimal Bayesian designs in a regression setting.

In the next section, the role of the fixed prior parameters is examined in the context of estimating the proportion of diabetics in a population of adult women. For simplicity, the analysis is based on a single balanced RSS. Chapter 4 introduces judgment post-stratification, in which judgment ranking is done *after* measurement, thereby relieving the statistician of the burden of allocating measurement units to judgment ranks.

3.4 Application: Diabetes in Pima Indian Women

In this section, the roles of the prior mass and variance parameters are explored in a simulation study using a fixed sample from a data set on the prevalence of diabetes among Pima Indian women collected by the US National Institute of Diabetes and Digestive and Kidney Diseases. (The data set `Pima.te` is part of the `MASS` package in the statistical software `R`.) The parameter of interest is the proportion of diabetic women in the population. Body Mass Index (BMI) is used as the concomitant variable for judgment ranking, since women with high BMI appear more likely to be diabetic in a cursory examination of the data. A balanced RSS of size $n = 60$ consisting of $m = 20$ cycles and using set size $K = 3$ is obtained and remains fixed throughout this simulation study. The within-rank sample proportions are given by $\hat{p}_{[1]} = 0.3$, $\hat{p}_{[2]} = 0.4$, and $\hat{p}_{[3]} = 0.45$, and the overall sample proportion is given by $\hat{p} = 0.3833$. (The true population proportion is $p = 0.3283$.) For simplicity, the components of the hyperprior mean $\boldsymbol{\mu}_0$ are set to the logits of $\hat{p}_{[1]}$, $\hat{p}_{[2]}$, and $\hat{p}_{[3]}$, and $\boldsymbol{\Sigma}_0$ is taken to be $10^{-4}\mathbf{I}_3$, thereby concentrating the prior distribution of $\boldsymbol{\mu}$ around (logits of) the usual frequentist estimates.

The total variability in the hyperprior also includes the covariance matrix $\boldsymbol{\Sigma}$. In fact, a simple pre-integration could be carried out to eliminate the parameter $\boldsymbol{\mu}$, leaving the DDP with an effective $N\left(\boldsymbol{\mu}_0, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_0\right)$ hyperprior. (See MacEachern, 1998, for more on pre-integration and other techniques for improving MCMC convergence and mixing.) For the purposes of this simulation study, the covariance matrix $\boldsymbol{\Sigma}$ is given the form $v\left[\left(\frac{K}{K+1}\right)^{|i-j|}\right]_{i,j=1}^{3}$, where v takes on the successive values 0.01, 0.1, and 0.5. When translated back to the $0-1$ scale, these values of v produce hyperpriors that assign most of their mass within 0.05 of the $\hat{p}_{[r]}$'s (for $v = 0.01$), to ones that spread out over virtually the entire unit interval (for $v = 0.5$). In addition, the mass parameter M is assigned the values 1, 10, and 20.

For each (M, v) combination, the MCMC algorithm of Section 3.2 is run with 1000 burn-in iterations and 10,000 saved posterior draws. The estimated posterior means of $p_{[1]}$, $p_{[2]}$, $p_{[3]}$, and $p = \frac{1}{3}\sum_{r=1}^{3} p_{[r]}$ (rounded to 4 decimal digits) are calculated from the MCMC output and are given in Table 3.1.

	$v = 0.01$			
	$p_{[1]}$	$p_{[2]}$	$p_{[3]}$	p
$M = 1$	0.2795	0.3640	0.4171	0.3535
$M = 10$	0.2796	0.3566	0.4053	0.3471
$M = 20$	0.2834	0.3593	0.4093	0.3489

	$v = 0.1$			
	$p_{[1]}$	$p_{[2]}$	$p_{[3]}$	p
$M = 1$	0.2513	0.3231	0.3681	0.3142
$M = 10$	0.2468	0.3102	0.3598	0.3056
$M = 20$	0.2507	0.3144	0.3636	0.3096

	$v = 0.5$			
	$p_{[1]}$	$p_{[2]}$	$p_{[3]}$	p
$M = 1$	0.2256	0.2874	0.3342	0.2824
$M = 10$	0.2197	0.2641	0.2987	0.2609
$M = 20$	0.2266	0.2727	0.3113	0.2702

Table 3.1: Estimated posterior means of $p_{[1]}$, $p_{[2]}$, $p_{[3]}$, and $p = \frac{1}{3}\sum_{r=1}^{3} p_{[r]}$ for all combinations of $M = 1$, 10, 20 and $v = 0.01$, 0.1, 0.5 for the example in Section 3.4.

The main impact of increasing the mass parameter M is increasing the number of clusters k. In fact, the mean number of clusters ranges from about 5 when $M = 1$, to about 21 when $M = 10$ and 28 when $M = 20$. This is in line with the intuitive implications of Equations 2.8 and 3.3 in which the probabilities of starting new clusters increase with M. It is also consistent with the findings of De Iorio et al. [2004] who show that k is stochastically increasing in M. In fact, larger values of M would produce estimates that differ little from the fully parametric version of the model (with no intervening Dirichlet process), whereas smaller values of M would force observations from different judgment ranks to cohabitate in the same clusters, leading to more dependence between the posterior distributions of the $p_{[r]}$'s. Hence, all else being equal, estimates obtained for $M = 1$ should be preferred to those obtained under larger values of M.

Notice that the Bayesian estimates of p are all smaller than the frequentist estimate $\hat{p} = 0.3833$. Moreover, the most accurate Bayes estimates of p appear to be those for $v = 0.1$. The intuitive explanation of this is that since the frequentist estimate differs somewhat from the true value of p, a very small value of v (such as 0.01) will concentrate the prior on the $p_{[r]}$'s very close to the frequentist estimates $\hat{p}_{[r]}$ and thus restrict the posterior's ability to place mass elsewhere. By contrast, a larger value of v (such as 0.5) may lead to a prior that is far too diffuse. Thus, placing v in the "middle ground" balances the statistician's desire for an "informative" prior with skepticism about the choice of $\boldsymbol{\mu}_0$.

Chapter 4

Judgment Post-stratification: Multiple Rankers and the Role of Set Size

4.1 Judgment Post-Stratification

In Chapter 1, RSS was introduced as a means of combining imperfect judgment rankings with measurements for statistical inference. Crucially, judgment ranking was always performed before units were chosen for measurement. That is, population units were first allocated to sets (of fixed set size K, say), and then presented to the ranker. Working on one set at a time, the ranker assigned its units their judgment ranks, from 1 to K. Finally, a subset of these judgment-ranked units was selected for measurement, in accordance with a pre-determined allocation scheme. In balanced RSS (Section 1.3), all judgment ranks are equally represented among the measurement units. However, design considerations (such as Neyman allocation, c.f. Section 3.3) can sometimes prompt the experimenter to choose an unbalanced allocation scheme (Section 1.6).

Judgment ranking, however, need not always precede the selection of measurement units. MacEachern et al. [2004] introduce judgment post-stratification as a procedure which starts out with a SRS of units from the population. Each unit is measured and then placed in a set with $K - 1$ other (unmeasured) population units. The units in each set are then judgment ranked. The end result, after all the sets are ordered, is much the same as with RSS, a sample consisting of measurements and their associated judgment ranks (except that a judgment post-stratified sample is far more likely to be unbalanced). The main difference is that the data were not collected under RSS *as a design*. This feature may be attractive to practitioners. Consider, for example, the case of investigators who wish to reap the benefits of RSS but are wary of employing a design unfamiliar to subject matter journal editors and reviewers. By using judgment post-stratification, the investigators can easily ignore the judgment ranking information (should the need arise) and revert to SRS-based analysis.

MacEachern et al. propose the average of the within-rank sample means as an estimator of the population mean, and study its properties under both RSS and judgment post-stratification. Wang et al. [2008] propose an improved version of MacEachern et al.'s estimator which uses an isotonized version of the within-rank sample means. This is intended to account for the fact that judgment order statistic distributions are usually stochastically ordered, which in turn forces monotonicity on their means. Frey and Ozturk (in press) derive additional constraints satisfied by judgment rank post-strata which are not satisfied by ordinary strata. These constraints can be used to obtain better small-sample estimates of the judgment order statistic distribution functions. Du and MacEachern [2008] use judgment post-stratification in an experimental design setup to estimate a contrast parameter (the difference between control and treatment effects).

An estimator based on judgment-ranked data will have different frequentist properties depending on whether the data arise from RSS or judgment post-stratification.

From a Bayesian standpoint, however, it should not matter whether a sample consisting of measurements and their associated judgment ranks originated from a RSS design or judgment post-stratification. (See the discussion of the role of the likelihood principle in survey sampling in Section 2.1.) As a result, the Bayesian methods developed in Chapters 2 and 3 for RSS can also be applied to judgment post-stratified data. Moreover, judgment post-stratification allows investigators to conveniently side-step the thorny problem of optimal allocation for Bayesian nonparametric models (Section 3.3).

4.2 Imprecise Rankings and Multiple Rankers

The subjectivity of judgment ranking opens up a possibility not mentioned so far, namely, that the ranker may express varying degrees of confidence in the assigned ranks. MacEachern et al. [2004] develop judgment post-stratification as a means of allowing such imprecise rankings. That is, instead of being forced to assign the ranks 1 through K to the members of a set, the ranker may assign a probability distribution to the ranks. For example, consider a set of size $K = 3$ in which the ranker has confidently identified the minimum, but is unable to differentiate between the remaining two units. Instead of being forced to assign the ranks 2 and 3 to these two units (a choice that may differ little from a coin toss), the ranker may assign their ranks the probability distribution $\left(0, \frac{1}{2}, \frac{1}{2}\right)$, which means that the ranker believes that both units are equally likely to have ranks 2 and 3.

Another advantage for judgment post-stratification over RSS is that it provides a mechanism for combining rankings from multiple rankers. Suppose an investigator wishes to elicit input from two rankers. If they disagree over the ranking of a particular set or wish to provide imprecise rankings, drawing a (balanced or unbalanced) RSS based on their rankings may be problematic, since they may differ on which unit of a ranked set should be measured. By fixing the measured units beforehand, judgment post-stratification avoids this problem. MacEachern et al. suggest using a convex combination of the rankings as a way of combining the judgment ranking opinions of multiple rankers. Wang et al. [2006] define concomitants of multivariate order statistics and obtain some of their theoretical properties, to derive estimators of the population mean that combine ranking information from several auxiliary variables.

The Bayesian methods of Chapters 2 and 3 are based on Model 2.5, which assumes a non-random covariate. That is, the model was not intended to accommodate "randomness" in the judgment ranks, which tie observations to specific coordinates of the $\boldsymbol{\theta}_j^*$ vectors. For two rankers, one may express the problem as a two-way random effects ANOVA model [De Iorio et al., 2004], taking care to impose the right constraints on the hyperprior, in addition to the standard identifiability constraints.

4.3 Quality of Ranking and the Role of Set Size

Chen et al. [2006a] conduct an empirical simulation study to examine the quality of ranking and the role of set size in the context of estimating a population proportion. They find that judgment ranking based on multiple logistic regression (combining several concomitant variables) is more accurate than ranking based on a single concomitant. Moreover, "the ranking errors increase progressively as the set size increases, which, of course, has a negative effect on the precision of a RSS estimator. The larger set size itself, on the other hand, will increase the precision of a RSS estimator if the rankings are perfect. In general, the combined effect of these opposing factors still leads to overall improvement in precision with increasing set size [Chen et al., 2005]."

Providing Bayesian answers to questions about the quality of ranking and the role of set size is more difficult, since it involves evaluating models rather than estimators. There are many factors that enter the picture, including the data (both the measurements and the judgment ranks), the choice of set size and sample size, and the fixed parameters in the model (not to mention the model itself). Naturally, one tries to fix as many of these variables as possible in order to examine the effects of changing one aspect of the problem, such as increasing the set size or improving the accuracy of ranking. However, things are not always so clear-cut. For example, when studying the role of the set size K, should the mass parameter M of the DDP in Model 2.5 be held fixed as the set size is changed, or should it vary with the set size? This is an important question since M is the mass assigned to a space of dimension K.

Frequentist properties of statistical procedures are evaluated by averaging over a large number of samples (all of the same size). Thus, any peculiar features specific to a particular sample will be washed out in the long run. For Bayesian models, however, the sample remains fixed, and the model's properties are studied by averaging over the unknown model parameters. This makes it very difficult to compare Bayesian models estimated from different samples. Any attempt to answer questions about the quality of judgment ranking and the role of set size falls under the banner of model comparison. To help control for the many unknowns in this problem, the simulation study in the following section keeps the sample measurements fixed, and uses judgment post-stratification to allocate them to sets of different set sizes by rankers with varying accuracy.

4.3.1 Simulation Study Using the Normal Distribution

This computationally intensive simulation study was carried out with the aim of simultaneously exploring the effect of both the quality of ranking and set size. The "sample measurements," which were fixed from the outset, consist of a SRS of size $n = 100$, Y_1, ..., Y_{100}, from the standard normal distribution. The sample size was chosen to allow for a wider variety of possible set sizes, although the simulation can be replicated with a smaller sample size. The set sizes used in the study are $K = 2$, 3, 4, 5, 8, and 10.

The normal distribution provides a simple and efficient method for judgment ranking with different levels of accuracy. Using properties of the bivariate normal distribution, one can simply create a concomitant variable having correlation coefficient $\rho \in (0, 1)$ with the variable of interest. Smaller values of ρ lead to judgment ranking that is hardly any better than random, whereas larger values of ρ make ranking more accurate. The values of ρ used in this study are 0.2, 0.5, and 0.8. For each value of ρ, a concomitant variable X is generated from the $N\left(\rho y,\, 1 - \rho^2\right)$ distribution, and for each particular set size K, each measurement unit is ranked according to its X value in a set of size K, with $K - 1$ hypothetical "other population units," i.e., draws from X. The end result is a judgment post-stratified sample, $\left(r_1, Y_1\right)$, ..., $\left(r_{100}, Y_{100}\right)$. Model 2.7 is fit to the data with $M = 0.1$, $\mathbf{\Sigma} = 3\left[\left(\frac{K}{K+1}\right)^{|i-j|}\right]_{i,j=1}^{K}$, $\boldsymbol{\mu}_0 = \rho\left(E\left(Z_{1:K}\right), ..., E\left(Z_{K:K}\right)\right)'$, $\mathbf{\Sigma}_0 = diag\,(25, ..., 25)$, $a = 2$, and $b = 0.3$. The estimated posterior mean $\hat{F}$ of the population CDF is obtained from the MCMC output via Equation 2.15.

There are many ways of assessing the fit of Bayesian nonparametric models. (See, for example, Mukhopadhyay and Gelfand, 1997.) This simulation study only considered the posterior dispersion about $\hat{F}$, for each (ρ, K) combination, estimated by the posterior Integrated Mean Square Error (IMSE),

$$\int_{-\infty}^{\infty} E\left[\left(\hat{F}\left(y\right) - F\left(y\right)\right)^{2} |data\right] dy,$$

where the inner expectation was estimated point-wise from the MCMC output, and the integral was crudely approximated via the Trapezoidal Rule. Finally, to account for variability of the judgment ranking process itself, the judgment ranking was carried out 10 times and the resulting IMSE values were averaged.

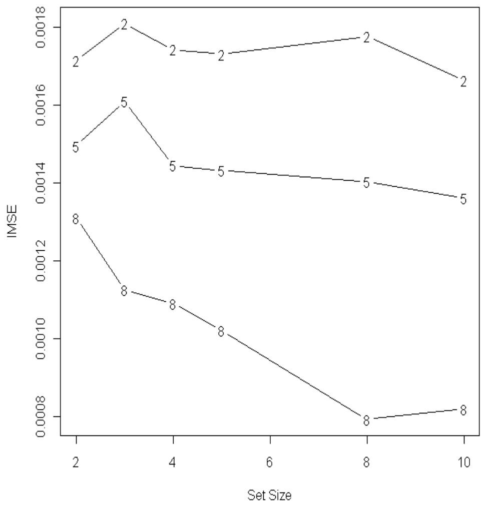

Figure 4.1: Approximate estimated posterior IMSE for set sizes K = 2, 3, 4, 5, 8, and 10 and $\rho = 0.2$, 0.5, and 0.8, averaged over 10 judgment rankings. The plotting symbol is 10ρ.

Figure 4.1 plots the average approximate estimated posterior IMSE against the set size K. The plotting symbol used is 10ρ. Overall, the magnitude of the IMSE is small, indicating a strong concentration of the posterior about its mean. This is only to be expected since the sample size used is fairly large. Moreover, when ranking is not much better than random, the IMSE does not vary much with set size. On the other hand, when ranking is very accurate, the IMSE decreases with set size, implying that the posterior becomes more concentrated. (Compare to the frequentist properties of the RSS mean in Section 1.4.) It is worth noting that replicating this simulation study usually leads to different plots, but the overall trends remain largely the same.

4.3.2 Application to the Auditing Data Set

This section revisits the auditing application of Section 1.4 and its main goal of estimating the mean audited value of a population of accounting books. Recall that book value is a very accurate concomitant ranking variable in this example, except in the presence of massive accounting fraud. For the purposes of this application, a SRS of size $n = 100$ audited accounting books is held fixed. (The sample mean audited value is \$371.53, whereas the population mean audited value is about \$375.57.) The accounting books in the sample are judgment post-stratified in sets of size K, which takes on the successive values 2, 4, and 8. Table 4.1 reports the frequentist estimates (within-rank sample means) and Bayesian estimates (estimated posterior means) of the K judgment order statistic means $\mu_{[1]}$, ..., $\mu_{[K]}$, as well as the frequentist estimate of μ (the average of the within-rank sample means) and its Bayesian estimate, all rounded to two decimal digits.

$K = 2$	μ	$\mu_{[1]}$	$\mu_{[2]}$
Frequentist	375.67	306.66	444.69
Bayesian	375.52	306.67	444.38

$K = 4$	μ	$\mu_{[1]}$	$\mu_{[2]}$	$\mu_{[3]}$	$\mu_{[4]}$
Frequentist	381.29	267.11	368.68	429.25	460.14
Bayesian	381.27	267.30	368.38	429.31	460.09

$K = 8$	μ	$\mu_{[1]}$	$\mu_{[2]}$	$\mu_{[3]}$	$\mu_{[4]}$	$\mu_{[5]}$	$\mu_{[6]}$	$\mu_{[7]}$	$\mu_{[8]}$
Frequentist	378.70	213.07	295.79	337.20	395.40	416.59	447.65	417.76	506.11
Bayesian	378.78	213.22	295.82	337.45	395.48	416.72	447.69	417.76	506.12

Table 4.1: Frequentist and Bayesian estimates of the judgment order statistic means and the population mean for a SRS of $n = 100$ accounting books judgment post-stratified into sets of size $K = 2$, 4, and 8.

For the Bayesian estimates, the fixed parameters of the MCMC were set to $\boldsymbol{\mu}_0 = \left(\bar{Y}_{[1]}, ..., \bar{Y}_{[K]}\right)'$, $M = 1$, $\boldsymbol{\Sigma} = 100\left[\left(\frac{K}{K+1}\right)^{|i-j|}\right]_{i,j=1}^{K}$, $\boldsymbol{\Sigma}_0 = 100 \cdot \mathbf{I}_K$, $a = 2$, and $b = 0.01$. Notice that, for this particular application, the frequentist and Bayesian estimates are all remarkably close. Moreover, this remains the case for different specifications of the prior parameters, suggesting that there is little to be gained from resorting to these Bayesian methods in this particular instance.

Chapter 5

Ranked Set Sampling with a Concomitant Variable

5.1 Concomitants of Order Statistics and Judgment Ranking

The examples of the previous chapters illustrate the two main sources of the imperfect ranking information which is used in RSS: subjective ranking, as in the examples of Section 1.1 (the termites example) and Subsection 2.6.2 (median household income by state), and ranking by a concomitant variable, as in the examples of Sections 1.4 (the auditing application) and 3.4 (diabetes among Pima Indian women). In the case of judgment ranking by a concomitant variable, the concomitant may even be the output of a model that combines several auxiliary variables, such as logistic regression [Chen et al., 2006b]. For the purposes of this chapter, it is enough to assume that the concomitant variable X has a continuous distribution (to avoid ties in ranking) and that the variable of interest Y "tends to increase" with X. For simplicity, Y will also be assumed to follow a continuous distribution (although it is important to note that this does not imply that the joint distribution of (X, Y) is also continuous). This section reviews some basic results from the theory of concomitants of order statistics and examines their implications for RSS, including the choice of set size and allocation.

Since X is used for ranking sets of population units, it is important that X be easily obtained (if not already available) for all population units. For example, in a medical setting, measurement of the variable of interest Y may require intrusive or expensive procedures, but X may be based on easy-to-obtain factors, such as age, Body Mass Index (BMI), and blood pressure. To expand upon an argument already begun in Section 2.3, consider a generic ranked set of K population units, represented symbolically by the K pairs $\left(X_{(1)}, Y_{[1]}\right)$, ..., $\left(X_{(K)}, Y_{[K]}\right)$. Here, $X_{(1)} < ... < X_{(K)}$ are necessarily known (how could the set have been ranked otherwise?), but none of the Y values are known before measurement. It is well known that the random variables $Y_{[1]}$, ..., $Y_{[K]}$ are conditionally independent given $x_{(1)}$, ..., $x_{(K)}$, and that the conditional density of $Y_{[r]}$ given $x_{(1)}$, ..., $x_{(K)}$ is merely the population conditional density of Y given $X = x_{(r)}$, namely $f\left(y_{[r]}|x_{(r)}\right)$ [Yang, 1977, David and Nagaraja, 2003, page 145]. (Notice the discrepancy between the RSS literature, which labels X as the concomitant variable for the quantity of interest Y, and the order statistics literature, in which $Y_{[r]}$ is called the concomitant of the order statistic $X_{(r)}$.)

If a statistician draws a RSS of size n under a (balanced or unbalanced) design that measures only one unit from each ranked set, then the measurements $Y_{[r_1]1}$, ..., $Y_{[r_n]n}$ are independent but not identically distributed (INID) conditional on the X values. They are also marginally independent since they come from independent sets. Moreover, the statistician has amassed a SRS of $n \times K$ X's, which can be used to estimate the marginal distribution of X. (Note that the marginal distribution of X plays the role of a "mixing distribution" when averaging over the conditional distributions of Y given different X values to recover features of the marginal distribution of Y.) In particular, the set of X values at which a corresponding Y is measured constitute a RSS following the same design and drawn with perfect ranking. Furthermore, since the Y values from a single ranked set are conditionally independent, it is also possible to employ a design that calls for measurement of more than one unit per ranked set [Wang et al., 2004] without introducing dependence among the measured Y values, when inference is carried out conditionally on the X values. Regardless of the design, the end result is a sample consisting of $(X,\, Y)$ pairs as well as unpaired X's, in which the Y's are conditionally independent, and the entire set of X's is a SRS from the marginal distribution of X. Thus, when modeling the evolution of the conditional distributions of Y over the range of X, the set size K is irrelevant (as long as the sample size n remains fixed), and so are the specifics of the design. A balanced RSS, for example, would spread out the measurements over the spectrum of the concomitant, whereas a SRS is more likely to be a clump of adjacent points.

As argued in Section 2.3, "an essential feature of inference for RSS is the ability to learn about the (conditional) distribution of Y across the range of the concomitant." The next section extends Model 2.5 by replacing the multivariate normal hyperprior of the DDP with a Gaussian Stochastic Process (GSP). Thus, the K-dimensional vector $\boldsymbol{\theta}$ is replaced with a curve $\theta\,(x)$. In addition, this model allows the user to "fill in the blanks" and obtain posterior estimates of the conditional distribution of Y even at those X values where no measurements were taken.

5.2 A Model for RSS with a Concomitant Variable

This section proposes a model for the conditional distribution of Y given $X = x$. By conditional independence of the Y's given the X's, it is enough to consider the $(X,\, Y)$ pairs in the sample and disregard the unpaired X's for the moment. Let $(x_1,\, Y_1)\,,\, ...,\, (x_n, Y_n)$ denote the sample measurements and their associated values of the concomitant variable, and suppose these have been ordered so that $x_1 <\, ... <\, x_n$.

Also, let $\mathbf{x} = (x_1, ..., x_n)'$ and $\mathbf{X} = (\mathbf{1}_n, \mathbf{x})$. The base measure of the DDP is taken to be a GSP with mean function $\beta_0 + \beta_1 x$, constant variance σ_θ^2, and correlation function $R(\cdot, \cdot)$. Finally, the model specification is completed with $\boldsymbol{\beta} = (\beta_0, \beta_1)' \sim N_2\left(\boldsymbol{\beta}^{(0)}, \sigma_\beta^2 \mathbf{I}_2\right)$. To summarize, the model is given by

$$\begin{aligned} Y_i|\theta_i(x), \sigma_Y^2 &\sim N\left(\theta_i(x_i), \sigma_Y^2\right), i = 1, ..., n \qquad (5.1) \\ \theta_1(x), ..., \theta_n(x)|P &\sim P \\ P|\boldsymbol{\beta} &\sim DP\left(M, GSP\left(\beta_0 + \beta_1 x, \sigma_\theta^2 R(\cdot, \cdot)\right)\right) \\ \boldsymbol{\beta} &\sim N\left(\boldsymbol{\beta}^{(0)}, \sigma_\beta^2 \mathbf{I}_2\right) \\ \sigma_Y^2 &\sim IG(a, b) \end{aligned}$$

The Gibbs sampler of Section 2.5 is easily modified to fit this model, using the usual technique of dealing with Gaussian Stochastic Processes via multivariate normal distributions. To that end, let k once again denote the number of distinct curves $\theta_i(x)$, namely, $\theta_1^*(x), ..., \theta_k^*(x)$, and define the random variables $s_1, ..., s_n$ by $s_i = j$ iff $\theta_i(x) = \theta_j^*(x)$. As before, let n_j denote the number of curves $\theta_i(x)$ equal to $\theta_j^*(x)$. Finally, let $\boldsymbol{\theta}_i = \theta_i(\mathbf{x})$ and $\boldsymbol{\theta}_j^* = \theta_j^*(\mathbf{x})$.

The Gibbs sampler for model 5.1 consists of the following steps:

- Step 0: Set the values of the fixed parameters M, σ_θ^2, $R(\cdot, \cdot)$, $\boldsymbol{\beta}^{(0)}$, σ_β^2, a, and b. Initialize all the other parameters to “plausible” values.

- Step 1: Update $(s_i, \boldsymbol{\theta}_i)|rest$. Excluding $\boldsymbol{\theta}_i$, there are k^- clusters left, containing $n_1^-, ..., n_{k^-}^-$ members. Following the Polya urn scheme, $\boldsymbol{\theta}_i$ can join one of the existing k^- clusters, or it can start a new cluster. In fact, it joins cluster j, i.e.

$(s_i,\, \boldsymbol{\theta}_i) = \left(j,\, \boldsymbol{\theta}_j^*\right)$, with probability

$$q_j \propto \frac{n_j^-}{M+n-1} \varphi\left(y_i | \theta_j^*\left(x_i\right), \sigma_Y^2\right),$$

for $j = 1, ..., k^-$, or it starts a new cluster, i.e. $(s_i,\, \boldsymbol{\theta}_i) = (k^- + 1,\, \boldsymbol{\theta}_{new}^*)$, with probability

$$\begin{aligned} q_0 &\propto \frac{M}{M+n-1} \int_{-\infty}^{\infty} \varphi\left(y_i | \theta, \sigma_Y^2\right)\, \varphi\left(\theta | \beta_0 + \beta_1 x_i, \sigma_\theta^2\right)\, d\theta \\ &= \frac{M}{M+n-1} \varphi\left(y_i | \beta_0 + \beta_1 x_i, \sigma_Y^2 + \sigma_\theta^2\right). \end{aligned}$$

In the latter case, $\boldsymbol{\theta}_{new}^*$ is drawn from the distribution with density proportional to $\varphi\left(y_i | d_i \boldsymbol{\theta}_{new}^*, \sigma_Y^2\right) \cdot \varphi\left(\boldsymbol{\theta}_{new}^* | \mathbf{X}\boldsymbol{\beta}, \sigma_\theta^2 \mathbf{R}\right)$, where d_i is a $1 \times n$ vector with 1 in the i^{th} position and 0 elsewhere, and $\mathbf{R} = \left[R\left(x_i, x_j\right)\right]_{i,j=1}^n$. This density is that of the normal distribution with mean

$$\left(\frac{1}{\sigma_Y^2} d_i' d_i + \frac{1}{\sigma_\theta^2} \mathbf{R}^{-1}\right)^{-1} \left(\frac{y_i}{\sigma_Y^2} d_i' + \frac{1}{\sigma_\theta^2} \mathbf{R}^{-1} \mathbf{X}\boldsymbol{\beta}\right)$$

and covariance matrix $\left(\frac{1}{\sigma_Y^2} d_i' d_i + \frac{1}{\sigma_\theta^2} \mathbf{R}^{-1}\right)^{-1}$. Repeat this step for $i = 1, ..., n$.

- Step 2: Update $\boldsymbol{\theta}_1^*, ..., \boldsymbol{\theta}_k^* | rest$. The $\boldsymbol{\theta}_j^*$'s are independent, and

$$[\boldsymbol{\theta}_j^* | rest] \propto \prod_{i=1,\, s_i=j}^{n} \varphi\left(y_i | \theta_j^*\left(x_i\right), \sigma_Y^2\right) \cdot \varphi\left(\boldsymbol{\theta}_j^* | \mathbf{X}\boldsymbol{\beta}, \sigma_\theta^2 \mathbf{R}\right),$$

which turns out to be the normal density with mean

$$\left(\frac{1}{\sigma_Y^2} \boldsymbol{B} + \frac{1}{\sigma_\theta^2} \mathbf{R}^{-1}\right)^{-1} \left(\frac{1}{\sigma_Y^2} \boldsymbol{B}\boldsymbol{y} + \frac{1}{\sigma_\theta^2} \mathbf{R}^{-1} \mathbf{X}\boldsymbol{\beta}\right)$$

and covariance matrix $\left(\frac{1}{\sigma_Y^2} \boldsymbol{B} + \frac{1}{\sigma_\theta^2} \mathbf{R}^{-1}\right)^{-1}$, where $\boldsymbol{B} = \mathrm{diag}\left(\mathbf{1}\left(s_1 = j\right), ..., \mathbf{1}\left(s_n = j\right)\right)$ and $\boldsymbol{y} = \left(y_1, ..., y_n\right)'$. Repeat this step for $j = 1, ..., k$.

- Step 3: Update $\boldsymbol{\beta}|rest$. A familiar calculation shows that

$$[\boldsymbol{\beta}|rest] \propto \prod_{j=1}^{k} \varphi\left(\boldsymbol{\theta}_j^*|\mathbf{X}\boldsymbol{\beta}, \sigma_\theta^2\mathbf{R}\right) \cdot \varphi\left(\boldsymbol{\beta}|\boldsymbol{\beta}^{(0)}, \sigma_\beta^2\mathbf{I}_2\right),$$

  which is proportional to the density of the normal distribution with mean

$$\left(\frac{k}{\sigma_\theta^2}\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} + \frac{1}{\sigma_\beta^2}\mathbf{I}_2\right)^{-1}\left(\frac{1}{\sigma_\theta^2}\mathbf{X}'\mathbf{R}^{-1}\sum_{j=1}^{k}\boldsymbol{\theta}_j^* + \frac{1}{\sigma_\beta^2}\boldsymbol{\beta}^{(0)}\right)$$

  and variance $\left(\frac{k}{\sigma_\theta^2}\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} + \frac{1}{\sigma_\beta^2}\mathbf{I}_2\right)^{-1}$.

- Step 4: Update $\sigma_Y^2|rest$. From

$$[\sigma_Y^2|rest] \propto \prod_{i=1}^{n} \varphi\left(y_i|\theta_i\left(x_i\right), \sigma_Y^2\right) \cdot \left(\sigma_Y^2\right)^{-(a+1)} \exp\left(-\frac{1}{b\sigma_Y^2}\right),$$

  it is easy to see that the posterior distribution of σ_Y^2 is $IG\left(a',\, b'\right)$, where $a' = a + \frac{n}{2}$ and $\frac{1}{b'} = \frac{1}{b} + \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \theta_i\left(x_i\right)\right)^2$.

- Step 5: Repeat steps 1 - 4 until convergence, and from then on, until T posterior draws are obtained.

As before, one can replace some updates with Rao-Blackwellized versions that merely require computation of the posterior means (rather than random number generation).

To obtain a predictive density for Y at $X = x_0$, let $\boldsymbol{\psi}$ denote the unknown model parameters, and note that

$$\begin{aligned} p\left(y|x_0,\, data\right) &= E\left[p\left(y|x_0,\, \boldsymbol{\psi},\, data\right)\right] \\ &\approx \frac{1}{T}\sum_{t=1}^{T} p\left(y|x_0,\, \boldsymbol{\psi}^{(t)}\right), \end{aligned}$$

where

$$\begin{aligned} p\left(y|x_0,\, \boldsymbol{\psi}^{(t)}\right) = \frac{1}{M+n}\sum_{j=1}^{k^{(t)}} n_j^{(t)} p\left(y|x_0,\, \theta_j^{*(t)}\left(\mathbf{x}\right),\, \sigma_Y^{2(t)}\right) \\ + \frac{M}{M+n}\varphi\left(y|\beta_0^{(t)} + \beta_1^{(t)}x_0,\, \sigma_Y^{2(t)} + \sigma_\theta^2\right). \end{aligned}$$

Next, note that

$$\begin{pmatrix} \theta_j^{*(t)}(x_0) \\ \theta_j^{*(t)}(\mathbf{x}) \end{pmatrix} \sim N_{n+1}\left(\begin{pmatrix} 1\, x_0 \\ \mathbf{X} \end{pmatrix} \boldsymbol{\beta}^{(t)},\, \sigma_\theta^2 \begin{pmatrix} 1 & \mathbf{r}_0' \\ \mathbf{r}_0 & \mathbf{R} \end{pmatrix}\right),$$

where $\mathbf{r}_0 = \left(R\left(x_0, x_1\right),\, ...,\, R\left(x_0, x_n\right)\right)'$. Recall that this implies that

$$\theta_j^{*(t)}(x_0) \,|\theta_j^{*(t)}(\mathbf{x}) \sim N\left(m_j^{(t)}(x_0), V(x_0)\right),$$

where

$$m_j^{(t)}(x_0) = \beta_0^{(t)} + \beta_1^{(t)} x_0 + \mathbf{r}_0' \mathbf{R}^{-1}\left(\theta_j^{*(t)}(\mathbf{x}) - \mathbf{X}\boldsymbol{\beta}^{(t)}\right)$$

and $V(x_0) = \sigma_\theta^2\left(1 - \mathbf{r}_0'\mathbf{R}^{-1}\mathbf{r}_0\right)$. Therefore,

$$\begin{aligned} p\left(y|x_0,\, \theta_j^{*(t)}(\mathbf{x}),\, \sigma_Y^{2(t)}\right) &= \int_{-\infty}^{\infty} \varphi\left(y|\theta_j^{*(t)}(x_0), \sigma_Y^{2(t)}\right) \cdot \\ &\qquad \varphi\left(\theta_j^{*(t)}(x_0)\,|m_j^{(t)}(x_0), V(x_0)\right)\, d\theta_j^{*(t)}(x_0) \\ &= \varphi\left(y|m_j^{(t)}(x_0), \sigma_Y^{2(t)} + V(x_0)\right). \end{aligned}$$

Hence,

$$\begin{aligned} p\left(y|x_0,\, \boldsymbol{\psi}^{(t)}\right) = \frac{1}{M+n}\sum_{j=1}^{k^{(t)}} n_j^{(t)} \varphi\left(y|m_j^{(t)}(x_0), \sigma_Y^{2(t)} + V(x_0)\right) \\ + \frac{M}{M+n}\varphi\left(y|\beta_0^{(t)} + \beta_1^{(t)}x_0,\, \sigma_Y^{2(t)} + \sigma_\theta^2\right), \end{aligned}$$

and the predictive density for Y at x_0 is given by

$$\hat{f}(y|x_0) = \frac{1}{T}\sum_{t=1}^{T} p\left(y|x_0,\, \boldsymbol{\psi}^{(t)}\right). \tag{5.2}$$

Moreover, the conditional mean of Y given x_0 may be estimated by

$$\begin{aligned} \hat{E}\left[Y|x_0\right] &= \int_{-\infty}^{\infty} y\hat{f}(y|x_0)\, dy \\ &= \frac{1}{T}\sum_{t=1}^{T}\left[\beta_0^{(t)} + \beta_1^{(t)}x_0 + \frac{1}{M+n}\mathbf{r}_0'\mathbf{R}^{-1}\sum_{j=1}^{k^{(t)}} n_j^{(t)}\left(\theta_j^{*(t)}(\mathbf{x}) - \mathbf{X}\boldsymbol{\beta}^{(t)}\right)\right]. \end{aligned} \tag{5.3}$$

5.3 Application: Weights and Prices of Cars

The methods derived in the previous section are applied to a data set consisting of the weights (in thousands of pounds) and prices (in thousands of dollars) of 93 car models. (The data set `Cars93` is available in the `R` package `MASS`, and more details about it can be found in Lock, 1993.) Treating price as the variable of interest (Y) and weight as the concomitant variable (X), a balanced RSS of size $n = 30$ with $m = 15$ cycles and set size $K = 2$ is obtained. A cursory examination of the data shows that there is a positive heteroskedastic relationship between weight and price; that is, as car weight increases, so too do the prices and the variability in price. The purpose of this application is to compare the Ordinary Least Squares (OLS) regression line and the estimator 5.3 as estimators of the population regression function $E\left[Y|x\right]$. These are given in Figure 5.1.

The unconditional covariance (under Model 5.1) between the Y values at two generic points x_1 and x_2 is given by

$$Cov\left(Y\left(x_1\right), Y\left(x_2\right)\right) = \frac{\sigma_\theta^2}{M+1} R\left(x_1, x_2\right),$$

implying that a small value for M increases the amount of information extracted from the sample (x, y) pairs. The correlation function is given the form $R\left(x_1, x_2\right) = \exp\left[-\gamma\left(x_1 - x_2\right)^2\right]$, where the choice of γ is driven primarily by the need to make the correlation matrix $\mathbf{R}$ invertible. The estimator 5.3 plotted in Figure 5.1 was derived for the specific values $M = 0.1$, $\sigma_\theta^2 = 10$, $\gamma = 25$, $\sigma_\beta^2 = 10,000$, $a = 2$, and $b = 0.01$. In addition, $\boldsymbol{\beta}^{(0)}$ was set to the OLS estimates for the regression of y_1, ..., y_n on x_1, ..., x_n. By trial and error, it was found that the estimator 5.3 is very close to the regression line when σ_θ^2 is small, but very serpentine when it is large.

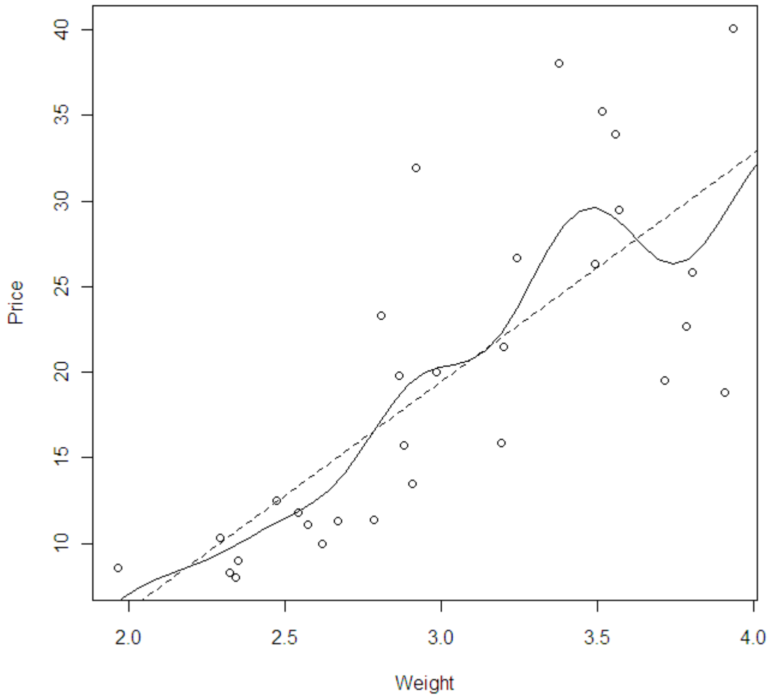

Figure 5.1: Plot of the estimator 5.3 (solid curve), the OLS regression line (dashed line), and the RSS of size $n = 30$ for the application of Section 5.3.

Finally, note that while one would expect $E[y|x]$ to be monotone in this setting, this estimator does take a small downward plunge for weights around 3500 pounds, where it clearly mistakes the increasing spread in the data for a dip in its target function.

Chapter 6

Conclusions and Future Research

The research process that culminated with the ideas presented in this dissertation began with three seemingly disparate points of inquiry:

1. Devising effective sequential sampling procedures for RSS:

   There are many reasons why a statistician would choose to carry out sequential sampling. These may range from the purely theoretical (inference carried out under abstract loss criteria and stopping rules) to real-life applications (such as an auditor using previously-collected data to net a larger number of fraudulent accounting records while not exceeding budget constraints). Regardless of the motivation behind it, one can easily envision a place for sequential sampling procedures in RSS by asking the question: given a sample of measurements and their associated judgment ranks, does one have any preference regarding which ranked units to measure from future sets? Answering this question is clearly tied to the specifics of the application at hand and, while this dissertation does not address sequential sampling procedures for RSS, it does provide a meaningful starting point by showing how to calculate predictive distributions for measurements from future ranked sets.

2. Estimation of infinite-dimensional parameters:

   When the variable of interest is absolutely continuous, estimating the judgment order statistic distribution functions $F_{[r]}$ and densities $f_{[r]}$ is no straightforward task. On the one hand, one may estimate them using only the measurements from judgment rank r (for example, by a kernel density estimate for $f_{[r]}$ or the empirical distribution function for $F_{[r]}$, c.f. Stokes and Sager, 1988), but then one would be treating the judgment order statistic distributions as though they were independent populations and not making full use of the information provided by judgment ranking. At the other extreme, one could assume that judgment ranking is perfect, in which case the judgment order statistic distributions reduce to their true order statistic counterparts. The latter in turn are completely determined by the population distribution function and density, which can be estimated using the entire sample [Kvam and Samaniego, 1994]. The approach proposed in this dissertation neither treats the judgment order statistic distributions as independent, nor does it force stringent assumptions upon the judgment ranking process. Instead, it uses hierarchical Bayesian modeling to maintain independence between the data while embedding dependence between the judgment order statistic distributions (using the DDP). This construction was shown in Chapter 2 to be equivalent to a Bayesian random-effects ANOVA with a nonparametric prior on the random effects structure, which is analogous to the frequentist ANOVA interpretation of RSS from Chapter 1. Furthermore, this construction allows information to be shared freely among the judgment ranks.

3. Skepticism of common assumptions:

> As discussed in Section 2.2, many estimators and procedures in the RSS literature are derived under some very restrictive assumptions, such as perfect ranking or absolute continuity of the population distribution, among others. This dissertation seeks a more comprehensive understanding of RSS that accepts the inevitable errors in judgment ranking and holds for both discrete and continuous data.

The discussion in Section 2.3 is a very accurate depiction of the thought process that led to the results of the subsequent section. The adoption of this Bayesian approach may be met with two criticisms. The first is the subjectivity of the prior specification, implying that two reasonable statisticians can arrive at varying inferences by specifying their own priors. This criticism is easily dismissed in the RSS setting, since two subjective rankers may also assign different judgment ranks to the same set of units, leading to different inferences. The second, more substantial, criticism is the computational burden of MCMC models. Even with today's powerful computers, the algorithms presented in this dissertation need at least several minutes to complete their run. The time expended on writing and checking the code, trying different values of the fixed model parameters, and examining model diagnostics may be substantial, and all of these steps need to be carried out before actual posterior inference is performed.

There remain some interesting questions which are not resolved in this dissertation but which may be addressed in light of its results. Most of these fall under the banner of Bayesian design, and they include:

1. the choice of sample size, n, which depends heavily on the cost of measurement.

2. the choice of set size, K, which depends on the cost and quality of ranking.

3. optimal allocation of measurement units to judgment ranks (c.f. Section 3.3).

4. sequential sampling schemes for RSS.

Finally, one more possible area of research is the pursuit of large-sample (asymptotic) properties of the posterior distribution and whether they are at all useful in simplifying computations. Naturally, this question is only worth asking for Model 2.5 and its derivatives when the set size K (and the dimension of the DDP) is fixed.

Bibliography

B. Amzal, F. Y. Bois, E. Parent, and C. P. Robert. Bayesian-Optimal Design via Interacting Particle Systems. *Journal of the American Statistical Association*, 101: 773–785, 2006.

V. Barnett and K. Moore. Best Linear Unbiased Estimates in Ranked-set Sampling with Particular Reference to Imperfect Ordering. *Journal of Applied Statistics*, 24: 697–710, 1997.

D. Basu. Role of the Sufficiency and Likelihood Principles in Sample Survey Theory. *Sankhya: The Indian Journal of Statistics, Series A*, 31:441–454, 1969.

D. Blackwell and J. B. MacQueen. Ferguson Distributions via Polya Urn Schemes. *The Annals of Statistics*, 1:353–355, 1973.

L. L. Bohn and D. A. Wolfe. The Effect of Imperfect Judgment Rankings on Properties of Procedures Based on the Ranked-set Samples Analog of the Mann-Whitney-Wilcoxon Statistic. *Journal of the American Statistical Association*, 89:168–176, 1994.

U.S. Census Bureau. Current Population Survey, 2006 to 2009 Annual Social and Economic Supplements. Retrieved on December 1, 2009, from `http://www.census.gov/hhes/www/income/statemedfaminc.html`, 2009.

K. Chaloner. Optimal Bayesian Experimental Design for Linear Models. *The Annals of Statistics*, 12:283–300, 1984.

K. Chaloner and K. Larntz. Optimal Bayesian Design Applied to Logistic Regression Experiments. *Journal of Statistical Planning and Inference*, 21:191–208, 1989.

K. Chaloner and I. Verdinelli. Bayesian Experimental Design: A Review. *Statistical Science*, 10:273–304, 1995.

H. Chen, E. A. Stasny, and D. A. Wolfe. Ranked Set Sampling for Efficient Estimation of a Population Proportion. *Statistics in Medicine*, 24:3319–3329, 2005.

H. Chen, E. A. Stasny, and D. A. Wolfe. An Empirical Assessment of Ranking Accuracy in Ranked Set Sampling. *Computational Statistics & Data Analysis*, 51: 1411–1419, 2006a.

H. Chen, E. A. Stasny, and D. A. Wolfe. Unbalanced Ranked Set Sampling for Estimating a Population Proportion. *Biometrics*, 62:150–158, 2006b.

H. A. David and H. N. Nagaraja. *Order Statistics, Third Edition.* John Wiley & Sons, Inc., Hoboken, New Jersey, 2003.

M. De Iorio, P. Muller, G. L. Rosner, and S. N. MacEachern. An ANOVA Model for Dependent Random Measures. *Journal of the American Statistical Association*, 99: 205–215, 2004.

T. R. Dell and J. L. Clutter. Ranked Set Sampling Theory with Order Statistics Background. *Biometrics*, 28:545–555, 1972.

N. R. Draper and I. Guttman. Some Bayesian Stratified Two-Phase Sampling Results. *Biometrika*, 55:131–139, 1968.

J. Du and S. N. MacEachern. Judgment Post-Stratification for Designed Experiments. *Biometrics*, 64:345–354, 2008.

W. A. Ericson. Optimum Stratified Sampling Using Prior Information. *Journal of the American Statistical Association*, 60:750–771, 1965.

W. A. Ericson. Bayesian Models in Sampling Finite Populations. *Journal of the Royal Statistical Society, Series B*, 31:195–233, 1969.

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

J. Frey and O. Ozturk. Constrained Estimation Using Judgment Post-Stratification. *Annals of the Institute of Statistical Mathematics*. To appear.

J. C. Frey. New Imperfect Rankings Models for Ranked Set Sampling. *Journal of Statistical Planning and Inference*, 137:1433–1445, 2007.

A. E. Gelfand and A. F. M. Smith. Sampling-based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

N. M. Gemayel, E. A. Stasny, and D. A. Wolfe. Optimal Ranked Set Sampling Estimation Based on Medians from Multiple Set Sizes. *Journal of Nonparametric Statistics*. To appear.

S. Guha. Posterior Simulation in the Generalized Linear Mixed Model with Semiparametric Random Effects. *Journal of Computational and Graphical Statistics*, 17:410–425, 2008.

L. K. Halls and T. R. Dell. Trial of Ranked-Set Sampling for Forage Yields. *Forest Science*, 12:22–26, 1966.

M. Hamada, H. F. Martz, C. S. Reese, and A. G. Wilson. Finding Near-Optimal Bayesian Experimental Designs via Genetic Algorithms. *The American Statistician*, 55:175–181, 2001.

R. W. Howard, S. C. Jones, J. K. Mauldin, and R. H. Beal. Abundance, Distribution, and Colony Size Estimates for Reticulitermes spp. (Isoptera: Rhinotermitidae) in Southern Mississippi. *Environmental Entomology*, 11:1290–1293, 1982.

J. Huang. Asymptotic Properties of the NPMLE of a Distribution Function Based on Ranked Set Samples. *The Annals of Statistics*, 25:1036–1049, 1997.

S. Jain and R. M. Neal. Splitting and Merging Components of a Nonconjugate Dirichlet Process Mixture Model. *Bayesian Analysis*, 2:445–472, 2007.

A. Kaur, G. P. Patil, A. K. Sinha, and C. Taillie. Ranked Set Sampling: An Annotated Bibliography. *Environmental and Ecological Statistics*, 2:25–54, 1995.

J. Kiefer and J. Wolfowitz. Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters. *The Annals of Mathematical Statistics*, 27:887–906, 1956.

J. Kohlschmidt. *Ranked Set Sampling: A Look at Allocation Issues and Missing Data Complications*. PhD thesis, The Ohio State University, 2009.

P. H. Kvam and F. J. Samaniego. Nonparametric Maximum Likelihood Estimation Based on Ranked Set Samples. *Journal of the American Statistical Association*, 89: 526–537, 1994.

M. Lavine. The 'Bayesics' of Ranked Set Sampling. *Environmental and Ecological Statistics*, 6:47–57, 1999.

D. V. Lindley. On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics*, 27:986–1005, 1956.

D. V. Lindley. *Bayesian Statistics - A Review*. SIAM, Philadelphia, 1972.

A. Y. Lo. Bayesian Statistical Inference for Sampling a Finite Population. *The Annals of Statistics*, 14:1226–1233, 1986.

A. Y. Lo. A Bayesian Bootstrap for a Finite Population. *The Annals of Statistics*, 16:1684–1695, 1988.

R. H. Lock. 1993 New Car Data. *Journal of Statistics Education*, 1(1), 1993.

S. L. Lohr. Optimal Bayesian Design of Experiments for the One-Way Random Effects Model. *Biometrika*, 82:175–186, 1995.

S. N. MacEachern. Computational Methods for Mixture of Dirichlet Process Models. In D. D. Dey, P. Muller, and D. Sinha, editors, *Practical Nonparametric and Semiparametric Bayesian Statistics*, pages 23–44. Springer, New York, 1998.

S. N. MacEachern. Dependent Nonparametric Processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, 1999.

S. N. MacEachern. Dependent Dirichlet Processes. Technical report, The Ohio State University, 2000.

S. N. MacEachern and P. Muller. Estimating Mixtures of Dirichlet Process Models. *Journal of Computational and Graphical Statistics*, 7:223–238, 1998.

S. N. MacEachern, O. Ozturk, D. A. Wolfe, and G. V. Stark. A New Ranked Set Sample Estimator of Variance. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 64:177–188, 2002.

S. N. MacEachern, E. A. Stasny, and D. A. Wolfe. Judgment Post-Stratification with Imprecise Rankings. *Biometrics*, 60:207–215, 2004.

B. K. Mallick and A. E. Gelfand. Generalized Linear Models with Unknown Link Functions. *Biometrika*, 81:237–245, 1994.

P. McCullagh and J. A. Nelder. *Generalized Linear Models, Second Edition.* Chapman & Hall/CRC, 1999.

G. A. McIntyre. A Method for Unbiased Selective Sampling, Using Ranked Sets. *Australian Journal of Agricultural Research*, 3:385–390, 1952.

G. Meeden and S. Vardeman. A Noninformative Bayesian Approach to Interval Estimation in Finite Population Sampling. *Journal of the American Statistical Association*, 86:972–980, 1991.

S. Mukhopadhyay and A. E. Gelfand. Dirichlet Process Mixed Generalized Linear Models. *Journal of the American Statistical Association*, 92:633–639, 1997.

P. Muller. Simulation-Based Optimal Design. *Bayesian Statistics*, 6:459–474, 1999.

R. M. Neal. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 9:249–265, 2000.

D. Nelson and G. Meeden. Using Prior Information about Population Quantiles in a Finite Population Setting. *Sankhya: The Indian Journal of Statistics, Series A*, 60:426–445, 1998.

O. Ozturk. Nonparametric Maximum-Likelihood Estimation of Within-Set Ranking Errors in Ranked Set Sampling. *Journal of Nonparametric Statistics*. To appear.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods, Second Edition*. Springer, New York, 2005.

D. Rubin. The Bayesian Bootstrap. *The Annals of Statistics*, 9:130–134, 1981.

J. Sethuraman. A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, 4: 639–650, 1994.

H. Solomon and S. Zacks. Optimal Design of Sampling from Finite Populations: A Critical Review and Indication of New Research Areas. *Journal of the American Statistical Association*, 65:653–677, 1970.

S. L. Stokes. Estimation of Variance Using Judgment Ordered Ranked Set Samples. *Biometrics*, 36:35–42, 1980a.

S. L. Stokes. Inferences on the Correlation Coefficient in Bivariate Normal Populations from Ranked Set Samples. *Journal of the American Statistical Association*, 75:989–995, 1980b.

S. L. Stokes and T. W. Sager. Characterization of a Ranked-Set Sample with Application to Estimating Distribution Functions. *Journal of the American Statistical Association*, 83:374–381, 1988.

K. Takahasi and K. Wakimoto. On Unbiased Estimates of the Population Mean Based on the Sample Stratified by Means of Ordering. *Annals of the Institute of Statistical Mathematics*, 20:1–31, 1968.

J. T. Terpstra and E. J. Nelson. Optimal Rank Set Sampling Estimates for a Population Proportion. *Journal of Statistical Planning and Inference*, 127:309–321, 2005.

X. Wang, L. Stokes, J. Lim, and M. Chen. Concomitants of Multivariate Order Statistics with Application to Judgment Poststratification. *Journal of the American Statistical Association*, 101:1693–1704, 2006.

X. Wang, J. Lim, and L. Stokes. A Nonparametric Mean Estimator for Judgment Poststratified Data. *Biometrics*, 64:355–363, 2008.

Y. Wang, Z. Chen, and J. Liu. General Ranked Set Sampling with Cost Considerations. *Biometrics*, 60:556–561, 2004.

M. West, P. Muller, and M. D. Escobar. Hierarchical Priors and Mixture Models, with Applications in Regression and Density Estimation. In P. R. Freeman and A. F. M. Smith, editors, *Aspects of Uncertainty: A Tribute to D. V. Lindley*, pages 363–386. Wiley, New York, 1994.

D. A. Wolfe. Ranked Set Sampling: An Approach to More Efficient Data Collection. *Statistical Science*, 19:636–643, 2004.

S. S. Yang. General Distribution Theory of the Concomitants of Order Statistics. *The Annals of Statistics*, 5:996–1002, 1977.

S. Zacks. Bayes Sequential Designs of Fixed Size Samples from Finite Populations. *Journal of the American Statistical Association*, 64:1342–1349, 1969.

S. Zacks. Bayesian Design of Single and Double Stratified Sampling for Estimating Proportions in Finite Populations. *Technometrics*, 12:119–130, 1970.