Estimating Multinomial Cell Probabilities Using Normalized Beta Kernels

by

Phongsack Manivong

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfilment of the requirements of the degree of

Master of Science

Department of Statistics

University of Manitoba

Winnipeg, Manitoba, Canada

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

*****

COPYRIGHT PERMISSION

Estimating Multinomial Cell Probabilities Using Normalized Beta Kernels

By

Phongsack Manivong

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of Manitoba in partial fulfillment of the requirement of the degree

Of

Master of Science

Abstract

The focus of this thesis is on estimating multinomial cell probabilities in the context of sparse, ordered data, in particular, using the normalized beta kernel estimator (NBKE). The NBKE is a local smoothing estimator that uses non-negative weights and that takes advantage of the natural ordering inherent to certain types of data. It is flexible and self-adapting and leads to estimated probabilities that form a proper probability distribution. Furthermore, it is an asymptotically unbiased and normal estimator that is free of boundary bias.

Specifically, this thesis begins with a general discussion on probability estimation, smoothing, and kernel estimation for discrete or categorical data. Secondly, many of the properties and pitfalls of such estimators are discussed. Then the desirable properties of the NBKE are examined through visual and mathematical proofs, and with a simulation study. The final chapter of this thesis is an illustrative example using the NBKE on data from a medical survey on inflammatory bowel disease (IBD).

Key Words: Kernel estimation; Normalized beta kernel estimator; NBKE; Smoothing; Bias-variance tradeoff; Boundary bias; Sparseness; Non-negative weights, Multinomial cell probabilities.

Acknowledgments

Thank you to Dr. Alex Leblanc for his guidance and patience.

Thank you to Dr. Lisa Lix and Dr. Charles Bernstein for providing the data.

Thank you to NSERC for providing funding.

Contents

List of Figures

Chapter 1

Introduction

The impetus behind this Master's thesis is a medical survey on inflammatory bowel disease (IBD). One covariate of interest is the type of IBD as diagnosed by a subject's physician. There are three main types, Crohn's disease, Ulcerative Colitis and Ulcerative Proctitis with the first two being the most prevalent. One outcome measure is the number of symptom flares (outbursts) subjects experienced within the six months prior to the study. Subjects were allowed to respond with values ranging from 0 to a maximum of 99. Thus, under the reasonable assumption that subjects are independent, symptom flares can be viewed as observations from a multinomial distribution with proportion parameters $P_0, P_1, P_2, \ldots, P_{99}$. And, consequently, the focus of this thesis will be on estimating multinomial cell probabilities, in particular, using the normalized beta kernel estimator (NBKE).

Another important objective of this thesis is to determine if there is a statistically significant difference in the distribution of symptom flares between the group classified with Crohn's disease and the group classified as Ulcerative Colitis. There are a total of 247 testable subjects of interest in the IBD survey, fairly

evenly divided between the two IBD groups. There were very few subjects classified as having Ulcerative Proctitis or some other form of IBD. (See the Data Appendix on the post processing done to the data.)

Under the null hypothesis, subjects in the two IBD groups come from the same homogeneous population. Figure 1.1 plots the observed cell probabilities for the two IBD groups. These preliminary results show that there could possibly be differences, particularly in the upper-tail with counts for the Ulcerative Colitis group being more clustered about 0. Observations for cells greater than 40 could be considered extreme cases. Although this thesis will focus mainly on methods of estimating multinomial cell probabilities, we will look at the result of including and removing these cases in Chapter 6.

Figure 1.1 clearly shows that the assumption of an approximate normal distribution for the number of sympton flares would not be appropriate for either group. Thus, a simple analysis using ANOVA or ANCOVA would not suffice for this problem. Furthermore, these bar charts reveal the sparseness of the data. The problem of sparseness in contingency tables occurs when the number of cells is relatively large in comparison to the number of observations. This leads to many categories or cells having low counts, namely zeroes and ones. A simple χ^2 goodness-of-fit test for homogeneity would also not be adequate for this problem as this test is not adequate for sparse data. (Symptom flares could also be treated as being from a Poisson or zero-inflated Poisson distribution, but that is beyond the scope of this thesis.)

One solution to the problem of sparseness is to restrict observations into fewer categories. By doing this, however, we lose some information and statistical tests become less powerful. Another issue with grouping the data into fewer categories

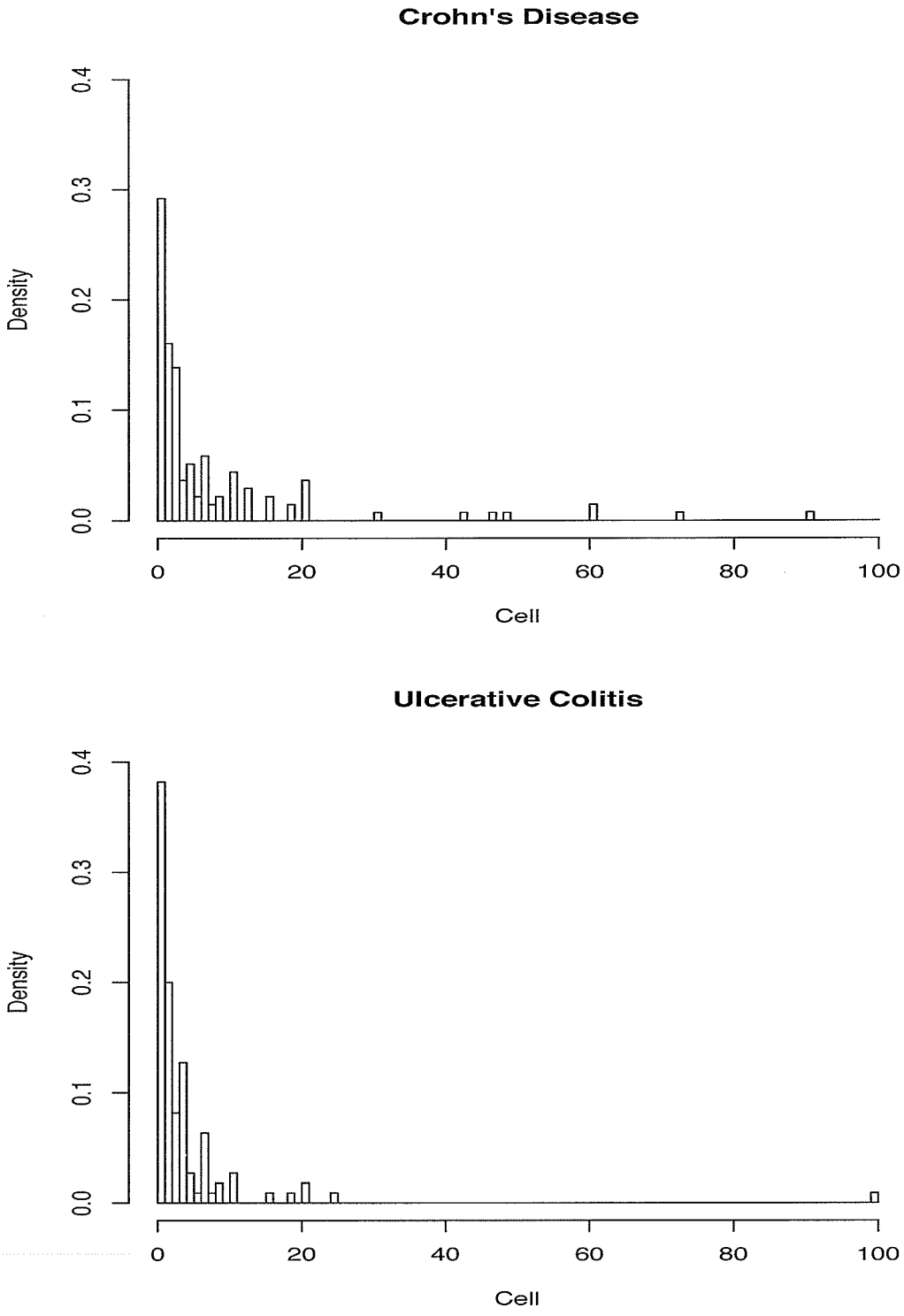

Figure 1.1: Bar charts of proportions of symptom flares for the two IBD groups.

is that the method of grouping is quite arbitrary. Results from the same type of test can vary greatly depending on the grouping configuration. Also, the upper cut-off within a group is often more similar to the lower cut-off of the next group than its own lower cut-off. For example, if the range of one group is 11-20 and the range of the next group is 21-30, the value 20 is more similar to 21 than 11, which is intuitively undesirable.

Another solution, which we pursue here, is to adjust the observed cell probabilities, also known as maximum likelihood estimates (MLE), by using weights. One method of adjustment is to use a so-called kernel estimator. A kernel is the underlying function that allocates weights to cells or observations. A higher level of adjustment or "smoothing" denotes a greater departure from MLE probabilities. When most of the weight is assigned to the cell of interest and its immediate neighbours, there is less smoothing. When weights are more spread out, there is greater smoothing.

The focus of this thesis is on estimating multinomial cell probabilities. Specifically, in Chapter 2, we present a review on nonparametric probability estimators for ordinal discrete or categorical data. In Chapter 3, we introduce the normalized beta kernel estimator (NBKE) and derive some of its basic properties. In Chapter 4, we discuss the properties of the bias and variance, both generally and asymptotically. We also discuss the bias-variance tradeoff inherent to practically all smoothing estimators, and offer some practical considerations when dealing with finite samples. In Chapter 5, we perform a simulation study to compare and contrast the effectiveness of the NBKE against the MLE and the geometric kernel estimator (GKE) which we will later discuss. In the final chapter, we analyze the data that was the motivation behind this thesis.

Chapter 2

Probability Estimation for Discrete Data

When estimating probabilities, whether for continuous or discrete data, there is a need for a measure that specifies what possible values a variable can take, and the likelihood of obtaining such values. Together, these determine the distribution of a variable. Within the realm of discrete or categorical data, the probability mass function (pmf) is such a measure for the distribution.

Moreover, there is also a need for measures of accuracy and precision of an estimator. The former describes, on average, how different an estimate (or that is, its expected value) is from a true parameter, and is defined as the bias. The latter describes how similar estimates are in general, and is defined as the variance. A measure that encapsulates both the bias and the variance is the mean squared error (MSE).

For now, let us concentrate on the distribution of multinomial data. A histogram, or bar chart, is commonly used as a visual representation of the distribution of a variable. For continuous data, a histogram groups data into bins. The width of the bin determines the range of a group. The height determines a group's overall frequency or relative proportion. The bin width in conjunction with the range of the histogram determines its overall shape and smoothness. Wider bins translate to having a larger smoothing effect, and so, flatten the histogram. The position of the first and last bins can also affect the number of peaks and troughs that appear in the density curve. Thus, data from a multimodal distribution can appear to have only one peak; see Silverman [12] for an illustration using data on geyser eruption times.

In the case of multinomial data, the bin width and range of the bar chart are fixed where each bar represents a cell or category. The bar chart, of course, can be condensed into fewer bars (or cells). This is what would be required when conducting a χ^2 test for sparse data, when many cells have low counts, namely zeros and ones. One of the goals of this thesis is to explore a method that eliminates the need to collapse data into fewer cells. This will be emphasized later in Chapter 6, when we look at the IBD data in greater detail.

The parametric approach to probability estimation assumes that the data follows a known form of distribution. For example, data assumed to follow a Poisson distribution would have a mean that is equal to its variance and would have a specific form of pmf depending on only one parameter. A nonparametric approach requires fewer restrictions or assumptions for the purpose of estimation, but typically involves many more parameters to be estimated.

A common nonparametric method to estimate cell probabilities, or proportions, for discrete data is to use unadjusted observed cell proportions. Note that these are the maximum likelihood estimates (MLE) of the true underlying cell probabilities. Let X_i be an observation from a multinomial distribution with probabilities $P_0, P_1, P_2, \ldots, P_m$. Also, let N be the total number of observations and N_k represent the number of observations equal to k. The MLE of P_k, the probability for cell $k = 0, 1, 2, \ldots, m$ is,

$$\begin{aligned} \hat{P}_k &= \frac{1}{N}\sum_{i=1}^{N} \mathrm{I}(X_i = k) \\ &= \frac{N_k}{N}, \end{aligned} \tag{2.1}$$

where $\mathrm{I}(X_i = k)$ is an indicator function. This method, however, is problematic for data with many cells having low or zero frequencies.

Good and Gaskins [6] were the first to introduce the maximum penalized likelihood estimator (MPLE), initially in the continuous setup, then later in the discrete setup. For the MPLE, a larger number of cells results in a greater penalty. See Simonoff [13] for a further expansion on the discrete setup.

Simonoff [14] also details several basic methods that account for zero counts. One such estimator, he describes as a 'flattening' one, adds a constant λ to each cell. The form of the 'flattening' estimator is very much like the MLE (2.1). It is defined as,

$$\tilde{P}_k(\lambda) = \frac{N_k + \lambda}{N + \lambda(m+1)}. \tag{2.2}$$

Note that (2.2) can also be expressed as a Bayesian shrinkage estimator. In this form, the estimator becomes a weighted average of the MLE and the discrete uniform distribution which gives the same probability of $1/(m+1)$ to all cells. Recall that shrinkage estimators pull individual estimates toward the overall mean. The shrinkage estimator is defined as,

$$\tilde{P}_k(\epsilon) = (1-\epsilon)\hat{P}_k + \frac{\epsilon}{m+1}$$

where $\epsilon = \lambda(m+1)/\left[N + \lambda(m+1)\right]$.

When $\lambda = 0$, the shrinkage estimator is equivalent to the MLE. A larger value for λ gives more weight to the uniform distribution, and thus, implies a higher level of smoothing. As λ approaches infinity, ϵ aproaches 1, and the shrinkage estimator converges to the uniform distribution.

Several choices of λ were considered by Simonoff: 1, 1/2, $1/(m+1)$ and $\sqrt{N}/(m+1)$. To illustrate the effects of varying the level of smoothing, however, we use $\lambda = 0, 1, 10,$ and 50. Figures 2.1 and 2.2 plot fitted probability curves for the two IBD groups. In both figures, we see the pronounced reduction of the peak at the boundary. It is also important to note that for $\lambda = 10$ (or in general, large λ), the curves for the two IBD groups are almost indistinguishable. At $\lambda = 50$, the shrinkage estimator is already almost equivalent to the uniform estimator. This is a precursor to the bias-variance tradeoff dilemma that is inherent to all smoothing estimators. This issue will be discussed in fuller detail in later sections.

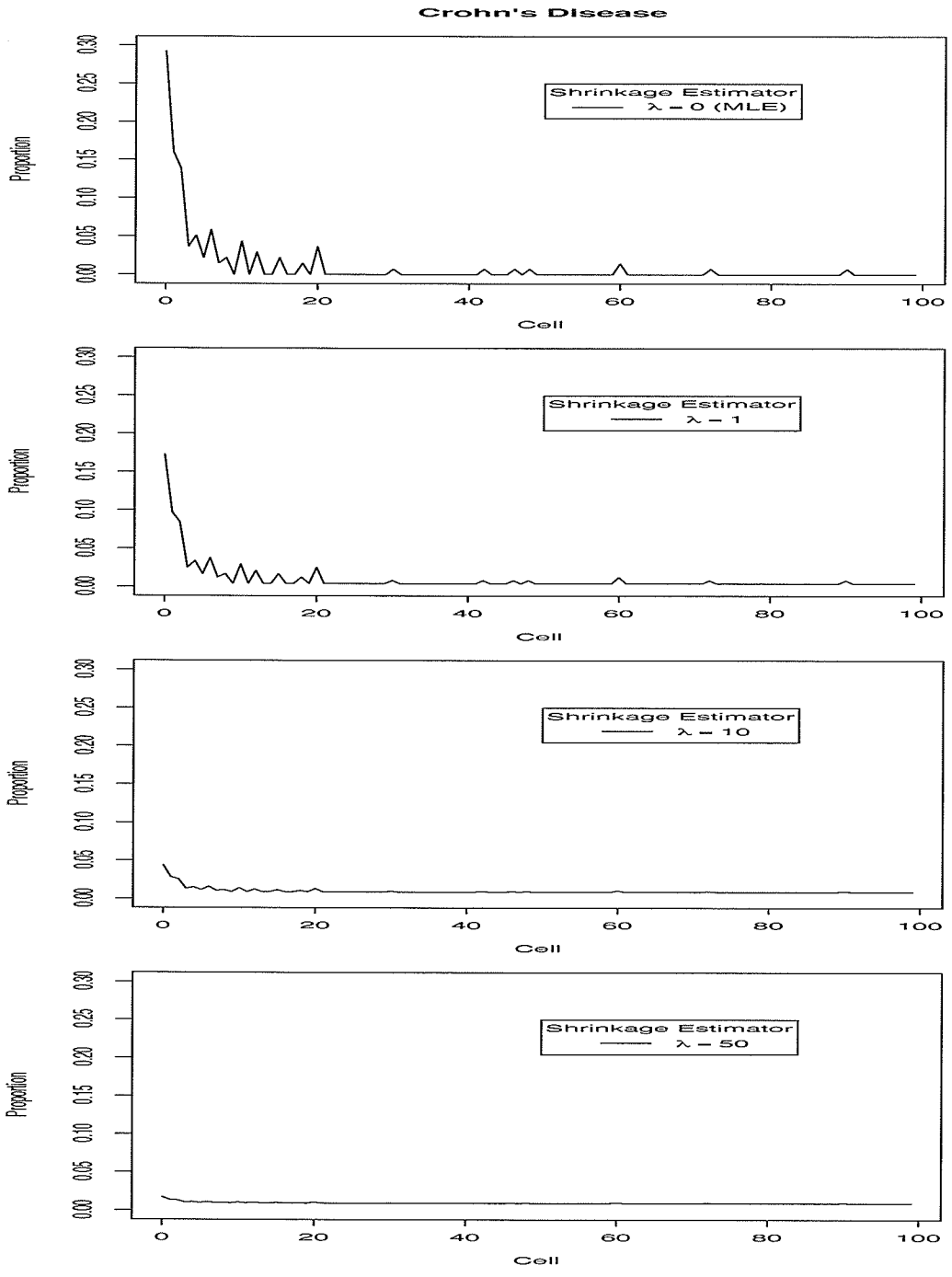

Figure 2.1: Fitted probability curves for Crohn's disease IBD group using the shrinkage estimator. From top to bottom, $\lambda = 0$,1,10, and 50.

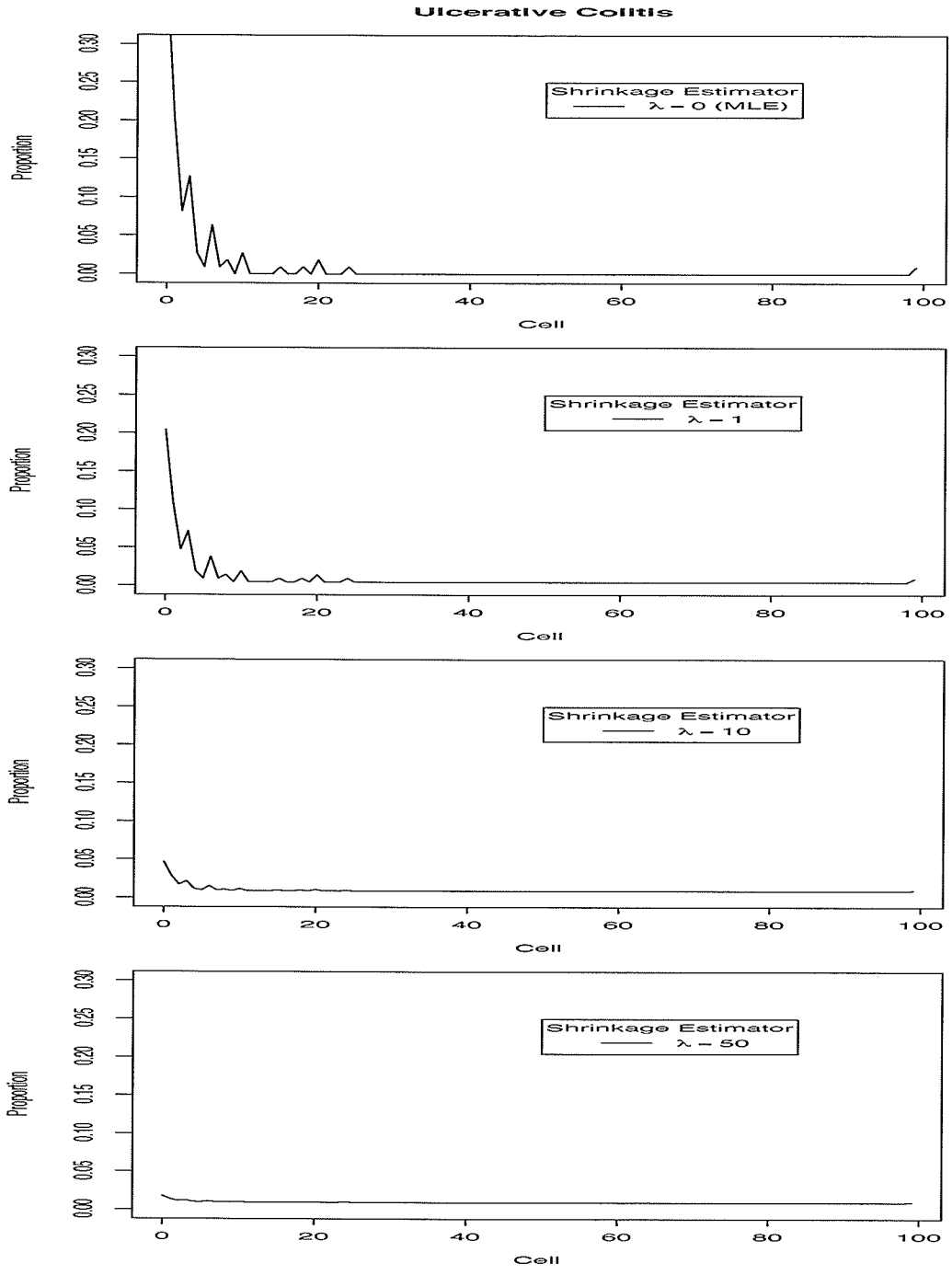

Figure 2.2: Fitted probability curves for Ulcerative Colitis IBD group using the shrinkage estimator. From top to bottom, $\lambda = 0$,1,10, and 50.

Another possible solution to the problem of sparseness is to smooth the data by using a kernel estimator. In essence, a kernel-weighted estimated probability is a linear combination, or weighted average, of all the observations or observed cell probabilities. A general form for a kernel estimator for discrete data can be expressed as,

$$\hat{P}_{W,k} = \sum_{l=0}^{m} W_k(l) \frac{N_l}{N}. \tag{2.3}$$

$\hat{P}_{W,k}$ denotes the probability estimate of cell k, $\frac{N_l}{N}$ is the observed proportion of observations falling into cell l, and $W_k(l)$ is the weight assignted to cell l, when estimating the probability of cell k, for $k = 0, 1, 2, \ldots, m$ and $l = 0, 1, 2, \ldots, m$.

Note that (2.3) reduces to the MLE probability estimates when

$$W_k(l) = \begin{cases} 1 & \text{for } l = k, \\ 0 & \text{otherwise.} \end{cases}$$

In other words, no weight is given to other cells when estimating P_k.

It is also important to note the difference between a simple scalar-weighted probability estimator (rarely used), and a kernel-weighted one. A scalar-weighted estimator can be expressed as,

$$\hat{P}_{S,k} = S_k \frac{N_k}{N},$$

for some sequence of constansts $0 \leq S_k \leq 1$. This is more similar to the shrinkage estimator as smaller weights pull $\hat{P}_{S,k}$ closer to 0 and weights closer to 1 pull $\hat{P}_{S,k}$ closer to the MLE probability. Also, with this method, categories with zero frequencies will continue to have probability estimates of zero. With kernel estimators, however, zero frequencies do not necessarily lead to zero probability estimates because estimated probabilities are based on a weighted average of several cell probabilities, not just one.

The level of smoothing is determined by the allocation of weights to cells or categories. Most smoothing techniques allocate larger weights to neighbouring cells. A higher level of smoothing is achieved as the weights assigned to each cell approach equality. Figure 2.3 shows the NBKE weight functions for cells 10 and 50 at various levels of smoothing. The parameter c, which will be discussed in greater detail in Chapter 3, specifies the level of smoothing with a higher value indicating less smoothing. We can see that the range of cells with non-neglible weight widens and the peak decreases as the level of smoothing increases. A challenge to using any smoothing technique is determining the appropriate level of smoothing for a given situation. For an introductory discussion on kernel estimation and smoothing parameter selection, see Schucany [11].

A local smoothing estimator is one that assigns weights primarily to the cell of interest and its neighbours. This is practical when neighbouring cells are assumed to be similar, which is often the case of ordinal, numeric data. One situation where this assumption is appropriate is if data values are truncated or grouped from continuous values. For example, height measured in centimetres can be rounded up or grouped into discrete categories. A property inherent to local smoothing techniques is the flattening of local extrema. Since full weight is not given to the cell of interest, but distributed among neighbouring cells, the magnitude of a local minimum and maximum is lessened. In other words, the height of a modal area is decreased and the depth of a valley area is reduced. In Figure 2.3, if either cell 10 or 50 were a local extrema, and probability estimates were plotted, we would see the mitigation of the height or depth with higher smoothing. See Aerts et al. [1] for an example and discussion on local polynomial fitting for sparse multinomial data.

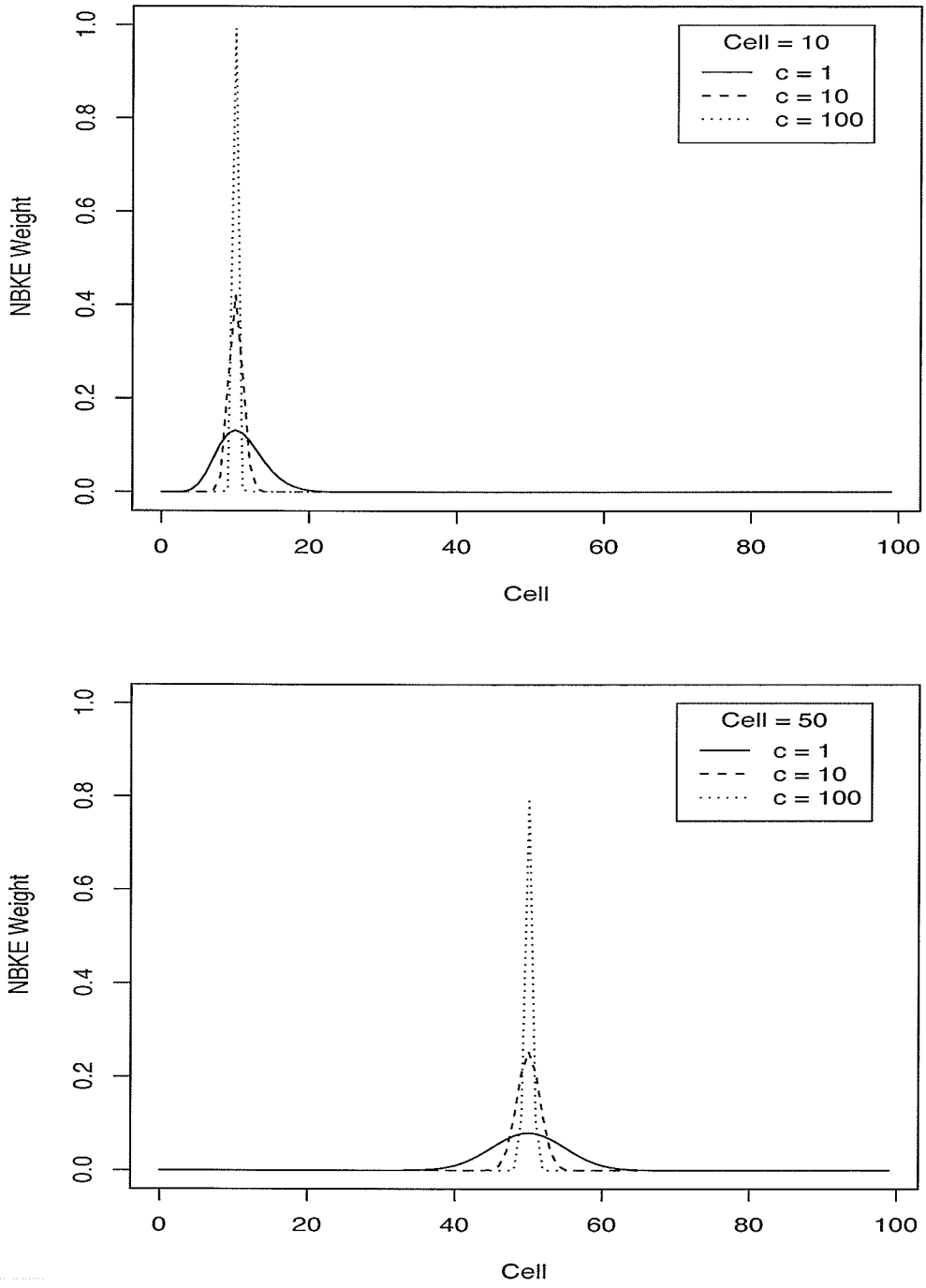

Figure 2.3: NBKE weight functions for cells 10 (top) and 50 (bottom) for a multinomial distribution with 100 cells, at $c = 1, 10,$ and 100.

Burman [3] summarizes two main reasons for smoothing:

1. Smoothed estimators (which include kernel estimators) are often superior to the MLE under squared error loss due to the bias-variance tradeoff.

2. The occurrence of zero cell counts can be problematic when estimating or testing in some parametric and nonparametric models.

For now, it would perhaps be better to address the general issues encountered when using kernel estimators. This will lead to the next chapter where the normalized beta kernel estimator (NBKE) is discussed more in depth, specifically detailing how this estimator does not exhibit some of the problems of other kernel estimators, but exhibits other desirable properties. For a comparative study of kernel-based estimators for categorical data, and a discussion on their implementation for missing or incomplete data, see Titterington [15].

Often, in the case of continuous data, the kernel is symmetric. For details on density estimation for continuous data, see Silverman [12]. In general, whether for continuous or discrete data, most would argue that it is reasonable to allocate equal weight to cells that are equal-distanced from the cell of interest, especially for neighbouring cells. This is a problem, however, when estimating cell probabilities near the boundary, or more specifically, cells 0 and m, as there are not an equal number of cells that are greater than and smaller than the cell of interest. The term 'bandwidth' is often used to define the range of a weighting function. Kernel estimators with a bandwidth that allocates non-negligible weights to cells outside the support of data will exhibit boundary bias. In this case, weights are assigned to cells that do not exist! It is important to note, however, that this issue is not exclusive to symmetric kernels.

An example of such an estimator is the geometric kernel estimator (GKE) as dicussed by Wang and Ryzin [16]. When estimating the probability of cell k, the weighting function of cell l is defined as,

$$W_{s,k}(l) = 0.5^{1-I(k=l)}(1-s)s^{|k-l|} \tag{2.4}$$

where $0 < s < 1$ is the smoothing parameter, with values closer to 0 denoting less smoothing. For each k, the weights sum up to 1 for all integer values of $|k-l|$. This is a problem, however, if the total number of cells or categories (m) is small, as the sum of the weights will not add up to 1. In general, when weights do not sum up to 1 (either across k or l), this can lead to probability estimates that do not sum up to 1 which is not a desirable property! We will compare the GKE with the NBKE in a simulation study later in Chapter 5.

A reflection technique can be used to redistribute the weight from the area beyond the boundary back to the scope of the data. This increases the weight assigned to the boundary cell and its neighbours, but does not necessarily solve the problem of boundary bias completely. Another undesirable result is that the resulting estimates do not form a proper probability distribution. This problem, however, is not limited to kernel estimators using data reflection. Kernel estimators that allow negative weights can also lead to probability estimates that are negative. Thus, these too can result in estimates not forming a proper probability distribution.

Rajagopalan and Lall [10] developed a kernel estimator to account for data exhibiting characteristics of multiple geometric distributions. In their case, the data were highly concentrated about the origin and had a long tail. Their motivation

was estimating the distribution for the length (in days) of continuous rainfall, and dry periods.

Burman [3] introduced a kernel estimator for smoothing sparse contingency tables that requires an important assumption of underlying smoothness in the density. Burman explains this assumption by defining the probability of cell k as the area under a density curve between $k/(m+1)$ and $(k+1)/(m+1)$ with a support ranging from $0/(m+1)$ to $(m+1)/(m+1)$ (or more specifically, 0 to 1). Burman also stipulates that there is no boundary bias when $P_0 = 0$ or $P_m = 0$.

The focus of this thesis is on the normalized beta kernel estimator (NBKE). It too is a kernel estimator that is most advantageous when used on discrete data that is naturally ordered and that may exhibit sparseness, but does not require the same stipulation as Burman for there to be no boundary bias. Chapters 3 and 4 focus on the general and asymptotic properties of the NBKE such as the bias-variance tradeoff and how this affects the choice for the level of smoothing. In Chapter 5, through a similation study we will see the effectiveness of the NBKE over the MLE and GKE. In Chapter 6, using the IBD data as a case study, the NBKE will be applied to sparse, ordered multinomial data with a high concentration about zero, and a long tail.

Chapter 3

The Normalized Beta Kernel Estimator

The normalized beta kernel estimator (NBKE) is a nonparametric estimator for discrete data adapted from a continuous setup, see Chen [4]. It is nonparametric because the estimated pmf (of the data) is not assumed to belong to a specific parametric family. Beta weighting functions are just tools used to improve upon MLE estimates which are simply observed proportions. As previously mentioned, the NBKE does not exhibit problems common to other kernel estimators for discrete data, but more importantly, it exhibits other desirable properties. Recall that its weighting scheme requires an important assumption that there is an underlying smoothness and order to the overall distribution to be estimated. The assumption implies that neighboring cells are similar to each other so allocating the most weight to the cell of interest and its immediate neighbours makes sense.

3.1 The Binomial Kernel

We start with the simplest version of the NBKE using binomial probability weights. It will become clear in the next section how this estimator based on binomial weights is a specific case of the NBKE. Let $X_1, X_2, \ldots, X_N$ be observations from a multinomial distribution with proportion parameters $P_0, P_1, P_2, \ldots, P_m$ and size parameter N. When weights are assigned to individual observations, the binomial kernel estimator of the k-th cell (or category) probability is defined as,

$$\hat{P}_{B,k} = \frac{1}{N}\sum_{i=1}^{N}\binom{m}{k}\left(\frac{x_i}{m}\right)^k\left(1-\frac{x_i}{m}\right)^{m-k} = \frac{1}{N}\sum_{i=1}^{N}B_{k,m}(x_i), \tag{3.1}$$

where the function $B_{k,m}(x_i)$ denotes the binomial weight of the i-th observation, for $i = 1, 2, \ldots, N$ and $k = 0, 1, \ldots, m$. Note that these weights are all non-negative and correspond to the binomial distribution with the probability of success $x_i/m \in [0, 1]$.

Note also that although there are N observations, there are only $m + 1$ distinct weights. So, it is possible to rewrite this estimator in a form similar to equation (3.1), but with weights expressed in terms of the cells. The binomial kernel estimator simplifies to,

$$\begin{aligned}\hat{P}_{B,k} &= \sum_{l=0}^{m}\frac{1}{N}\sum_{i=1}^{N}\mathrm{I}(l = x_i)B_{k,m}(x_i)\\ &= \sum_{l=0}^{m}\frac{N_l}{N}\binom{m}{k}\left(\frac{l}{m}\right)^k\left(1-\frac{l}{m}\right)^{m-k}\\ &= \sum_{l=0}^{m}\frac{N_l}{N}B_{k,m}(l),\end{aligned} \tag{3.2}$$

where $\mathrm{I}(l = x_i)$ is an indicator function and $B_{k,m}(l)$ is the binomial weight of the l-th cell when estimating the probability of the k-th cell.

Proposition 3.1.1. *When estimating the probability of the k-th cell, binomial kernel weights are maximized at* $l = k$.

In other words, when the weights are plotted, the peak is located at cell $l = k$. Hence, estimating P_k using (3.1) makes sense as observations close to k contribute more to this weighted sum of binomial probabilities. Similarly, estimating P_k using (3.2) makes sense as cells close to k are allocated more weight. Also recall the assumption that there is an underlying smoothness to the distribution of multinomial probabilities. This means that neighbouring cells are similar to each other; so it is reasonable to assign larger weights to cells closest to k.

We now present the proof for maximal weight at $l = k$ for the special case of the binomial kernel. We will later formally state this property as a theorem and provide an asymptotic expression of the weight for the general NBKE.

Proof of Proposition 3.1.1:

Recall the binomial weight function in (3.2),

$$B_{k,m}(l) = \binom{m}{k} \left(\frac{l}{m}\right)^k \left(1 - \frac{l}{m}\right)^{m-k}.$$

We will prove the above proposition for three different cases: (1) $k = 0$, (2) $k = m$, and (3) $0 < k < m$.

For the first case, the weight function becomes,

$$B_{0,m}(l) = (1 - \frac{l}{m})^m.$$

As $B_{0,m}(l)$ is a decreasing function of l, it is obvious that $B_{0,m}(l)$ is maximized at $l = 0$, and hence, at $l = k$.

For the second case, the weight function becomes,

$$B_{m,m}(l) = (\frac{l}{m})^m.$$

As $B_{m,m}(l)$ is an increasing function of l, it is obvious that $B_{m,m}(l)$ is also maximized at $l = m$, where $k = m$ in this case.

For the third case, in order to verify that $l = k$ is a global maximum over $0 \leq l \leq m$, we need to consider the boundaries. For this, note that,

$$B_{k,m}(0) = \binom{m}{k} \left(\frac{0}{m}\right)^k \left(1 - \frac{0}{m}\right)^{m-k} = 0 \tag{3.3}$$

and

$$B_{k,m}(m) = \binom{m}{k} \left(\frac{m}{m}\right)^k \left(1 - \frac{m}{m}\right)^{m-k} = 0. \tag{3.4}$$

Now, it easy to to see that $B_{k,m}(l) > 0$ for $0 < l < m$, implying that the maximum will be reached away from the boundaries. So let us consider the natural logarithm of $B_{k,m}(l)$ for $0 < l < m$,

$$\log B_{k,m}(l) = \log \binom{m}{k} + k \log l - k \log m + (m-k) \left[\log (m-l)\right] - (m-k) \log m.$$

The first derivative of $\log B_{k,m}(l)$ with respect to l is,

$$\frac{\partial \log B_{k,m}(l)}{\partial l} = \frac{k}{l} - \frac{(m-k)}{m-l}.$$

Equating the above to zero and solving for l, we get that $l = k$.

Next, consider the second derivative of $\log B_{k,m}(l)$,

$$\frac{\partial^2 \log B_{k,m}(l)}{\partial^2 l} = -\frac{k}{l^2} - \frac{(m-k)}{(m-l)^2} < 0$$

for all $0 < l < m$ which tells us that the weight is maximized at $l = k$. Therefore, for any value of $k = 0, 1, \ldots, m$, $B_{k,m}(l)$ is maximized at $l = k$. □

Note, (3.3) and (3.4) tell us that when we are estimating the probability of an interior cell, 0 weight is given to both boundary cells.

Proposition 3.1.2. *The estimated probabilities $\hat{P}_{B,k}$ for $k = 0, 1, \ldots, m$ are non-negative and sum up to 1.*

Proof of Proposition 3.1.2:

First, note that,

$$\sum_{k=0}^{m} B_{k,m}(l) = \sum_{k=0}^{m} \binom{m}{k} \left(\frac{l}{m}\right)^k \left(1 - \frac{l}{m}\right)^{m-k} = 1. \tag{3.5}$$

Then, using property (3.5), we have,

$$\begin{aligned} \sum_{k=0}^{m} \hat{P}_{B,k} &= \sum_{k=0}^{m} \sum_{l=0}^{m} \frac{N_l}{N} B_{k,m}(l) \\ &= \sum_{l=0}^{m} \frac{N_l}{N} \sum_{k=0}^{m} B_{k,m}(l) \\ &= \sum_{l=0}^{m} \frac{N_l}{N} [1] = \frac{N}{N} = 1. \end{aligned}$$

□

The property of having estimated probabilities sum to one is required for unbiased estimation. To obtain this property, the weights must sum to one across some index (in our case, k). It is important to note, however, that having weights sum to one does not always guarantee that estimated probabilities will sum to one. We will see an example of this in the next section as we explore the importance of the weighting scheme.

3.2 The NBKE

The normalized beta kernel estimator (NBKE) is essentially a generalization of the binomial kernel estimator introduced in the previous section. It is an estimator that allows an infinite range for the level of smoothing. More importantly, it is normalized so that the complete collection of estimated probabilities have the preferable feature of summing to 1.

Let $X_1, X_2, \ldots, X_N$ be observations from a multinomial distribution with proportion parameters $P_0, P_1, P_2, \cdots, P_m$ and size parameter N. For $c \geq 0$ and $0 \leq t \leq m$, let the beta kernel function be defined as,

$$B_{c,j,m}(t) = \frac{\Gamma(cm+1)}{\Gamma(cj+1)\Gamma(cm-cj+1)} \left(\frac{t}{m}\right)^{cj} \left(1-\frac{t}{m}\right)^{c(m-j)} \tag{3.6}$$

where m is the upper boundary cell, j is an index for cells and $\Gamma(z+1) = z!$ for positive integer values of z. Beta kernel functions will serve as tools to compute initial kernel weights, and eventually, normalized kernel weights. Note that beta kernel functions are non-negative, and so, a sum of kernel weights is also non-negative.

The beta kernel function in (3.6) can also be expressed as,

$$B_{c,j,m}(t) = \binom{cm}{cj} \left(\frac{t}{m}\right)^{cj} \left(1-\frac{t}{m}\right)^{c(m-j)}. \tag{3.7}$$

The second form of the beta kernel function in (3.7) is useful for integer values of c. For the NBKE, a larger value of c translates to less smoothing. (This is often not the case with smoothing parameters for other kernel estimators.)

When weights are assigned to individual observations, the general form of the NBKE of the k-th cell probability is defined as,

$$\hat{P}_k(c) = \frac{1}{N}\sum_{i=1}^{N} W_{c,k,m}(x_i), \tag{3.8}$$

where $W_{c,k,m}(x_i)$ denotes the NBKE weight of the i-th observation when estimating the k-th cell (or category) probability, for $i = 1, 2, \ldots, N$ and $k = 0, 1, \ldots, m$. In terms of beta kernel functions, $W_{c,k,m}(x_i)$ can be expressed as,

$$W_{c,k,m}(x_i) = \frac{B_{c,k,m}(x_i)}{\sum_{j=0}^{m} B_{c,j,m}(x_i)}. \tag{3.9}$$

Note that $B_{c,k,m}(x_i)$ in (3.9) can be viewed as the non-normalized beta kernel weight of the i-th observation, or for lack of a better word, an unscaled kernel weight. $\sum_{j=0}^{m} B_{c,j,m}(x_i)$ is the sum of all possible beta kernel weights of the i-th observation. (There are $m+1$ different beta kernel weights for the i-th observation, one for each category.) Thus, $W_{c,k,m}(x_i)$ can be interpreted as the fraction of weight for the i-th observation used for estimating P_k, with respect to the sum, hence the term normalized beta kernel.

Similarly to (3.2), the estimator in (3.8) can also be expressed in terms of the cells,

$$\hat{P}_k(c) = \sum_{l=0}^{m} \frac{N_l}{N} W_{c,k,m}(l) \tag{3.10}$$

where $W_{c,k,m}(l)$ is the NBKE weight of the l-th cell (or category), for $l = 0, 1, \ldots, m$, when estimating the probability of the k-th cell. It is defined as,

$$W_{c,k,m}(l) = \frac{B_{c,k,m}(l)}{\sum_{j=0}^{m} B_{c,j,m}(l)}. \tag{3.11}$$

Similarly, $B_{c,k,m}(l)$ can be viewed as the non-normalized (or unscaled) beta kernel weight of the l-th cell when estimating P_k. Also, $\sum_{j=0}^{m} B_{c,j,m}(l)$ is the sum of all possible beta kernel weights of the l-th cell, and $W_{c,k,m}(l)$ can be interpreted as the fraction of weight for the l-th cell used for estimating P_k. Table 3.1 is an $(m+1)\times(m+1)$ grid of all possible unscaled beta kernel weights, with k, the main cell of interest indexed along the rows, and l, the index of general cells indexed across the columns. The normalization method used in this paper is with respect to the sum of a column.

Table 3.1: Table of non-normalized beta kernel weights $B_{c,k,m}(l)$.

Cell	0	1	$\cdots$	l	$\cdots$	$m-1$	m
0	1	$B_{c,0,m}(1)$	$\cdots$	$B_{c,0,m}(l)$	$\cdots$	$B_{c,0,m}(m-1)$	0
1	0	$B_{c,1,m}(1)$	$\cdots$	$B_{c,1,m}(l)$	$\cdots$	$B_{c,1,m}(m-1)$	0
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
k	0	$B_{c,k,m}(1)$	$\cdots$	$B_{c,k,m}(l)$	$\cdots$	$B_{c,k,m}(m-1)$	0
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$m-1$	0	$B_{c,m-1,m}(1)$	$\cdots$	$B_{c,m-1,m}(l)$	$\cdots$	$B_{c,m-1,m}(m-1)$	0
m	0	$B_{c,m,m}(1)$	$\cdots$	$B_{c,m,m}(l)$	$\cdots$	$B_{c,m,m}(m-1)$	1

Although, perhaps counterintuitive, it is important to note that the weights $W_{c,k,m}(l)$ do not necessarily sum to unity, when added over l, for any fixed k. In other words, the weights used for estimating P_k do not, in general, sum to one. This implies that (3.10) cannot be interpreted as a convex combination of the observed cell frequencies.

In contrast, however,

$$\sum_{k=0}^{m} W_{c,k,m}(l) = 1,$$

for any fixed l because of the method of normalization. In other words, when the beta kernel weights in Table 3.1 are normalized, the resulting weights sum to one within a column. We have already seen this property used to prove that the estimated probabilities obtained from binomial kernel estimator sum up to one. It will be used again when we prove, for the general case, that NBKE estimated probabilities sum up to one.

Some readers may be inclined to think that the normalization is with respect to the wrong sum, that is, that the normalization should be with respect to the sum of a row, when refering to Table 3.1. Specifically, when estimating P_k, the weight assigned to individual observations could be defined as,

$$H_{c,k,m}(x_i) = \frac{B_{c,k,m}(x_i)}{\sum_{j=0}^{m} B_{c,k,m}(j)}.$$

Or, when weights are assigned to cells,

$$H_{c,k,m}(l) = \frac{B_{c,k,m}(l)}{\sum_{j=0}^{m} B_{c,k,m}(j)}. \tag{3.12}$$

In this case, when estimating P_k, the beta kernel weight of cell l is divided by the sum of beta kernel weights of all cells. An advantage of this version of the normalization is that the proof of maximal weight at $l = k$ would be easier to obtain. Since the denominator of $H_{c,k,m}(l)$ does not contain l, we could treat it as a constant, and modify the proof for Proposition 3.1.1 of the binomial case. The disadvantage, however, is that the estimated multinomial probabilities do not, in general, sum up to one. This is obviously not a desirable property.

To see this, let $\tilde{P}_k(c)$ denote the probability estimate of cell k using the second version of the normalization in (3.12). Then,

$$\begin{aligned}\sum_{k=0}^{m} \tilde{P}_k(c) &= \sum_{k=0}^{m}\sum_{l=0}^{m} \frac{N_l}{N} H_{c,k,m}(l) \\ &= \sum_{k=0}^{m} \frac{1}{N}\sum_{l=0}^{m} N_l H_{c,k,m}(l). \end{aligned} \tag{3.13}$$

In order for (3.13) to equal 1, the following must hold true,

$$\sum_{l=0}^{m} N_l H_{c,k,m}(l) = N_k.$$

This is not always true, however, unlike the normalization method in (3.11).

Now, it is obvious that the beta kernel functions defined as in (3.6) and (3.7) are non-negative. Hence, the NBKE weights as defined in (3.9) and (3.11) are non-negative. A useful property of the NBKE, then, is that probability estimates are non-negative, but more importantly form, a proper probability distribution. In other words, the probability estimates sum up to one. A formal theorem and proof are given in the next section.

Moreover, the NBKE weight function naturally varies for each cell, but its scope is fixed between the minimum and maximum cells (the domain of the data). This means that the shape of the kernel changes automatically without having to change any smoothing parameter, and that no weight is assigned to non-exisiting cells. This is in contrast with most kernel estimators, particularly the ones with symmetric weight functions as mentioned in Chapter 2.

Figure 3.1 shows the NBKE weight functions of some interior cells for $c = 1$, $c = 10$, and $c = 100$, where cells range from 0 to 99, as in the IBD study. Values of 1 and 10 for c were chosen to represent high and moderate levels of smoothing. A value of 100 for c was chosen to represent an estimate close to the MLE, or in other words, minimal smoothing. It is clear that for cells closer to the boundary, weight functions are more asymmetric and their peaks are higher. Asymmetry, however, is less obvious at lower levels of smoothing (i.e. for larger values of c).

It is also important to note that the range of cells with non-neglible weight is narrower for cells closer to the boundary. This means that for the same value of c, interior cells receive greater smoothing and boundary cells receive lesser smoothing. This is a property we will witness through a simulation study in Chapter 5.

Because the NBKE weight function is self-adapting, its peak is always located at the cell of interest. In other words, when estimating P_k, the largest weight is assigned to cell k. (We prove this property for $c = 1$ and for asymptotic c, but we do not prove it for $c < 1$.) This might explain the absence of boundary bias. This will be formally expressed in Theorem (3.3.1). We can see in Figure 3.1 that for interior cells, the largest weight (peak) is located at the cell of interest regardless of the degree of smoothing. This is also true for boundary cells as can be seen in Figure 3.2. Furthermore, no weight is assigned to cells beyond the boundary. Thus, because of the NBKE's self-adapting and local smoothing properties and because kernel weights become more concentrated as the smoothing parameter c increases, the NBKE is asymptotically unbiased with respect to c. This is also formally expressed as a theorem in the following chapter.

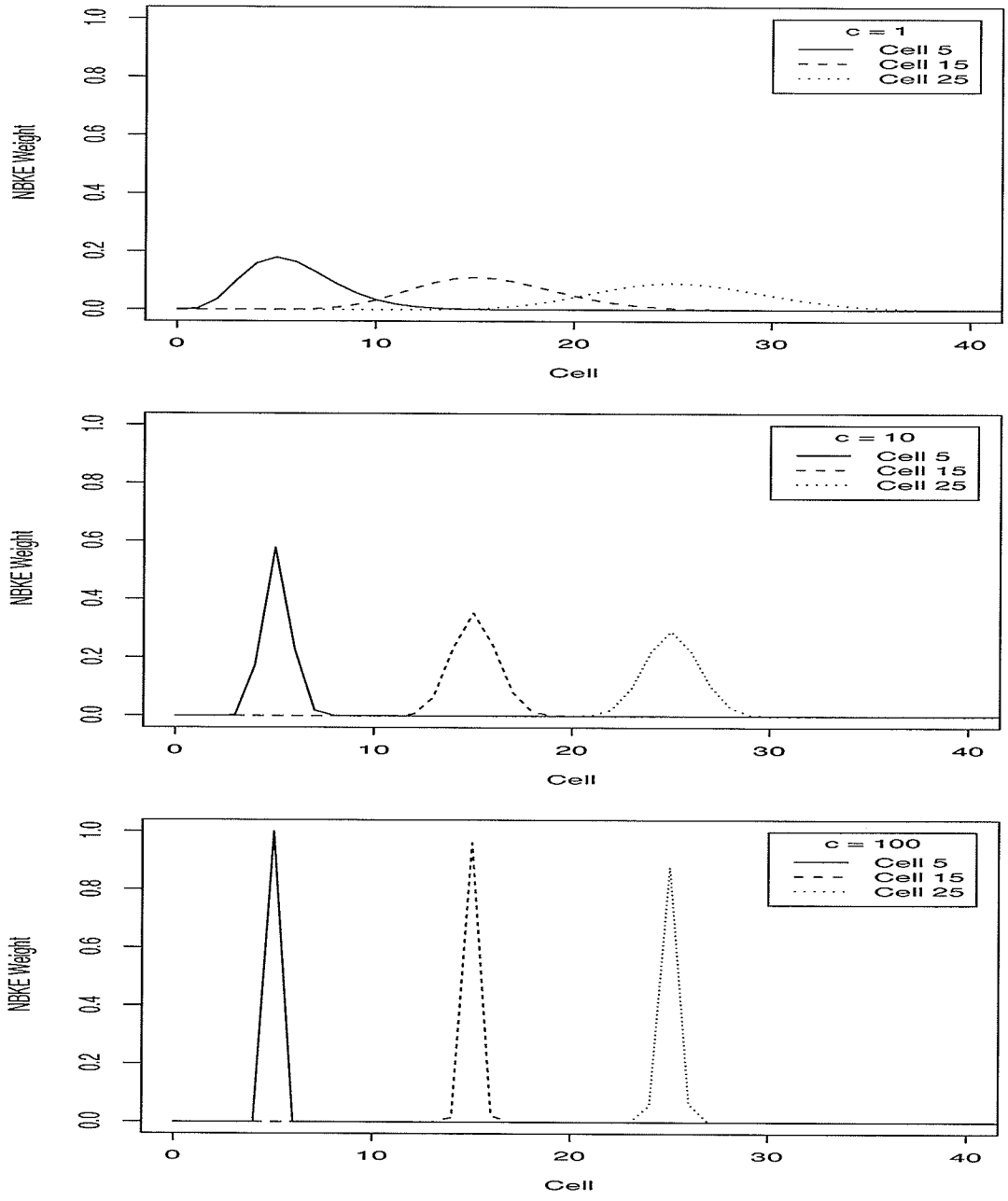

Figure 3.1: Self-adapting NBKE weight functions, $c = 1, 10$ and 100.

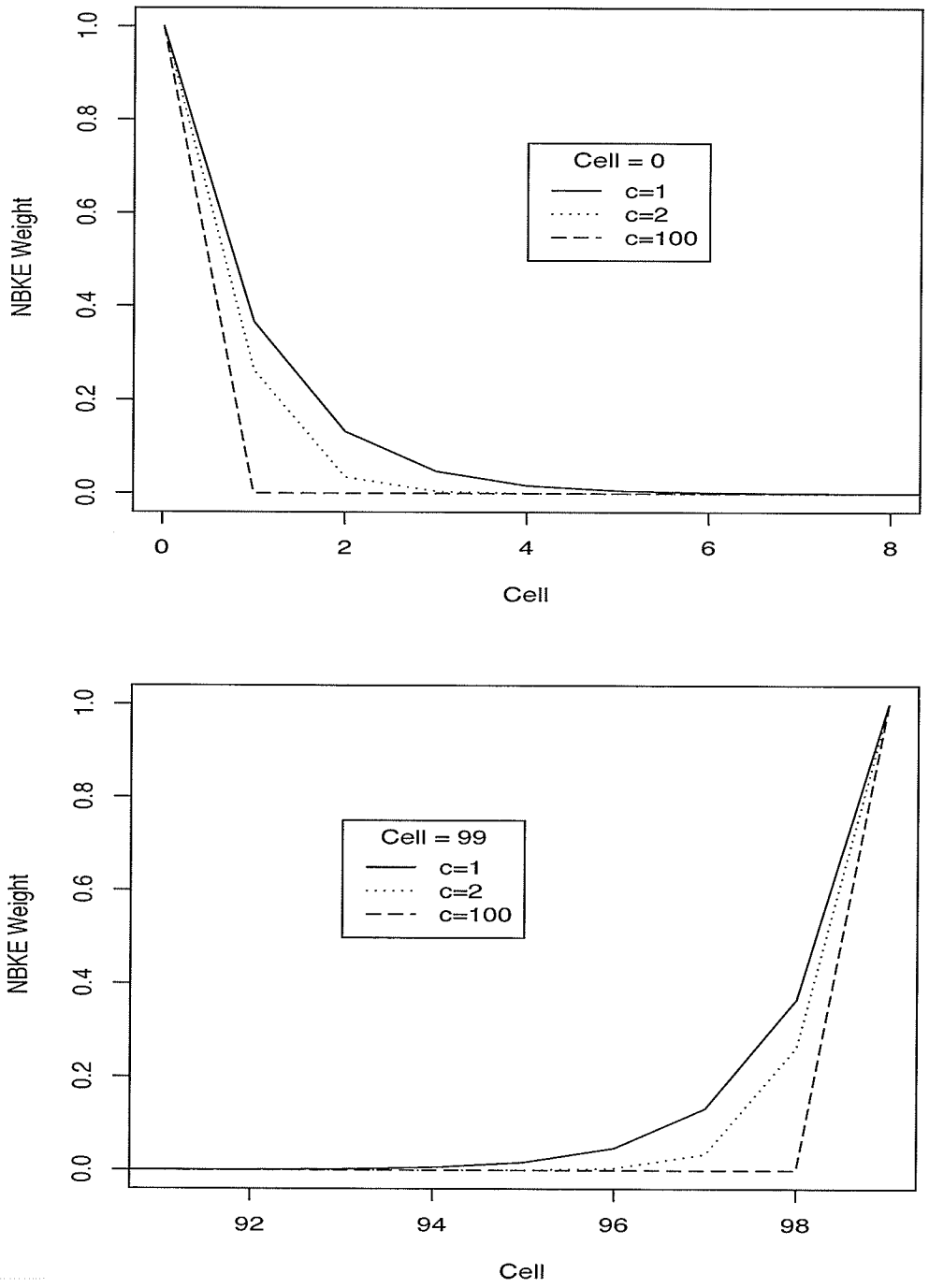

Figure 3.2: Various levels of smoothing for lower (top) and upper (bottom) boundary cells, $k = 0$ and $k = 99$, respectively.

Also, note that when $c = 0$, the NBKE takes the form of a uniform estimator. This is the greatest level of smoothing possible. When estimating any P_k, all cells are given equal weight, namely $1/(m+1)$. The resulting estimate will then be $\hat{P}_k(0) = 1/(m+1)$ for all cells.

An inherent property of practically all smoothing techniques is the existence of a tradeoff between bias and variance. Although the NBKE will be proven to be asymptotically unbiased, it is biased in general, in particular at the local extrema. Recall that we use local smoothing under the assumption that neighbouring cells are similar, and that there is an overall smoothness to the distribution. In modal areas, the NBKE tends to underestimate and in valley regions, the NBKE tends to overestimate. Dong and Simonoff [5] postulate that for many discrete kernel estimators, probability estimates for cells about the boundary can exhibit the most volatilty.

In our simulation study in Chapter 5, we will see that for the NBKE, the degree of bias differs for modal regions. Thus, choosing the degree of smoothing is very important. This is a reflection of the bias-variance tradeoff first mentioned in Chapter 2. The decision whether to accept larger systematic error (bias) versus random error (variance) has to be made based on the needs of the situation. This tradeoff, the Mean Squared Error (MSE) and the Mean Sum of Squared Error (MSSE) of the normalized beta kernel estimator are also discussed in Chapter 4.

3.3 Basic Properties of the NBKE

In this section, we formally present some useful theoretical results previously mentioned in Section 1 of this chapter.

As can be seen in Figures 3.1, and 3.2, the peak is located at cell k regardless of the level of smoothing. This implies that the k-th cell has the largest contribution when estimating P_k, which is, of course, a desirable property. We now formally state this in the form of a theorem.

Theorem 3.3.1. *When estimating P_k, the probability for cell k, the largest weight is assigned to cell k.*

Theorem 3.3.2. *When estimating P_k,*

$$\lim_{c\to\infty} W_{c,k,m}(l) = \begin{cases} 1 & \text{if } k = l, \\ 0 & \text{if } k \neq l. \end{cases}$$

This property is useful for confirming two main properties. The first is that the NBKE estimated probabilities are asymptotically equivalent (with respect to the smoothing parameter c) to the MLE, or, in other words, the observed proportions. The second is Theorem 3.3.1. When estimating P_k, as c increases (or with less smoothing) the weight allocated to cell k approaches 1 and all other weights approach 0. We know this condition already holds for the binomial kernel, or, when $c = 1$. We, however, leave the case for $0 < c < 1$ open for future studies.

Proof of Theorem 3.3.2:

First, consider,

$$\begin{aligned} W_{c,k,m}(l) &= \frac{B_{c,k,m}(l)}{\sum_{j=0}^{m} B_{c,j,m}(l)} \\ &= \frac{B_{c,k,m}(l)/B_{c,l,m}(l)}{\sum_{j=0}^{m} B_{c,j,m}(l)/B_{c,l,m}(l)} \\ &= \frac{B_{c,k,m}(l)/B_{c,l,m}(l)}{1+\sum_{j\neq l} B_{c,j,m}(l)/B_{c,l,m}(l)}. \end{aligned} \tag{3.14}$$

We now need to show that,

$$\lim_{c\to\infty} \frac{B_{c,j,m}(l)}{B_{c,l,m}(l)} = \begin{cases} 1 & \text{if } j = l \\ 0 & \text{if } j \neq l. \end{cases}$$

It is trivial to show that the above limit holds when $j = l$. The proof, in the case of $j \neq l$, will be derived separately for (1) $j > l$ and (2) $j < l$. For these proofs, Lemma B.1 is required. (See the Technical Appendix.)

For these two cases, first consider,

$$\begin{aligned} \frac{B_{c,j,m}(l)}{B_{c,l,m}(l)} &= \frac{\frac{\Gamma(cm+1)}{\Gamma(cj+1)\Gamma(cm-cj+1)}\left(\frac{l}{m}\right)^{cj}\left(1-\frac{l}{m}\right)^{c(m-j)}}{\frac{\Gamma(cm+1)}{\Gamma(cl+1)\Gamma(cm-cl+1)}\left(\frac{l}{m}\right)^{cl}\left(1-\frac{l}{m}\right)^{c(m-l)}} \\ &= \frac{\Gamma(cl+1)\Gamma(cm-cl+1)}{\Gamma(cj+1)\Gamma(cm-cj+1)}\left(\frac{l}{m}\right)^{c(j-l)}\left(\frac{m-l}{m}\right)^{-c(j-l)} \\ &= \frac{\Gamma(cl+1)\Gamma(cm-cl+1)}{\Gamma(cj+1)\Gamma(cm-cj+1)}(cl)^{c(j-l)}\left[\frac{1}{c(m-l)}\right]^{c(j-l)}. \end{aligned} \tag{3.15}$$

Now, for the case where $l < j \leq m$, expanding (3.15), we get the following

$$\begin{aligned}\frac{B_{c,j,m}(l)}{B_{c,l,m}(l)} &= \frac{\Gamma(cl+1)\Gamma(cm-cj)(cm-cj+1)\cdots[cm-cj+cj-cl]}{\Gamma(cl+1)(cl+1)\cdots[cl+(cj-cl)]\Gamma(cm-cj)} \\ &\quad \times (cl)^{c(j-l)}\left[\frac{1}{c(m-l)}\right]^{c(j-l)} \\ &= \left[\frac{(cl)^{c(j-l)}}{(cl+1)\cdots(cj)}\right]\left[\frac{(cm-cj+1)\cdots(cm-cl)}{[c(m-l)]^{c(j-l)}}\right]. \end{aligned} \tag{3.16}$$

The first term on the right-hand side of the previous equality can be bounded as,

$$\left[\frac{(cl)^{c(j-l)}}{(cl+1)\cdots(cj)}\right] \leq \left[\frac{cl}{cl+1}\right]^{c(j-l)} = \left[1-\frac{1}{cl+1}\right]^{c(j-l)} \leq \left[1-\frac{1}{cl}\right]^{c(j-l)},$$

so that,

$$\lim_{c\to\infty}\left[\frac{(cl)^{c(j-l)}}{(cl+1)\cdots(cj)}\right] \leq \lim_{c\to\infty}\left[1-\frac{1}{cl}\right]^{c(j-l)} = e^{-(j-l)/l}. \tag{3.17}$$

Using Lemma B.1, the second term of (3.16) can be bounded as

$$\begin{aligned} &\left[\frac{(cm-cj+1)\cdots(cm-cj+cj-cl)}{[c(m-l)]^{c(j-l)}}\right] \\ \leq\ & \frac{\left[cm-cj+\frac{(cj-cl)+1}{2}\right]^{c(j-l)}}{[c(m-l)]^{c(j-l)}} \\ \leq\ & \left\{\frac{cm-cj+\frac{(cj-cl)}{2}}{[c(m-l)]}\left[1+\frac{1}{2(cm-cj+\frac{cj-cl}{2})}\right]\right\}^{c(j-l)} \\ =\ & \left[\frac{2m-(j+l)}{2m-2l}\right]^{c(j-l)}\left[1+\frac{1}{c(2m-j-l)}\right]^{c(j-l)}. \end{aligned}$$

From this, we have,

$$\begin{aligned} &\lim_{c\to\infty}\left[\frac{(cm-cj+1)\cdots(cm-cj+cj-cl)}{[c(m-l)]^{c(j-l)}}\right] \\ \leq\ & \lim_{c\to\infty}\left[\frac{2m-(j+l)}{2m-2l}\right]^{c(j-l)} \times \lim_{c\to\infty}\left[1+\frac{1}{c(2m-j-l)}\right]^{c(j-l)}. \end{aligned} \tag{3.18}$$

Now, note that in (3.18),

$$\lim_{c\to\infty}\left[1+\frac{1}{c(2m-j-l)}\right]^{c(j-l)} = e^{(j-l)/(2m-j-l)},$$

and $j > l$ implies that,

$$\lim_{c\to\infty}\left[\frac{2m-(j+l)}{2m-2l}\right]^{c(j-l)} = 0.$$

Therefore from (3.16), (3.17) and (3.18), for $j > l$, we get,

$$\lim_{c\to\infty}\frac{B_{c,j,m}(l)}{B_{c,l,m}(l)} \leq \lim_{c\to\infty}\left[\frac{2m-(j+l)}{2m-2l}\right]^{c(j-l)} \times \left[e^{(j-l)/(2m-j-l)}\right] \times e^{-(j-l)/l} = 0.$$

For the case where $0 \leq j < l$, expanding (3.15), we get the following,

$$\begin{aligned}\frac{B_{c,j,m}(l)}{B_{c,l,m}(l)} &= \frac{\Gamma(cj+1)(cj+1)\cdots[cj+(cl-cj)]\Gamma(cm-cl+1)}{\Gamma(cj+1)\Gamma(cm-cl+1)(cm-cl+1)\cdots[(cm-cl+cl-cj)]} \\ &\quad \times (cl)^{c(j-l)}\left[\frac{1}{c(m-l)}\right]^{c(j-l)} \\ &= \left[\frac{(cm-cl)^{c(l-j)}}{(cm-cl+1)\cdots(cm-cj)}\right]\left[\frac{(cj+1)\cdots(cj+cl-cj)}{(cl)^{c(l-j)}}\right]. \quad (3.19)\end{aligned}$$

The first term on the right-hand side of the above equality can be bounded as

$$\begin{aligned}\left[\frac{(cm-cl)^{c(l-j)}}{(cm-cl+1)\cdots(cm-cj)}\right] &\leq \left[\frac{cm-cl}{cm-cl+1}\right]^{c(l-j)} \\ &\leq \left[1-\frac{1}{cm-cl+1}\right]^{c(l-j)} \\ &\leq \left[1-\frac{1}{cm-cl}\right]^{c(l-j)},\end{aligned}$$

so that,

$$\lim_{c\to\infty}\left[\frac{(cm-cl)^{c(l-j)}}{(cm-cl+1)\cdots(cm-cj)}\right] \leq \lim_{c\to\infty}\left[1-\frac{1}{cm-cl}\right]^{c(l-j)} = e^{-(l-j)/(m-l)}. \quad (3.20)$$

Using Lemma B.1 again, the second term can be bounded as,

$$\begin{aligned}\left[\frac{(cj+1)\cdots(cj+cl-cj)}{(cl)^{c(l-j)}}\right] &\leq \frac{\left[cj+\frac{(cl-cj)+1}{2}\right]^{c(l-j)}}{[cl]^{c(l-j)}} \\ &\leq \left\{\frac{cj+\frac{(cl-cj)}{2}}{(cl)}\left[1+\frac{1}{cj+\frac{(cl-cj)}{2}}\right]\right\}^{c(l-j)} \\ &\leq \left[\frac{j+l}{2l}\right]^{c(l-j)}\left[1+\frac{2}{c(j+l)}\right]^{c(l-j)}.\end{aligned}$$

From this, we have,

$$\begin{aligned}&\lim_{c\to\infty}\left[\frac{(cj+1)\cdots(cj+cl-cj)}{(cl)^{c(l-j)}}\right] \\ \leq\ &\lim_{c\to\infty}\left[\frac{j+l}{2l}\right]^{c(l-j)} \times \lim_{c\to\infty}\left[1+\frac{2}{c(j+l)}\right]^{c(l-j)}. \qquad (3.21)\end{aligned}$$

However,

$$\lim_{c\to\infty}\left[1+\frac{2}{c(j+l)}\right]^{c(l-j)} = e^{(l-j)/(j+l)},$$

and, $j<l$ implies that,

$$\lim_{c\to\infty}\left[\frac{j+l}{2l}\right]^{c(l-j)} = 0.$$

Therefore, from (3.19), (3.20) and (3.21), for $j<l$, we get,

$$\lim_{c\to\infty}\frac{B_{c,j,m}(l)}{B_{c,l,m}(l)} \leq \lim_{c\to\infty}\left[\frac{j+l}{2l}\right]^{c(l-j)} \times \left[e^{(l-j)/(j+l)}\right] \times e^{-(l-j)/(m-l)} = 0.$$

Since m is fixed, the above two cases lead to the following result,

$$\lim_{c\to\infty}\sum_{j\neq l}\frac{B_{c,j,m}(l)}{B_{c,l,m}(l)} = 0.$$

From this, we can conclude that,

$$\lim_{c\to\infty} W_{c,k,m}(l) = \begin{cases} 1 & \text{if } k=l \\ 0 & \text{if } k\neq l. \end{cases}$$

□

We now provide a general proof for a property that was mentioned earlier in Section 3.1.

Theorem 3.3.3. *The estimated normalized beta kernel probabilties using (3.10) for $k = 0, 1, \ldots, m$ sum up to 1.*

Proof of Theorem 3.3.3:

First note that,

$$\begin{aligned}\sum_{k=0}^{m} W_{c,k,m}(l) &= \sum_{k=0}^{m} \frac{B_{c,k,m}(l)}{\sum_{j=0}^{m} B_{c,j,m}(l)} \\ &= \frac{\sum_{k=0}^{m} B_{c,k,m}(l)}{\sum_{j=0}^{m} B_{c,j,m}(l)} = 1.\end{aligned}$$

From (3.8), we have,

$$\begin{aligned}\sum_{k=0}^{m} \hat{P}_k(c) &= \sum_{k=0}^{m}\sum_{l=0}^{m} \frac{N_l}{N} W_{c,k,m}(l) \\ &= \sum_{l=0}^{m} \frac{N_l}{N} \sum_{k=0}^{m} W_{c,k,m}(l) \\ &= \sum_{l=0}^{m} \frac{N_l}{N} [1] \\ &= \frac{N}{N} = 1.\end{aligned}$$

□

We have now introduced the NBKE and looked at some of its basic properties. In the next chapter, we focus on the general and asymptotic properties of the bias and variance. We also discuss the tradeoff between the two which is a property inherent to most, if not all, smoothing estimators. We will also offer practical considerations for selecting the level of smoothing for finite samples that account for the bias-variance tradeoff.

Chapter 4

Bias, Variance and Practical Considerations

In this chapter, we look at the expected value and variance of $\hat{P}_k(c)$ when m is fixed. Another focus of the chapter is to look at the asymptotic properties of $\hat{P}_k(c)$ with respect to the smoothing parameter c. By doing so, we confirm that the properties first alluded to in Chapters 2 and 3 do hold.

Recall that $X_1, X_2, \ldots, X_N$ are observations from a multinomial distribution with proportion parameters $P_0, P_1, P_2, \cdots, P_m$ and size parameter N. Furthermore, recall Theorem 3.3.2 states that,

$$\lim_{c\to\infty} W_{c,k,m}(l) = \mathrm{I}(l = k) = \begin{cases} 1 & \text{if } k = l, \\ 0 & \text{if } k \neq l. \end{cases}$$

4.1 Expected Value and Bias of $\hat{P}_k(c)$

Let us start by finding a general expression for the expected value of the estimator $\hat{P}_k(c)$, the NBKE of P_k. We have,

$$\begin{aligned} \mathrm{E}[\hat{P}_k(c)] &= \mathrm{E}\left[\sum_{l=0}^{m} W_{c,k,m}(l)\frac{N_l}{N}\right] \\ &= \sum_{l=0}^{m} \frac{W_{c,k,m}(l)}{N}\mathrm{E}[N_l] \\ &= \sum_{l=0}^{m} \frac{W_{c,k,m}(l)}{N}[NP_l] \\ &= \sum_{l=0}^{m} W_{c,k,m}(l)P_l, \end{aligned} \tag{4.1}$$

for any cell k.

More formally, the bias of $\hat{P}_k(c)$ can be expressed as:

$$\begin{aligned} \mathrm{Bias}[\hat{P}_k(c)] &= P_k - \mathrm{E}[\hat{P}_k(c)] \\ &= P_k - \sum_{l=0}^{m} W_{c,k,m}(l)P_l. \end{aligned} \tag{4.2}$$

Corollary 4.1.1. *For fixed m, the NBKE is asymptotically unbiased with respect to the smoothing parameter c, that is, for all $k = 0, \ldots, m$,*

$$\lim_{c\to\infty} \mathrm{E}[\hat{P}_k(c)] = P_k.$$

Proof of Corollary 4.1.1:

As a consequence of Thereom 3.3.2 and (4.1),

$$\begin{aligned}\lim_{c\to\infty} \mathrm{E}[\hat{P}_k(c)] &= \sum_{l=0}^{m} \lim_{c\to\infty} W_{c,k,m}(l) P_l \\ &= \sum_{l=0}^{m} \mathrm{I}(l=k) P_l \\ &= P_k\end{aligned}$$

for any value of $k = 0, \ldots, m$. □

Note that this guarantees asymptotic unbiasedness as c approaches infinity, so that the reduction of the bias here is not necessarily directly dependent on the sample size, but it is dependent on the smoothing parameter c. Of course, c could be chosen as a function of the sample size. In other words, if $c = c_N$ such that $c_N \to \infty$ as $N \to \infty$, then,

$$\lim_{N\to\infty} \mathrm{E}[\hat{P}_k(c)] = P_k.$$

Here, a large c implies less smoothing. This is desired because as sparseness decreases (or as the sample size increases), less smoothing should be required. This is in contrast to other kernel estimators where typically, the smoothing parameter $h = h_N \to 0$ as $N \to \infty$. In our case, a large value of the smoothing parameter implies less smoothing.

Also, note that the previous result does not guarantee that there is no boundary effect which we will see in a later section.

4.2 Variance of $\hat{P}_k(c)$

To find the variance of $\hat{P}_k(c)$, we begin with a result that will shortly prove useful.

Proposition 4.2.1. *$\hat{P}_k(c)-\mathrm{E}[\hat{P}_k(c)]$ can be expressed as an average of independent identically distributed (i.i.d.) observations. Specifically,*

$$\hat{P}_k(c) - \mathrm{E}[\hat{P}_k(c)] = \frac{1}{N}\sum_{i=1}^{N} Y_{i,k}(c) = \bar{Y}_k(c),$$

where,

$$Y_{i,k}(c) = h_c(X_i) = \sum_{l=0}^{m} W_{c,k,m}(l)\left[\mathrm{I}(X_i = l) - P_l\right].$$

This result implies asymptotic normality when c and m are fixed, since the variance of $Y_{i,k}(c)$ is finite. This issue will be revisited in Section 4.4.

Proof of Proposition 4.2.1:

First, we have

$$\begin{aligned}
\hat{P}_k(c) - \mathrm{E}[\hat{P}_k(c)] &= \sum_{l=0}^{m} W_{c,k,m}(l)\left[\frac{N_l}{N}\right] - \sum_{l=0}^{m} W_{c,k,m}(l)P_l \\
&= \sum_{l=0}^{m} W_{c,k,m}(l)\left[\frac{N_l}{N} - P_l\right] \\
&= \sum_{l=0}^{m} W_{c,k,m}(l)\left[\sum_{i=1}^{N}\frac{\mathrm{I}(X_i = l)}{N} - P_l\right] \\
&= \frac{1}{N}\sum_{l=0}^{m} W_{c,k,m}(l)\sum_{i=1}^{N}\left[\mathrm{I}(X_i = l) - P_l\right].
\end{aligned}$$

Now, the order of summation is changed so that,

$$\begin{aligned}
\frac{1}{N}\sum_{l=0}^{m}\sum_{i=1}^{N} W_{c,k,m}(l)\left[\mathrm{I}(X_i = l) - P_l\right] &= \frac{1}{N}\sum_{i=1}^{N}\left\{\sum_{l=0}^{m} W_{c,k,m}(l)\left[\mathrm{I}(X_i = l) - P_l\right]\right\} \\
&= \frac{1}{N}\sum_{i=i}^{N} Y_{i,k}(c) = \bar{Y}_k(c).
\end{aligned}$$

□

Now, we determine the expectations of $Y_{i,k}(c)$ and $Y_{i,k}^2(c)$. These will also prove to be useful in the proof that follows.

Proposition 4.2.2. *For any value of i, k and c,*

$$\mathrm{E}[Y_{i,k}(c)] = 0.$$

Proof of Proposition 4.2.2:

This is quite straightforward. Since $\mathrm{I}(X_i = l)$ is Bernoulli (P_l), we have that

$$\begin{aligned}
\mathrm{E}[Y_{i,k}(c)] &= \mathrm{E}\left\{\sum_{l=0}^{m} W_{c,k,m}(l)[\mathrm{I}(X_i = l) - P_l]\right\} \\
&= \sum_{l=0}^{m} W_{c,k,m}(l)\left\{\mathrm{E}[\mathrm{I}(X_i = l)] - P_l\right\} \\
&= \sum_{l=0}^{m} W_{c,k,m}(l)[P_l - P_l] = 0.
\end{aligned}$$

□

Naturally, this was expected given the form of the equation in Proposition 4.2.1 and the definition of $Y_{i,k}(c)$.

Proposition 4.2.3.

$$\mathrm{E}[Y_{i,k}^2(c)] = \sum_{l=0}^{m} W_{c,k,m}(l)^2[P_l(1 - P_l)] - \sum_{l=0}^{m}\sum_{j\neq l} W_{c,k,m}(l)W_{c,k,m}(j)P_lP_j.$$

Proof of Proposition 4.2.3:

$$\begin{aligned}
\mathrm{E}[Y_{i,k}^2(c)] &= \mathrm{E}\left\{\left[\sum_{l=0}^{m} W_{c,k,m}(l)[\mathrm{I}(X_i = l) - P_l]\right]^2\right\} \\
&= \mathrm{E}\left\{\sum_{l=0}^{m}\left\{W_{c,k,m}(l)[\mathrm{I}(X_i = l) - P_l]\right\}^2\right\} + \\
&\quad \mathrm{E}\left\{\sum_{l=0}^{m}\sum_{j\neq l} W_{c,k,m}(l)[\mathrm{I}(X_i = l) - P_l]\cdot W_{c,k,m}(j)[\mathrm{I}(X_i = j) - P_j]\right\}.
\end{aligned}$$

Now, let the terms from the previous expression be denoted as A and B, respectively. A reduces to the following expression,

$$\begin{aligned} A &= \sum_{l=0}^{m} W_{c,k,m}(l)^2 \mathrm{E}[\mathrm{I}(X_i = l)^2 - 2P_l \mathrm{I}(X_i = l) + {P_l}^2] \\ &= \sum_{l=0}^{m} W_{c,k,m}(l)^2 \mathrm{E}[\mathrm{I}(X_i = l) - 2P_l \mathrm{I}(X_i = l) + {P_l}^2] \\ &= \sum_{l=0}^{m} W_{c,k,m}(l)^2 [P_l - 2P_l(1 \cdot P_l) + {P_l}^2] \\ &= \sum_{l=0}^{m} W_{c,k,m}(l)^2 [P_l(1 - P_l)], \end{aligned}$$

while, B reduces to the following expression,

$$\begin{aligned} B &= \sum_{l=0}^{m} \sum_{j \neq l} W_{c,k,m}(l) W_{c,k,m}(j) \times \\ &\quad \mathrm{E}[\mathrm{I}(X_i = l) \cdot \mathrm{I}(X_i = j) - \mathrm{I}(X_i = l) P_j - \mathrm{I}(X_i = j) P_l + P_l P_j] \\ &= \sum_{l=0}^{m} \sum_{j \neq l} W_{c,k,m}(l) W_{c,k,m}(j) [0 - P_l P_j - P_j P_l + P_l P_j] \\ &= -\sum_{l=0}^{m} \sum_{j \neq l} W_{c,k,m}(l) W_{c,k,m}(j) [P_l P_j]. \end{aligned}$$

Thus, we have,

$$\mathrm{E}[Y_{i,k}^2(c)] = \sum_{l=0}^{m} W_{c,k,m}(l)^2 [P_l(1 - P_l)] - \sum_{l=0}^{m} \sum_{j \neq l} W_{c,k,m}(l) W_{c,k,m}(j) P_l P_j.$$

□

Using the previous two results, we can now obtain the following theorem.

Theorem 4.2.4. *The variance of the NBKE probability estimator is*

$$\mathrm{Var}[\hat{P}_k(c)] = \frac{1}{N} \left\{ \sum_{l=0}^{m} W_{c,k,m}(l)^2 [P_l(1 - P_l)] \right\} - \frac{1}{N} \left\{ \sum_{l=0}^{m} \sum_{j \neq l} W_{c,k,m}(l) W_{c,k,m}(j) P_l P_j \right\}$$

Proof of Theorem 4.2.4:

Using Proposition 4.2.1, we have

$$\begin{aligned} \mathrm{Var}[\hat{P}_k(c)] &= \mathrm{Var}\left\{\hat{P}_k(c) - \mathrm{E}[\hat{P}_k(c)]\right\} \\ &= \mathrm{Var}[\bar{Y}_k(c)] = \frac{\mathrm{Var}[Y_{i,k}(c)]}{N}, \end{aligned}$$

where the $Y_{i,k}(c)$ are i.i.d. Using Proposition 4.2.2, we can further simplify this expression to

$$\begin{aligned} \mathrm{Var}[\hat{P}_k(c)] &= \frac{1}{N}\left\{\mathrm{E}[Y_{i,k}^2(c)] - \mathrm{E}[Y_{i,k}(c)]^2\right\} \\ &= \frac{1}{N}\mathrm{E}[Y_{i,k}^2(c)]. \end{aligned}$$

Finally, from Proposition 4.2.3, we have

$$\begin{aligned} \mathrm{Var}[\hat{P}_k(c)] &= \frac{1}{N}\left\{\sum_{l=0}^{m} W_{c,k,m}(l)^2[P_l(1-P_l)]\right\} \\ &\quad -\frac{1}{N}\left\{\sum_{l=0}^{m}\sum_{j\neq l} W_{c,k,m}(l)W_{c,k,m}(j)P_lP_j\right\} \end{aligned} \tag{4.3}$$

□

Corollary 4.2.5.

$$\lim_{c\to\infty} \mathrm{Var}[\hat{P}_k(c)] = \frac{P_k(1-P_k)}{N}.$$

The proof of this is straightforward in that we must proceed as with Corollary 4.1.1 and apply Theorem 3.3.2 on equation (4.3), the general expression of $\mathrm{Var}[\hat{P}_k(c)]$. This result implies that estimates for cells with probabilities close to 0.5 will have larger volatility. Conversely, cells with extremely low or high probabilities will have smaller volatility.

4.3 Order of the Bias and Variance

In Section 4.1 and 4.2, we looked at the general forms of the bias and variance and their asymptotic properties. We have shown that when c or $c = c_N$ approaches infinity, the NBKE probabilities become equivalent to the MLE probabilities. We do not know, however, about the gains and compromises with respect to the bias-variance tradeoff of using the NBKE over the MLE for small samples. Furthermore, we do not know if the order of the bias and the order of the departure from the MLE-variance is the same for all cells for a particular value of c. This is useful to know as practical situtations never involve the asymptote. Thus, we look at defining more precise expressions for the bias and variance. We start by deriving more precise expressions for the weights $W_{c,k,m}(l)$.

Proposition 4.3.1. *The weight of cell l, when estimating P_k can be expressed as,*

$$W_{c,k,m}(l) = \begin{cases} 1 - Q_c & l = k \\ Q_c^{|l-k|} & l \neq k \end{cases}$$

where $Q_c = O[1 - \frac{1}{2m}]^c$ is a positive quantity less than 1 that approaches 0 as $c \to \infty$. (Note that the maximum weight of any cell is 1.)

Theorem 4.3.2. *For fixed m, the absolute order for the bias of $\hat{P}_k(c)$ can be expressed as,*

$$\text{Bias}(\hat{P}_k) = P_k - \sum_{l=0}^{m} W_{c,k,m}(l) P_l = O\left[1 - \frac{1}{2m}\right]^c.$$

The proof of this theorem is a direct application of Proposition 4.3.1 to (4.2). Note how the bias is uniform with respect to k. Hence, there is no boundary effect. Also note that for larger c, the bias is smaller.

Theorem 4.3.3. *For fixed m, the variance of $\hat{P}_k(c)$ can be expressed as,*

$$\text{Var}[\hat{P}_k(c)] = \frac{1}{N}P_k(1-P_k) - \frac{Q_c}{N}.$$

where, again, $Q_c = O[1-\frac{1}{2m}]^c$ is a positive quantity less than 1 that approaches 0 as $c \to \infty$.

Note how the order of the extra term in the above variance expression is also uniform with respect to k. And, as previously mentioned in Proposition 4.2.5, the variance is larger for probabilities closer to 0.5 and smaller for values closer to 0 or 1.

Also, note that when c approaches infinity, or in other words, with less smoothing, the quantity Q_c decreases so the variance of our estimator increases. Smaller bias, however, requires a larger c value, hence the bias-variance tradeoff.

Proof of Proposition 4.3.1:

Recall Theorem 3.3.2 states that when estimating P_k,

$$\lim_{c\to\infty} W_{c,k,m}(l) = \begin{cases} 1 & \text{if } k = l, \\ 0 & \text{if } k \neq l. \end{cases}$$

We will use some of the steps used in that proof for this one. So, also note that for given k, and m, expanding (3.14) further gives the following result:

$$W_{c,k,m}(l) = \frac{B_{c,k,m}(l)}{B_{c,l,m}(l)} \left[\frac{1}{1 + \sum_{j>l} \frac{B_{c,j,m}(l)}{B_{c,l,m}(l)} + \sum_{j<l} \frac{B_{c,j,m}(l)}{B_{c,l,m}(l)}} \right]. \tag{4.4}$$

Once again, we will consider $\frac{B_{c,j,m}(l)}{B_{c,l,m}(l)}$ separately for the two cases: $j > l$ and $j < l$. Now, recall the expression (3.15),

$$\frac{B_{c,j,m}(l)}{B_{c,l,m}(l)} = \frac{\Gamma(cl+1)\Gamma(cm-cl+1)}{\Gamma(cj+1)\Gamma(cm-cj+1)} (cl)^{c(j-l)} \left[\frac{1}{c(m-l)}\right]^{c(j-l)},$$

which is valid for any value of j and l.

For $j > l$ ($j = l+1, \ldots, m$), (3.15) expands to (3.16), or more specifically,

$$\frac{B_{c,j,m}(l)}{B_{c,l,m}(l)} = \left[\frac{(cl)^{c(j-l)}}{(cl+1)\cdots(cj)}\right]\left[\frac{(cm-cj+1)\cdots(cm-cl)}{[c(m-l)]^{c(j-l)}}\right].$$

From (3.17) and (3.18), we know that,

$$\frac{B_{c,j,m}(l)}{B_{c,l,m}(l)} \leq \left[1-\frac{1}{cl}\right]^{c(j-l)}\left[1+\frac{1}{c(2m-j-l)}\right]^{c(j-l)}\left[\frac{2m-(j+l)}{2m-2l}\right]^{c(j-l)}.$$

Since $c > 0$ and $j > l$, the first term on the right-hand side of the inequality can be bounded as,

$$\left[1-\frac{1}{cl}\right]^{c(j-l)} < 1.$$

The second term is a strictly increasing function of c converging to $e^{(j-l)/(2m-j-l)}$, and hence can be bounded as,

$$\left[1+\frac{1}{c(2m-j-l)}\right]^{c(j-l)} \leq e^{(j-l)/(2m-j-l)} < e.$$

To handle the third part, note that,

$$\left[\frac{2m-(j+l)}{2m-2l}\right] = \left[1-\frac{j-l}{2m-2l}\right] \leq \left[1-\frac{1}{2m-2l}\right] \leq \left[1-\frac{1}{2m}\right].$$

Therefore, we have that

$$\frac{B_{c,j,m}(l)}{B_{c,l,m}(l)} < e\left[1-\frac{1}{2m}\right]^{c(j-l)} = O\left[1-\frac{1}{2m}\right]^{c(j-l)} = Q_c^{j-l}. \tag{4.5}$$

Note that when $l = m$ (for fixed k and m), only the case of $j < l$ needs to be considered.

Similarly, for $j < l$ ($j = 0, \ldots, l-1$), (3.15) expands to (3.19),

$$\frac{B_{c,j,m}(l)}{B_{c,l,m}(l)} = \left[\frac{(cm-cl)^{c(l-j)}}{(cm-cl+1)\cdots(cm-cj)}\right]\left[\frac{(cj+1)\cdots(cj+cl-cj)}{(cl)^{c(l-j)}}\right].$$

From (3.20) and (3.21), we know that,

$$\frac{B_{c,j,m}(l)}{B_{c,l,m}(l)} \leq \left[1-\frac{1}{c(m-l)}\right]^{c(l-j)}\left[1+\frac{2}{c(j+l)}\right]^{c(l-j)}\left[\frac{j+l}{2l}\right]^{c(l-j)}.$$

The first term on the right-hand side of the inequality can be bounded as,

$$\left[1-\frac{1}{c(m-l)}\right]^{c(l-j)} < 1.$$

The second term is a strictly increasing function of c converging to $e^{2(l-j)/(j+l)}$, and hence, can be bounded as,

$$\left[1+\frac{2}{c(j+l)}\right]^{c(l-j)} \leq e^{2(l-j)/(j+l)} < e^2.$$

For the third part, note that

$$\left[\frac{j+l}{2l}\right] = \left[1-\frac{l-j}{2l}\right] \leq \left[1-\frac{1}{2l}\right] \leq \left[1-\frac{1}{2m}\right].$$

Therefore, we see that

$$\frac{B_{c,j,m}(l)}{B_{c,l,m}(l)} < e^2\left[1-\frac{1}{2m}\right]^{c(l-j)} = O\left[1-\frac{1}{2m}\right]^{c(l-j)} = Q_c^{l-j} \tag{4.6}$$

Note that when $l = 0$ (for fixed k and m), only the case of $j > l$ needs to be considered.

Let $\alpha = \left[1 - \frac{1}{2m}\right]$. The expressions in (4.5) and (4.6) then lead to the following results,

$$\sum_{j>l} \frac{B_{c,j,m}(l)}{B_{c,l,m}(l)} < e \sum_{j>l} \alpha^{c(j-l)},$$

and,

$$\sum_{j<l} \frac{B_{c,j,m}(l)}{B_{c,l,m}(l)} < e^2 \sum_{j<l} \alpha^{c(l-j)}.$$

Now consider $\sum_{j>l} \alpha^{c(j-l)}$ where $\alpha < 1$. Let $s = j - l$.

$$\begin{aligned}
\sum_{j>l} \alpha^{c(j-l)} &= \sum_{s=1}^{m-l} \alpha^{cs} \\
&= \alpha^c + \alpha^{2c} + \cdots + \alpha^{c(m-l)} \\
&\leq \alpha^c + (m-l-1)\alpha^{2c} \\
&= \alpha^c + O(\alpha^{2c})
\end{aligned}$$

Similarly, consider $\sum_{j<l} \alpha^{c(l-j)}$ where $\alpha < 1$. Let $t = l - j$.

$$\begin{aligned}
\sum_{j<l} \alpha^{c(l-j)} &= \sum_{t=1}^{l} \alpha^{ct} \\
&= \alpha^c + \alpha^{2c} + \cdots + \alpha^{cl} \\
&\leq \alpha^c + (l-1)\alpha^{2c} \\
&= \alpha^c + O(\alpha^{2c}).
\end{aligned}$$

From the above results, since $\sum_{j \neq l} \frac{B_{c,j,m}(l)}{B_{c,l,m}(l)}$ is a positive quantity, it can then be bounded as,

$$\sum_{j \neq l} \frac{B_{c,j,m}(l)}{B_{c,l,m}(l)} \leq (e + e^2)\alpha^c + O(\alpha^c) = Q_c.$$

Using Lemma B.2 (see Appendix B for details) and the previous result,

$$\begin{aligned}\frac{1}{1+\sum_{j\neq l}\frac{B_{c,j,m}(l)}{B_{c,l,m}(l)}} &= 1-\sum_{j\neq l}\frac{B_{c,j,m}(l)}{B_{c,l,m}(l)}+O\left[\sum_{j\neq l}\frac{B_{c,j,m}(l)}{B_{c,l,m}(l)}\right]^2 \\ &= 1-Q_c.\end{aligned}$$

Note that the left-hand side of the above equation is a fraction less than or equal to 1.

Using (4.5) and (4.6) and the above result, the weights $W_{c,k,m}(l)$ can now be expressed as,

$$\begin{aligned}W_{c,k,m}(l) &= \frac{B_{c,k,m}(l)}{B_{c,l,m}(l)}\left\{\frac{1}{1+\sum_{j\neq l}\frac{B_{c,j,m}(l)}{B_{c,l,m}(l)}}\right\} \\ &= \frac{B_{c,k,m}(l)}{B_{c,l,m}(l)}\{1-Q_c\} \\ &= \begin{cases}1-Q_c & l=k \\ Q_c^{(l-k)}(1-Q_c) & l>k \\ Q_c^{(k-l)}(1-Q_c) & l<k.\end{cases}\end{aligned}$$

Thus,

$$W_{c,k,m}(l) = \begin{cases}1-Q_c & l=k \\ Q_c^{|l-k|} & l\neq k.\end{cases}$$

where $Q_c = O[1-\frac{1}{2m}]^c$ is a positive quantity less than 1 that approaches 0 as $c\to\infty$. □

Proof of Theorem 4.3.3:

Expanding on equation (4.3), we get the following:

$$\mathrm{Var}[\hat{P}_k(c)] = \frac{1}{N}\sum_{l=0}^{m} W_{c,k,m}(l)^2[P_l(1-P_l)] - \frac{1}{N}\sum_{l=0}^{m} W_{c,k,m}(l)P_l \sum_{j\neq l} W_{c,k,m}(j)P_j.$$

Applying Proposition 4.3.1, the first term on the right-hand side of the above expression simplifies to the following,

$$\begin{aligned}
& \frac{1}{N}W_{c,k,m}(k)^2[P_k(1-P_k)] + \frac{1}{N}\sum_{l\neq k} W_{c,k,m}(l)^2[P_l(1-P_l)] \\
= \; & \frac{1}{N}\left\{1-Q_c\right\}^2 P_k(1-P_k) + \frac{1}{N}\left\{\sum_{l\neq k} Q_c^{2|l-k|}P_l(1-P_l)\right\} \\
= \; & \frac{1}{N}\left\{1-2Q_c+Q_c^2\right\} P_k(1-P_k) + \frac{1}{N}\left\{\sum_{l\neq k} Q_c^{2|l-k|}P_l(1-P_l)\right\} \\
= \; & \frac{1}{N}P_k(1-P_k) - \frac{Q_c}{N} + \frac{Q_c^2}{N} + \frac{Q_c^2}{N}.
\end{aligned}$$

The second term simplifies to the following,

$$\begin{aligned}
& -\frac{1}{N}W_{c,k,m}(k)P_k\sum_{j\neq k} W_{c,k,m}(j)P_j - \frac{1}{N}\sum_{l\neq k} W_{c,k,m}(l)P_l\sum_{j\neq k} W_{c,k,m}(j)P_j \\
= \; & -\frac{1}{N}\left\{1+Q_c\right\} P_k\sum_{j\neq k} Q_c^{|j-k|}P_j - \frac{1}{N}\left\{\sum_{l\neq k} Q_c^{|l-k|}P_l\sum_{j\neq l} Q_c^{|j-k|}P_j\right\} \\
= \; & -\frac{Q_c}{N} + \frac{Q_c^2}{N} - \frac{Q_c^2}{N}.
\end{aligned}$$

Since the lowest order of Q is c for both terms, $\mathrm{Var}[\hat{P}_k(c)]$ can be expressed as,

$$\mathrm{Var}[\hat{P}_k(c)] = \frac{1}{N}P_k(1-P_k) - \frac{Q_c}{N}.$$

where $0 < Q_c < 1$. □

4.4 Asymptotic Normality

Consider the following three asymptotic cases for the NBKE, when m is fixed:

1. the smoothing parameter c approaches infinity while the sample size is fixed,
2. the smoothing parameter c is fixed while the sample size approaches infinity,
3. both the smoothing paramter c and the sample size approach infinity.

The Central Limit Theorem, of course, can be applied to only the second and third cases to show that the NBKE is asymptotically normal.

Theorem 4.4.1. *The NBKE is asymptotically normal. Specifically,*

1. *If $c \in \Re$ is fixed, then,*

$$\hat{P}_k(c) - E[\hat{P}_k(c)] \dot{\sim} Normal\left\{0, \sigma_k^2(c)\right\},$$

where $\sigma_k^2(c)$ is the expression of the variance in Theorem 4.2.4.

2. *If $c = c_N \to \infty$ and $\sqrt{N}(1 - \frac{2}{m})^c \to 0$ as $N \to \infty$, then,*

$$\hat{P}_k(c) \dot{\sim} Normal\left\{P_k, \frac{P_k(1-P_k)}{N}\right\}.$$

Recall that the NBKE weighting scheme requires an important assumption that there is an underlying smoothness and order to the overall distribution. The assumption of underlying smoothing is not used mathematically in any of these asymptotic results as we will prove shortly for fixed m. It is more of a conceptual construct that is most appropriate for data that is sparse, particularly when m, the number of cells approaches infinity. In this case, it will even be more necessary to borrow information from neighbouring cells. We, however, will leave this issue open for future studies.

Proof of Theorem 4.4.1:

For the first case, we need only apply Proposition 4.2.1 which states that $\hat{P}_k(c) - \mathrm{E}[\hat{P}_k(c)]$ can be expressed as an average of iid observations. For the second case, the proof will parallel that of Babu et al [2] and Leblanc [7]. It requires the Central Limit Theorem for double arrays. For this, we start by showing that if $c = c_N \to \infty$ as $N \to \infty$, then,

$$\hat{P}_k(c) - \mathrm{E}[\hat{P}_k(c)] \dot{\sim} Normal\left\{0, \frac{P_k(1-P_k)}{N}\right\}.$$

To prove this, recall that Proposition 4.2.1 states that,

$$\hat{P}_k(c) - \mathrm{E}[\hat{P}_k(c)] = \frac{1}{N}\sum_{i=1}^{N} Y_{i,k}(c) = \bar{Y}_k(c).$$

Let $s_c^2 = \mathrm{E}[Y_{i,k}^2(c)]$. Then the wanted result holds if and only if the following Lindeberg condition is satisfied for every $\varepsilon > 0$ as $N \to \infty$,

$$\frac{\mathrm{E}\left\{Y_{i,k}^2(c)\mathrm{I}[|Y_{i,k}(c)| > \varepsilon s_c\sqrt{N}]\right\}}{s_c^2} \to 0. \tag{4.7}$$

Using Theorem 3.3.2 on Proposition 4.2.3, we have,

$$\lim_{c\to\infty} s_c^2 = P_k(1-P_k).$$

We can also bound $|Y_{i,k}(c)|$ as,

$$\begin{aligned}
|Y_{i,k}(c)| &= |\sum_{l=0}^{m} W_{c,k,m}(l)\left[\mathrm{I}(X_i = l) - P_l\right]| \\
&\leq |\sum_{l=0}^{m} W_{c,k,m}(l)\mathrm{I}(X_i = l)| + |\sum_{l=0}^{m} W_{c,k,m}(l)P_l| \\
&\leq \max[W_{c,k,m}(l)] + \sum_{l=0}^{m} W_{c,k,m}(l)P_l.
\end{aligned}$$

Simplifying the previous expression further, we get,

$$|Y_{i,k}(c)| \leq 1 + \sum_{l=0}^{m} W_{c,k,m}(l) P_l \leq 1 + \sum_{l=0}^{m} P_l \leq 2$$

for any c. Thus, (4.7) holds when c and $N \to \infty$, implying in turn that

$$\hat{P}_k(c) - \mathrm{E}[\hat{P}_k(c)] \stackrel{\cdot}{\sim} Normal\left\{0, \frac{P_k(1-P_k)}{N}\right\},$$

or,

$$\frac{\sqrt{N}\left\{\hat{P}_k(c) - \mathrm{E}[\hat{P}_k(c)]\right\}}{\sqrt{P_k(1-P_k)}} \stackrel{\cdot}{\sim} Normal(0,1).$$

First, note that

$$\begin{aligned} \hat{P}_k(c) - \mathrm{E}[\hat{P}_k(c)] &= \hat{P}_k(c) - P_k + P_k - \mathrm{E}[\hat{P}_k(c)] \\ &= \hat{P}_k(c) - P_k + \mathrm{Bias}[\hat{P}_k(c)], \end{aligned}$$

or,

$$\hat{P}_k(c) - P_k = \hat{P}_k(c) - \mathrm{E}[\hat{P}_k(c)] - \mathrm{Bias}[\hat{P}_k(c)].$$

Now, note that

$$\frac{\sqrt{N}[\hat{P}_k(c) - P_k]}{\sqrt{P_k(1-P_k)}} = \frac{\sqrt{N}\left\{\hat{P}_k(c) - \mathrm{E}[\hat{P}_k(c)]\right\}}{\sqrt{P_k(1-P_k)}} - \frac{\sqrt{N}\mathrm{Bias}[\hat{P}_k(c)]}{\sqrt{P_k(1-P_k)}}.$$

Hence, if $\sqrt{N}\mathrm{Bias}[\hat{P}_k(c)] \to 0$, or equivalently, $\sqrt{N}(1-\frac{2}{m})^c \to 0$, then,

$$\hat{P}_k(c) \stackrel{\cdot}{\sim} Normal\left\{P_k, \frac{P_k(1-P_k)}{N}\right\}.$$

□

4.5 Practical Considerations

In the previous sections, we discussed the order of the bias and variance and their asymptotic properties with respect to the smoothing parameter c. We also showed that the NBKE is asymptotically normal. These results, however, are not useful in obtaining an appropriate or practical value of c for a given sample. Thus, in this section, we focus on a data-driven technique for selecting the level of smoothing and discuss some of the implications regarding the aforementioned bias-variance tradeoff.

One measure often used to describe the precision of a point estimator is the Mean Squared Error (MSE). It describes the expected degree of departure of an estimator from its target value. Recall, for cell $k = 0, 1, \ldots, m$,

$$\begin{aligned} \mathrm{MSE}[\hat{P}_k(c)] &= \mathrm{E}\left\{\left[\hat{P}_k(c) - P_k\right]^2\right\} \\ &= \left\{\mathrm{Bias}[\hat{P}_k(c)]\right\}^2 + \mathrm{Var}[\hat{P}_k(c)]. \end{aligned}$$

Increasing the value of c, or equivalently smoothing less, reduces the bias of $\hat{P}_k(c)$. Reducing the bias, however, occurs at the cost of increasing the variance. Hence, one hopes to choose a c value that strikes a balance between the bias and variance. Furthermore, working with mulitnomial data adds to that complexity because of the need to estimate several parameters simultaneously. For the NBKE, the same level of smoothing has to be applied to all cells. (Recall, however, that the weight function varies according to the cell of interest, so smoothing is adaptive in that sense.) It is then only natural to consider a more global measure of accuracy such as the Mean Sum of Squared Error (MSSE).

In some studies, it may be sufficient to use a subjective or trial-and-error method for choosing the level of smoothing. For comparative studies, or for the sake of efficiency, however, it may be more appropriate to use a standardized or objective method for choosing the level for smoothing. For an introduction to different techniques for choosing the level of smoothing for continuous data, see Silverman [12]. Park and Marron [9] compare three different data-driven approaches for choosing the level of smoothing in their paper: least squares cross-validation, biased cross-validation, and a plug-in method.

One method for choosing the level of smoothing, that is adaptable to the discrete case, is least squares cross-validation. The objective is to minimize the MSSE, or rather, an unbiased estimate of the MSSE since it is an unknown quantity. The least squares component involves minimizing the MSSE, with respect to the smoothing parameter c. The cross-validation component involves estimating a series of probabilities where each probability is based on the removal of a single observation, also known as the leave-one-out estimated probabilities. Because least-squares cross-validation is data-driven and fairly easily implemented, it is the method selected for optimal smoothing in this thesis.

Consider the following expression for MSSE as a function of c,

$$\begin{aligned}
\text{MSSE}(c) &= \text{E}\left\{\sum_{k=0}^{m}\left[\hat{P}_k(c) - P_k\right]^2\right\} \\
&= \text{E}\left[\sum_{k=0}^{m}\hat{P}_k(c)^2\right] - 2\text{E}\left[\sum_{k=0}^{m}\hat{P}_k(c)P_k\right] + \sum_{k=0}^{m}P_k^2 \\
&= \text{E}\left[\sum_{k=0}^{m}\hat{P}_k(c)^2\right] - 2\text{E}\left\{\text{E}_Y\left[\hat{P}_Y(c)\right]\right\} + \sum_{k=0}^{m}P_k^2, \qquad (4.8)
\end{aligned}$$

where Y follows a multinomial distribution with probabilities $P_0, P_1, \ldots, P_m$, and is independent of $X_1, \ldots, X_N$, and E_Y denotes an expectation with respect to Y.

Let $G(c)$ be the portion of (4.8) involving smoothing parameter c. Obviously,

$$G(c) = \mathrm{E}\left[\sum_{k=0}^{m} \hat{P}_k(c)^2\right] - 2\mathrm{E}\left\{\mathrm{E}_Y\left[\hat{P}_Y(c)\right]\right\}$$

and,

$$\mathrm{MSSE}(c) = G(c) + \sum_{k=0}^{m} P_k^2.$$

Minimizing $G(c)$ is then equivalent to minimizing MSSE(c). Now, let

$$g(c) = \sum_{k=0}^{m} \hat{P}_k(c)^2 - 2\mathrm{E}_Y\left[\hat{P}_Y(c)\right] \tag{4.9}$$

and note that $g(c)$ is an unbiased estimator of $G(c)$ that cannot itself be calculated from the observed data. Let $\hat{P}_{x_i,-i}(c) = \hat{P}_c(X_i = x_i|N_{-i})$ be the leave-one-out estimated probability of observing the value $X_i = x_i$, based on the estimated NBKE probabilities calculated by leaving the observed X_i out of the sample.

Applying the leave-one-out technique to (4.9), we further estimate $g(c)$ using the following,

$$\hat{g}(c) = \sum_{k=0}^{m} \hat{P}_k(c)^2 - \frac{2}{N}\sum_{i=1}^{N} \hat{P}_{x_i,-i}(c).$$

This can be further written as,

$$\hat{g}(c) = \sum_{k=0}^{m} \hat{P}_k(c)^2 - \frac{2}{N}\sum_{i=1}^{N}\sum_{k=0}^{m} I(X_i = k)\bar{P}_k(c),$$

where

$$\bar{P}_k(c) = \sum_{l=0}^{m} \frac{\bar{N}_l}{N-1} W_{c,k,m}(l),$$

and,

$$\bar{N}_l = \begin{cases} N_l & \text{for } l \neq k, \\ N_k - 1 & \text{for } l = k. \end{cases}$$

We can simplify this further to obtain the following expression for $g(c)$,

$$\begin{aligned}\hat{g}(c) &= \sum_{k=0}^{m} \hat{P}_k(c)^2 - \frac{2}{N}\sum_{k=0}^{m}\sum_{i=1}^{N} I(X_i = k)\bar{P}_k(c) \\ &= \sum_{k=0}^{m} \hat{P}_k(c)^2 - \frac{2}{N}\sum_{k=0}^{m} N_k\bar{P}_k(c). \end{aligned} \tag{4.10}$$

The final expression in (4.10) requires computing new probabilities for all cells by sequentially deleting each observation. This, of course, can be extremely time consuming! To avoid this problem, we can further express (4.10) in terms of the original observed NBKE probabilities and weights. If we do this, we have,

$$\begin{aligned}\hat{g}(c) &= \sum_{k=0}^{m} \hat{P}_k(c)^2 - \frac{2}{N}\sum_{k=0}^{m} N_k\left[\sum_{l=0}^{m}\frac{N_l}{N-1}W_{c,k,m}(l) - \frac{1}{N-1}W_{c,k,m}(k)\right] \\ &= \sum_{k=0}^{m} \hat{P}_k(c)^2 - \frac{2}{N-1}\sum_{k=0}^{m} N_k\sum_{l=0}^{m}\frac{N_l}{N}W_{c,k,m}(l) + \frac{2}{N(N-1)}\sum_{k=0}^{m} N_kW_{c,k,m}(k) \\ &= \sum_{k=0}^{m} \hat{P}_k(c)^2 - \frac{2}{N-1}\sum_{k=0}^{m} N_k\hat{P}_k(c) + \frac{2}{N(N-1)}\sum_{k=0}^{m} N_kW_{c,k,m}(k), \end{aligned} \tag{4.11}$$

which does not depend on the leave-one-out estimated probabilities. Then, for a given observed sample, optimal smoothing is obtained using the value of c that minimizes (4.11).

Chapter 5

Simulation Study

A simulation study was performed to compare and contrast the effectiveness of the NBKE against the GKE and MLE at various levels of sparseness. Another interest is to compare the effectiveness of the NBKE against the GKE (both with optimal smoothing), particularly for estimating boundary cell probabilities as the GKE is an estimator that can produce boundary bias. Two simulation scenarios were considered, each with three different sample sizes to adjust for the level of sparseness. Both sets were based on multinomial distributions exhibiting characteristics of the IBD survey data, and mainly for the type of data where the NBKE is most appropriate. Thus, each simulation consisted of sparse, ordinal data. Furthermore, we apply the assumption of a smooth underlying density curve. In other words, probabilities of neighbouring cells are similar to each other. Also, data points are highly concentrated about zero and form one or two small modes in other cell-regions.

In the first simulation scenario, there are 25 cells ($k = 0, 1, \ldots, 24$) with sample sizes $N = 50, 125,$ and 1000. In the second simulation case, there are 100 cells ($k = 0, 1, \ldots, 99$) with sample sizes $N = 200, 500,$ and 2500. For each scenario and sample size combination, 1000 sets of data were simulated using R Software (Version 2.8.1).

For each set of simulated data, probability estimates were computed using three methods: the NBKE with optimal smoothing, the GKE with optimal smoothing and the MLE. In Figures 5.1 and 5.2, probabilities are plotted for the first set of simulated data for each sample size in the 25-cell and 100-cell scenarios, respectively. Both figures show general smoothness for the curves representing the probability estimates using optimal smoothing with the NBKE, regardless of the number of cells and sample size. The curves for the MLE probabilities, however, can vary greatly in the level of smoothness. Smoothness seems to be inversely related to the number of cells, but positively related to sample size. Also, greater volatility appears about modal regions. The level of smoothness of the GKE appears to fall between the NBKE and MLE. Thus, per any given simulation (or dataset) using the NBKE appears to produce estimates reflecting the true shape of smooth, multinomial data better than the GKE and MLE. On the other hand, there are some indications of the bias-variance tradeoff in effect for the NBKE and GKE. This effect is much more pronounced in the figures based on the aggregate of 1000 simulations. This will be discussed shortly.

Figures 5.3, 5.4, 5.5 and 5.6 show probability curves of the first five sets of simulated data for the NBKE and GKE. They are included to indicate the level of variability among the simulated datasets. (Probability curves for the MLE were omitted due to excessive volatilty.) They suggest that increasing the sample

size reduces the overall level of variability, as expected. And again, we can that the curves for the NBKE are generally smoother and less volatile than the GKE. Looking at these figures more closely, we see that for the NBKE, there is greater volatility and bias at the lower boundary than other regions. These effects are even more apparent for the GKE, particularly for the 100-cell scenario. Perhaps, the lower boundary cells exhibit more variability because probabilities in that region are closer to 0.5, as previously proposed in Theorem 4.2.5. Also, although the NBKE is asymptotically unbiased with respect to the smoothing parameter c, the c value from optimal smoothing may not be large enough to eliminate or reduce the bias to negligible levels. These preliminary findings are confirmed with subsequent figures that encapsulate each 1000-set simulation combination.

Figures 5.7, 5.8, 5.9 and 5.10 show the NBKE and GKE cross-validation functions for the first five simulation runs. Recall that the smoothing parameter s for the GKE ranges between 0 and 1 with lower values indicating less smoothing. This is the opposite to the NBKE smoothing parameter c which ranges from 0 to infinity. We can see that increasing the sample size lead to larger optimal values of c or smaller optimal values of s which means that less smoothing is required, as expected again. They also suggest that the cross-validation function flattens out more rapidly with larger sample size.

The distributions of the optimal smoothing parameter c are shown in Figures 5.11 and 5.12. These figures have modes at higher values as sample size increases. This further supports the finding that larger c values are required for optimal smoothing as sample sizes increases. Figures 5.13 and 5.14 consist of histograms of all the optimal s values for the GKE which clearly show that less smoothing is required for larger sample sizes. More precisely, this means for both the NBKE

and GKE, less smoothing is needed for large sample sizes which intuitively makes sense. Note that both the 25-cell and 100-cell scenarios have similar ratios of cell size to sample size. We can see that more smoothing is required as the number of cells increase.

For each cell-sample size combination consisting of 1000 simulations, the empirical mean and 95% confidence interval were determined for each cell proportion (or probability). For example, to determine the lower confidence limit for the cell $k = 10$, we used the 2.5% quantile and for the upper confidence limit, we used the 97.5% quantile. Means and confidence intervals were computed for all three estimators. Means for all three estimators are fairly similar to each other. Similar results are expected since the simulated data are not necessarily sparse.

In Figures 5.15 and 5.16 we can see that confidence bands for the NBKE are much narrower than the MLE, even for very large sample sizes, and thus the NBKE can provide more precise estimates than the MLE. The NBKE, however, does produce some bias at local extrema, as expected. Peak regions are slightly lower and valley regions are slightly higher. Boundary cells appear somewhat biased or exhibit greater variability. (It is also important to note that these are not global confidence bands at the 95% level. They are actually obtained separately for each cell.)

Figures 5.17 and 5.18 confirm that there is less variability overall for both the NBKE and GKE as sample size increases. We do, however, see that the GKE has greater lower-boundary bias and slightly more variability than the NBKE in most regions. For the GKE, the lower-boundary bias can be attributed to it being a symmetric kernel. Weights also allocated to cells beyond the domain so they do not sum up to 1 within the support.

It is also interesting to note the absence of upper-boundary bias. Dong and Simonoff [5] propose that estimators in general are unbiased at the boundary if the true probabilities are zero which intuitively makes sense. In other words, if either $P_0 = 0$ or $P_m = 0$, then there is no boundary bias at the respective region. Hence, there is no significant upper-boundary bias as $P_m \approx 0$.

SSE was computed for each simulation to compare the precision of the three estimators. (SSE is defined as the sum of the squared difference between the estimated probability and the true probability across all cells.) In Figures 5.19 and 5.20, SSE values decrease as sample size increases, as expected. These figures also show that SSE values for the NBKE tend to be lower than values for the GKE and MLE. For each sample size in the 25-cell scenario, about 75% or more of the NBKE SSE values are smaller than at least the 50% quantile of the MLE SSE. This contrast is even more striking for the scenario with 100 cells. Almost all of the NBKE SSE values are smaller than *minimum* MLE SSE! Even with a very large sample size (where the cell size to sample size ratio is 1:25), precision is greatly increased with data driven optimal smoothing. Overall, we see that the NBKE has slightly more consistently precise estimates.

To summarize, NBKE estimates are smoother than both the GKE and MLE estimates, or in other words, the NBKE fitted curves better reflect the true density curve. NBKE estimates have smaller variance as indicated by the smaller confidence band. Both the NBKE and GKE, however, experience some bias about local extrema, with the GKE experiencing particularly large bias about the lower boundary. As expected, less smoothing is required for larger samples sizes, and more smoothing is required for multinomial data with a larger number of cells.

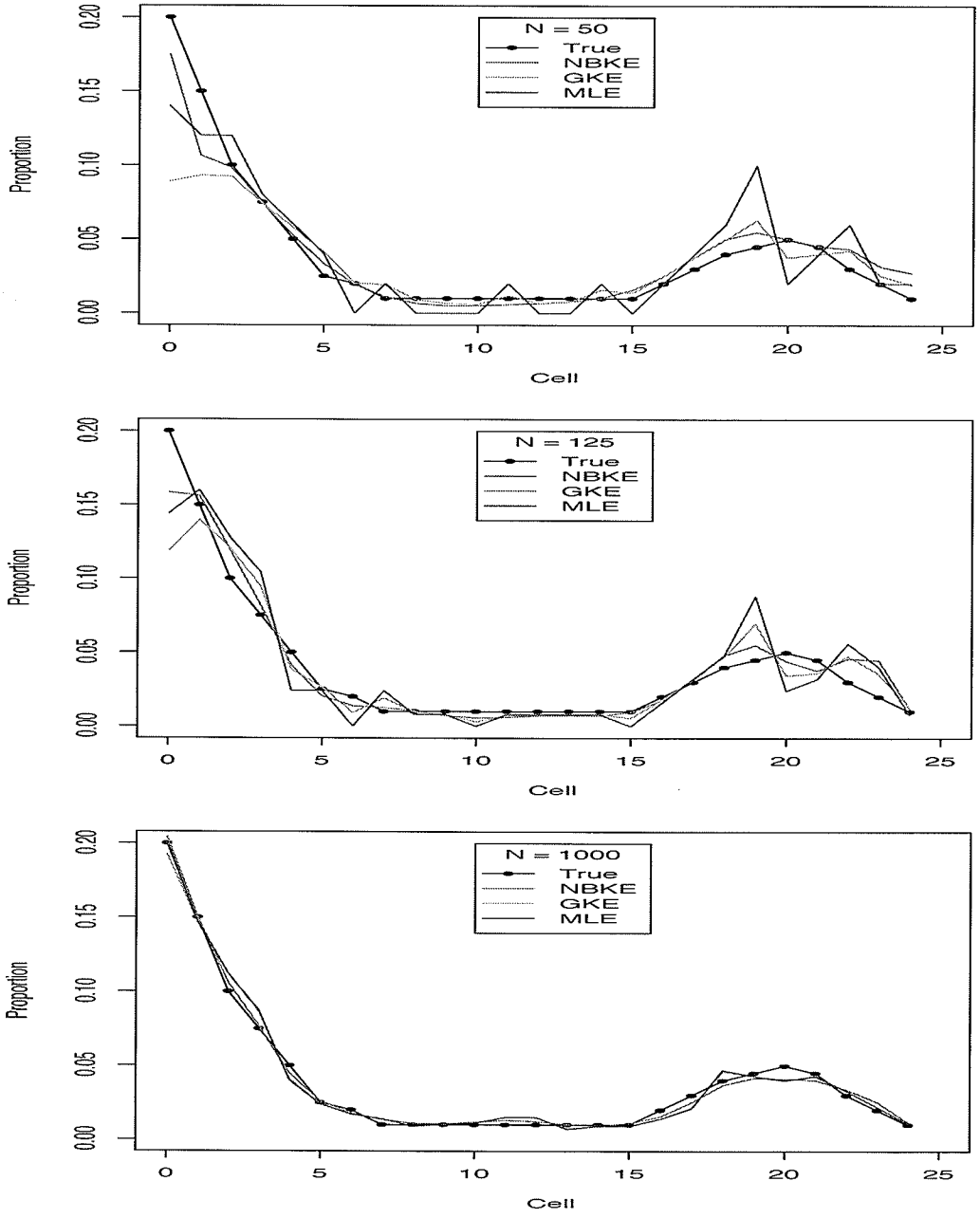

Figure 5.1: Probabilities for the true distribution and estimates using the NBKE and GKE with optimal smoothing and MLE for the first set of simulated data, based on the scenario with 25 cells. From top to bottom, $N = 50, 125,$ and 1000.

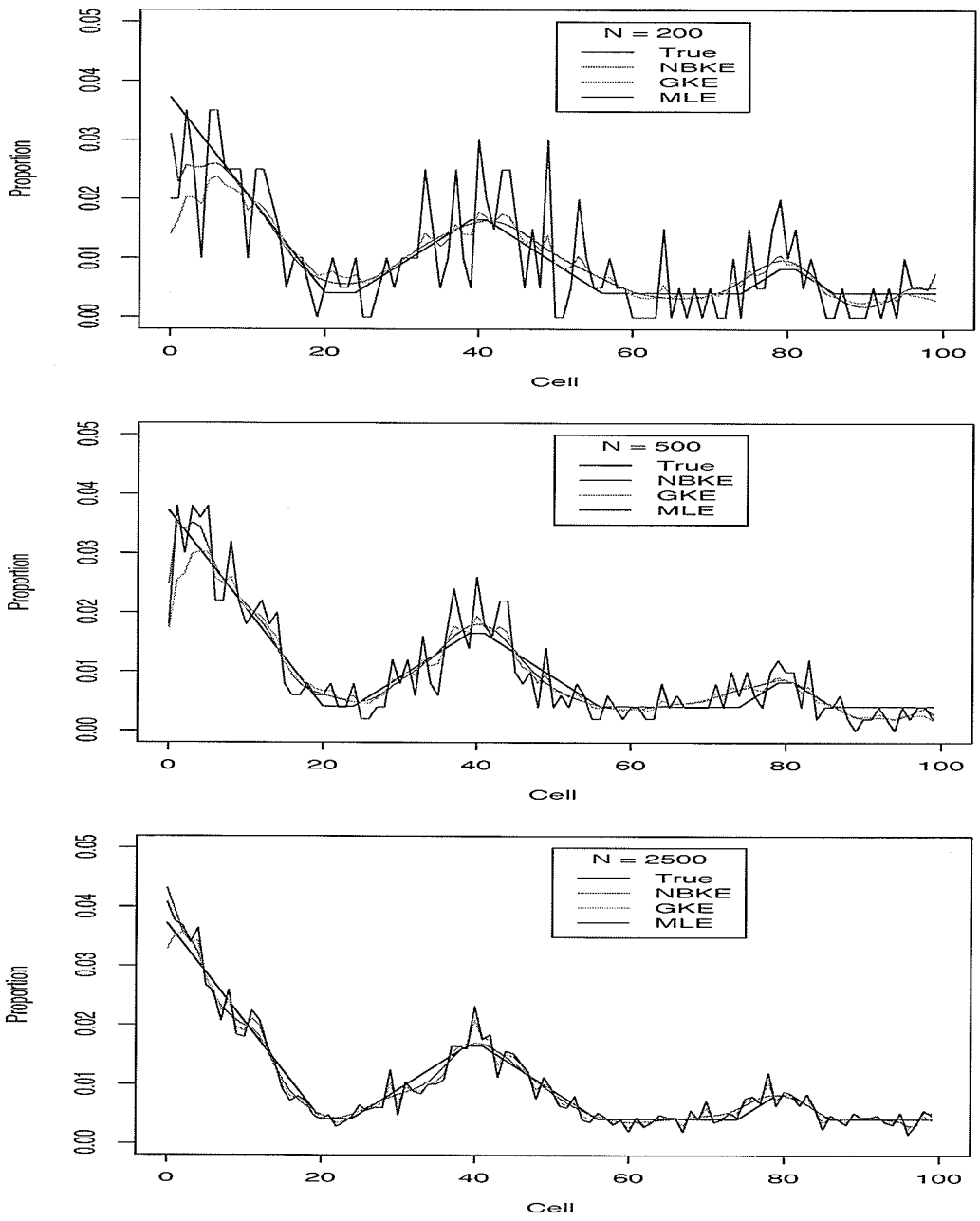

Figure 5.2: Probabilities for the true distribution and estimates using the NBKE and GKE with optimal smoothing and MLE for the first set of simulated data, based on the scenario with 100 cells. From top to bottom, $N = 200, 500,$ and 2500.

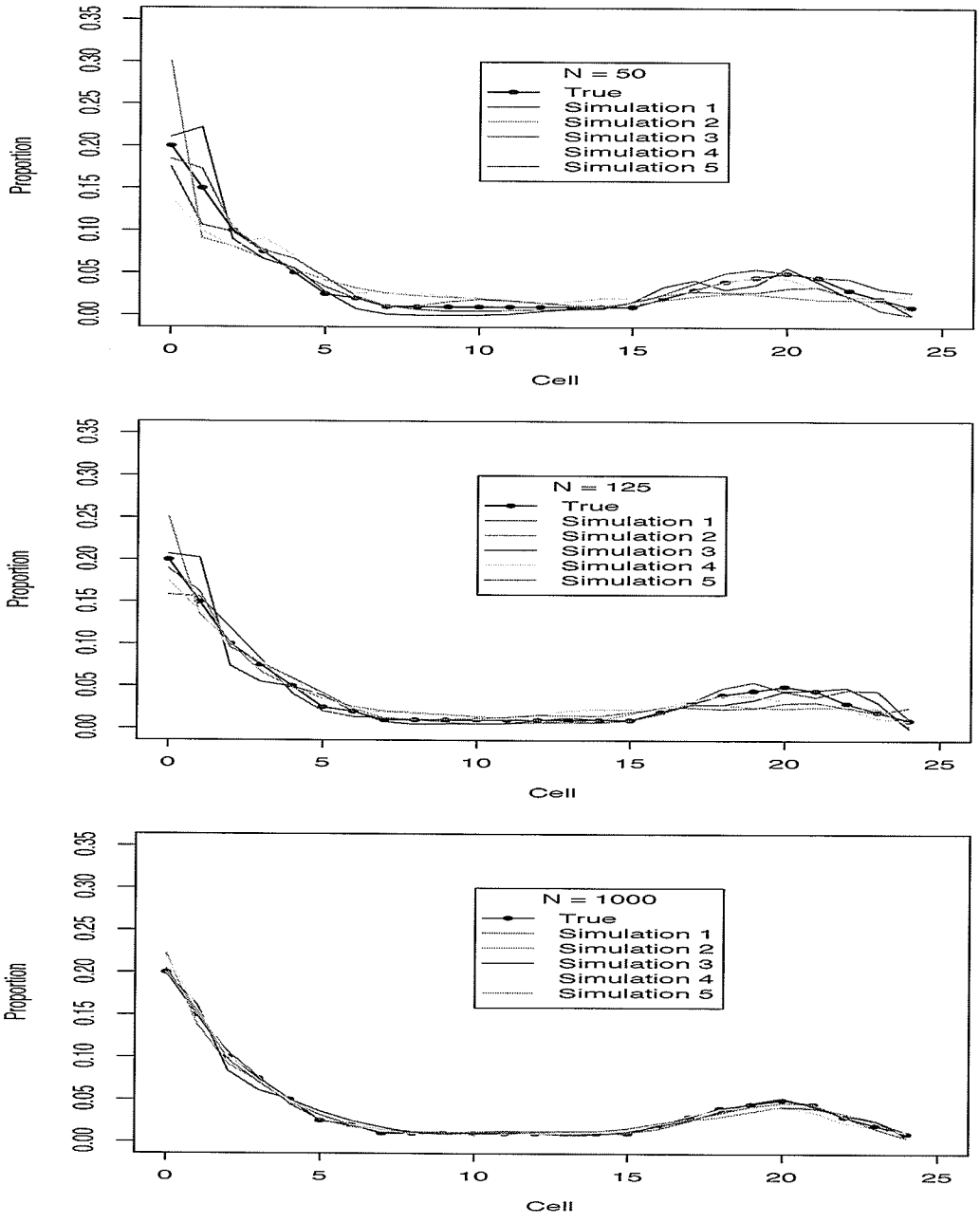

Figure 5.3: Probability estimates using the NBKE with optimal smoothing for the first five sets of simulated data, based on the scenario with 25 cells. From top to bottom, $N = 50, 125,$ and 1000.

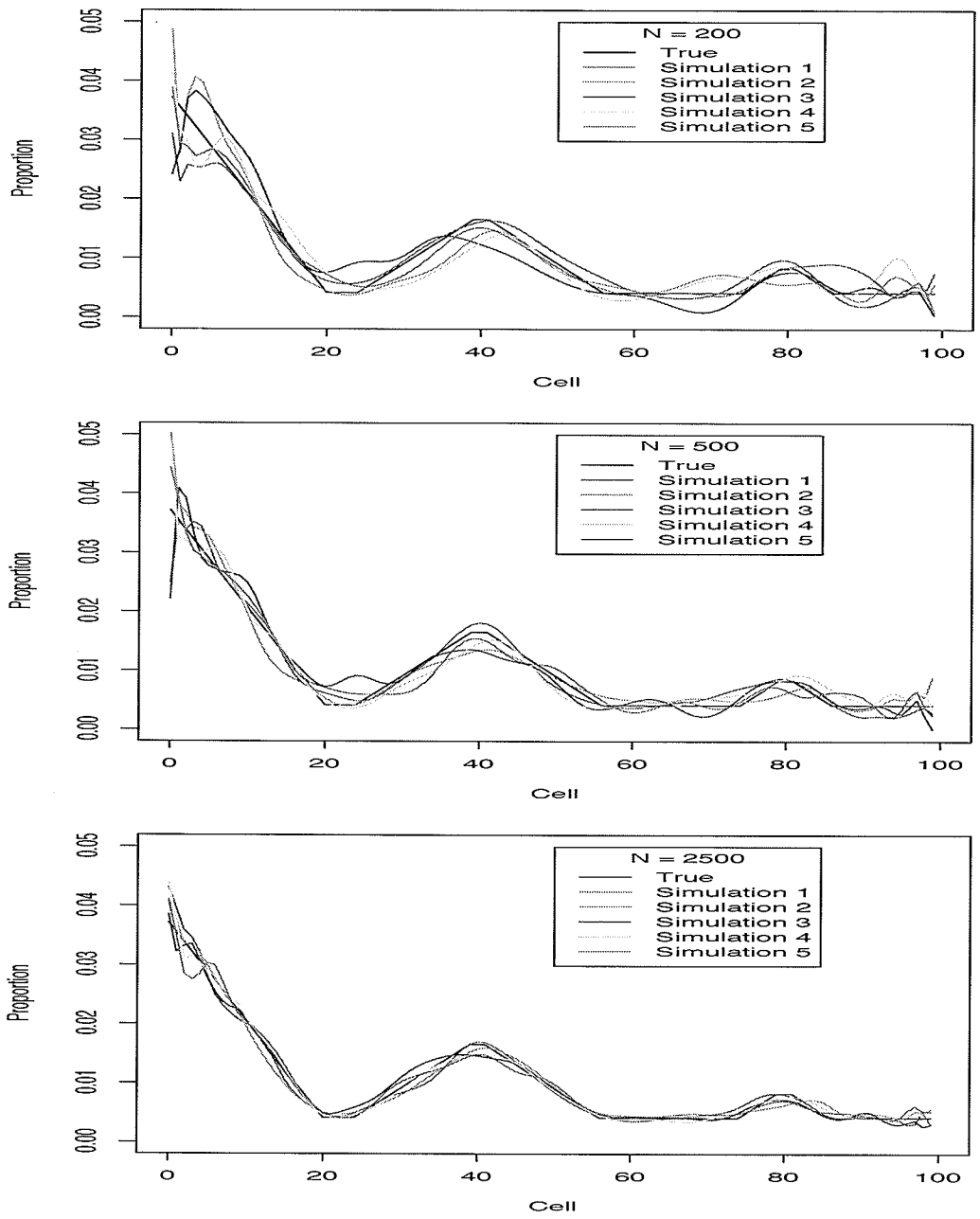

Figure 5.4: Probability estimates using the NBKE with optimal smoothing for the first five sets of simulated data, based on the scenario with 100 cells. From top to bottom, $N = 200, 500,$ and 2500.

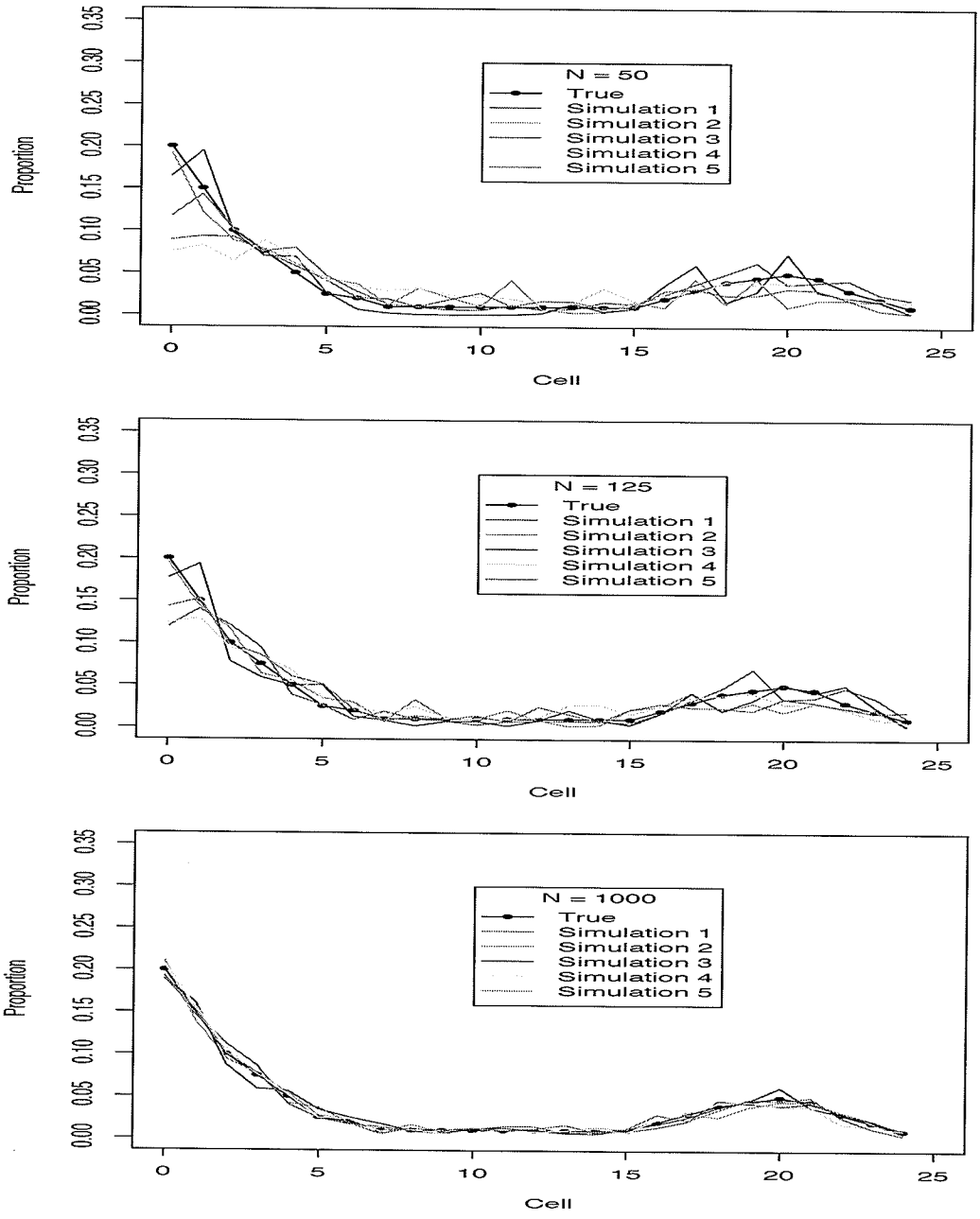

Figure 5.5: Probability estimates using the GKE with optimal smoothing for the first five sets of simulated data, based on the scenario with 25 cells. From top to bottom, $N = 50, 125,$ and 1000.

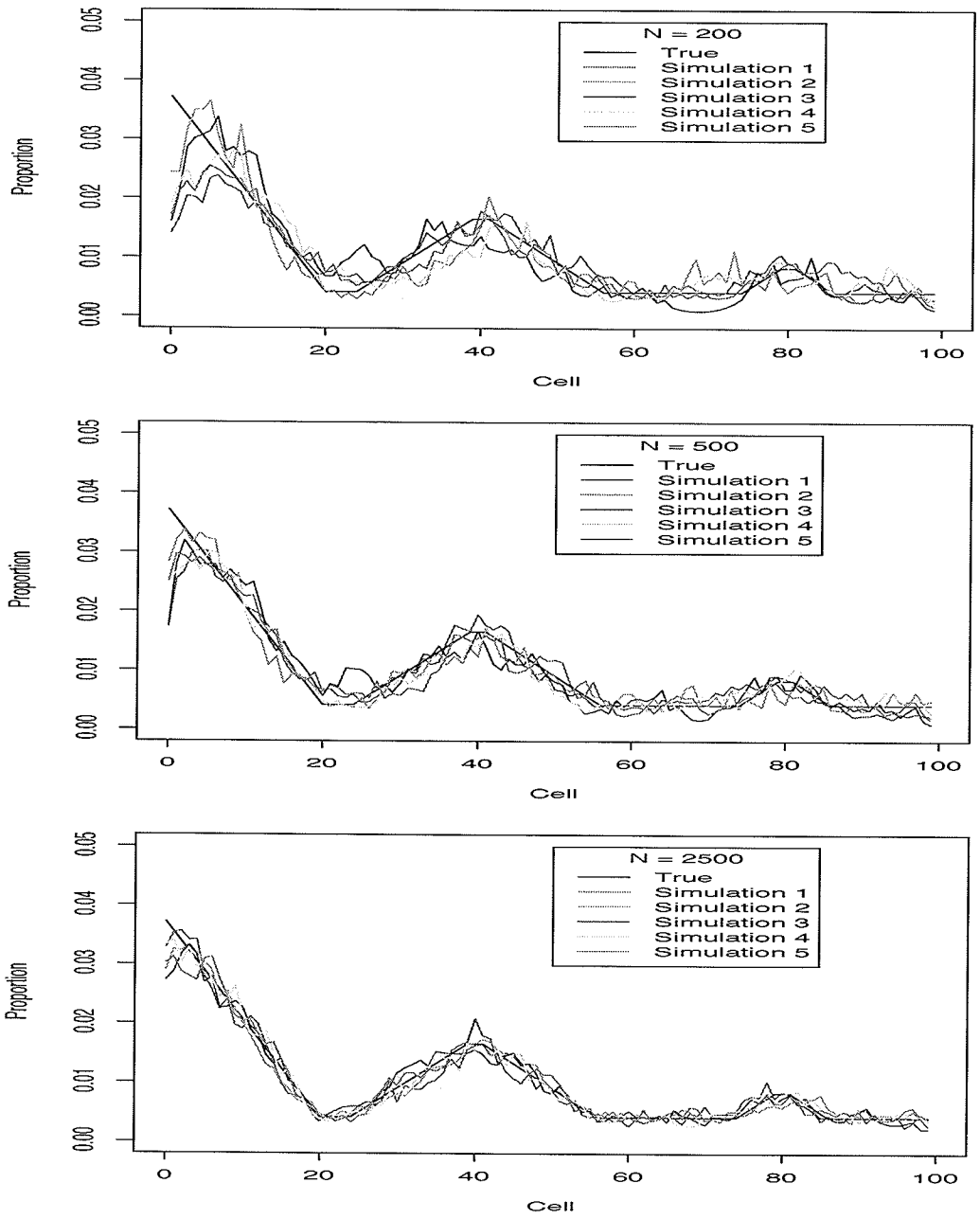

Figure 5.6: Probability estimates using the GKE with optimal smoothing for the first five sets of simulated data, based on the scenario with 100 cells. From top to bottom, $N = 200, 500,$ and 2500.

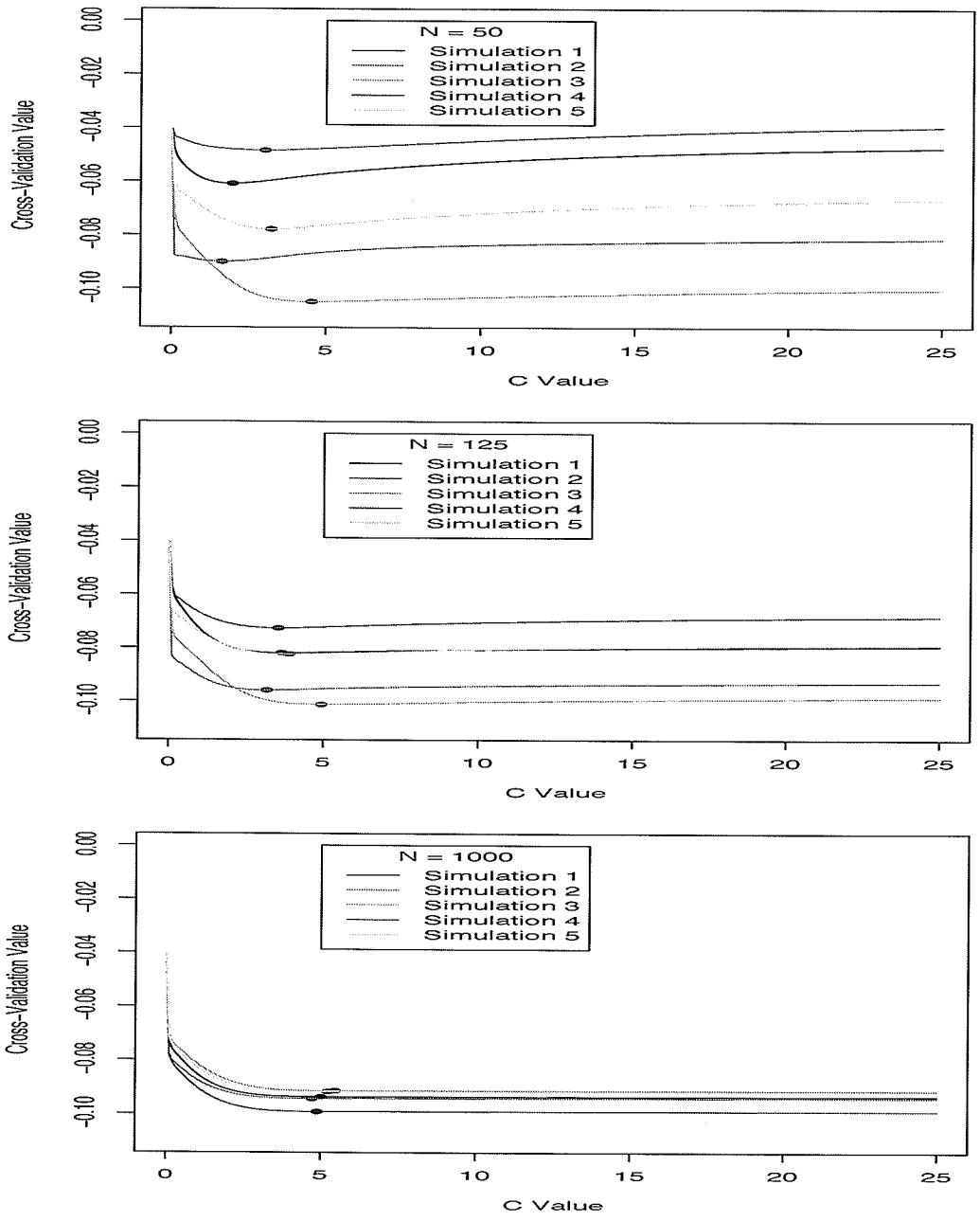

Figure 5.7: Cross-validation functions (with optimal c) for the NBKE for the first five sets of simulated data, based on the scenario with 25 cells. From top to bottom, $N = 50, 125,$ and 1000.

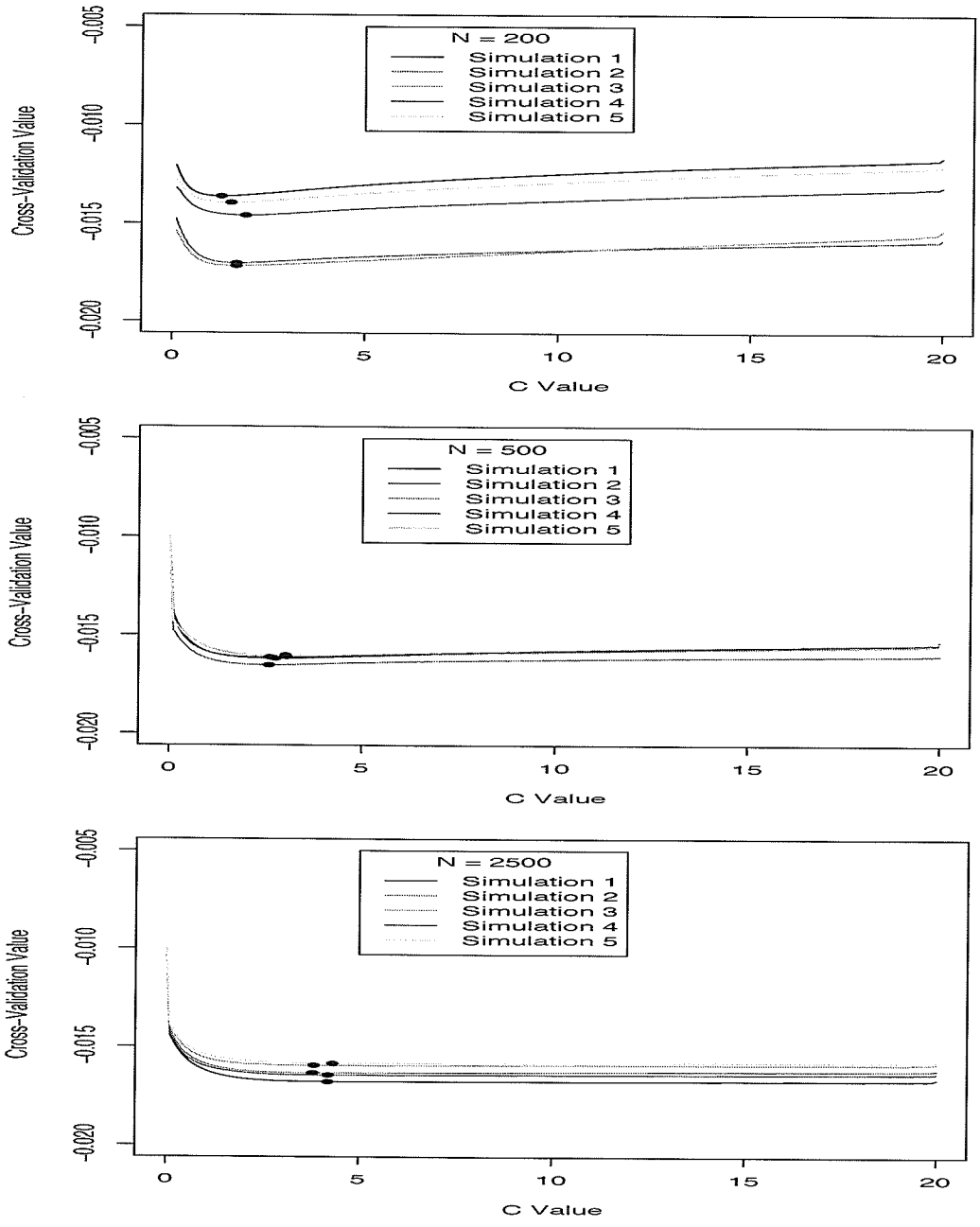

Figure 5.8: Cross-validation functions (with optimal c) for the NBKE for the first five sets of simulated data, based on the scenario with 100 cells. From top to bottom, $N = 200, 500,$ and 2500.

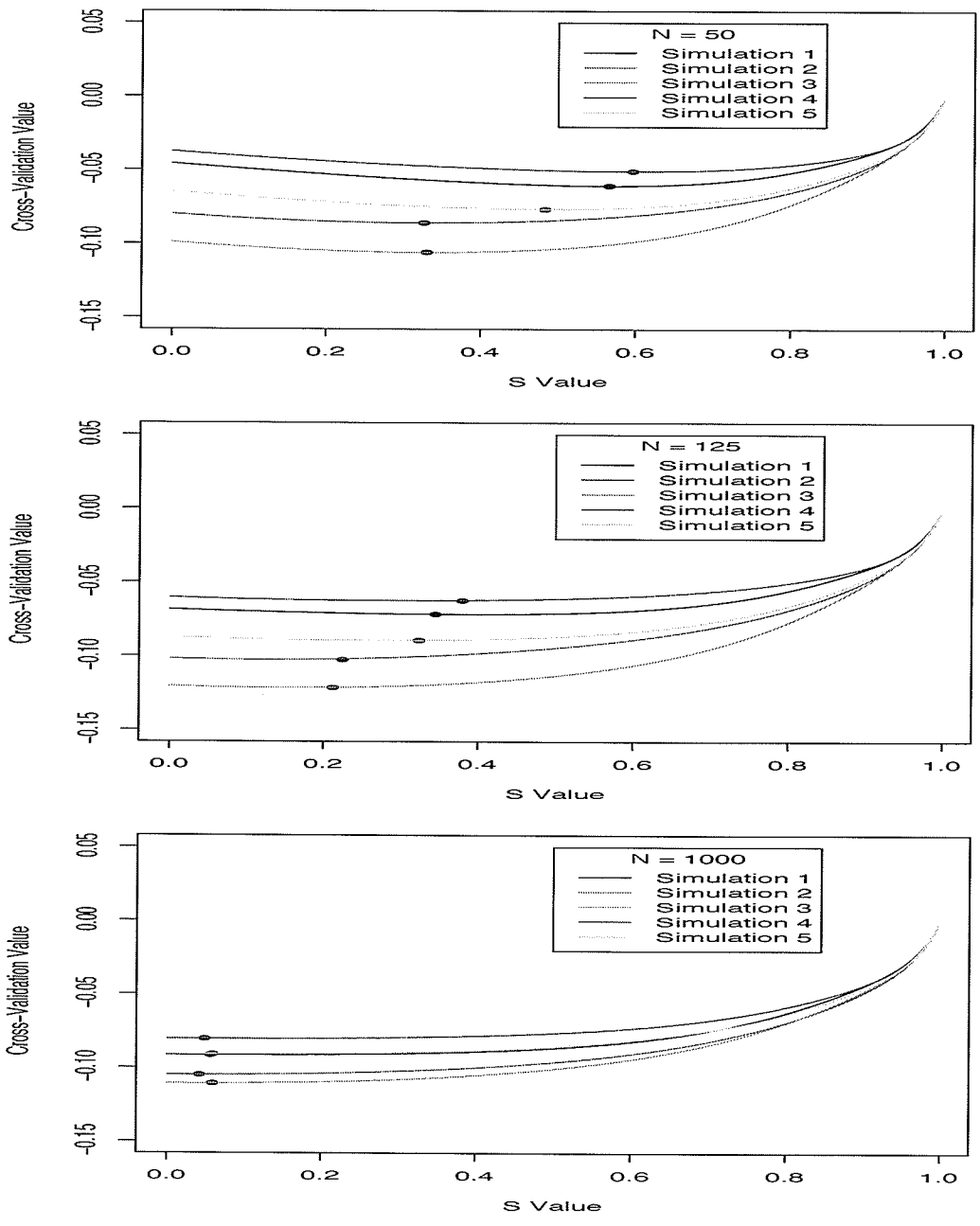

Figure 5.9: Cross-validation functions (with optimal s) for the GKE for the first five sets of simulated data, based on the scenario with 25 cells. From top to bottom, $N = 50, 125,$ and 1000.

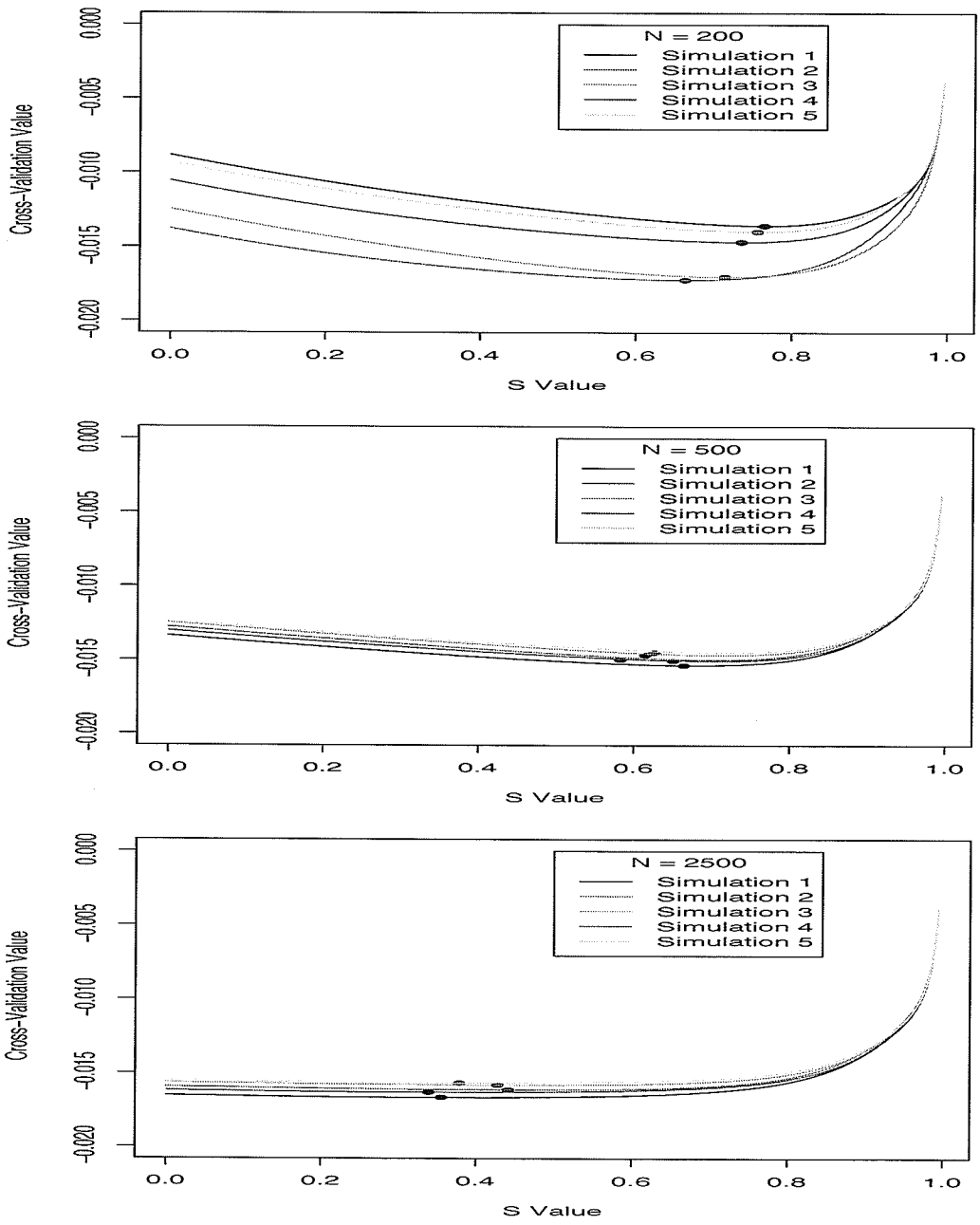

Figure 5.10: Cross-validation functions (with optimal s) for the GKE for the first five sets of simulated data, based on the scenario with 100 cells. From top to bottom, $N = 200, 500,$ and 2500.

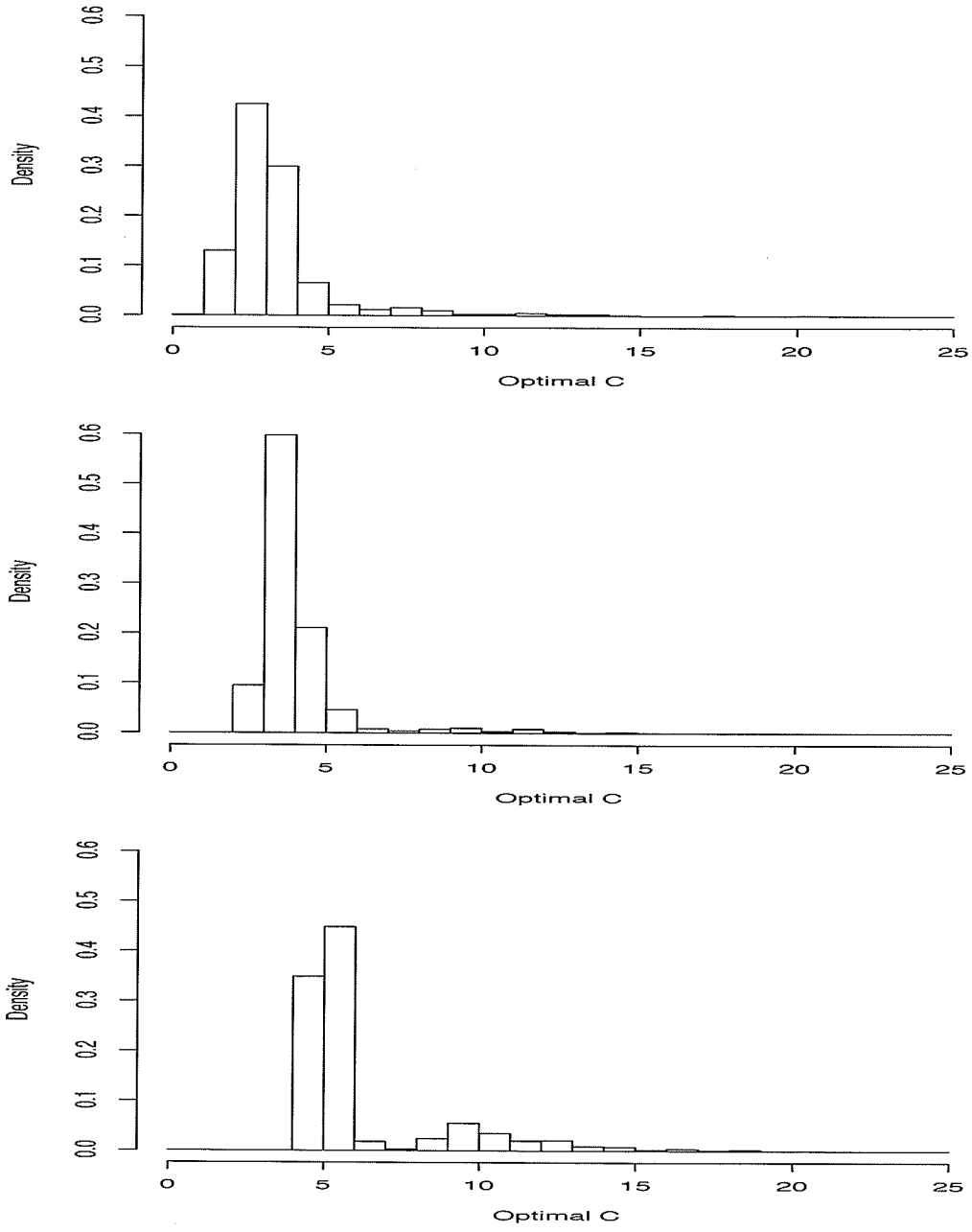

Figure 5.11: Density histograms of optimal smoothing parameter c values for the NBKE, based on the scenario with 25 cells. From top to bottom, $N = 50, 125,$ and 1000.

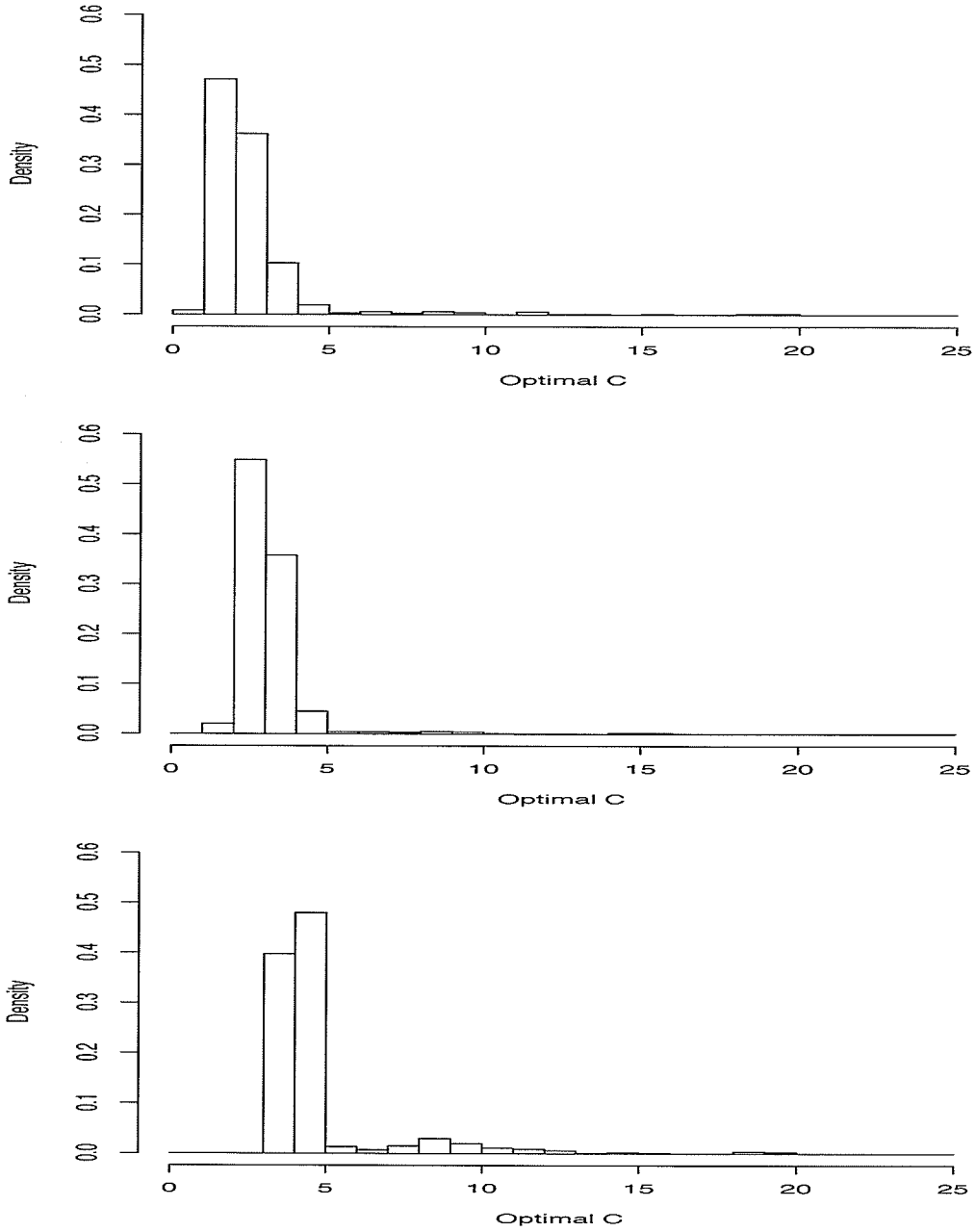

Figure 5.12: Density histograms of optimal smoothing parameter c values for the NBKE, based on the scenario with 100 cells. From top to bottom, $N = 200, 500,$ and 2500.

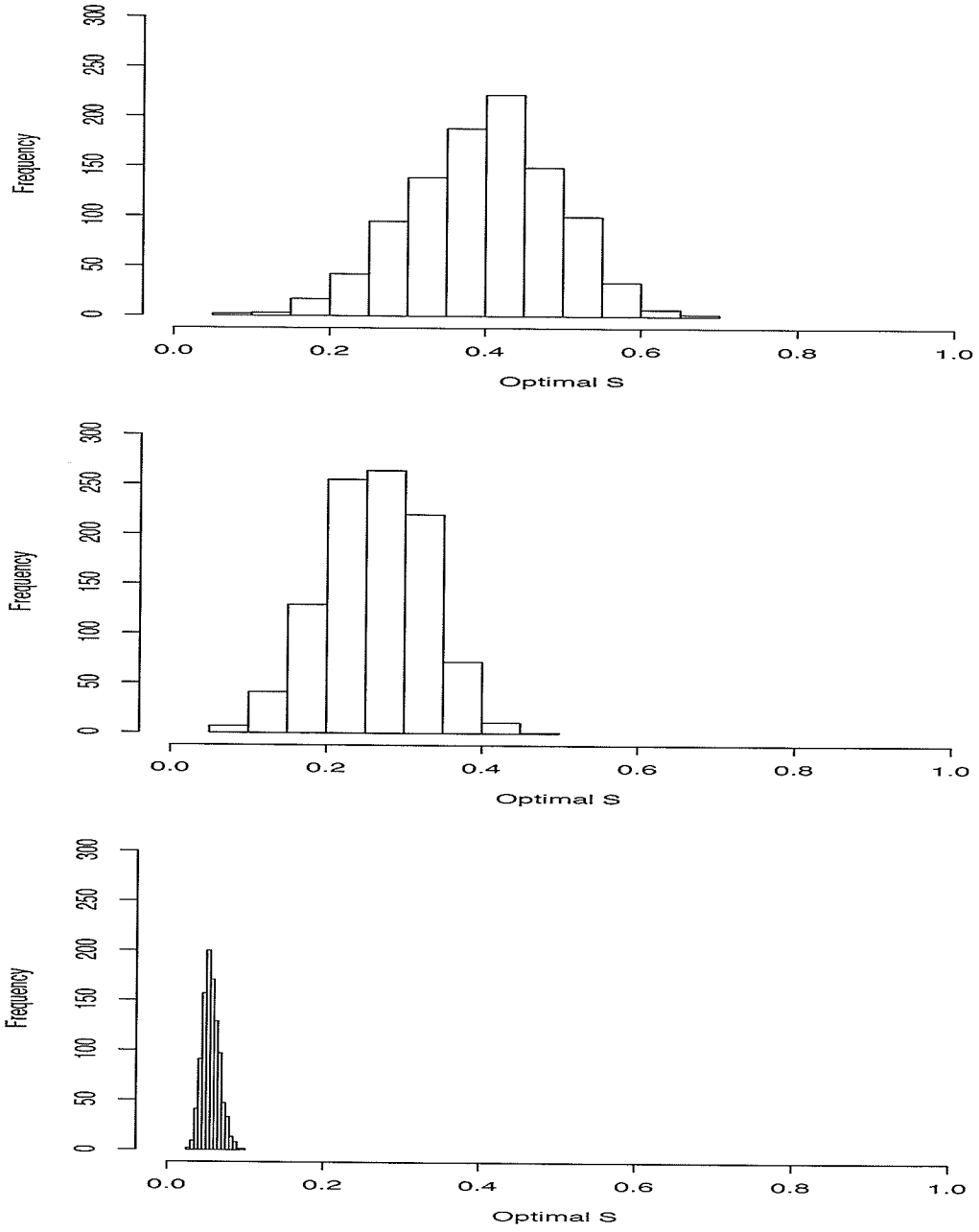

Figure 5.13: Frequency histograms of optimal smoothing parameter s values for the GKE, based on the scenario with 25 cells. From top to bottom, $N = 50, 125$, and 1000.

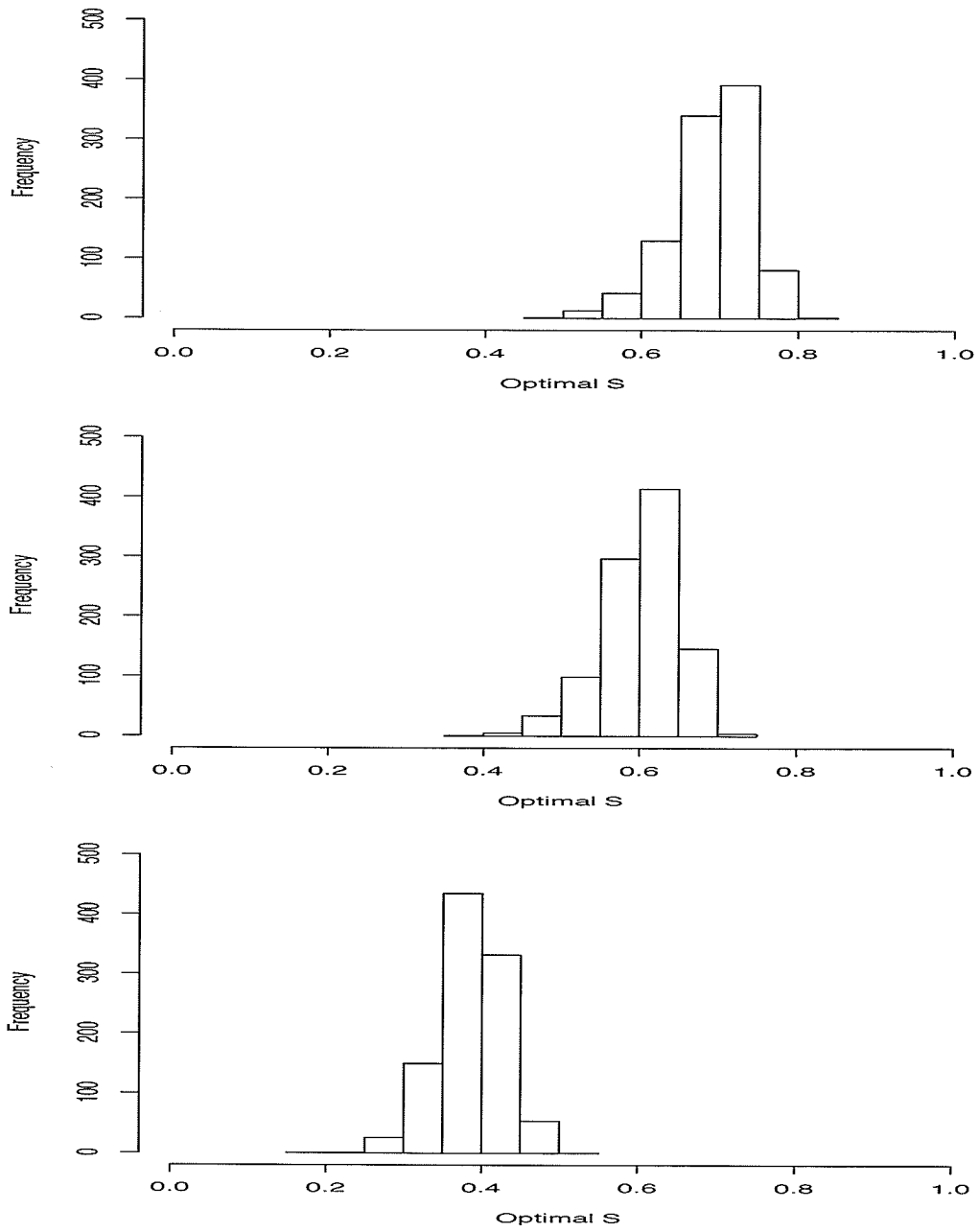

Figure 5.14: Frequency histograms of optimal smoothing parameter s values for the GKE, based on the scenario with 100 cells. From top to bottom, $N = 200, 500,$ and 2500.

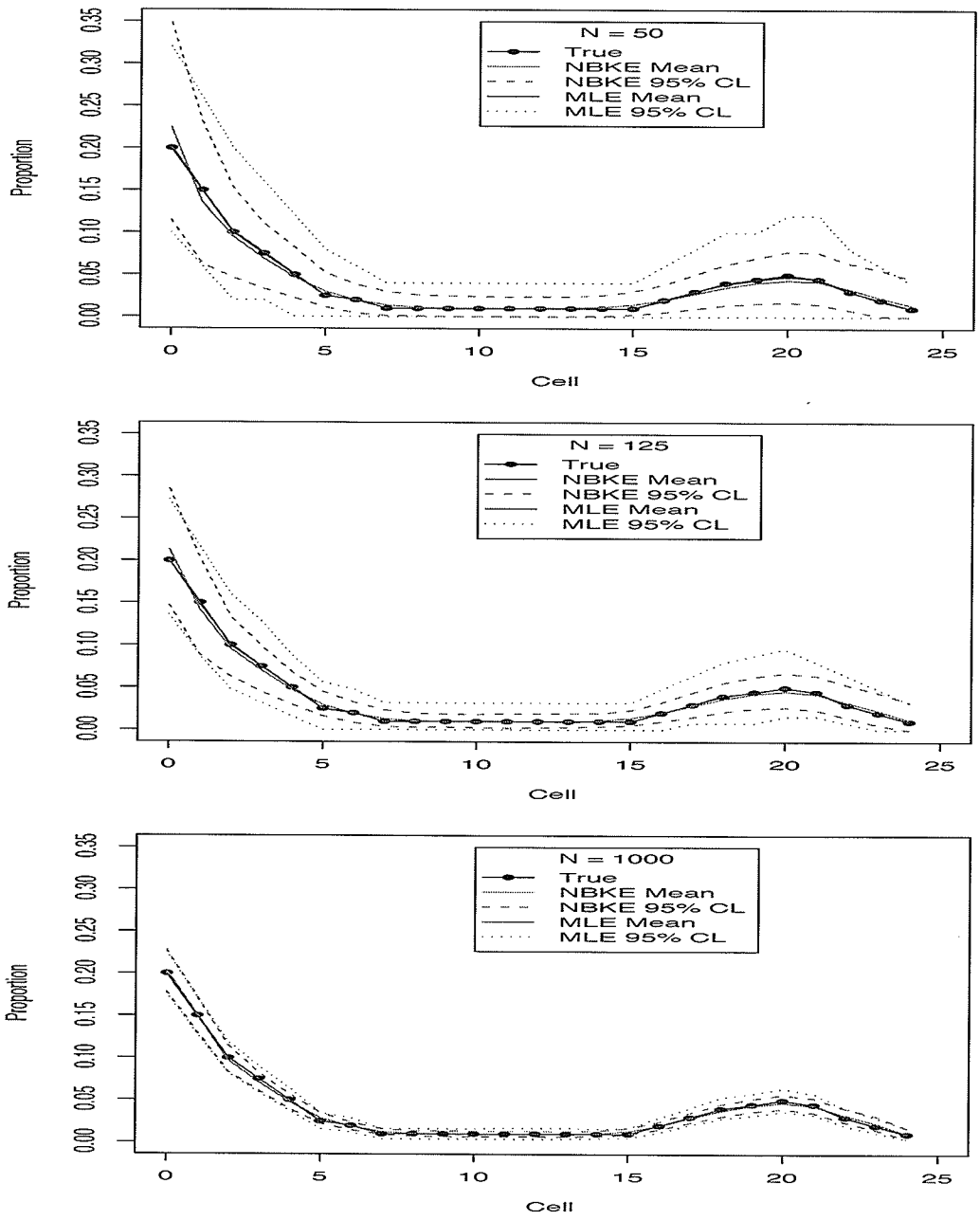

Figure 5.15: Empirical means and 95% CIs for the NBKE with optimal smoothing and MLE, based on the scenario with 25 cells. From top to bottom, $N = 50, 125,$ and 1000.

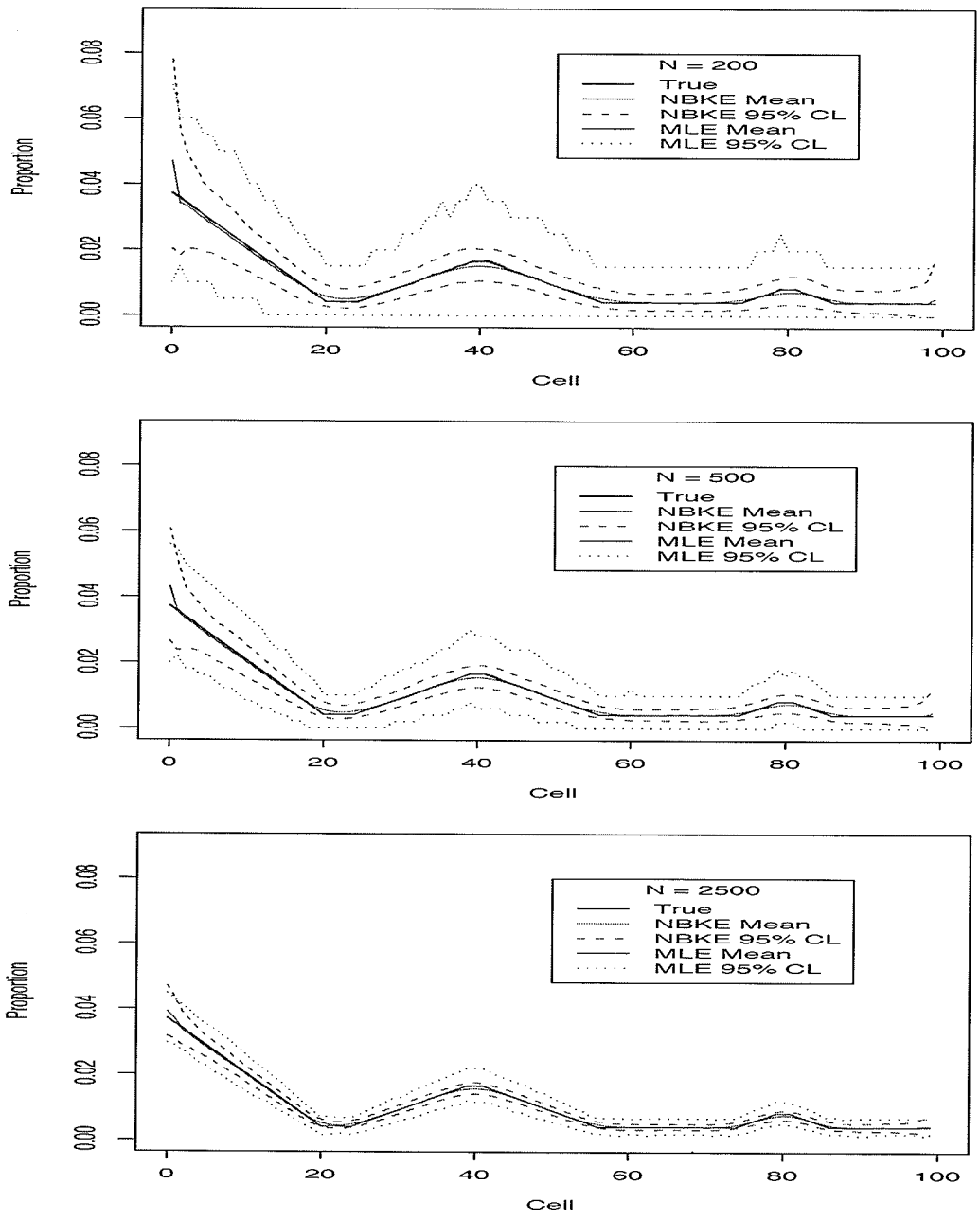

Figure 5.16: Empirical means and 95% CIs for the NBKE with optimal smoothing and MLE, based on the scenario with 100 cells. From top to bottom, $N = 200, 500,$ and 2500.

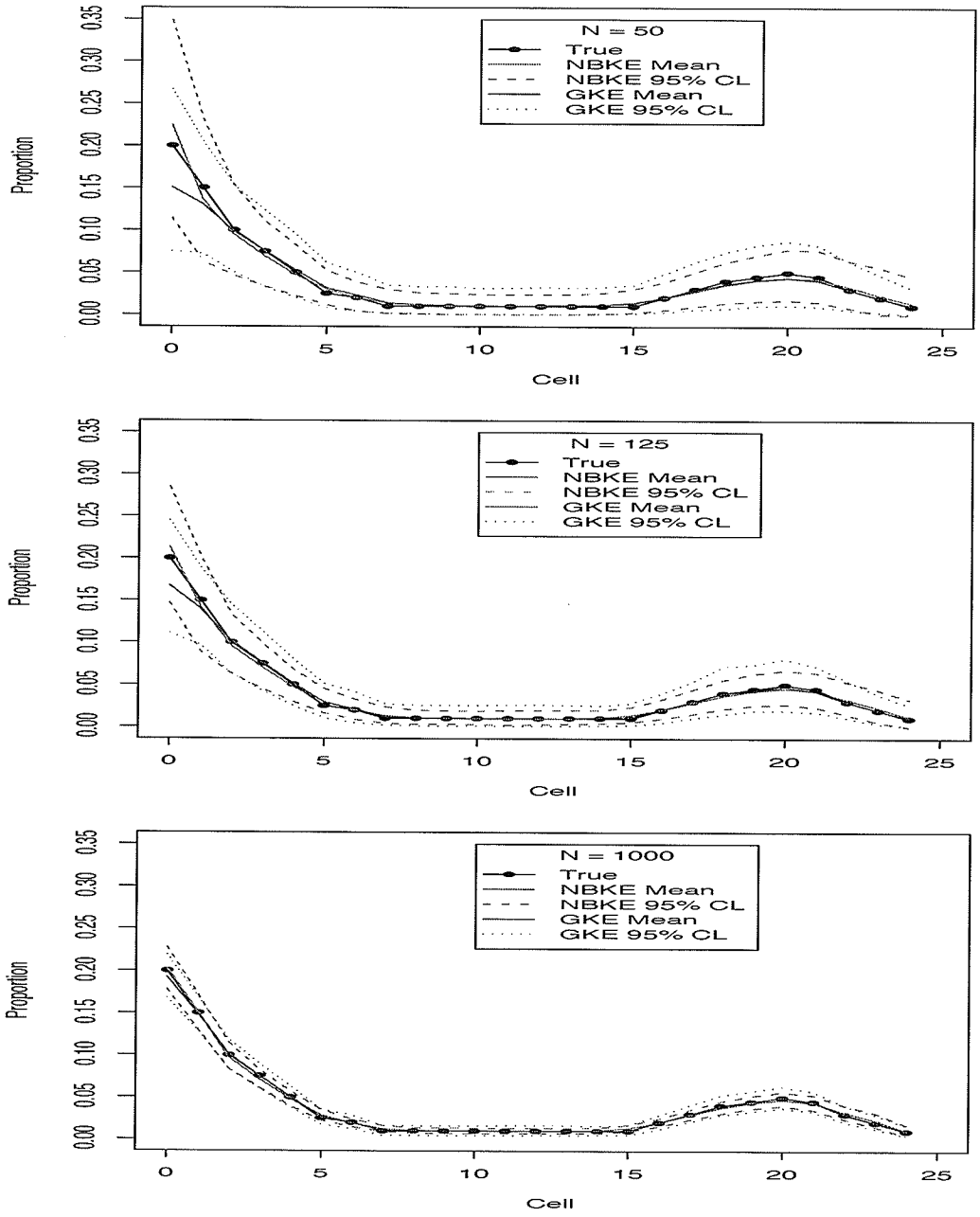

Figure 5.17: Empirical means and 95% CIs for the NBKE and GKE with optimal smoothing. From top to bottom, $N = 50, 125,$ and 1000.

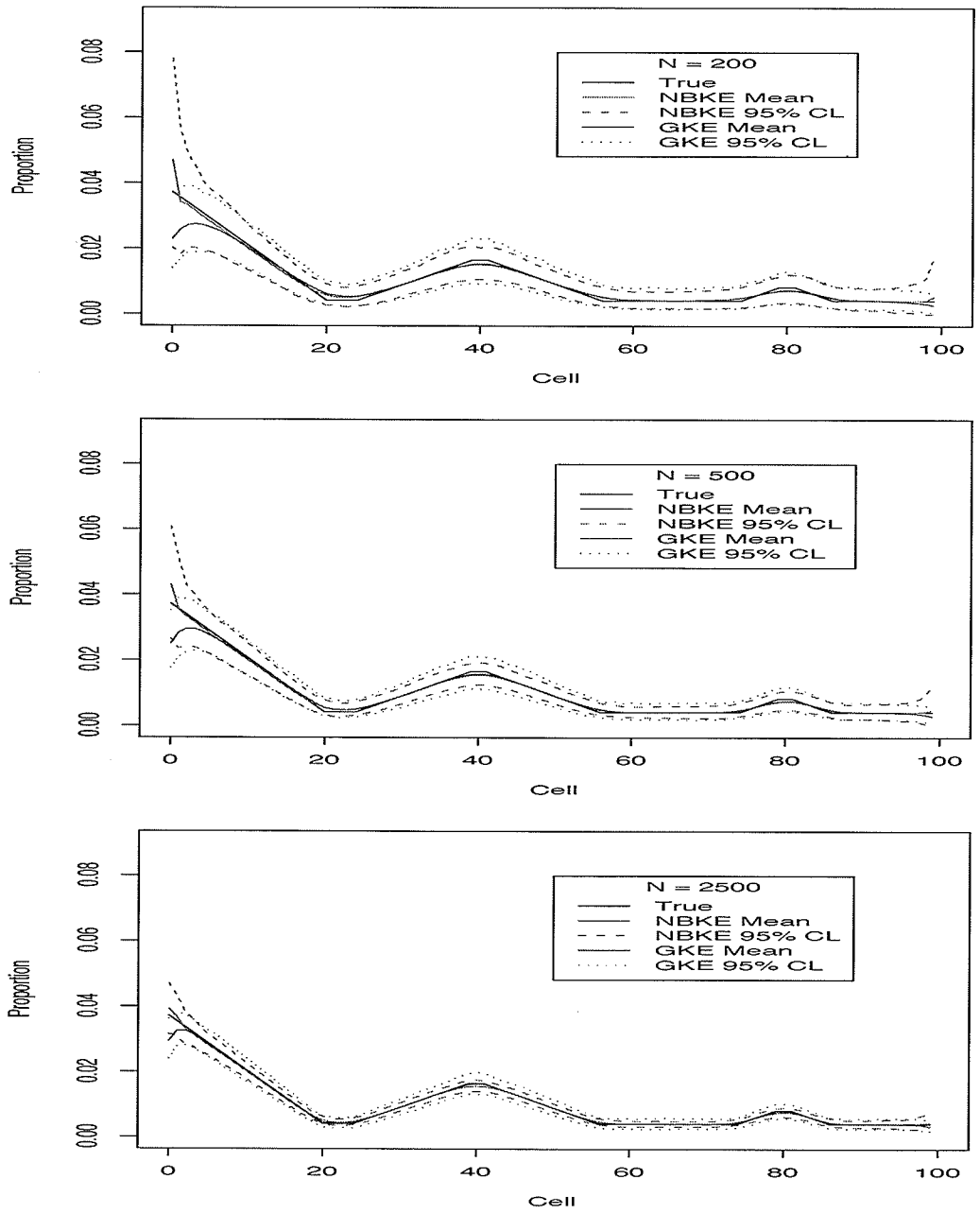

Figure 5.18: Empirical means and 95% CIs for the NBKE and GKE with optimal smoothing. From top to bottom, $N = 200, 500,$ and 2500.

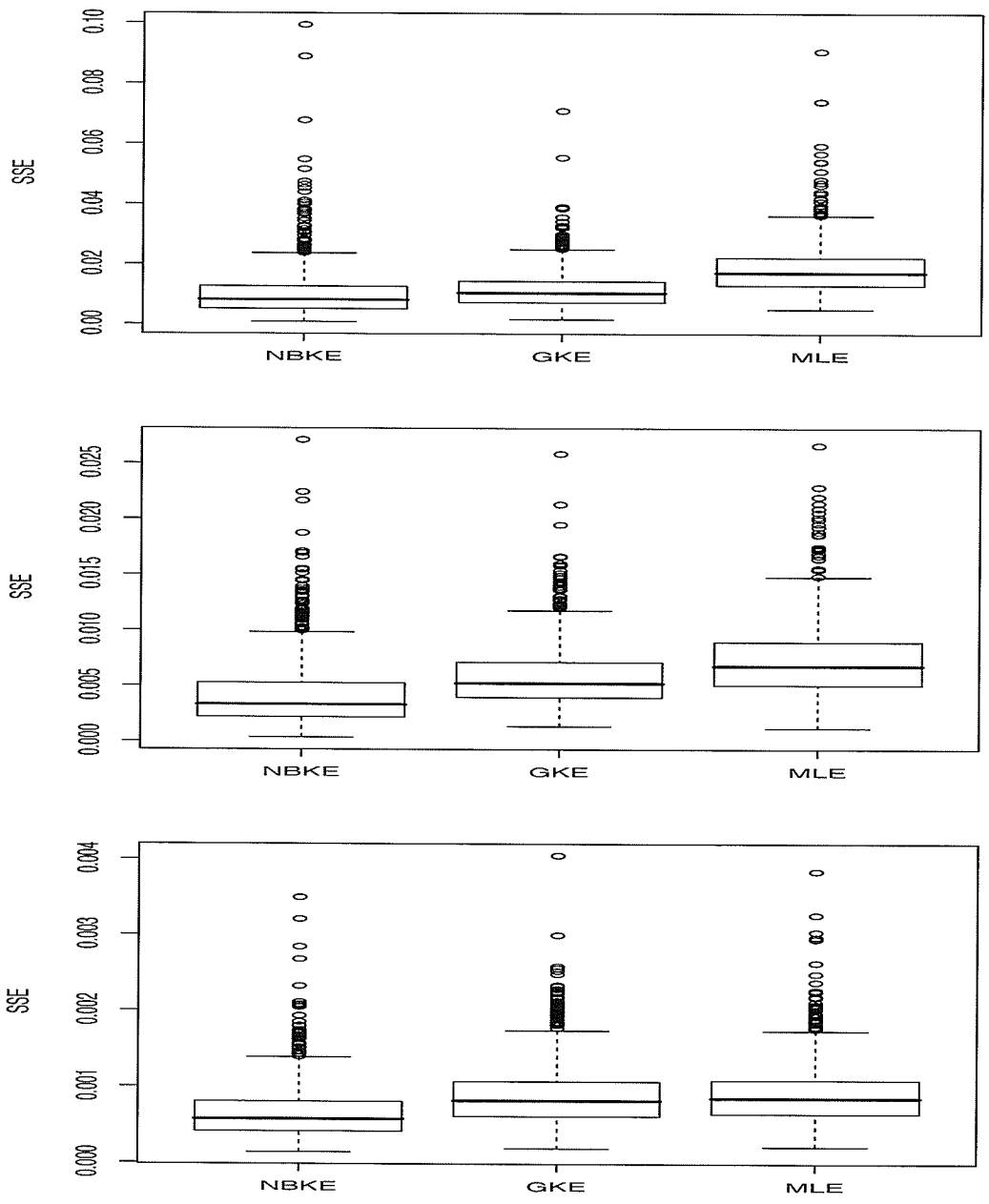

Figure 5.19: SSE Box-plots for the NBKE and GKE with optimal smoothing and MLE, based on the scenario with 25 cells. From top to bottom, $N = 50, 125,$ and 1000.

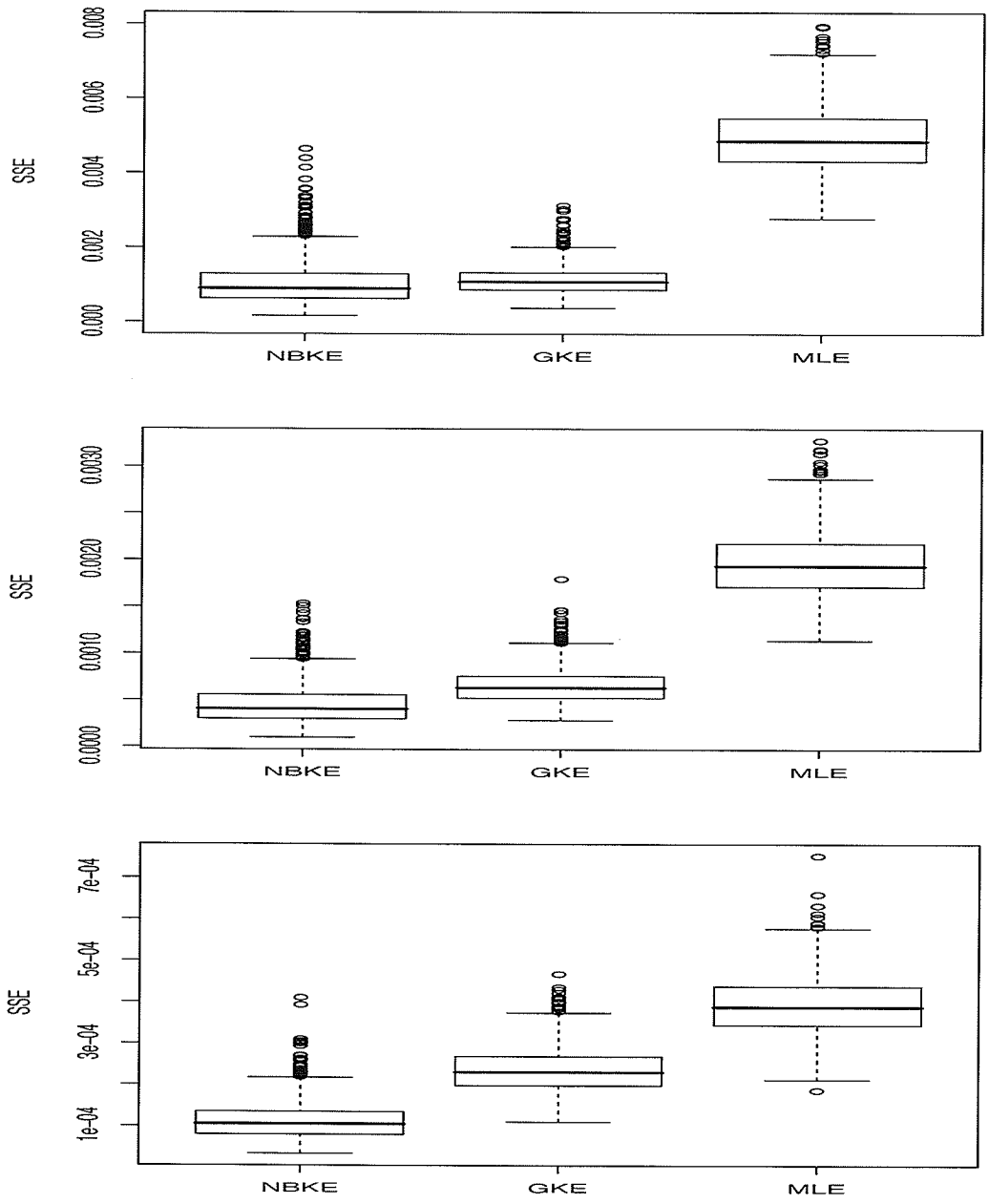

Figure 5.20: SSE Box-plots for the NBKE and GKE with optimal smoothing and MLE, based on the scenario with 100 cells. From top to bottom, $N = 200, 500,$ and 2500.

Chapter 6

Case Study: IBD Data

In the IBD survey, there are two main IBD groups of interest. Between the two groups, there are a total of 247 subjects; 137 diagnosed with Crohn's disease and the other 110 diagnosed with Ulcerative Colitis. (See the Data Appendix on the post processing done to the data.) One objective of the study was to determine whether or not a statistically significant difference exists between the two groups with respect to the number of symptom flares experienced within the last six months. Figure 6.1 shows the fitted curves for the proportions of subjects with a given number of IBD symptom flares, using the MLE approach and using the NBKE approach with optimal smoothing. As previously stated in Chapter 1, the most likely area for the difference between the two groups is in the upper tail. This difference, however, is less apparent in the graph with smoothed curves.

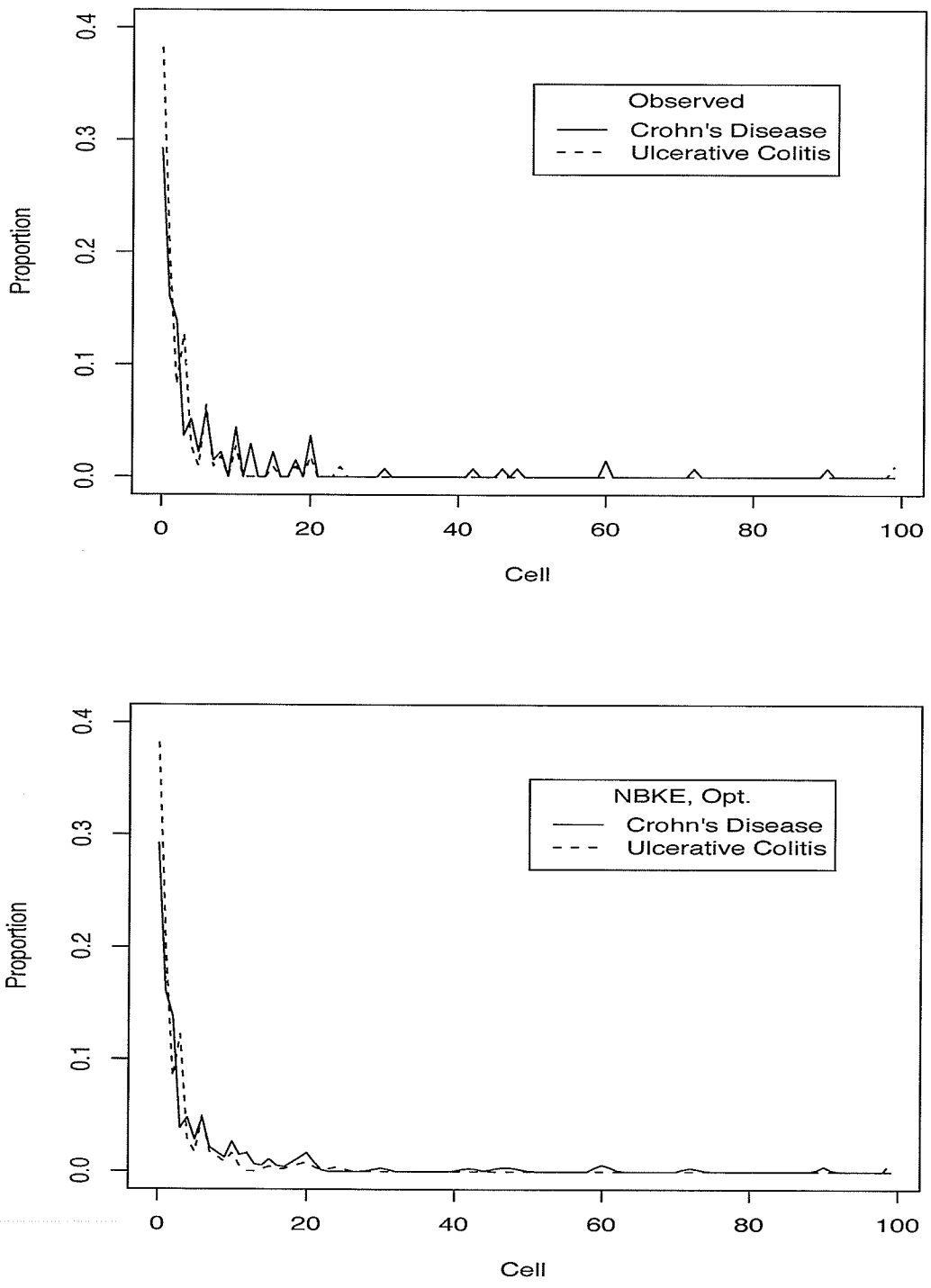

Figure 6.1: Fitted curves of proportions of symptom flares for the two IBD groups.

Specifically, we are interested in testing the following hypothesis,

$$H_0 : \vec{P}_{CD} = \vec{P}_{UC} \text{ vs. } H_A : \vec{P}_{CD} \neq \vec{P}_{UC} \tag{6.1}$$

where $\vec{P}_{CD}$ and $\vec{P}_{UC}$ are the vectors of cell probabilities for the Crohn's Disease and Ulcerative Colitis IBD groups, respectively.

Under the null hypothesis, both groups belong to the same population. However, due to the small ratio of the sample size to the number of cells, low cell frequencies, and the fact that MLE estimates are not recommended, a Chi-Square test of homogeneity is not appropriate if applied to the data in its current state. A common approach would be to group cells together so there are fewer categories. As first mentioned in Chapter 1, there are some potential pitfalls in doing this. We lose information and tests for detecting differences between the two groups become less powerful. Also, defining groups is quite arbitrary, so tests results are potentially highly dependent on the grouping method. For the sake of completeness, however, we will first try this method using three grouping configurations to show how much results can vary. We will later compare these results to what is obtained with our suggested methodology relying on a permutation test.

For the first Chi-Square test, we combined the 100 cells into 7 categories so that the expected count for each category is approximately 5 or more. Table 6.1 shows the 2×7 contingency table with 6 degrees of freedom. The value of the test statistic is 11.3578 and the p-value is 0.07793. Based on this configuration, there is no statistically significant difference between the two IBD groups at $\alpha = 5\%$.

For the second Chi-Square test, we combined the 100 cells into 6 categories so that there were at least 5 observations in each category. Table 6.2 shows the 2×6 contingency table with 5 degrees of freedom. The value of the test statistic

is 11.3449 and the p-value is 0.04495. Based on this configuration, there is a statistically significant difference between the two IBD groups at $\alpha = 5\%$. Of course, this does not necessarily mean there is a practical difference between the two IBD groups.

Most of the survey participants reported having fewer than 40 symptom flares within the last six months. The Ulcerative Colitis group actually has only 1 observation greater than 30 at cell 99. Now, suppose we focus on a subset with a low or moderately high number of symptom flares. Or in a different perspective, we treat the observations in cells beyond 40 as extreme (but still possible) cases. By doing this, we still capture over 96% of the original sample. Table 6.3 shows the 2×6 contingency table with 5 degrees of freedom for the third Chi-Square test. The value of the test statistic is then 8.2363 and the p-value is 0.14370. Based on this test, there is no statistically significant difference between the two IBD groups at $\alpha = 10\%$. This brings us back to our original hypothesis that the difference, if any, is most likely to occur in the upper region. It could be argued, however, that this isn't a practical difference as it affects less than 4% of the subjects.

We now consider testing (6.1) using the NBKE. We do this to avoid having to condense the data into fewer categories. For this case study, we used permutation tests, but bootstrapping could also be used. For the permutation tests, observations from both groups were combined together and tests for their differences were based on random permutations (re-arrangements) of the data. There is a total of 247!/(137!110!) possible re-arrangements, so that, due to time and computational constraints, only 1000 of these were randomly selected. For each re-arrangement, estimated probabilities were computed for each group using the NBKE with optimal smoothing.

These tests were performed using two different scenarios. The first scenario uses all the available data, while the second scenario focuses on participants reporting fewer than 40 IBD symptom flares which is similar to the grouping configuration used previously in the third Chi-Square test. (Note that were are we considering cells 0 to 39.)

Three empirical tests to detect differences in probabilities between the groups were considered: (1) sum of squared differences, (2) maximum absolute difference, and (3) sum of absolute differences. Specifically, for the first test, a sum of the squared difference between corresponding cell probabilities from each IBD group is calculated in each permutation. For the second test, a maximum absolute difference between corresponding cell probabilities from each IBD group is calculated in each permutation. For the third test, a sum of the absolute difference between corresponding cell probabilities from each IBD group is calculated in each permutation. The p-value is then the probability of having a test statistic from a permutation that is larger than the observed statistic from the original sample. Empirical 90%, and 95% confidence intervals (or rather upper-bounds since these are one-sided tests) were also computed. (They are empirical in the sense that they are not based on any specific parametric distribution.)

The following are the results of the 1000 permutations. Figures 6.2, 6.3, and 6.4 are histograms of the aforementioned three types of differences. In Table 6.4, where tests are performed using all 100 cells, there is only a statistically significant difference for the third test at the 5% level of significance. Tests using the absolute max difference are known to be not very powerful. From that standpoint, the much larger p-value was to be expected. The first test suggests that there might be a difference between the two IBD groups (since the p-value is very close to 0.05) and

this is using all the categories. Recall, however, that for the Chi-Square test, we had to collapse the data into 6 categories in order to be able to detect a difference.

Table 6.5 shows the results of the permutation tests using only 40 cells. At the 5% level of significance, there is no statistically significant difference for any of the permutation tests. These results agree with the Chi-square test based on 40 cells as well.

Hence, if we consider all 100 cells, there is some evidence of a statistical difference in distribution of symptom flares between the two IBD groups (either at 5% or 10% level of significance). Once we consider the smaller subset, however, there is weak evidence to suggest that there could be a statistically significant difference. This suggests that most of the statistical difference can be attributed to the extreme upper values which account for less than 4% of the total sample. Therefore, although there may be a statistical difference between the two IBD group, this does not necessarily translate to a practical difference, at least with respect to the number of symptom flares.

Table 6.1: 2×7 contingency table for Grouping Configuration 1.

Groups	0	1	2	3-5	6-10	11-20	21-99	Row Total
Crohn's Disease	40	22	19	15	19	14	8	137
Ulcerative Colitis	42	22	9	18	13	4	2	110
Column Total	82	44	28	33	32	18	10	247

Table 6.2: 2×6 contingency table for Grouping Configuration 2.

Groups	0	1	2	3-5	6-10	11-99	Row Total
Crohn's Disease	40	22	19	15	19	22	137
Ulcerative Colitis	42	22	9	18	13	6	110
Column Total	82	44	28	33	32	28	247

Table 6.3: 2×6 contingency table for Grouping Configuration 3.

Groups	0	1	2	3-5	6-10	11-40	Row Total
Crohn's Disease	40	22	19	15	19	15	130
Ulcerative Colitis	42	22	9	18	13	5	109
Column Total	82	44	28	33	32	20	239

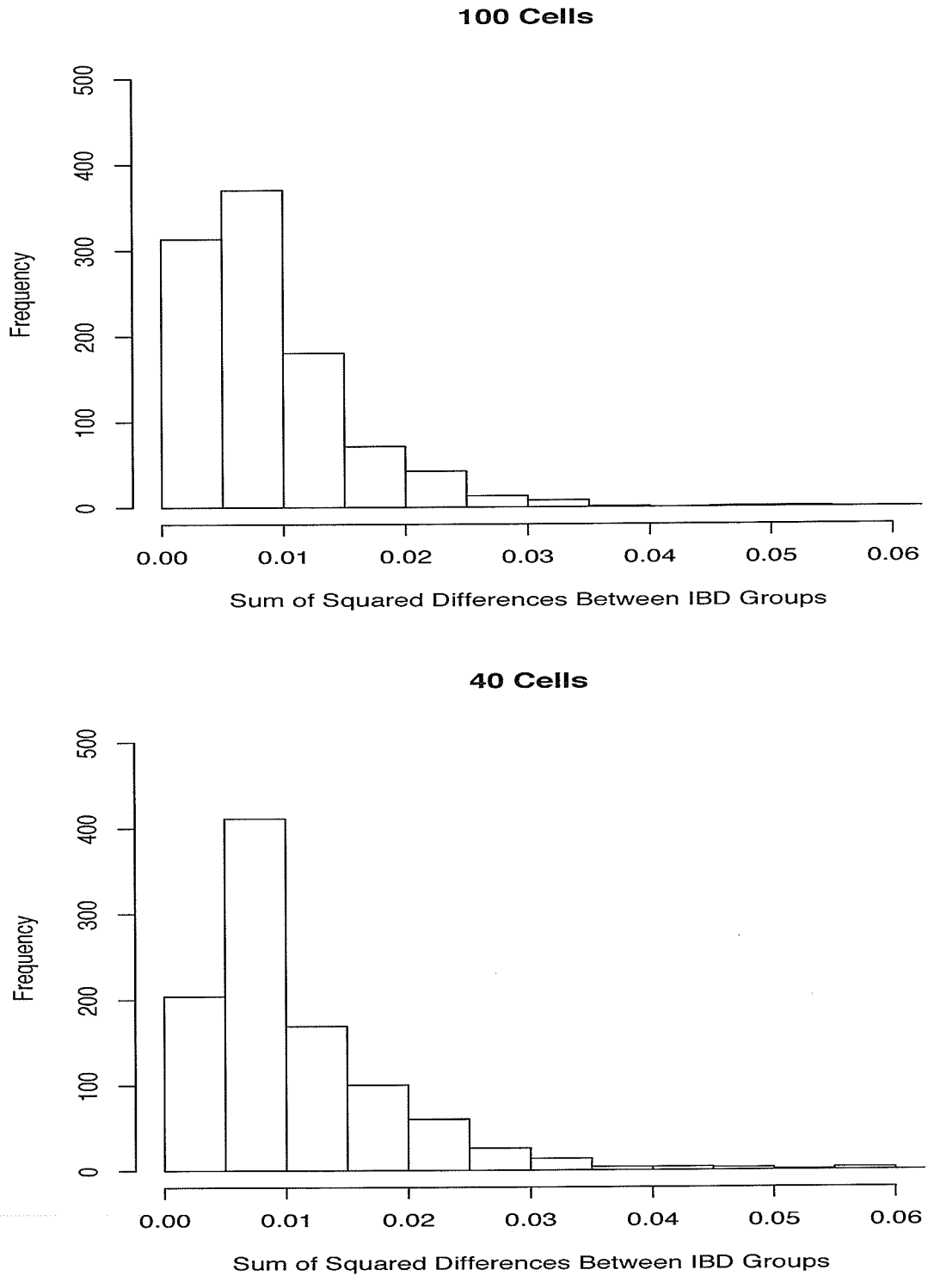

Figure 6.2: Histograms of sum of squared differences between IBD groups.

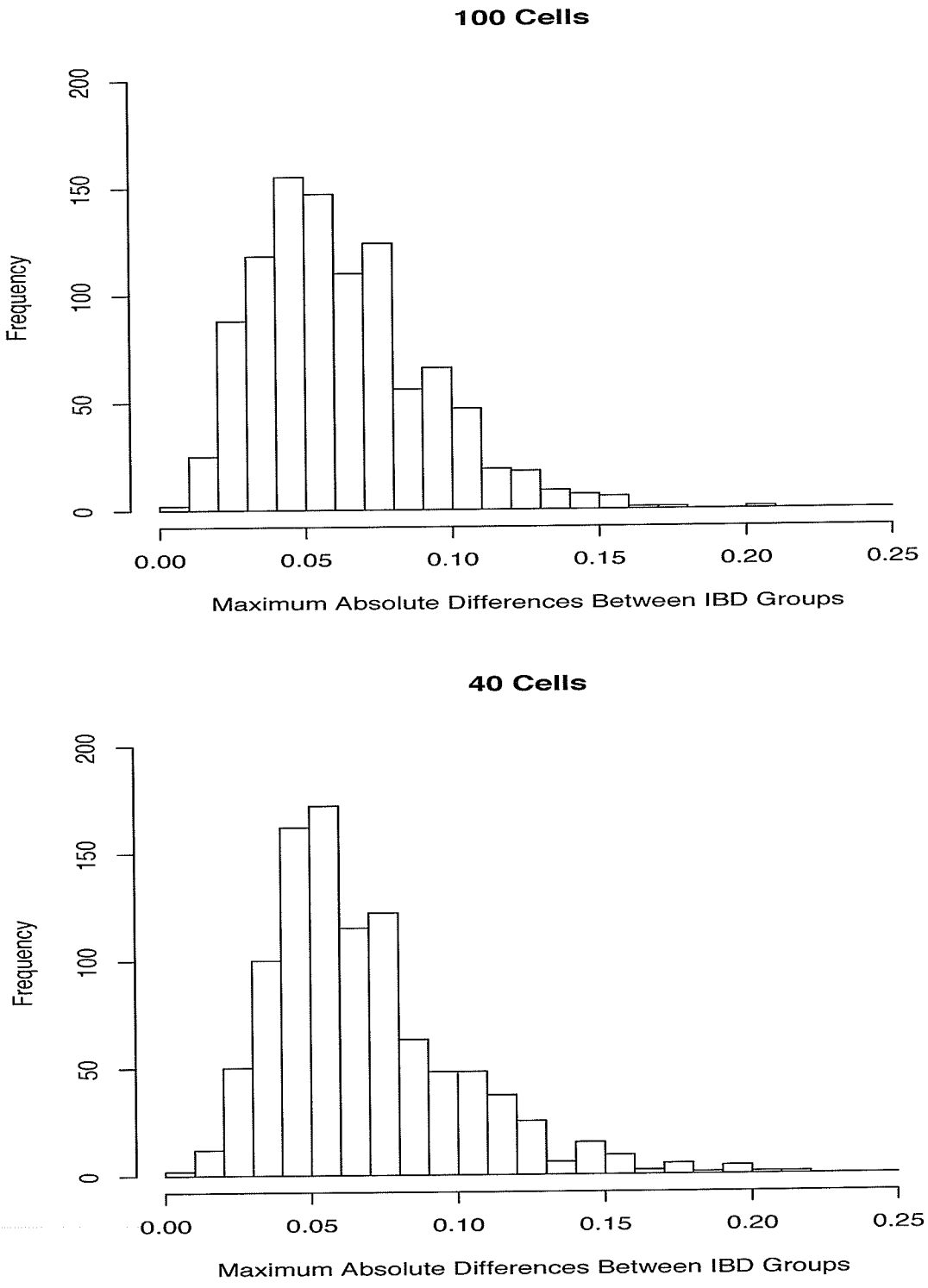

Figure 6.3: Histograms of maximum absolute difference between IBD groups.

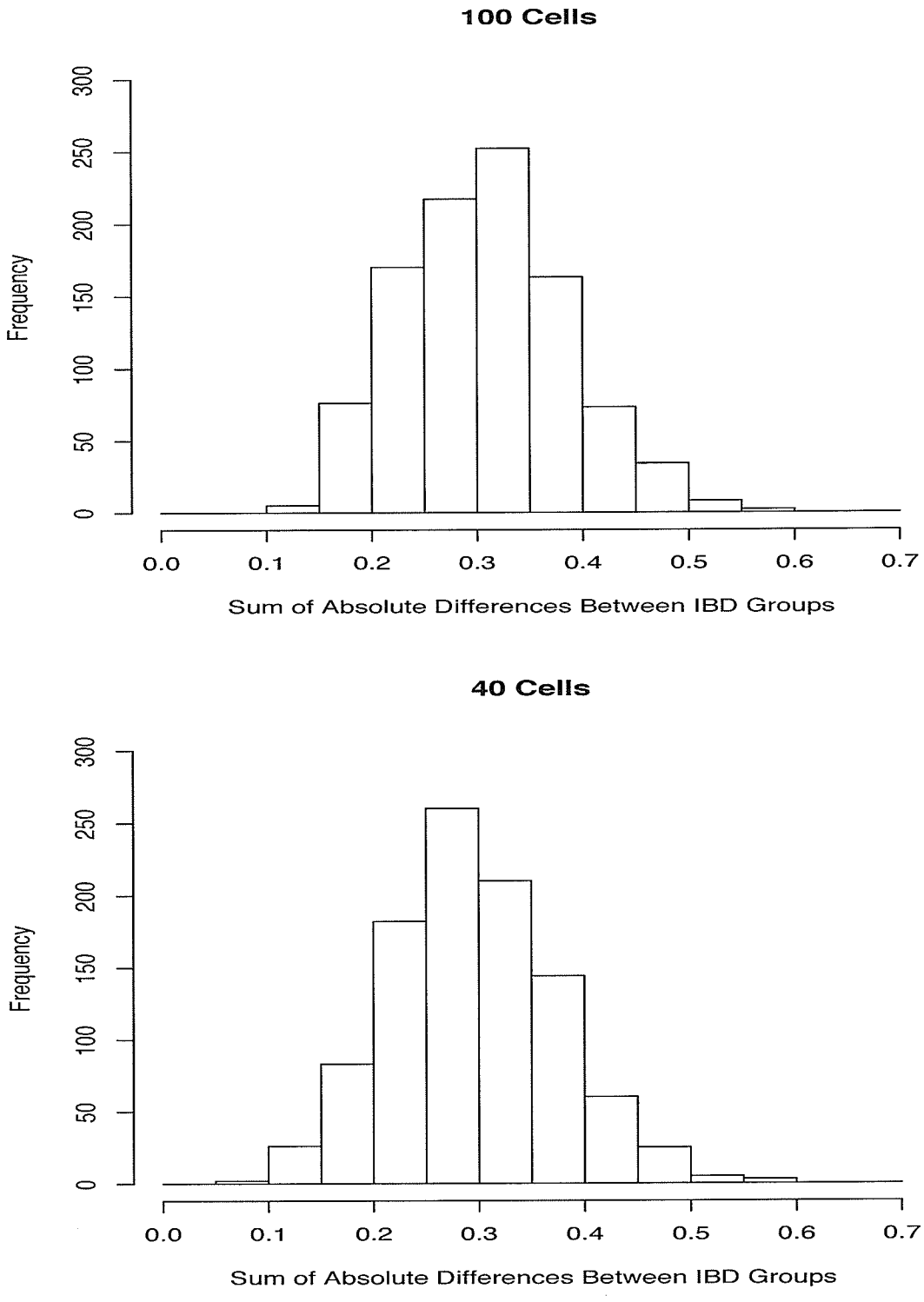

Figure 6.4: Histograms of sum of absolute differences between IBD groups.

Table 6.4: Permutation tests of difference based on 1000 random re-arrangements using 100 cells.

Difference	Observed	RT p-value	90% Quantile	95% Quantile
Squared Sum	0.02112431	0.052	0.01702809	0.02149444
Absolute Max	0.08984738	0.176	0.10203655	0.11728899
Absolute Sum	0.46407519	0.029	0.40889484	0.44231978

Table 6.5: Permutation tests of difference based on 1000 random re-arrangements using 40 cells.

Difference	Observed	RT p-value	90% Quantile	95% Quantile
Squared Sum	0.01950796	0.119	0.02127143	0.02560525
Absolute Max	0.08266466	0.253	0.11102016	0.12708662
Absolute Sum	0.40474644	0.084	0.39505111	0.42808815

Appendix A

Data Appendix

Although the main outcome of interest is the number of symptom flares within the past six months, two survey responses were considered to verify the reliability of the data. (1) The number of symptoms flares in the past six months. (2) The time to the last symptom flare. If the value for the number of flares is missing, and the last flare occurred 6 months or later, the number of flares is set to 0. Some respondents contradicted themselves, so these records were not kept:

1. Individuals with a missing value for the number of flares, but reported to have experienced the last symptom flare within 6 months.

2. Individuals who claimed to have experienced their last symptom flare more than six months before the survey, but reported a non-zero value for the number of flares experienced within six months.

3. Individuals who claimed to have experienced their last symptom flare within 6 months, but also reported of experiencing 0 symptom flares within six months.

Appendix B

Mathematical Lemmas

Lemma B.1. *The product of r consecutive numbers (each 1 unit apart) is less than the average of the first and last term to the power of r. In other words, for* $y > 0$

$$(y+1)(y+2)\cdots(y+r) < \left(y + \frac{1+r}{2}\right)^r.$$

Proof of Lemma B.1: A result in Lorch (1984) [8] states that for $r > 2$, and for $y = 0, 1, \ldots, 2$,

$$\frac{\Gamma(y+r)}{\Gamma(y+1)} < \left(y + \frac{r}{2}\right)^{r-1}.$$

We generalize Lorch's result for non-integer values of y, and derive the following, valid for $r > 1$:

$$\begin{aligned} \frac{\Gamma(y+r+1)}{\Gamma(y+1)} &< \left(y + \frac{r+1}{2}\right)^r \\ \text{or, } (y+1)(y+2)\cdots(y+r) &< \left(y + \frac{r+1}{2}\right)^r \end{aligned}$$

□

Lemma B.2. *$\frac{1}{1+x}$ and $\frac{1}{1-x}$ can be expressed as the following:*

$$\begin{aligned} \frac{1}{1+x} &= 1 - x + O(x^2), \text{ and,} \\ \frac{1}{1-x} &= 1 + x + O(x^2). \end{aligned}$$

where $|x| < 1$.

Proof of Lemma B.2:

We start by expanding the expression for $\frac{1}{1+x}$ into the following,

$$\frac{1}{1+x} = 1 - x + \frac{x^2}{1+x}.$$

Now, note that for $x > 0$,

$$\frac{x^2}{1+x} \leq x^2.$$

Thus, we have,

$$\frac{1}{1+x} = 1 - x + O(x^2).$$

Similarly, we can expand the expression $\frac{1}{1-x}$ as,

$$\frac{1}{1-x} = 1 + x + \frac{x^2}{1-x}.$$

Now, note that as long as $x \to 0$,

$$\frac{x^2}{1-x} = O(x^2).$$

Thus, we have,

$$\frac{1}{1-x} = 1 + x + O(x^2).$$

□

Bibliography

[1] Aerts, M., Augustyns, I. and Janssen, P. (1997), Smoothing Sparse Multinomial Data Using Local Polynomial Fitting, *Nonparametric Statistics*, Vol. 8, p. 127-147. (Cited on page 12.)

[2] Babu, G.J., Canty, A.J., Chaubey, Y.P. (2002), Application of Bernstein Polynomials for Smooth Estimation of a Distribution and Density Function, *Journal of Statistical Planning and Inference*, Vol. 105, p. 377-392. (Cited on page 52.)

[3] Burman, P.P. (1987), Smoothing Sparse Contingency Tables, *Sankhyā: The Indian Journal of Statistics*, Vol. 49, p. 24-36. (Cited on pages 14 and 16.)

[4] Chen, S.X. (1996), Beta Kernel Estimators for Density Functions, *Computational Statistics and Data Analysis*, Vol. 31, p. 131-145. (Cited on page 17.)

[5] Dong, J. and Simonoff, J.S. (1994), The Construction and Properties of Boundary Kernels for Smoothing Sparse Multinomials, *Journal of Computational and Graphical Statistics*, Vol. 3, p. 57-66. (Cited on pages 30 and 62.)

[6] Good, I.J. and Gaskins, R.A. (1971), Nonparametric Roughness Penalties for Probability Densitys, *Biometrika*, Vol. 58, p. 255-277. (Cited on page 7.)

[7] Leblanc, A. (2009), On Estimating Distribution Functions Using Bernstein Polynomials, *Submitted for publication.* (Cited on page 52.)

[8] Lorch, Lee (1984), Inequalities for Ultraspherical Polynomials and the Gamma Function, *Journal of Approximation Theory*, Vol. 40, p. 115-120. (Cited on page 95.)

[9] Park B.U. and Marron, J.S. (1990), Comparison of Data-Driven Bandwidth Selectors, *Journal of the American Statistical Association*, Vol. 85, p. 66-72. (Cited on page 55.)

[10] Rajagopalan, B. and Lall, U. (1995), A Kernel Estimator For Discrete Distributions, *Nonparametric Statistics*, Vol. 4, p. 409-426. (Cited on page 15.)

[11] Schucany, W.R. (2004), Kernel Smoothers: An Overview of Curve Estimators for the First Graduate Course in Nonparametric Statistics, *Statistical Science*, Vol. 19, p. 663-675. (Cited on page 12.)

[12] Silverman, B.W (1986), Density Estimation for Statistics and Data Analysis, Chapman & Hall/CRC, Boca Raton. (Cited on pages 6, 14 and 55.)

[13] Simonoff, J.S. (1983), A Penalty Function Approach to Smoothing Large Sparse Contingency Tables, *The Annals of Statistics*, Vol. 11, p. 208-218. (Cited on page 7.)

[14] Simonoff, J.S. (1995), Smoothing Categorical Data. *Journal of Statistical Planning and Inference*, Vol. 47, p. 41-69. (Cited on page 7.)

[15] Titterington, D.M. (1980), A Comparative Study of Kernel-Based Density Estimates for Category Data, *Technometrics*, Vol. 22, p. 259-268. (Cited on page 14.)

[16] Wang, M. and Ryzin J.V. (1981), A Class of Smooth Estimators for Discrete Distributions. *Biometrika*, Vol. 68, p. 301-309. (Cited on page 15.)

Index