Dissimilarity and Optimal Sampling in Urn Ensembles

by

Jerrad Hampton

B.S., University of Colorado, 2008

M.S., University of Colorado, 2010

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Applied Mathematics

2012

This thesis entitled:
Dissimilarity and Optimal Sampling in Urn Ensembles
written by Jerrad Hampton
has been approved for the Department of Applied Mathematics

_____

Prof. Manuel E. Lladser

_____

Prof. Rob Knight

_____

Prof. Jem Corcoran

_____

Prof. Vanja Dukic

_____

Prof. Juan Restrepo

Date _____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Hampton, Jerrad (Ph.D., Applied Mathematics)

Dissimilarity and Optimal Sampling in Urn Ensembles

Thesis directed by Prof. Manuel E. Lladser

We study an ensemble of urns with unknown compositions inferred from initial samples with replacement from each urn. This model fits diverse situations. For instance, in microbial ecology studies each urn represents an environment, each ball within an urn corresponds to an individual bacterium, and a ball's color represents its taxonomic label. In a different context, each urn could represent a random RNA pool and each colored ball a possible solution to a particular binding site problem over that pool.

The main parameter of this study is dissimilarity, which we define as the probability that a draw from one urn is not seen in a sample of size k from a possibly different urn. We estimate this parameter with a U-statistic, shown to be the uniformly minimum variance unbiased estimator (UMVUE) of dissimilarity over a range for k determined by initial sample sizes. Furthermore, despite the non-Markovian nature of our estimator when applied sequentially over k, we provide conditions that guarantee uniformly consistent estimates of variances via a jackknife method, and show uniform convergence in probability as well as approximately normal marginal distributions.

We apply our U-statistics and a restricted exponential regression to extrapolate dissimilarity over a range beyond that determined by initial sample sizes, which we use to identify an allocation of draws for subsequent sampling that minimizes a measure of pair-wise dissimilarities over the whole ensemble. This is motivated by the challenge faced by microbiome projects worldwide to effectively allocate additional samples for a more robust and reliable estimation of UniFrac distances between pairs of environments. Similar methods are applied to measures of sample quality of the ensemble derived from α-diversity and coverage. We test our methods against simulated data, where we compare optimal and inferred draw allocations when considering these three measures, and analyze 16S ribosomal RNA data from the Human Microbiome Project.

Dedication

To $\emptyset$, and Ω.

Acknowledgements

I am grateful to Manuel Lladser for quality mentoring and the large amount of time invested in my development, and for his efforts in teaching a number of classes critical to my skill development. I thank Rob Knight and his lab for exposing me to state of the art science to which I could contribute. Jem Corcoran has been invaluable in the many classes she has taught, and has helped in grasping many concepts. I would like to thank the members of my committee, including the above, as well as Vanja Dukic, and Juan Restrepo for their interest in my work. Antonio Gonzalez assisted in processing the 16S rRNA data into the count form applicable for the methods presented here. Susan Pryor has been invaluable for insuring that graduate requirements are satisfied in a complete and timely fashion. My parents Archie and Rachel Hampton have always supported and encouraged my mathematical interests. This work has been funded by NSF DMS grant #0805950; NIH grant HG4872; a University of Colorado Graduate School Fellowship, and the Crohns and Colitis Foundation of America. Finally I would like to thank all my teachers, friends and family for their support.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

1.1 Urn Model

Consider $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ to be independent sequences of independent and identically distributed (i.i.d.) discrete random variables with probability mass functions $\mathbb{P}_x$ and $\mathbb{P}_y$, respectively. Without loss of generality we assume that $\mathbb{P}_x$ and $\mathbb{P}_y$ are supported over a subset of $\mathbb{N} = \{1, 2, 3, \ldots\}$, and think of outcomes from these distributions as "colors": i.e. we speak of color-1, color-2, etc.

The probability mass of color-i under a distribution $\mathbb{P}_z$ over $\mathbb{N}$ is denoted as $\mathbb{P}_z(i)$. Furthermore, we view any such $\mathbb{P}_z$ as the composition of a certain urn-z composed of colored balls, from which we sample with replacement, and let I_z denote the support of $\mathbb{P}_z$. In accordance with the above notation, Z_k denotes the color of the k-th ball drawn with replacement from urn-z; in particular, distinct draws are always independent.

1.2 Motivation

This model is broadly applicable to any study of two unknown discrete distributions. For example, each urn could represent a random RNA pool and each colored ball a possible solution to a particular binding site problem over that pool [62]. On the other hand, we may envision a game theoretic study where each urn represents a player and each colored ball a particular strategy to be deployed [47]. As another example, a marketing study could set each urn as a demographic, each ball as an individual, and the color of a ball may represent consumer preferences or marketing

exposure [37]. In a different study, the urn could represent currency in circulation during antiquity, each ball a particular coin, and each color a die variety [19].

Though applicable to the above situations, the practical applications of this work have focused on microbial ecology studies where each urn represents an environment, each ball within an urn corresponds to an individual bacterium, and a color represents a taxonomic label. Specifically, we seek to measure sample quality and allocate sampling resources in a way that leads to robust, reliable estimation of the UniFrac distance between environments [41, 42], particularly using data collected for the Human Microbiome Project [51, 50, 63, 15, 16] and Earth Microbiome Project [22].

1.3 Parameters of Interest

Various parameters of the distributions in the urn model are of use in determining sample quality. We study two parameters of a single urn. First, $\psi_z(k)$ tracks the average proportion of urn-z which is unobserved in a sample of size k from urn-z, which assists in quantifying the proportion of an urn likely to be exposed by additional samples. The other parameter of one urn, $\phi_z(k)$, tracks the average number of observed colors in a sample of size k from urn-z. This assists in quantifying the number of colors one might see in additional samples from that urn. Since urn-z is usually fixed or understood in context, we suppress the subscript z in $\psi_z(k)$ and $\phi_z(k)$ and usually write $\psi(k)$ and $\phi(k)$ respectively.

In contrast with the above parameters, we extensively study $\theta_{x,y}(k)$, which tracks the average proportion of urn-x which is unobserved in a sample of size k from urn-y. This parameter assists in quantifying the proportion of urn-x which we expect to observe in additional samples from urn-y. Since urn-x and urn-y are usually fixed or understood in context, we often suppress the subscript in $\theta_{x,y}(k)$ and write $\theta(k)$.

We estimate each of $\psi(k), \phi(k)$, and $\theta(k)$ with a U-statistic [13, 26, 29, 24, 55], giving the UMVUE in each case. We address the consistency of these estimates, as well as their asymptotic distribution. Further, we estimate the variance of each U-statistic by a jackknife approach [2, 18, 57, 58], and show these estimates to be asymptotically consistent. We discuss these methods in

detail for the estimation of $\phi(k)$ and $\psi(k)$ in Chapter 2, and for estimation of $\theta(k)$ in Chapter 3. We expand upon results in the literature by viewing our statistics not as point estimators for fixed k, but as estimators over a range of k, growing with the size of samples used for estimation. This introduces non-Markovian dependencies that we address with a projection method [25] in Sections 2.4 and 3.4.

1.3.1 $\psi(k)$: An Unobserved Probability Parameter

We may consider the probability mass of the colors unobserved in a sample from urn-z as a measure of the quality of the sample from that urn. A quality sample will leave unobserved only colors which represent a small proportion of the urn. Using either a realized or an average sample, we can measure the difference between full sampling and partial sampling by coverage probability, the total probability mass which the sample represents in urn-z, or unobserved probability, the probability mass of colors unobserved in the sample [13, 19, 23, 38, 40, 45, 52, 61]. For this study we use an average sample and let

$$\psi(k) := \mathbb{P}(Z_{k+1} \notin \{Z_1, \ldots, Z_k\}) = \sum_{i \in I_z} \mathbb{P}_z(i)(1 - \mathbb{P}_z(i))^k. \tag{1.1}$$

That is $\psi(k)$ is the average unobserved probability relative to k-samples. The UMVUE for $\psi(k)$ [13, 61] has been established in the literature, which we use as our estimator for $\psi(k)$. We review results and present new results about this estimator in Chapter 2.

1.3.2 $\phi(k)$: An α-Diversity Parameter

In a different sense, we may consider urn-z to be well sampled when the unobserved colors after a sample are few in number. Let $|\cdot|$ denote set cardinality, and define

$$\phi(k) := \mathbb{E}\left(|\{Z_1, \ldots, Z_k\}|\right) = |I_z| - \sum_{i \in I_z} (1 - \mathbb{P}_z(i))^k; \tag{1.2}$$

$$\phi(\infty) := \lim_{k \to \infty} \phi(k) = |I_z|. \tag{1.3}$$

That is $\phi(\infty)$ is the number of colors present in urn-z, a quantity which in ecological literature is referred to as α-diversity [65]. This quantity is difficult to estimate accurately as an arbitrarily large set of colors may exist in an arbitrarily small probability mass of $\mathbb{P}_z$. In contrast, $\phi(k)$, the average number of colors seen in k-samples, which we refer to as the average α-diversity of the urn relative to k-samples, is better posed for estimation, for example by $|\{Z_1, \ldots, Z_k\}|$. While these quantities and their estimation have received attention under diverse methods in varied contexts [5, 12, 10, 11, 31, 46], we focus on the uniformly minimum variance unbiased estimator (UMVUE) of $\phi(k)$ as studied in Chapter 2.

1.3.3 $\theta(k)$: A Dissimilarity Probability Parameter

Note that ψ and ϕ are functions in k of one urn, while comparison of urns depends on accuracy in identifying colors which are unique in urn-x, urn-y, or are common to both. For example in the context of microbial ecology studies, the unweighted UniFrac [41, 42] distance between environments is calculated from the phylogenetic tree with leaf nodes given by the taxonomic labels of bacteria present in either environment. Internal nodes are given taxonomic labels which are inferred from the leaf nodes. The distance between a node and its ancestor is given by a branch length, derived from the evolutionary distance between the two taxonomic labels. We identify a branch length as unique if the child node of that branch or any of its descendants are present in only one sample, and identify a branch length as shared if the child node of that branch or any of its descendants may be found in both samples. The UniFrac distance is the ratio of the sum of unique branch lengths over the sum of branch lengths, unique and shared, in the tree.

In estimation of the UniFrac distance, the labels seen in a sample are assumed to represent the environment. As a result, identification in a sample of new taxonomic labels unique to that environment tends to increase the estimated UniFrac distance, while identifying labels present in both environments tends to decrease this distance. As the addition of leaf nodes changes the structure of the phylogenetic tree, particularly the branch lengths and interior nodes, the addition of leaf nodes produces a complicated effect on the distance estimate.

Hence, in the language of urns, an accurate estimation of distance between urns depends on an accurate understanding of shared and distinct colors. Although $\psi(k)$ and $\phi(k)$ are useful for understanding I_x or I_y, information about $I_x \cap I_y$ is not captured by $\psi(k)$ or $\phi(k)$. To contrast these parameters of individual urns is the main parameter of interest in this thesis, the proportion of balls in urn-x with a color that is absent relative to k-samples from urn-y. While this is a function of two urns, it is not symmetric if the roles of the urns are interchanged. Specifically, let

$$\theta(k) := \mathbb{P}\left(X_1 \notin \{Y_1, \ldots, Y_k\}\right) = \sum_{i \in I_x} \mathbb{P}_x(i)(1 - \mathbb{P}_y(i))^k; \tag{1.4}$$

$$\theta(\infty) := \lim_{k \to \infty} \theta(k) = \sum_{i:\, \mathbb{P}_y(i)=0} \mathbb{P}_x(i). \tag{1.5}$$

The parameter $\theta(\infty)$ measures the probability mass of colors in urn-x which are unique from urn-y. On the other hand, $\theta(k)$ is a measure of the effectiveness of k-samples from urn-y to determine uniqueness in urn-x. This motivates us to refer to $\theta(\infty)$ as the dissimilarity of urn-x from urn-y, and $\theta(k)$ as the average dissimilarity of urn-x relative to k-draws from urn-y. Related to these quantities, $1 - \theta(\infty)$ measures the probability mass of colors in urn-x which are represented in urn-y, while $1 - \theta(k)$ is a measure of the effectiveness of k-samples from urn-y to determine this overlap with urn-x. Estimating $\theta(\infty)$ is a difficult task as arbitrarily small perturbations of urn-y can potentially lead to large changes in $\theta(\infty)$. To demonstrate, let $\mathbb{P}_x(1) = \mathbb{P}_x(2) = 0.5$, and $\mathbb{P}_y(1) = \epsilon, \mathbb{P}_y(2) = 1 - \epsilon$. If $\epsilon = 0$, then $\theta(\infty) = 0.5$, while for $\epsilon > 0$, $\theta(\infty) = 0$. In contrast with this discontinuity, for fixed k, $\theta(k)$ depends continuously on $(\mathbb{P}_x, \mathbb{P}_y)$ e.g. under the metric $d\big((\mathbb{P}_x, \mathbb{P}_y), (\mathbb{P}'_x, \mathbb{P}'_y)\big) = \|\mathbb{P}_x - \mathbb{P}'_x\| + \|\mathbb{P}_y - \mathbb{P}'_y\|$, where $\|\nu\| = \frac{1}{2}\sum_i |\nu(i)|$ denotes the total variation distance of the signed measure ν. Motivated by this continuous dependence on urn-x and urn-y, we consider the unbiased estimation of $\theta(k)$ in Chapter 3.

1.4 The Urn Ensemble

Consider now an ensemble of r urns, and let $X_j(i)$ denote the ith draw with replacement from urn-j, with X_j a generic draw from urn-j. Suppose that we have a sample of n_j draws from each urn-j from which we infer information about our ensemble. We consider the parameters given

by

$$\phi_j(k) := \mathbb{E}(|\{X_j(1), \cdots, X_j(k)\}|); \tag{1.6}$$

$$\psi_j(k) := \mathbb{P}(X_j \notin \{X_j(1), \cdots, X_j(k)\}); \tag{1.7}$$

$$\theta_j(k) := \frac{1}{r-1} \sum_{i=1, i\neq j}^{r} \theta_{i,j}(k), \tag{1.8}$$

where

$$\theta_{i,j}(k) := \mathbb{P}(X_i \notin \{X_j(1), \cdots, X_j(k)\}).$$

Note that as a function of k, $\phi_j(k)$ increases to $\phi(\infty)$ whereas $\psi_j(k)$ and $\theta_j(k)$ decrease to their limiting values. Note also that ϕ_j and ψ_j depend only on urn-j. In contrast, the parameter $\theta_j(k)$ averages the $\theta_{i,j}(k)$ over each urn-i. Equivalently, $\theta_j(k)$ is the expected value of the dissimilarity of a randomly chosen urn-i in the ensemble, with $i \neq j$, relative to k draws from urn-j. The parameter function θ_j therefore depends on the entire ensemble of urns. Let λ_j be one of the parameters defined in (1.6), (1.7), or (1.8). We estimate $\lambda_j(k)$ with the appropriate U-statistic $\hat{\lambda}_j(k)$ for $j = 1 : r$, over a range of k between 1 and n_j or $n_j - 1$.

Notice the similarity in form of ϕ_j, ψ_j, and θ_j as a function of k implied by (1.1), (1.2), and (1.4). In particular as a function of k each is a linear combination of terms of the form $(1-p)^k$. Using restrictions derived from (1.1), (1.2), and (1.4), we form a regression function $\hat{\lambda}_j^R(k)$ from the estimates, $\hat{\lambda}_j(k)$, using methods discussed in Section 4.1. By extrapolation along these regression functions we obtain estimates for $\lambda_j(k)$ beyond the original constraints on k. In Section 4.2, we present a method which uses these $\hat{\lambda}_j^R(k)$ to assign a subsequent sampling effort of m draws, with m_j draws to urn-j, chosen to optimize a function of the $\hat{\lambda}_j^R(n_j + m_j)$, S, which is monotonic in the m_j, such that $|\partial S/\partial m_j|$ is a decreasing function of m_j, and for $i \neq j$

$$\frac{\partial^2 S}{\partial m_i \partial m_j} = 0.$$

These methods for sample allocation are discussed in detail in Section 4.3.

Chapter 2

U-Statistics for One Urn Paramters

In this chapter we review and present new results involving U-statistics for estimation of $\psi(k)$ and $\phi(k)$ as given by (1.1) and (1.2), respectively; in particular the methods of this chapter apply to one-urn problems. The main contributions of this thesis are presented in Sections 2.4 and 2.5.

As there is a single urn we drop references to its label. Further, since $\phi(1) = 1$ independently of the urn's distribution, in the remainder of this chapter we only consider estimation of $\phi(k)$ for $k \geq 2$, and it is implied that the theorems that follow apply to $\psi(k)$ for $k \geq 1$ and $\phi(k)$ for $k \geq 2$.

Consider $Z_1, \ldots, Z_n$ to be a sample of i.i.d. discrete random variables with probability mass function $\mathbb{P}_z$, supported on I_z. From this sample we estimate each of $\psi(k) := \mathbb{P}(Z_{k+1} \notin \{Z_1, \ldots, Z_k\})$ and $\phi(k) := \mathbb{E}|\{Z_1, \ldots, Z_k\}|$ by an optimal unbiased estimator; in particular, an estimator which has the least variance of all estimators which are unbiased for any $\mathbb{P}_z \in \mathcal{D}$, where $\mathcal{D}$ is defined to be the set of discrete distributions over a finite subset of $\mathbb{N}$.

2.1 U-Statistic Definition

U-statistics are formed by a symmetrization performed on a simpler statistic referred to as a kernel statistic [26, 29]. We refer to our general kernel statistic by $h(Z_1, \ldots, Z_k)$. For our purposes these kernel statistics are allowed to vary over k. Let $[\![\cdot]\!]$ refer to the Iverson Bracket notation for the indicator function, returning 1 if the event in the bracket is true, and 0 otherwise. Our kernel

statistics are given by

$$h_\psi(Z_1, \ldots, Z_{k+1}) := [\![Z_{k+1} \notin \{Z_1, \ldots, Z_k\}]\!]; \tag{2.1}$$

$$h_\phi(Z_1, \ldots, Z_k) := |\{Z_1, \ldots, Z_k\}|. \tag{2.2}$$

The kernel statistic in (2.1) has been previously established as the kernel for the UMVUE for $\psi(k)$ [13]. Note that

$$\mathbb{E}(h_\psi(Z_1, \ldots, Z_{k+1})) = \psi(k); \qquad \mathbb{E}(h_\phi(Z_1, \ldots, Z_k)) = \phi(k).$$

The U-statistic is defined as a symmetric estimator formed by averaging this kernel statistic over all sub-samples of size k [26, 29]. Intuitively, this symmetry reduces variance in the estimator, which we verify mathematically in Section 2.3.

Let $S_{k,n}$ be the set of one-to-one functions from $\{1, \ldots, k\}$ into $\{1, \ldots, n\}$. Note that $|S_{k,n}| = \binom{n}{k}k!$. The U-statistics based on (2.1) and (2.2) are given by

$$\hat{\psi}(k) := \frac{1}{|S_{k+1,n}|} \sum_{\sigma \in S_{k+1,n}} [\![Z_{\sigma(k+1)} \notin \{Z_{\sigma(1)}, \ldots, Z_{\sigma(k)}\}]\!]; \tag{2.3}$$

$$\hat{\phi}(k) := \frac{1}{|S_{k,n}|} \sum_{\sigma \in S_{k,n}} |\{Z_{\sigma(1)}, \ldots, Z_{\sigma(k)}\}|. \tag{2.4}$$

The quality of each of these statistics that defines them as U-statistics is the symmetry with relation to the data. In particular, reordering data does not change the value returned by the statistic [26, 29]. This form of the statistic is convenient for mathematical analysis [24, 55], but difficult for practical computation of the statistic.

2.2 Computationally Convenient Forms of U-Statistics

2.2.1 Summary Statistics

Forms convenient for the computation of the statistics given in (2.3) and (2.4) require effectively summarizing the data. Note that the quantities we estimate do not depend on specifically observed colors, but do depend on the relative proportions of colors in the urn. This suggests that

our summary of the data should not depend on the labels of colors in the sample, but instead on how often each color has been observed. Motivated by this observation, let for each $j \in \mathbb{N}$,

$$R(j) := \text{number of indices } i = 1 : n \text{ such that color} \\ Z_i \text{ occurs exactly } j\text{-times in } Z_1, \ldots, Z_n. \tag{2.5}$$

Stated in other ways, $R(j)$ is the number of draws with a color seen j times in the sample, or $R(j)$ is j times the number of colors seen j times in the sample. Note that $\sum_j R(j) = n$.

2.2.2 Estimation of $\psi(k)$

Here we write (2.3) in a computationally convenient form using the $R(j)$ through the use of combinatoric identities. The following form is equivalent to the form first investigated by Starr [61], which was equated to the U-statistic form by Clayton and Frees [13].

Lemma 2.1. *[13] The U-statistic estimate for $\psi(k)$ in (2.3) may be written as*

$$\hat{\psi}(k) = \sum_{j=1}^{n-k} \frac{\binom{n-j}{k} R(j)}{\binom{n-1}{k} n}. \tag{2.6}$$

Proof. We count the number of $\sigma \in S_{k+1,n}$, such that $Z_{\sigma(k+1)}$ is not seen in $Z_{\sigma(1)}, \ldots, Z_{\sigma(k)}$. Suppose that $Z_{\sigma(k+1)}$ is a draw that contributes to $R(j)$. Of the $j-1$ other draws of the same color, the proportion of indicators in (2.3) that has none of these draws present in $Z_{\sigma(1)}, \ldots, Z_{\sigma(k)}$ is $\binom{n-j}{k} / \binom{n-1}{k}$. The proportion of indicators which place a draw contributing to $R(j)$ in the $\sigma(k+1)$ position is $R(j)/n$. It follows from (2.3) that

$$\hat{\psi}(k) = \sum_{j=1}^{n} \frac{\binom{n-j}{k} R(j)}{\binom{n-1}{k} n}.$$

Note that $\binom{n-j}{k}$ is zero when $j > n-k$. This corresponds to a color seen more than $(n-k)$-times being present in each subsample of size k. The form in (2.6) follows. □

2.2.3 Estimation of $\phi(k)$

Here we write (2.4) in a form convenient for computation.

Lemma 2.2. *The U-statistic estimate for $\phi(k)$ in (2.4) may be written as*

$$\hat{\phi}(k) = \sum_{j=1}^{n-k} \frac{R(j)}{j} \sum_{i=1}^{j} \frac{\binom{n-j}{k-i}\binom{j}{i}}{\binom{n}{k}} + \sum_{j=n-k+1}^{n} \frac{R(j)}{j}. \tag{2.7}$$

Proof. We may rewrite (2.4) such that

$$\hat{\phi}(k) = \frac{1}{|S_{k,n}|} \sum_{c \in \{Z_1,\ldots,Z_n\}} \sum_{\sigma \in S_{k,n}} [\![c \in \{Z_{\sigma(1)}, \ldots, Z_{\sigma(k)}\}]\!].$$

Note that $R(j)/j$ is the number of colors seen j times in the sample. A color c seen more than $(n-k)$-times in the sample $Z_1, \ldots, Z_n$ must be seen in each sub-sample of size k, which explains the last term in (2.7). On the other hand, if a certain color c is seen j-times in the sample with $j \leq (n-k)$, then there are $k!\binom{j}{i}\binom{n-j}{k-i}$ functions σ such that c occurs exactly i-times in the sub-sample $Z_{\sigma(1)}, \ldots, Z_{\sigma(k)}$. Summing over i gives the number of σ such that c is present in the sub-sample as

$$k! \sum_{i=1}^{j} \binom{j}{i}\binom{n-j}{k-i},$$

which explains the first-term on the right-hand side of (2.7). This demonstrates the equivalence of (2.4) and (2.7). □

2.3 Uniformly Minimum Variance Unbiased Estimation

We now show that these U-statistics are UMVUE in our setting, reproducing the arguments of Halmos, Clayton and Frees [26, 13]. Recall that $\mathcal{D}$ denotes the set of all probability distributions that are finitely supported over $\mathbb{N}$. Note that our kernel statistics (2.1) and (2.2) are unbiased statistics for $\psi(k)$ and $\phi(k)$ respectively. Our next result asserts that no estimator may be unbiased uniformly in $\mathcal{D}$ when using less data. Using the terminology of Halmos, Clayton and Frees [26, 13], we say that $\psi(k)$ is $(k+1)$-homogeneous, while $\phi(k)$ is k-homogeneous. Lemmas 2.3 through 2.7 and Theorem 2.8 follow the method of proof by Halmos [26].

Lemma 2.3. *Let $m \geq 1$. If $g(Z_1, \ldots, Z_m)$ is unbiased for $\psi(k)$ for all $\mathbb{P}_x \in \mathcal{D}$, then $m \geq k+1$. If $g(Z_1, \ldots, Z_m)$ is unbiased for $\phi(k)$, then $m \geq k$.*

Proof. This result for $\psi(k)$ has been established [13] but is presented here for completeness. Consider in $\mathcal{D}$ probability distributions of the form $\mathbb{P}_x(1) = 2u/3$, $\mathbb{P}_x(2) = u/3$, and $\mathbb{P}_x(3) = (1-u)$, where $0 \leq u \leq 1$ is an arbitrary real number. Clearly, $\mathbb{E}[g(Z_1, \ldots, Z_m)]$ is a linear combination of polynomials of degree m in u and as a result, it is a polynomial of degree at most m in u. As $\psi(k) = 2u/3(1-2u/3)^k + u/3(1-u/3)^k + (1-u)u^k$ (see (1.1)) has degree $k+1$ in u, if $g(Z_1, \ldots, Z_m)$ is unbiased for $\psi(k)$, then we conclude that $k+1 \leq m$. Furthermore, $\phi(k) = 3 - (1-2u/3)^k - (1-u/3)^k - u^k$ (see (1.2)) is a polynomial of degree k for $k > 1$. We conclude that if $g(Z_1, \ldots, Z_m)$ is unbiased for $\phi(k)$, then $k \leq m$. □

We now define the appropriate notion of symmetry to show that the U-statistics given in (2.3) and (2.4) are UMVUEs. In what follows, we say that a function $f : \mathbb{N}^n \to \mathbb{R}$ is **n-symmetric** when

$$f(z_1, \ldots, z_n) = f(z_{\sigma(1)}, \ldots, z_{\sigma(n)}),$$

for all $z_1, \ldots, z_n \in \mathbb{N}$ and permutation σ of $1, \ldots, n$. Alternatively, f is n-symmetric if and only if it may be viewed as a function of $z_{(1\ldots n)}$, the order statistics $z_{(1)}, \ldots, z_{(n)}$.

A statistic of $(Z_1, \ldots, Z_n)$ is called n-symmetric when it may be represented in the form $f(Z_1, \ldots, Z_n)$ for some n-symmetric function f. Note that it is immediate from (2.3) and (2.4) that $\hat{\psi}(k)$ and $\hat{\phi}(k)$ are n-symmetric. The next result asserts that the variance of any asymmetric unbiased estimator of $\psi(k)$ or $\phi(k)$ may be reduced by a corresponding symmetric unbiased estimator. The proof is based on the well-known fact that conditioning does not increase the variance of a statistic while preserving its mean.

Lemma 2.4. *An asymmetric unbiased estimator of $\psi(k)$ or $\phi(k)$ has a larger variance than a corresponding n-symmetric unbiased estimator.*

Proof. This result for $\psi(k)$ has been established [13] but is presented here for completeness. Let $\mathcal{F}$ denote the sigma-field generated by the random vector $Z_{(1\ldots n)}$ and suppose that the statistic $T = f(Z_1, \ldots, Z_n)$ is unbiased for $\psi(k)$. As $\hat{\psi}(k)$ is n-symmetric and bounded by 1, we may assume without loss of generality that $\mathbb{E}[T^2] < +\infty$. In particular, $U = \mathbb{E}[T \,|\, \mathcal{F}]$ is a well-defined statistic

and there is an n-symmetric function $g : \mathbb{N}^n \to \mathbb{R}$ such that $U = g(Z_1, \ldots, Z_n)$. Clearly, U is unbiased for $\psi(k)$ and n-symmetric. The result is now a direct consequence of Jensen's inequality for conditional expectations [32], which in our context asserts that $\mathbb{E}(U^2) \leq \mathbb{E}(T^2)$ with equality if and only if T is n-symmetric. The same argument holds for $\phi(k)$, noting that $\hat{\phi}(k)$ is n-symmetric and bounded by k. □

The above lemma implies that if an UMVUE for $\psi(k)$ or $\phi(k)$ exists then it must be n-symmetric. Next, we show that there is a unique symmetric and unbiased estimator for each of $\psi(k)$ and $\phi(k)$ which immediately implies that $\hat{\psi}(k)$ and $\hat{\phi}(k)$ are the unique UMVUEs.

In what follows, we say that a polynomial $Q(u_1, \ldots, u_m)$ is **k-homogeneous** when it is a linear combination of polynomials of the form $\prod_{i=1}^m u_i^{b_i}$ with $\sum_{i=1}^m b_i = k$.

Further, we say that Q satisfies the **partial vanishing condition** if $Q(u_1, \ldots, u_m) = 0$ whenever $u_1, \ldots, u_m \geq 0$, and $\sum_{i=1}^m u_i = 1$.

The first lemma is an intermediate step to show that a k-homogeneous polynomial which satisfies the partial vanishing condition is the zero polynomial.

Lemma 2.5. *[26] If Q is a k-homogeneous polynomial in the real variables $u_1, \ldots, u_m$ with $m \geq 1$ that satisfies the partial vanishing condition, then $Q(u_1, \ldots, u_m) = 0$ whenever $u_1, \ldots, u_m \geq 0$ and $\sum_{i=1}^m u_i > 0$.*

Proof. Fix $u_1, \ldots, u_m \geq 0$ such that $\sum_{i=1}^m u_i > 0$ and observe that

$$Q(u_1, \ldots, u_m) := \left(\sum_{i=1}^m u_i \right)^k Q\left(\frac{u_1}{\sum_{i=1}^m u_i}, \ldots, \frac{u_m}{\sum_{i=1}^m u_i} \right),$$

because Q is a k-homogeneous polynomial. The right-hand side above is zero because Q satisfies the partial vanishing condition. □

Lemma 2.6. *[26] Let Q be a k-homogeneous polynomial in the real variables $u_1, \ldots, u_m$ with $m \geq 1$. If Q satisfies the partial vanishing condition, then $Q = 0$ identically.*

Proof. We prove the lemma using induction on the total number of variables for all $k \geq 0$. If $m = 1$ then a k-homogeneous polynomial $Q(u_1)$ must be of the form cu_1^k, for an appropriate constant c.

As such a polynomial satisfies the partial-vanishing condition only when $c = 0$, the base case for induction is established.

Next, consider a k-homogeneous polynomial $Q(u_1, \ldots, u_m, u_{m+1})$ with $m \geq 1$, which satisfies the partial vanishing condition, and let d denote its degree with respect to the variable u_{m+1}. In particular, there are polynomials $Q_0, \ldots, Q_d$ in the variables $u_1, \ldots, u_m$ such that

$$Q(u_1, \ldots, u_m, u_{m+1}) = \sum_{i=0}^{d} Q_i(u_1, \ldots, u_m) u_{m+1}^i.$$

Now fix $u_1, \ldots, u_m \geq 0$ such that $\sum_{i=1}^m u_i > 0$. Because Q satisfies the partial vanishing condition, Lemma 2.5 implies that $\sum_{i=0}^d Q_i(u_1, \ldots, u_m) u_{m+1}^i = 0$ for all $u_{m+1} > 0$. In particular, for each i, $Q_i(u_1, \ldots, u_m) = 0$ whenever $u_1, \ldots, u_m \geq 0$ and $\sum_{i=1}^m u_i > 0$. Thus Q_i satisfies the partial vanishing condition. Since Q_i is a $(k-i)$-homogeneous polynomial, the inductive hypothesis implies that $Q_i = 0$ identically and hence $Q = 0$ identically. □

The next result shows that $\psi(k)$ and $\phi(k)$ each admit a unique symmetric and unbiased estimator. This proof depends on the variety of distributions considered, and uses the requirement that our estimator must be unbiased for any distribution chosen from $\mathcal{D}$.

Lemma 2.7. *[26] If f is an n-symmetric function such that $\mathbb{E}[f(Z_1, \ldots, Z_n)] = 0$ for all $\mathbb{P}_z \in \mathcal{D}$, then $f = 0$ identically.*

Proof. Consider a point $\vec{z} = (z_1, \ldots, z_n) \in \mathbb{N}^n$ and define m to be the cardinality of the set $\{z_1, \ldots, z_n\}$. In addition, let $z_1', \ldots, z_m'$ denote the distinct elements in the set $\{z_1, \ldots, z_n\}$ and define m_i to be the number of times that z_i' appears in this set. Furthermore, let $\mathbb{P}_z \in \mathcal{D}$ be a probability distribution such that $\mathbb{P}_z(\{z_1', \ldots, z_m'\}) = 1$ and set $p_i = \mathbb{P}_z(z_i')$.

Notice that $\mathbb{E}[f(\vec{Z}_n)]$ is an n-homogeneous polynomial in the real variables $p_1, \ldots, p_m$ that satisfies the partial vanishing condition. Due to Lemma 2.6 this polynomial is identically zero. However, because f is n-symmetric, the coefficient of $\prod_{j=1}^m p_j^{m_j}$ in $\mathbb{E}[f(\vec{Z}_n)]$ is

$$f(\vec{z}) \binom{n}{m_1; \ldots; m_m},$$

implying that $f(\vec{z}) = 0$. □

With this result we are able to prove that $\hat{\psi}(k)$ and $\hat{\phi}(k)$ are the unique UMVUEs for $\psi(k)$ and $\phi(k)$ respectively.

Theorem 2.8. *If $n \geq k+1$, then $\hat{\psi}(k)$ is the unique UMVUE of $\psi(k)$. Further, no unbiased estimator of $\psi(k)$ exists for $n < k+1$. Similarly, if $n \geq k$, then $\hat{\phi}(k)$ is the unique UMVUE of $\phi(k)$, and no unbiased estimator of $\phi(k)$ exists for $n < k$.*

Proof. This result for $\psi(k)$ has been established [13] but is presented here for completeness. From Lemma 2.4, if the UMVUE for $\psi(k)$ exists, then it must be n-symmetric. Suppose there are two n-symmetric functions such that $f(Z_1, \ldots, Z_n)$ and $g(Z_1, \ldots, Z_n)$ are unbiased for $\psi(k)$. Applying Lemma 2.7 to $(f-g)$ shows that $f = g$, and therefore $\psi(k)$ admits a unique symmetric and unbiased estimator. As $\hat{\psi}(k)$ is n-symmetric and unbiased for $\psi(k)$ it follows that $\hat{\psi}(k)$ is the unique UMVUE for $\psi(k)$. From Lemma 2.3, it follows that no unbiased estimator of $\psi(k)$ exists for $n < k+1$. The analogous argument shows that $\hat{\phi}(k)$ is the UMVUE for $\phi(k)$, and that no unbiased estimator of $\phi(k)$ exists for $n < k$. □

2.4 Projection Statistic Approach

We now seek to analyze our U-statistics to show consistency, identify an asymptotic distribution, and estimate variances. Note the significant amount of dependence between the terms in the sums from (2.3) and (2.4). This section and Section 2.5 contains the main contributions of this thesis for this chapter. Our approach to the remaining theorems relies on the method of projection by Grams and Serfling [24]. Accordingly, we define the projections of our U-statistics by

$$\hat{\psi}_P(k) := \psi(k) + \sum_{i=1}^{n} (\mathbb{E}(\hat{\psi}(k)|Z_i) - \psi(k)); \tag{2.8}$$

$$\hat{\phi}_P(k) := \phi(k) + \sum_{i=1}^{n} (\mathbb{E}(\hat{\phi}(k)|Z_i) - \phi(k)). \tag{2.9}$$

Note that in terms of mean squared error the projection is the best approximation to the U-statistic that is linearly dependent on each data point. This linear dependence on the data implies that the projection satisfies the hypothesis of classical theorems in probability such as the Law of Large

Numbers (LLN) and the Central Limit Theorem (CLT) [17]. We define our remainder statistic as the difference between the U-statistic and its projection, given by

$$R_\psi(k) := \hat{\psi}(k) - \hat{\psi}_P(k); \tag{2.10}$$

$$R_\phi(k) := \hat{\phi}(k) - \hat{\phi}_P(k). \tag{2.11}$$

The remainder is convenient as when the remainder is appropriately small, results which apply to the projection can also be shown to apply to the U-statistic.

We find it convenient to represent $\hat{\psi}(k)$ and $\hat{\phi}(k)$ using kernel statistics which are symmetric in the data. While (2.2) is already symmetric, we symmetrize the kernel statistic in (2.1). Let

$$g_\psi(Z_1, \ldots, Z_{k+1}) := \frac{1}{k+1} \sum_{i=1}^{k+1} \left[\left[Z_i \notin \bigcup_{j=1, j\neq i}^{k+1} \{Z_j\} \right]\right]; \tag{2.12}$$

$$g_\phi(Z_1, \ldots, Z_k) := |\{Z_1, \ldots, Z_k\}|. \tag{2.13}$$

Note that these kernel statistic forms take those found in (2.1) and (2.2) and make them symmetric in the data. While the kernel for $\hat{\phi}(k)$ is unchanged, $g_\psi(Z_1, \ldots, Z_{k+1})$ is formed by symmetrizing $h_\psi(Z_1, \ldots, Z_{k+1})$ to a $(k+1)$-symmetric function. As the U-statistic associated with $g_\psi(Z_1, \ldots, Z_{k+1})$ is n-symmetric and has mean $\psi(k)$, we have by Lemma 2.7 that $\hat{\psi}(k)$ is also the U-statistic associated with the kernel function in (2.12). The following variances written in terms of these symmetric kernel functions are used in writing the variances of our U-statistics, as well as the associated projections and remainders. Motivated by the analysis of Hoeffding [29], for $j \geq 0$ let

$$\xi_{\psi,j}(k) := \mathbb{V}(\mathbb{E}(g_\psi(Z_1, \ldots, Z_k)|Z_1, \ldots, Z_j)); \tag{2.14}$$

$$\xi_{\phi,j}(k) := \mathbb{V}(\mathbb{E}(g_\phi(Z_1, \ldots, Z_k)|Z_1, \ldots, Z_j)). \tag{2.15}$$

Above it is understood that the sigma-field generated by $(Z_1, \ldots, Z_j)$ when $j = 0$ is $\{\emptyset, \Omega\}$; in particular, $\xi_{\psi,0}(k) = \xi_{\phi,0}(k) = 0$. We make ample use of bounds on these variances given by

$$\xi_{\psi,j}(k) \leq 1; \qquad\qquad \xi_{\phi,j}(k) \leq k^2.$$

An approach by Hoeffding [29] allows us to write the variance of the U-statistics in (2.3) and (2.4) in terms of the coefficients ξ. Using (2.3) but with the symmetrized kernels, we recognize that $\hat{\psi}^2(k)$ is a sum of $\binom{n}{k+1}^2$ products of two indicators. By counting the pairs of indicators that share a sample of size j, we have that

$$\mathbb{V}(\hat{\psi}(k)) = \frac{\sum_{j=1}^{k+1} \binom{k+1}{j}\binom{n-k-1}{k+1-j}\xi_{\psi,j}(k+1)}{\binom{n}{k+1}}. \tag{2.16}$$

Similarly, using (2.4) we have that

$$\mathbb{V}(\hat{\phi}(k)) = \frac{\sum_{j=1}^{k} \binom{k}{j}\binom{n-k}{k-j}\xi_{\phi,j}(k)}{\binom{n}{k}}. \tag{2.17}$$

We may also calculate the variances of the projections in (2.8) and (2.9). We have that

$$\mathbb{E}(\hat{\psi}(k)|Z_1) = \frac{k+1}{n}\mathbb{E}(g_\psi(Z_1,\ldots,Z_{k+1})|Z_1) + \left(1-\frac{k+1}{n}\right)\psi(k);$$
$$\mathbb{E}(\mathbb{E}(\hat{\psi}(k)|Z_1) - \psi(k))^2 = \frac{(k+1)^2\xi_{\psi,1}(k+1)}{n^2}.$$

It follows that

$$\mathbb{V}(\hat{\psi}_P(k)) = \frac{(k+1)^2\xi_{\psi,1}(k+1)}{n}. \tag{2.18}$$

Similarly,

$$\mathbb{E}(\mathbb{E}(\hat{\phi}(k)|Z_i) - \phi(k))^2 = \frac{k^2\xi_{\phi,1}(k)}{n^2};$$
$$\mathbb{V}(\hat{\phi}_P(k)) = \frac{k^2\xi_{\phi,1}(k)}{n}. \tag{2.19}$$

We now present a lemma, relating the variance of the remainder to the variance of the U-statistic and projection.

Lemma 2.9. *The $\mathcal{L}^2$ norm of the remainder may be written in terms of the variance of a U-statistic and its projection as*

$$\mathbb{E}(R_\psi^2(k)) = \mathbb{V}(\hat{\psi}(k)) - \mathbb{V}(\hat{\psi}_P(k));$$
$$\mathbb{E}(R_\phi^2(k)) = \mathbb{V}(\hat{\phi}(k)) - \mathbb{V}(\hat{\phi}_P(k)).$$

Proof. Note that by (2.10) and (2.11),

$$\mathbb{E}(R_\psi^2) = \mathbb{V}(\hat{\psi}(k)) + \mathbb{V}(\hat{\psi}_P(k)) - 2\mathrm{Cov}(\hat{\psi}(k), \hat{\psi}_P(k));$$

$$\mathbb{E}(R_\phi^2) = \mathbb{V}(\hat{\phi}(k)) + \mathbb{V}(\hat{\phi}_P(k)) - 2\mathrm{Cov}(\hat{\phi}(k), \hat{\phi}_P(k)).$$

We wish to show that

$$\mathrm{Cov}(\hat{\psi}(k), \hat{\psi}_P(k)) = \mathbb{V}(\hat{\psi}_P(k)); \tag{2.20}$$

$$\mathrm{Cov}(\hat{\phi}(k), \hat{\phi}_P(k)) = \mathbb{V}(\hat{\phi}_P(k)). \tag{2.21}$$

Note that

$$\begin{aligned}
\mathrm{Cov}(\hat{\psi}(k), \hat{\psi}_P(k)) &= \sum_{i=1}^{n} \mathrm{Cov}(\hat{\psi}(k), \mathbb{E}(\hat{\psi}(k)|Z_i)), \\
&= \sum_{i=1}^{n} \mathrm{Cov}(\mathbb{E}(\hat{\psi}(k)|Z_i), \mathbb{E}(\hat{\psi}(k)|Z_i)), \\
&= \sum_{i=1}^{n} \mathbb{V}(\mathbb{E}(\hat{\psi}(k)|Z_i)), \\
&= \mathbb{V}\left(\sum_{i=1}^{n} \mathbb{E}(\hat{\psi}(k)|Z_i)\right), \\
&= \mathbb{V}(\hat{\psi}_P(k)).
\end{aligned}$$

This shows (2.20), and a similar calculation shows (2.21). □

The technical convenience afforded by this lemma is that the remainder can be controlled in $\mathcal{L}^2$ norm based solely on the difference between the variance of the U-statistic and its projection. The following lemma assists in measuring the magnitude of the second moment of the remainder, using the sums in (2.16), (2.17), (2.18) and (2.19).

Lemma 2.10. *Uniformly for* $k = 1 : \log\lfloor n \rfloor$,

$$\frac{\binom{k}{1}\binom{n-k}{k-1}}{\binom{n}{k}} = \frac{k^2}{n} + O\left(\frac{k^4}{n^2}\right); \tag{2.22}$$

$$\frac{\sum_{j=2}^{k} \binom{k}{j}\binom{n-k}{k-j}}{\binom{n}{k}} = O\left(\frac{k^4}{n^2}\right). \tag{2.23}$$

Proof. We have that

$$\frac{\binom{k}{1}\binom{n-k}{k-1}}{\binom{n}{k}} = \frac{k!^2(n-k)!^2}{(k-1)!^2(n-2k+1)!n!} = \frac{k^2}{n}\prod_{i=0}^{k-2}\frac{n-k-i}{n-1-i} = \frac{k^2}{n}\prod_{i=0}^{k-2}\left(1-\frac{k-1}{n-1-i}\right),$$

which noting that the implicit constant may be bounded uniformly for $k = \log\lfloor n\rfloor$, establishes (2.22). For $j = 0$ we perform a second order expansion to obtain that

$$\frac{\binom{n-k}{k}}{\binom{n}{k}} = \frac{(n-k)!^2}{n!(n-2k)!} = \prod_{i=0}^{k-1}\frac{n-k-i}{n-i} = \prod_{i=0}^{k-1}\left(1-\frac{k}{n-i}\right) = 1-\frac{k^2}{n}+O\left(\frac{k^3}{n^2}\right). \tag{2.24}$$

Note that $\binom{n}{k}$ is the number of ways of choosing k objects from a set of n, while $\binom{k}{j}\binom{n-k}{k-j}$ is the number of ways to choose j objects from a specific set of k, and choose $k-j$ from the remaining $n-k$. This observation gives that

$$\sum_{j=0}^{k}\binom{k}{j}\binom{n-k}{k-j} = \binom{n}{k},$$

which with (2.22) and (2.24), establishes (2.23), noting that the implicit constant may be bounded uniformly for $k = \log\lfloor n\rfloor$. □

With this we are able to bound the second moment of the remainder over a range of k. For a fixed k, we provide a bound relative to the variance of the projection.

Lemma 2.11. *The error in the approximation of the U-statistic by its projection obeys the following bounds with constants that hold uniformly in n:*

$$\sum_{k=1:\lfloor\log(n)\rfloor}\mathbb{E}(R_\psi^2(k)) = O\left(\frac{\log^5(n)}{n^2}\right); \tag{2.25}$$

$$\sum_{k=2:\lfloor\log(n)\rfloor}\mathbb{E}(R_\phi^2(k)) = O\left(\frac{\log^7(n)}{n^2}\right). \tag{2.26}$$

If $\mathbb{P}_z$ is not a uniform distribution, then for a fixed k,

$$\frac{\mathbb{E}(R_\psi^2(k))}{\mathbb{V}(\hat{\psi}_P(k))} = O\left(\frac{1}{n}\right); \tag{2.27}$$

$$\frac{\mathbb{E}(R_\phi^2(k))}{\mathbb{V}(\hat{\phi}_P(k))} = O\left(\frac{1}{n}\right). \tag{2.28}$$

Proof. By Lemma 2.9, (2.16), and (2.18) we have that

$$\mathbb{E}(R_\psi^2(k)) = \mathbb{V}(\hat{\psi}(k)) - \mathbb{V}(\hat{\psi}_P(k)) = \frac{\sum_{j=1}^{k+1} \binom{k+1}{j}\binom{n-k-1}{k+1-j}\xi_{\psi,j}(k+1)}{\binom{n}{k+1}} - \frac{(k+1)^2\xi_{\psi,1}(k+1)}{n}.$$

From $\xi_{\psi,j}(k) \leq 1$ and Lemma 2.10 it follows that

$$\mathbb{E}(R_\psi^2(k)) = O\left(\frac{k^4}{n^2}\right). \tag{2.29}$$

As the implied constants are bounded independently of k and n, (2.25) follows.

Similarly, noting that $\xi_{\phi,j}(k) \leq k^2$ gives

$$\mathbb{E}(R_\phi^2(k)) = \mathbb{V}(\hat{\phi}(k)) - \mathbb{V}(\hat{\phi}_P(k)) = \frac{\sum_{j=1}^{k} \binom{k}{j}\binom{n-k}{k-j}\xi_{\phi,j}(k)}{\binom{n}{k}} - \frac{k^2\xi_{\phi,1}(k)}{n} = O\left(\frac{k^6}{n^2}\right).$$

As the implied constants are again bounded independently of k and n, (2.26) follows.

If $\mathbb{P}_z$ is uniform, then $\xi_{\psi,1}(k) = 0$ as $\mathbb{E}(g_\psi(Z_1, \ldots, Z_{k+1})|Z_1) = \psi(k)$. If not, then $\xi_{\psi,1}(k) > 0$ and the projection statistic has a positive variance. From (2.18) and (2.29) we have for a fixed k that (2.27) holds. Similarly, when $\mathbb{P}_z$ is non-uniform then $\xi_{\phi,1}(k) > 0$, and (2.28) holds. □

2.4.1 Consistency

While the results of Sen [55] allow us to guarantee an almost sure convergence of our U-statistic to its mean for any specific k, the results of Lemma 2.11 allow us to show convergence in probability uniformly for a range of k that grows with n. Specifically, we are able to control the expected magnitude of the square of the remainder, uniformly for $k \leq \lfloor \log(n) \rfloor$. As $g_\psi(Z_1, \ldots, Z_k)$ has an almost sure limit of 0 as $k \to \infty$, we suspect convergence in probability that is uniform for $k = 1 : n-1$. Similarly, when the support of $\mathbb{P}_z$ is finite, we have that $g_\phi(Z_1, \ldots, Z_k)$ converges almost surely to $\phi(\infty) := |I_z| < \infty$, and suspect a uniform convergence in probability for $k = 2 : n$.

Theorem 2.12. *For any $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{k=1:n-1} |\hat{\psi}(k) - \psi(k)| > \epsilon\right) = 0. \tag{2.30}$$

Furthermore, when $|I_z| < \infty$, that is $\mathbb{P}_z$ is finitely supported,

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{k=2:n} |\hat{\phi}(k) - \phi(k)| > \epsilon\right) = 0. \tag{2.31}$$

Proof. Let $\hat{\psi}_n(k)$ be our U-statistic estimate for $\psi(k)$ using n draws from our urn, and similarly define $\hat{\phi}_n(k)$. Let $k_n = \lfloor \log(n) \rfloor$. Note that almost sure convergence implies convergence in probability. We show (2.30) for large and small k separately. Namely,

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{k=k_n:n-1} \left|\hat{\psi}_n(k) - \psi(k)\right| > \epsilon\right) = 0; \tag{2.32}$$

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{k=1:k_n} \left|\hat{\psi}_n(k) - \psi(k)\right| > \epsilon\right) = 0. \tag{2.33}$$

When $|I_z| < \infty$, we show (2.31) by

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{k=k_n:n} \left|\hat{\phi}_n(k) - \phi(k)\right| > \epsilon\right) = 0; \tag{2.34}$$

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{k=2:k_n} \left|\hat{\phi}_n(k) - \phi(k)\right| > \epsilon\right) = 0. \tag{2.35}$$

We now show (2.32). It follows by (2.6) and

$$\frac{\binom{n-j}{k}}{\binom{n-1}{k}n} \geq \frac{\binom{n-j}{k+1}}{\binom{n-1}{k+1}n},$$

that $\hat{\psi}_n(k) \geq \hat{\psi}_n(k+1)$ for all $k = 1 : n-1$. From this, (2.14), (2.16), and the Bounded Convergence Theorem [17] it follows that

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}\left(\max_{k=k_n:n-1} \hat{\psi}_n(k)\right)^2 &\leq \lim_{n\to\infty} \mathbb{E}\left(\hat{\psi}_n(k_n)\right)^2, \\ &\leq o(1) + \psi(k_n) = o(1).\end{aligned}$$

Thus as $\psi(k) \to 0$ as $k \to \infty$, (2.32) holds.

Next we show (2.34). We note that (2.13) satisfies

$$g_\phi(Z_1, \ldots, Z_k) = |\{Z_1, \ldots, Z_k\}| \leq |\{Z_1, \ldots, Z_{k+1}\}| = g_\phi(Z_1, \ldots, Z_{k+1}).$$

It follows that $\hat{\phi}_n(k) \leq \hat{\phi}_n(k+1)$ for $k = 2 : n-1$. Further, as $k \to \infty$, $g_\phi(Z_1, \ldots, Z_k)$ increases almost surely to $\phi(\infty)$. Hence

$$\begin{aligned}\max_{k=k_n:n} |\hat{\phi}_n(k) - \phi(k)| &\leq \max_{k=k_n:n} |\hat{\phi}_n(k) - \phi(\infty)| + \max_{k=k_n:n} |\phi(k) - \phi(\infty)|, \\ &\leq |\hat{\phi}_n(k_n) - \phi(\infty)| + |\phi(k_n) - \phi(\infty)|,\end{aligned}$$

which converges in probability to 0, establishing (2.34).

To address the convergence in (2.33) and (2.35), we proceed directly. As

$$\max_{k=1:k_n} |\hat{\psi}_n(k) - \psi(k)| \leq \max_{k=1:k_n} |\hat{\psi}_n(k) - \hat{\psi}_P(k)| + \max_{k=1:k_n} |\hat{\psi}_P(k) - \psi(k)|, \tag{2.36}$$

$$\max_{k=2:k_n} |\hat{\phi}_n(k) - \phi(k)| \leq \max_{k=2:k_n} |\hat{\phi}_n(k) - \hat{\phi}_P(k)| + \max_{k=2:k_n} |\hat{\phi}_P(k) - \phi(k)|, \tag{2.37}$$

we show that each term on the right-hand side tends to zero. We first show that the limit in probability of $\max_{k=1:k_n} |\hat{\psi}_P(k) - \psi(k)|$ is 0. Note that by (2.18),

$$\mathbb{E}\left(\max_{k=1:k_n} |\hat{\psi}_P(k) - \psi(k)|\right)^2 \leq \mathbb{E}\left(\sum_{k=1}^{k_n} |\hat{\psi}_P(k) - \psi(k)|\right)^2 \leq k_n \sum_{k=1}^{k_n} \mathbb{E}\left(\hat{\psi}_P(k) - \psi(k)\right)^2 = O\left(\frac{k_n^4}{n}\right),$$

and hence for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{k=1:k_n} |\hat{\psi}_P(k) - \psi(k)| > \epsilon\right) = 0. \tag{2.38}$$

The argument in the case of $\phi(k)$ is analogous. We have by (2.19) that

$$\mathbb{E}\left(\max_{k=2:k_n} |\hat{\phi}_P(k) - \phi(k)|\right)^2 \leq k_n \sum_{k=2}^{k_n} \mathbb{E}\left(|\hat{\phi}_P(k) - \phi(k)|\right)^2 = O\left(\frac{k_n^6}{n}\right),$$

and hence for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{k=2:k_n} |\hat{\phi}_P(k) - \phi(k)| > \epsilon\right) = 0. \tag{2.39}$$

We now show that

$$\lim_{n\to\infty} \max_{k=1:k_n} |\hat{\psi}_n(k) - \hat{\psi}_P(k)| \overset{\text{a.s.}}{=} 0; \tag{2.40}$$

$$\lim_{n\to\infty} \max_{k=2:k_n} |\hat{\phi}_n(k) - \hat{\phi}_P(k)| \overset{\text{a.s.}}{=} 0. \tag{2.41}$$

We have that

$$\mathbb{E}\left(\max_{k=1:k_n}|\hat{\psi}_n(k)-\hat{\psi}_P(k)|\right)^2 \leq \mathbb{E}\left(\sum_{k=1}^{k_n}|\hat{\psi}_n(k)-\hat{\psi}_P(k)|\right)^2 \leq k_n\sum_{k=1}^{k_n}\mathbb{E}\left(|\hat{\psi}_n(k)-\hat{\psi}_P(k)|\right)^2.$$

By (2.25) it follows that

$$k_n\sum_{k=1}^{k_n}\mathbb{E}\left(\hat{\psi}_n(k)-\hat{\psi}_P(k)\right)^2 = O\left(\frac{\log^6(n)}{n^2}\right).$$

In particular,

$$\sum_{n=1}^{\infty}\mathbb{E}\left(\max_{k=1:k_n}|\hat{\psi}_n(k)-\hat{\psi}_P(k)|\right)^2 < \infty.$$

Hence (2.40) holds by Chebyshev's inequality and the Borel-Cantelli lemma [17]. With (2.38) this shows (2.33), which completes the proof of (2.30).

An analogous argument using (2.26) shows (2.41), which with (2.39) shows (2.35). To finish the proof, (2.34) and (2.35) shows (2.31). □

2.4.2 Asymptotic Normality

Another key result of the decomposition of a U-statistic into its projection and remainder is that the projection is asymptotically normally distributed. In particular, when the remainder is small relative to the projection, we may prove that our U-statistic, appropriately normalized, tends in distribution to a normal random variable. This method of proof is similar to those given in the U-statistic literature [13, 29, 56] on the subject.

Theorem 2.13. *Let $Z \sim \mathcal{N}(0,1)$. If $\mathbb{P}_z$ is not a uniform distribution, then for fixed k,*

$$\lim_{n\to\infty}\frac{\hat{\psi}(k)-\psi(k)}{\sqrt{\mathbb{V}(\hat{\psi}(k))}} \overset{d}{=} Z; \tag{2.42}$$

$$\lim_{n\to\infty}\frac{\hat{\phi}(k)-\phi(k)}{\sqrt{\mathbb{V}(\hat{\phi}(k))}} \overset{d}{=} Z. \tag{2.43}$$

Proof. This result for $\psi(k)$ has been established [13] but we present our proof here for completeness. For a fixed k and $\mathbb{P}_z$ not uniform, note that $\hat{\phi}_P(k)$ as given in (2.8) is the sum of i.i.d. random

variables with strictly positive and finite variance, and thus $(\hat{\psi}_P(k) - \psi(k))/\sqrt{\mathbb{V}(\hat{\psi}_P(k))}$ is asymptotically $\mathcal{N}(0,1)$ by the CLT. It follows from Lemma 2.11 that (2.27) holds, which implies that for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{R_\psi(k)}{\sqrt{\mathbb{V}(\hat{\psi}_P(k))}}\right| > \epsilon\right) = 0.$$

Equivalently,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{\hat{\psi}(k) - \psi(k)}{\sqrt{\mathbb{V}(\hat{\psi}_P(k))}} - \frac{\hat{\psi}_P(k) - \psi(k)}{\sqrt{\mathbb{V}(\hat{\psi}_P(k))}}\right| > \epsilon\right) = 0.$$

From Lemma 2.9 and (2.27) it follows that

$$\lim_{n\to\infty} \frac{\mathbb{V}(\hat{\psi}(k))}{\mathbb{V}(\hat{\psi}_P(k))} = 1,$$

which shows (2.42). An analogous argument for ϕ shows (2.43). □

Though we do not present the case in more detail, the asymptotic distribution of the U-statistic when $\mathbb{P}_z$ is uniform is possible to analyze. A second order projection may be built by considering the U-statistic conditioned on each pair (Z_i, Z_j) for $i \neq j$. It has been shown that for $\mathbb{P}_z$ uniform, (2.3) converges to $\psi(k)$ such that $n(\hat{\psi}(k) - \psi(k))$ is asymptotically $c(Z^2 - 1)$, where Z is a standard normal random variable, and c depends on $|I_z|$ [13]. Similarly, we may expect the uniform case of $\mathbb{P}_z$ to give a similar result for the estimation of $\phi(k)$ [13, 56].

2.5 Variance Estimation

We are also able to utilize the projection decomposition in estimating the variance of $\hat{\psi}(k)$ and $\hat{\phi}(k)$. We have seen in the proof of Theorem 2.13, that $\mathbb{V}(\hat{\psi}(k))$ and $\mathbb{V}(\hat{\psi}_P(k))$ are asymptotically equivalent. This relation allows us to show that the associated delete-1 jackknife or take-one-away estimate of variance [18, 57, 58] for $\hat{\psi}(k)$ and $\hat{\phi}(k)$ gives an asymptotically accurate estimate of the variance for each of these U-statistics.

2.5.1 Jackknife Estimation of Variance

The jackknife estimate of variance is based on the measure of the difference in the estimate when it is calculated with one data point removed. Let $\hat{\psi}^i(k)$ be the U-statistic for $\psi(k)$, formed when Z_i is removed from the available data, and similarly define $\hat{\phi}^i(k)$. The jackknife estimate of variance [18, 57, 58] is given by

$$S^2_\psi(k) := \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\psi}^i(k) - \hat{\psi}(k)\right)^2; \qquad S^2_\phi(k) := \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\phi}^i(k) - \hat{\phi}(k)\right)^2.$$

2.5.2 Computationally Convenient Jackknife Estimation

We first rewrite these variance estimates in a computationally convenient form. Recall that $R(j)$ denotes the summary statistic defined in (2.5).

Theorem 2.14. *Let*

$$c_\psi(j,k,n) := \frac{\binom{n-j}{k}}{n\binom{n-1}{k}} \text{ for } j \le n-k; \qquad c_\phi(j,k,n) := \sum_{i=1}^{j}\frac{\binom{n-j}{k-i}\binom{j}{i}}{j\binom{n}{k}} \text{ for } j \le n-k;$$

$$c_\psi(j,k,n) := 0 \text{ for } j > n-k; \qquad c_\phi(j,k,n) := \frac{1}{j} \text{ for } j > n-k;$$

$$\hat{\psi}'(k) := \sum_{j=1}^{n} c_\psi(j,k,n-1)R(j); \qquad \hat{\phi}'(k) := \sum_{j=1}^{n} c_\phi(j,k,n-1)R(j).$$

Then,

$$S^2_\psi(k) = \frac{n-1}{n}\sum_{j=1}^{n} R(j)\left((j-1)c_\psi(j-1,k,n-1) - jc_\psi(j,k,n-1) + \hat{\psi}'(k) - \hat{\psi}(k)\right)^2; \quad (2.44)$$

$$S^2_\phi(k) = \frac{n-1}{n}\sum_{j=1}^{n} R(j)\left((j-1)c_\phi(j-1,k,n-1) - jc_\phi(j,k,n-1) + \hat{\phi}'(k) - \hat{\phi}(k)\right)^2. \quad (2.45)$$

Proof. The arguments for $\psi(k)$ and $\phi(k)$ are similar. Note from (2.6) and (2.7) that

$$\hat{\psi}(k) = \sum_{j=1}^{n} c_\psi(j,k,n)R(j); \qquad \hat{\phi}(k) = \sum_{j=1}^{n} c_\phi(j,k,n)R(j).$$

Note that removing a draw from the data which contributed to $R(j)$, decreases $R(j)$ by j, and increases $R(j-1)$ by $j-1$, leaving the remaining R-statistics unchanged.

Let R_i denote the R-statistics for $\hat{\psi}_i(k)$ when Z_i is removed. Note that if Z_i contributes to $R(j_i^\star)$, then $R_i(j_i^\star) = R(j_i^\star) - j_i^\star$, and $R_i(j_i^\star - 1) = R(j_i^\star - 1) + j_i^\star - 1$. We have therefore that

$$\begin{aligned} S_\psi^2(k) &= \frac{n-1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \big(c_\psi(j,k,n-1) R_i(j) - c_\psi(j,k,n) R(j) \big) \right)^2, \\ &= \frac{n-1}{n} \sum_{i=1}^{n} \Big((j_i^\star - 1) c_\psi(j_i^\star - 1, k, n-1) - j_i^\star c_\psi(j_i^\star, k, n-1) + \hat{\psi}'(k) - \hat{\psi}(k) \Big)^2. \end{aligned}$$

Since there are $R(j)$ draws which contribute to $R(j)$, (2.44) follows.

Similarly,

$$\begin{aligned} S_\phi^2(k) &= \frac{n-1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \big(c_\phi(j,k,n-1) R_i(j) - c_\phi(j,k,n) R(j) \big) \right)^2, \\ &= \frac{n-1}{n} \sum_{i=1}^{n} \Big((j_i^\star - 1) c_\phi(j_i^\star - 1, k, n-1) - j_i^\star c_\phi(j_i^\star, k, n-1) + \hat{\phi}'(k) - \hat{\phi}(k) \Big)^2, \end{aligned}$$

from which (2.45) follows. □

2.5.3 Consistency

We now show that for a fixed k the jackknife estimate of variance is asymptotically consistent. As the variance and estimate both tend to zero, we show this consistency under the appropriate normalization.

Theorem 2.15. *When $\mathbb{P}_z$ is not a uniform distribution, for a fixed k and any $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\left| \frac{S_\psi^2(k)}{\mathbb{V}(\hat{\psi}(k))} - 1 \right| > \epsilon \right) = 0; \tag{2.46}$$

$$\lim_{n\to\infty} \mathbb{P}\left(\left| \frac{S_\phi^2(k)}{\mathbb{V}(\hat{\phi}(k))} - 1 \right| > \epsilon \right) = 0. \tag{2.47}$$

Proof. In what follows $U_n \overset{p}{\sim} V_n$ means that U_n and V_n are asymptotically equivalent in probability, namely that $(U_n - V_n) \to 0$ in probability. Let $S_{k,n}^i$ be the set of one-to-one functions from $\{1, \ldots, k\}$

into the set $\{1,\ldots,n\}/\{i\}$, and define

$$\hat{\psi}^{i\prime}(k) := \frac{1}{|S^i_{k,n}|}\sum_{\sigma\in S^i_{k,n}} g_\psi(Z_i, Z_{\sigma(1)},\ldots,Z_{\sigma(k)});$$
$$\hat{\phi}^{i\prime}(k) := \frac{1}{|S^i_{k-1,n}|}\sum_{\sigma\in S^i_{k-1,n}} g_\phi(Z_i, Z_{\sigma(1)},\ldots,Z_{\sigma(k-1)}).$$

Using (2.3) and (2.4) we have that

$$\begin{aligned}
\hat{\psi}(k) &= \left(1-\frac{k+1}{n}\right)\hat{\psi}^i(k) + \frac{k+1}{n}\hat{\psi}^{i\prime}(k);\\
\hat{\psi}^i(k)-\hat{\psi}(k) &= \frac{k+1}{n}\left(\hat{\psi}^i(k)-\hat{\psi}^{i\prime}(k)\right);\\
\hat{\phi}(k) &= \left(1-\frac{k}{n}\right)\hat{\phi}^i(k) + \frac{k}{n}\hat{\phi}^{i\prime}(k);\\
\hat{\phi}^i(k)-\hat{\phi}(k) &= \frac{k}{n}\left(\hat{\phi}^i(k)-\hat{\phi}^{i\prime}(k)\right).
\end{aligned}$$

By (2.18), $n\mathbb{V}(\hat{\psi}_P(k)) = (k+1)^2\xi_{\psi,1}(k+1)$ and by (2.19), $n\mathbb{V}(\hat{\phi}_P(k)) = k^2\xi_{\phi,1}(k)$. It follows that

$$\frac{S^2_\psi(k)}{\mathbb{V}(\hat{\psi}_P(k))} = \frac{n-1}{n}\frac{1}{\xi_{\psi,1}(k+1)}\sum_{i=1}^n \frac{\left(\hat{\psi}^i(k)-\hat{\psi}^{i\prime}(k)\right)^2}{n}; \tag{2.48}$$
$$\frac{S^2_\phi(k)}{\mathbb{V}(\hat{\phi}_P(k))} = \frac{n-1}{n}\frac{1}{\xi_{\phi,1}(k)}\sum_{i=1}^n \frac{\left(\hat{\phi}^i(k)-\hat{\phi}^{i\prime}(k)\right)^2}{n}. \tag{2.49}$$

Let $A_n \overset{a.s.}{\sim} B_n$ mean that $(A_n - B_n)\to 0$ in an almost sure sense. We use a result from Sen [55] to show that

$$\left(\hat{\psi}^i(k)-\hat{\psi}^{i\prime}(k)\right)^2 \overset{a.s.}{\sim} \left(\psi(k)-\mathbb{E}(g_\psi(Z_i,\ldots,Z_{k+i})|Z_i)\right)^2; \tag{2.50}$$
$$\left(\hat{\phi}^i(k)-\hat{\phi}^{i\prime}(k)\right) \overset{a.s.}{\sim} \left(\phi(k)-\mathbb{E}(g_\phi(Z_i,\ldots,Z_{k+i-1})|Z_i)\right)^2, \tag{2.51}$$

for all i. Meanwhile, note that

$$\xi_{\psi,1}(k+1) = \mathbb{E}\left(\psi(k)-\mathbb{E}(g_\psi(Z_i,\ldots,Z_{k+i})|Z_i)\right)^2;$$
$$\xi_{\phi,1}(k) = \mathbb{E}\left(\phi(k)-\mathbb{E}(g_\phi(Z_i,\ldots,Z_{k+i-1})|Z_i)\right)^2.$$

Also note that $(\psi(k)-\mathbb{E}(g_\psi(Z_1,\ldots,Z_{k+1})|Z_1))^2$ and $(\psi(k)-\mathbb{E}(g_\psi(Z_2,\ldots,Z_{k+2})|Z_2))^2$, are uncorrelated, as are $(\phi(k)-\mathbb{E}(g_\phi(Z_1,\ldots,Z_k)|Z_1))^2$ and $(\phi(k)-\mathbb{E}(g_\phi(Z_2,\ldots,Z_{k+1})|Z_2))^2$. Since

$|\hat{\psi}^i(k) - \hat{\psi}^{i\prime}(k)| \leq 1$ and $|\hat{\phi}^i(k) - \hat{\phi}^{i\prime}(k)| \leq k$, we have by the Bounded Convergence Theorem [17] that

$$\begin{aligned}
\mathbb{V}\left(\frac{\sum_{i=1}^{n}\left(\hat{\psi}^i(k) - \hat{\psi}^{i\prime}(k)\right)^2}{n}\right) &= \frac{\mathbb{V}\left(\hat{\psi}^1(k) - \hat{\psi}^{1\prime}(k)\right)^2}{n} + \frac{n-1}{n}\mathrm{Cov}\left(\hat{\psi}^1(k) - \hat{\psi}^{1\prime}(k), \hat{\psi}^2(k) - \hat{\psi}^{2\prime}(k)\right), \\
&= o(1); \\
\mathbb{V}\left(\frac{\sum_{i=1}^{n}\left(\hat{\phi}^i(k) - \hat{\phi}^{i\prime}(k)\right)^2}{n}\right) &= \frac{\mathbb{V}\left(\hat{\phi}^1(k) - \hat{\phi}^{1\prime}(k)\right)^2}{n} + \frac{n-1}{n}\mathrm{Cov}\left(\hat{\phi}^1(k) - \hat{\phi}^{1\prime}(k), \hat{\phi}^2(k) - \hat{\phi}^{2\prime}(k)\right), \\
&= o(1),
\end{aligned}$$

as n tends to infinity. Hence, using Theorem 1.5.4 of Durrett [17] we have that

$$\sum_{i=1}^{n}\frac{\left(\hat{\psi}^i(k) - \hat{\psi}^{i\prime}(k)\right)^2}{n} \overset{p}{\sim} \xi_{\psi,1}(k+1); \qquad \sum_{i=1}^{n}\frac{\left(\hat{\phi}^i(k) - \hat{\phi}^{i\prime}(k)\right)^2}{n} \overset{p}{\sim} \xi_{\phi,1}(k),$$

which with (2.48), and (2.49) give that

$$\frac{S^2_{\psi}(k)}{\mathbb{V}(\hat{\psi}_P(k))} \overset{p}{\sim} \frac{\xi_{\psi,1}(k+1)}{\xi_{\psi,1}(k+1)} = 1; \qquad \frac{S^2_{\phi}(k)}{\mathbb{V}(\hat{\phi}_P(k))} \overset{p}{\sim} \frac{\xi_{\phi,1}(k)}{\xi_{\phi,1}(k)} = 1,$$

proving the theorem.

We finally show (2.50) and (2.51). We condition on the outcome of Z_i. In particular, if $Z_i = j$, then $\hat{\psi}^i(k)$, $\hat{\psi}^{i\prime}(k)$, $\hat{\phi}^i(k)$ and $\hat{\phi}^{i\prime}(k)$ are U-statistics in the data $Z_1, \ldots, Z_n$ after Z_i is removed (see Table 2.1). As the associated kernels are bounded and each Z_i is a discrete random variable, the hypotheses of Theorem 1 in Sen [55] hold, and

$$\begin{aligned}
\hat{\psi}^i(k) &\overset{a.s.}{\sim} \mathbb{E}(g_{\psi}(Z_1, \ldots, Z_{k+1})) = \psi(k); \\
\hat{\psi}^{i\prime}(k) &\overset{a.s.}{\sim} \mathbb{E}(g_{\psi}(Z_i, \ldots, Z_{k+i})|Z_i); \\
\hat{\phi}^i(k) &\overset{a.s.}{\sim} \mathbb{E}(g_{\phi}(Z_1, \ldots, Z_k)) = \phi(k); \\
\hat{\phi}^{i\prime}(k) &\overset{a.s.}{\sim} \mathbb{E}(g_{\phi}(Z_i \ldots, Z_{i+k-1})|Z_i),
\end{aligned}$$

which shows (2.50) and (2.51).

□

Table 2.1: **U-statistics and kernel functions.**

U-statistic	Kernel
$\hat{\psi}^i(k)$	$g_\psi(Z_1, \ldots, Z_{k+1})$
$\hat{\psi}^{i\prime}(k)$	$g_\psi(j, Z_1, \ldots, Z_k)$
$\hat{\phi}^i(k)$	$g_\phi(Z_1, \ldots, Z_k)$
$\hat{\phi}^{i\prime}(k)$	$g_\phi(j, Z_1 \ldots, Z_{k-1})$

Chapter 3

Dissimilarity Probability

In this chapter we present new results involving U-statistics to estimate $\theta(k)$ as given by (1.4); in particular, the methods in this chapter apply to two-urn problems. Let $X_1, \ldots, X_{n_x}$ and $Y_1, \ldots, Y_{n_y}$ be independent samples of independent and identically distributed (i.i.d.) discrete random variables with probability mass functions $\mathbb{P}_x$ and $\mathbb{P}_y$ representing the distribution of urn-x and urn-y, respectively. Without loss of generality we assume that $\mathbb{P}_x$ and $\mathbb{P}_y$ are supported over I_x and I_y respectively, each a subset of $\mathbb{N} = \{1, 2, 3, \ldots\}$. Note that based on our formulation, distinct draws from the urns are independent. From these draws we estimate $\theta(k)$ by an optimal unbiased estimator. In particular, the estimator has the least variance of all estimators which are unbiased uniformly for all $\mathbb{P}_x, \mathbb{P}_y \in \mathcal{D}$, where $\mathcal{D}$ is defined to be the set of discrete distributions over a finite subset of $\mathbb{N}$. We also show that this estimator is asymptotically normally distributed, and the jackknife estimate of its variance is consistent over a range of k, which under given conditions is the entire range available for unbiased estimation.

3.1 U-Statistic Definition in Two Distributions

Our U-statistic here is built around the kernel statistic, similarly to the one urn case. For our purposes this kernel is allowed to vary over k, specifically

$$h(X_1, Y_1, \ldots, Y_k) := [\![X_1 \notin \{Y_1, \ldots, Y_k\}]\!]. \tag{3.1}$$

Note that $\mathbb{E}(h(X_1, Y_1, \ldots, Y_k)) = \theta(k)$. Furthermore, this indicator is symmetric with regards to the X data and with regards to the Y data. The U-statistic is defined as a symmetric estimator

formed by averaging this kernel statistic over all permutations of data with respect to each urn [56]. This symmetry reduces variance in the estimator, as shown in Section 3.3.

Let S_{k,n_y} be the set of one-to-one functions from $\{1,\ldots,k\}$ into $\{1,\ldots,n_y\}$. Note that $|S_{k,n_y}| = \binom{n_y}{k}k!$. The U-statistic based on (3.1) is given by

$$\hat{\theta}(k) := \frac{1}{n_x|S_{k,n_y}|}\sum_{i=1}^{n_x}\sum_{\sigma\in S_{k,n_y}} [\![X_i \notin \{Y_{\sigma(1)},\ldots,Y_{\sigma(k)}\}]\!]. \tag{3.2}$$

As in the case of one-urn problems, this form is convenient for mathematical analysis [24, 55], but difficult for computation.

3.2 Computationally Convenient Form of the U-Statistic

3.2.1 Summary Statistics

A form convenient for the computation of the statistic given in (3.2) depends on an effective summary of the data. Note that the quantities we estimate do not depend on specifically observed colors, but do depend on the relative proportions of colors in each sample. This suggests that our summary statistics should not depend on the labels of colors in each sample, but instead on how often each color has been observed in each sample. Motivated by this observation, define for each $i, j \in \mathbb{N}$,

$$\begin{aligned} M(i,j) := \text{number of colors } c \text{ such that color } c \text{ occurs exactly} \\ i\text{-times in } X_1,\ldots,X_{n_x} \text{ and } j\text{-times in } Y_1,\ldots,Y_{n_y}. \end{aligned} \tag{3.3}$$

That is $M(i,j)$ is the number of colors seen i times in the sample from urn-x and j times in the sample from urn-y. While this level of precision is necessary for the jackknife estimates given in Section 3.5, for $\hat{\theta}(k)$ we may use the statistics given by

$$\begin{aligned} Q(j) := \text{number of indices } i = 1:n_x \text{ such that color} \\ X_i \text{ occurs exactly } j\text{-times in } Y_1,\ldots,Y_{n_y}. \end{aligned} \tag{3.4}$$

That is $Q(j)$ is the number of draws from urn-x with a color seen j times in the sample from urn-y. Note that $\sum_j Q(j) = n_x$, and that $Q(j) = \sum_i iM(i,j)$.

3.2.2 Estimation of $\theta(k)$

Here we write (3.2) in a computationally convenient form using the Q-statistics in (3.4). This form is similar to that presented in Section 2.2.2.

Lemma 3.1. *The U-statistic estimate for $\theta(k)$ may be written as*

$$\hat{\theta}(k) = \frac{1}{n_x\binom{n_y}{k}} \sum_{j=0}^{n_y-k} \binom{n_y - j}{k} Q(j).$$

Proof. Fix $1 \leq i \leq n_x$ and suppose that color X_i occurs j-times in $Y_1, \ldots, Y_{n_y}$. If $j > (n_y - k)$ then any choice of k draws from $Y_1, \ldots, Y_{n_y}$ contains X_i, hence $[\![X_i \notin \{Y_{\sigma(1)}, \ldots, Y_{\sigma(k)}\}]\!] = 0$, for all $\sigma \in S_{k,n_y}$. On the other hand, if $j \leq (n_y - k)$ then $\sum_{\sigma \in S_{k,n_y}} [\![X_i \notin \{Y_{\sigma(1)}, \ldots, Y_{\sigma(k)}\}]\!] = k!\binom{n_y - j}{k}$. Noting that $|S_{k,n_y}| = k!\binom{n_y}{k}$, we obtain:

$$\frac{1}{n_x|S_{k,n_y}|} \sum_{i=1}^{n_x} \sum_{\sigma \in S_{k,n_y}} [\![X_i \notin \{Y_{\sigma(1)}, \ldots, Y_{\sigma(k)}\}]\!] = \frac{1}{n_x\binom{n_y}{k}} \sum_{j=0}^{n_y-k} \binom{n_y - j}{k} Q(j),$$

completing the lemma. □

3.3 Uniformly Minimum Variance Unbiased Estimation

Our method of proof for the results in this section is similar to the one used by Halmos [26] for single distributions, which we extend here naturally to the setting of two distributions to show that the U-statistic given in (3.2) is the UMVUE. We first show that 1 draw from urn-x and k draws from urn-y is the minimal amount of data to estimate $\theta(k)$ unbiasedly.

Lemma 3.2. *If $g(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ is unbiased for $\theta(k)$ for all $\mathbb{P}_x, \mathbb{P}_y \in \mathcal{D}$, then $m \geq 1$ and $n \geq k$.*

Proof. Consider in $\mathcal{D}$ probability distributions of the form $\mathbb{P}_x(1) = u$, $\mathbb{P}_x(2) = (1-u)$, $\mathbb{P}_y(1) = v$ and $\mathbb{P}_y(2) = (1-v)$, where $0 \leq u, v \leq 1$ are arbitrary real numbers. Clearly, $\mathbb{E}[g(X_1, \ldots, X_m, Y_1, \ldots, Y_n)]$ is a linear combination of polynomials of degree m in u and n in v and as a result, it is a polynomial of degree at most m in u and n in v. Since $\theta(k) = u(1-v)^k + (1-u)v^k$ (see (1.4)) has degree 1

in u and k in v, and $g(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ is unbiased for $\theta(k)$, we conclude that $1 \leq m$ and $k \leq n$. □

In what follows, we say that a function $f : \mathbb{N}^{n_x+n_y} \to \mathbb{R}$ is (n_x, n_y)**-symmetric** when

$$f(x_1, \ldots, x_{n_x}; y_1, \ldots, y_{n_y}) = f(x_{\sigma(1)}, \ldots, x_{\sigma(n_x)}; y_{\sigma'(1)}, \ldots, y_{\sigma'(n_y)}),$$

for all $x_1, \ldots, x_{n_x}, y_1, \ldots, y_{n_y} \in \mathbb{N}$ and permutations σ and σ' of $1, \ldots, n_x$ and $1, \ldots, n_y$, respectively. Alternatively, f is (n_x, n_y)-symmetric if and only if it may be regarded a function of $(x_{(1\ldots n_x)}, y_{(1\ldots n_y)})$, where $x_{(1\ldots n_x)}$ and $y_{(1\ldots n_y)}$ correspond to the order statistics $x_{(1)}, \ldots, x_{(n_x)}$ and $y_{(1)}, \ldots, y_{(n_y)}$, respectively.

A statistic of $(X_1, \ldots, X_{n_x}, Y_1, \ldots, Y_{n_y})$ is called (n_x, n_y)-symmetric when it may be represented in the form $f(X_1, \ldots, X_{n_x}, Y_1, \ldots, Y_{n_y})$ for some (n_x, n_y)-symmetric function f. Note that it is immediate from (3.2) that $\hat{\theta}(k)$ is (n_x, n_y)-symmetric.

The next result asserts that the variance of any asymmetric unbiased estimator of $\theta(k)$ may be reduced by a corresponding symmetric unbiased estimator.

Lemma 3.3. *An asymmetric unbiased estimator of $\theta(k)$ has a larger variance than a corresponding (n_x, n_y)-symmetric unbiased estimator.*

Proof. Let $\mathcal{F}$ denote the sigma-field generated by the random vectors $(X_{(1\ldots n_x)}; Y_{(1\ldots n_y)})$ and suppose that the statistic $T = f(X_1, \ldots, X_{n_x}, Y_1, \ldots, Y_{n_y})$ is unbiased for $\theta(k)$. Since $\hat{\theta}(k)$ is (n_x, n_y)-symmetric and bounded, we may assume without loss of generality that $\mathbb{E}[T^2] < +\infty$. In particular, $U = \mathbb{E}[T \,|\, \mathcal{F}]$ is a well-defined statistic and there is an (n_x, n_y)-symmetric function $g : \mathbb{N}^{n_x+n_y} \to \mathbb{R}$ such that $U = g(X_1, \ldots, X_{n_x}; Y_1, \ldots, Y_{n_y})$. Clearly, U is unbiased for $\theta(k)$ and (n_x, n_y)-symmetric. The result is now a direct consequence of Jensen's inequality for conditional expectations, which in our context asserts that $\mathbb{E}(U^2) \leq \mathbb{E}(T^2)$ with equality if and only if T is (n_x, n_y)-symmetric. □

The above lemma implies that if an UMVUE for $\theta(k)$ exists then it must be (n_x, n_y)-symmetric. Next, we show that there is a unique symmetric and unbiased estimator of $\theta(k)$, which immediately implies that $\hat{\theta}(k)$ is the UMVUE.

In what follows, let k_1 and k_2 denote positive integers. We say that a polynomial $Q(u_1, \ldots, u_m; v_1, \ldots, v_n)$ is (k_1, k_2)**-homogeneous** when it is a linear combination of polynomials of the form $\prod_{i=1}^m u_i^{b_i} \prod_{j=1}^n v_j^{c_j}$, with $\sum_{i=1}^m b_i = k_1$ and $\sum_{j=1}^n c_j = k_2$. Furthermore, we say that Q satisfies the **partial vanishing condition** if $Q(u_1, \ldots, u_m; v_1, \ldots, v_n) = 0$ whenever $u_1, \ldots, u_m, v_1, \ldots, v_n \geq 0$, $\sum_{i=1}^m u_i = 1$, and $\sum_{i=1}^n v_i = 1$. The next lemma is an intermediate step to show that a (k_1, k_2)-homogeneous polynomial which satisfies the partial vanishing condition is the zero polynomial.

Lemma 3.4. *If Q is a (k_1, k_2)-homogeneous polynomial in the real variables $u_1, \ldots, u_m, v_1, \ldots, v_n$ with $m, n \geq 1$ that satisfies the partial vanishing condition, then $Q(u_1, \ldots, u_m; v_1, \ldots, v_n) = 0$ whenever $u_1, \ldots, u_m, v_1, \ldots, v_n \geq 0$, $\sum_{i=1}^m u_i > 0$, and $\sum_{i=1}^n v_i > 0$.*

Proof. Fix $u_1, \ldots, u_m, v_1, \ldots, v_n \geq 0$ such that $c_1(\vec{u}) := \sum_{i=1}^m u_i > 0$ and $c_2(\vec{v}) := \sum_{i=1}^n v_i > 0$, and observe that

$$Q(u_1, \ldots, u_m; v_1, \ldots, v_n) = c_1(\vec{u})^{k_1} c_2(\vec{v})^{k_2} Q\left(\frac{u_1}{\sum_{i=1}^m u_i}, \ldots, \frac{u_m}{\sum_{i=1}^m u_i}; \frac{v_1}{\sum_{i=1}^n v_i}, \ldots, \frac{v_n}{\sum_{i=1}^n v_i}\right),$$

because Q is a (k_1, k_2)-homogeneous polynomial. Notice now that the right-hand side above is zero because Q satisfies the partial vanishing condition. □

Lemma 3.5. *Let Q be a (k_1, k_2)-homogeneous polynomial in the real variables $u_1, \ldots, u_m, v_1, \ldots, v_n$ with $m, n \geq 1$. If Q satisfies the partial vanishing condition, then $Q = 0$ identically.*

Proof. We prove the lemma using structural induction on (m, n) for all $k_1, k_2 \geq 0$.

If $m = n = 1$, then a (k_1, k_2)-homogeneous polynomial $Q(u_1, v_1)$ must be of the form $cu_1^{k_1} v_1^{k_2}$ for an appropriate constant c. As such a polynomial satisfies the partial-vanishing condition only when $c = 0$, the base case for induction is established.

Next, consider a (k_1, k_2)-homogeneous polynomial $Q(u_1, \ldots, u_m; v_1, \ldots, v_n, v_{n+1})$ with $m, n \geq 1$ that satisfies the partial vanishing condition, and let d denote its degree with respect to the variable v_{n+1}. In particular, there are polynomials $Q_0, \ldots, Q_d$ in the variables $u_1, \ldots, u_m, v_1, \ldots, v_n$ such

that

$$Q(u_1,\ldots,u_m;v_1,\ldots,v_n,v_{n+1}) = \sum_{i=0}^{d} Q_i(u_1,\ldots,u_m;v_1,\ldots,v_n)v_{n+1}^i.$$

Now fix $u_1,\ldots,u_m,v_1,\ldots,v_n \geq 0$ such that $\sum_{i=1}^m u_i > 0$ and $\sum_{i=1}^n v_i > 0$. Because Q satisfies the partial vanishing condition, Lemma 3.4 implies that $\sum_{i=0}^d Q_i(u_1,\ldots,u_m;v_1,\ldots,v_n)v_{n+1}^i = 0$ for all $v_{n+1} > 0$. Particularly, for each i, $Q_i(u_1,\ldots,u_m;v_1,\ldots,v_n) = 0$ whenever $u_1,\ldots,u_m,v_1,\ldots,v_n \geq 0$, $\sum_{i=1}^m u_i > 0$, and $\sum_{i=1}^n v_i > 0$. Thus Q_i satisfies the partial vanishing condition. Since Q_i is a $(k_1, k_2 - i)$-homogeneous polynomial, the inductive hypothesis implies that $Q_i = 0$ identically and hence $Q = 0$ identically. The same argument shows that if $Q(u_1,\ldots,u_m,u_{m+1};v_1,\ldots,v_n)$ with $m, n \geq 1$ is a (k_1, k_2)-homogeneous polynomial that satisfies the partial vanishing condition, then $Q = 0$ identically, completing the inductive proof of the lemma. □

The next result implies that $\theta(k)$ has a unique symmetric and unbiased estimator. Its proof depends on the variety of distributions in $\mathcal{D}$, and uses the requirement that our estimator be unbiased for any pair of distributions chosen from $\mathcal{D}$.

Lemma 3.6. *If f is an (n_x, n_y)-symmetric function such that $\mathbb{E}[f(X_1,\ldots,X_{n_x},Y_1,\ldots,Y_{n_y})] = 0$ for all $\mathbb{P}_x, \mathbb{P}_y \in \mathcal{D}$, then $f = 0$ identically.*

Proof. Consider a point $\vec{z} = (x_1,\ldots,x_{n_x},y_1,\ldots,y_{n_y}) \in \mathbb{N}^{n_x+n_y}$ and define m_1 and m_2 as the cardinalities of the sets $\{x_1,\ldots,x_{n_x}\}$ and $\{y_1,\ldots,y_{n_y}\}$, respectively. Furthermore, let $x'_1,\ldots,x'_{m_1}$ denote the distinct elements in the set $\{x_1,\ldots,x_{n_x}\}$ and define $m_{1,i}$ to be the number of times that x'_i appears in this set. Furthermore, let $\mathbb{P}_x \in \mathcal{D}$ be a probability distribution such that $\mathbb{P}_x(\{x'_1,\ldots,x'_{m_1}\}) = 1$ and set $p_{1,i} = \mathbb{P}_x(x'_i)$. In a completely analogous manner define $y'_1,\ldots,y'_{m_2}$, $m_{2,i}$, $\mathbb{P}_y$ and $p_{2,i}$.

Notice that $\mathbb{E}[f(\vec{Z})]$ is a polynomial in the real variables $p_{1,1},\ldots,p_{1,m_1},p_{2,1},\ldots,p_{2,m_2}$ that satisfies the hypothesis of Lemma 3.5; in particular, this polynomial is identically zero. However, because f is (n_x, n_y)-symmetric, the coefficient of $\prod_{i=1}^2 \prod_{j=1}^{m_i} p_{i,j}^{m_{i,j}}$ in $\mathbb{E}[f(\vec{Z})]$ is

$$f(\vec{z})\binom{n_x}{m_{1,1};\ldots;m_{1,m_1}}\binom{n_y}{m_{2,1};\ldots;m_{2,m_2}},$$

implying that $f(\vec{z}) = 0$. □

Theorem 3.7. *If $n_x \geq 1$ and $n_y \geq k$ then $\hat{\theta}(k)$ is the unique uniformly minimum variance unbiased estimator (UMVUE) of $\theta(k)$. Further, no unbiased estimator of $\theta(k)$ exists for $n_y < k$.*

Proof. From Lemma 3.3, if the UMVUE for $\theta(k)$ exists then it must be (n_x, n_y)-symmetric. Suppose there are two (n_x, n_y)-symmetric functions such that $f(X_1, \ldots, X_{n_x}; Y_1, \ldots, Y_{n_y})$ and $g(X_1, \ldots, X_{n_x}; Y_1, \ldots, Y_{n_y})$ are unbiased for $\theta(k)$. Applying Lemma 3.6 to $(f - g)$ shows that $f = g$, and $\theta(k)$ therefore admits a unique symmetric and unbiased estimator. From (3.2), $\hat{\theta}(k)$ is (n_x, n_y)-symmetric and unbiased for $\theta(k)$, and hence is the UMVUE for $\theta(k)$. From Lemma 3.2, it follows that no unbiased estimator of $\theta(k)$ exists for $k > n_y$. □

3.4 Projection Statistic Approach

Next, we seek to analyze our U-statistics to show consistency, identify an asymptotic distribution, and estimate variances. Unfortunately, there is a significant amount of dependence between the terms in the sum from (3.2). In analogy to the methods presented in Section 2.4 we define the projection of $\hat{\theta}(k)$ as

$$\hat{\theta}_P(k) := \theta(k) + \sum_{i=1}^{n_x} (\mathbb{E}(\hat{\theta}(k)|X_i) - \theta(k)) + \sum_{j=1}^{n_y} (\mathbb{E}(\hat{\theta}(k)|Y_j) - \theta(k)), \tag{3.5}$$

and the remainder as

$$R_\theta(k) := \hat{\theta}(k) - \hat{\theta}_P(k). \tag{3.6}$$

Furthermore, motivated by the analysis of variance in Section 2.4, for $j \geq 0$ let

$$\xi_{0,j}(k) := \mathbb{V}(\mathbb{P}(X_1 \notin \{Y_1, \ldots, Y_k\}|Y_1, \ldots, Y_j)); \tag{3.7}$$

$$\xi_{1,j}(k) := \mathbb{V}(\mathbb{P}(X_1 \notin \{Y_1, \ldots, Y_k\}|X_1, Y_1, \ldots, Y_j)). \tag{3.8}$$

Above it is understood that the sigma-field generated by $(Y_1, \ldots, Y_j)$ when $j = 0$ is $\{\emptyset, \Omega\}$; in particular, $\xi_{0,0}(k) = 0$, for all $k \geq 1$. Furthermore, note that $\xi_{i,j}(k) \leq 1$ for all i, j, k.

To assess the asymptotic distribution of $\hat{\theta}(k)$ ruling out degenerate cases, we make use of the following assumptions. The following conditions are used in the remaining proofs of this chapter.

(a) $\theta(1) < 1$, or equivalently $I_x \cap I_y \neq \emptyset$, that is there is a color common to urn-x and urn-y.

(b) $\mathbb{P}_x$ and $\mathbb{P}_y$ are not degenerate, that is $|I_x|, |I_y| \geq 2$.

(c) $\mathbb{P}_y(\,\cdot\,|I_x \cap I_y)$ is not a uniform distribution.

(d) $|I_x \cap I_y|$ is finite.

Conditions (a) and (b) insure that $\hat{\theta}(k)$ is non-degenerate with regards to either draws from urn-x or urn-y. These with condition (c) are used to guarantee that the projection statistic in (3.5) is not degenerate. While pointwise convergence [24, 58, 55] of the following theorems may be shown by existing results, under (a)-(d) we are able to show that results hold uniformly for some $b > 1$ and $k = 1 : \lfloor \log_b(n_y) \rfloor$. The (e)-(f) assumptions that follow are used to strengthen convergence results to uniform convergence for $k = 1 : n_y$.

(e) $\theta(\infty) > 0$. This is equivalent to there being a color unique to urn-x.

(f) $n_x = \Theta(n_y)$. That is there exists finite constants $c_0, c_1 > 0$ such that $c_0 n_y < n_x < c_1 n_y$ as $n_x, n_y \to \infty$.

The following properties of $\xi_{i,j}(k)$ are useful in the remaining proofs.

Lemma 3.8. *In the general case we have that $\xi_{0,k}(k)$ satisfies*

$$\lim_{k \to \infty} \xi_{0,k}(k) = 0, \tag{3.9}$$

with $\xi_{0,j}(k) \leq \xi_{0,k}(k)$ for $j \leq k$. On the other hand, we may write $\xi_{1,j}(k)$ in precise terms as

$$\xi_{1,j}(k) = \theta(2k - j) - \theta^2(k). \tag{3.10}$$

Under assumptions (a)-(d),

$$c := \min_{i \in I_x \cap I_y} \mathbb{P}_y(i),$$

satisfies $0 < c < 1$. *It follows that when* $\theta(\infty) = 0$,

$$\xi_{1,0}(k) = \Theta((1-c)^{2k}). \tag{3.11}$$

When $\theta(\infty) > 0$,

$$\xi_{1,0}(k) = \theta(\infty) - \theta^2(\infty) + \Theta(1-c)^k, \tag{3.12}$$

Furthermore, regardless of the value of $\theta(\infty)$, *it follows that*

$$\xi_{0,1}(k) = \Theta((1-c)^{2k}); \tag{3.13}$$

$$\theta(k) - \theta(\infty) = \Theta((1-c)^k); \tag{3.14}$$

$$\xi_{0,k}(k) = O((1-c)^k); \tag{3.15}$$

$$\xi_{1,k}(k) - \xi_{1,0}(k) = \Theta((1-c)^k). \tag{3.16}$$

Proof. As conditioning on fewer draws, or equivalently conditioning on smaller sigma-fields, does not increase the variance of a bounded random variable, it follows that $\xi_{0,j}(k) \leq \xi_{0,k}(k)$ for $j \leq k$. Note that

$$\lim_{k\to\infty} \mathbb{P}(X_1 \notin \{Y_1, \ldots, Y_k\} \mid Y_1, \ldots, Y_k) = \lim_{k\to\infty} \sum_{i \notin \{Y_1,\ldots,Y_k\}} \mathbb{P}_x(i) = \theta(\infty),$$

where the convergence is in the almost sure sense. Hence, (3.9) follows by the dominated convergence theorem.

To show (3.10), observe that

$$\mathbb{P}(X_1 \notin \{Y_1, \ldots, Y_k\} \mid X_1, Y_1, \ldots, Y_j) = [\![X_1 \notin \{Y_1, \ldots, Y_j\}]\!](1 - \mathbb{P}_y(X_1))^{k-j}.$$

In particular, it follows that

$$\begin{aligned}
\mathbb{E}(\mathbb{P}(X_1 \notin \{Y_1, \ldots, Y_k\} \mid X_1, Y_1, \ldots, Y_j)^2) &= \mathbb{E}([\![X_1 \notin \{Y_1, \ldots, Y_j\}]\!](1 - \mathbb{P}_y(X_1))^{2k-2j}), \\
&= \mathbb{E}(\mathbb{P}(X_1 \notin \{Y_1, \ldots, Y_{2k-j}\} \mid X_1, Y_1, \ldots, Y_j)), \\
&= \theta(2k-j),
\end{aligned}$$

from which (3.10) follows.

We define

$$c := \min\{\mathbb{P}_y(i) | i \in I_x \cap I_y\};$$

$$I^\star := \{i | \mathbb{P}_y(i) = c\}.$$

That is c is the probability of observing the rarest color in $I_x \cap I_y$ with a draw from urn-y, which under conditions (a)-(d) satisfies $c > 0$. Furthermore, $I^\star$ is the set of rarest colors in urn-y which are also in urn-x. It follows that

$$\begin{aligned}\xi_{1,0}(k) &= \theta(2k) - \theta^2(k), \\ &= \sum_{i \in I_x} \mathbb{P}_x(i)(1 - \mathbb{P}_y(i))^{2k} - \left(\sum_{i \in I_x} \mathbb{P}_x(i)(1 - \mathbb{P}_y(i))^k \right)^2.\end{aligned}$$

We consider first the case when $\theta(\infty) = 0$, that is $I_x \subset I_y$. In this case

$$\begin{aligned}\xi_{1,0}(k) &= \sum_{i \in I_x \cap I_y} \mathbb{P}_x(i)(1 - \mathbb{P}_y(i))^{2k} - \left(\sum_{i \in I_x \cap I_y} \mathbb{P}_x(i)(1 - \mathbb{P}_y(i))^k \right)^2, \\ &= (1-c)^{2k} \left(\sum_{i \in I^\star} \mathbb{P}_x(i) - \left(\sum_{i \in I^\star} \mathbb{P}_x(i) \right)^2 \right) + o\left((1-c)^{2k} \right).\end{aligned}$$

Noting that condition (c) implies that the term multiplying $(1-c)^{2k}$ is positive establishes (3.11). In the case that $\theta(\infty) > 0$ we have that

$$\xi_{1,0}(k) = \theta(\infty) - \theta^2(\infty) - 2\theta(\infty)(1-c)^k \left(\sum_{i \in I^\star} \mathbb{P}_x(i) \right) + o(1-c)^k,$$

establishing (3.12).

On the other hand, note that

$$\xi_{0,1}(k) = \sum_{i \in I_y} \mathbb{P}_y(i) \left(\sum_{j \neq i} \mathbb{P}_x(j)(1 - \mathbb{P}_y(j))^{k-1} \right)^2 - \left(\sum_{i \in I_x} \mathbb{P}_x(i)(1 - \mathbb{P}_y(i))^k \right)^2.$$

It follows that

$$\begin{aligned}\xi_{0,1}(k) = (1-c)^{2k} &\left(\frac{(1 - |I^\star|c)}{(1-c)^2} \left(\sum_{i \in I^\star} \mathbb{P}_x(i) \right)^2 + \frac{c}{(1-c)^2} \sum_{i \in I^\star} \left(\sum_{j \neq i, j \in I^\star} \mathbb{P}_x(j) \right)^2 - \left(\sum_{i \in I^\star} \mathbb{P}_x(i) \right)^2 \right) \\ &+ o\left((1-c)^{2k} \right).\end{aligned}$$

Note that

$$\min_{\sum_{i=1}^n t_i = c} \sum_{i=1}^n t_i^2,$$

is achieved when each $t_i = c/n$. It follows that,

$$\begin{aligned}\sum_{j\in I^\star}\left(\sum_{j\neq i, j\in I^\star}\mathbb{P}_x(j)\right)^2 &\geq \sum_{j\in I^\star}\left(\frac{|I^\star|-1}{|I^\star|}\sum_{j\in I^\star}\mathbb{P}_x(j)\right)^2; \\ &= \frac{(|I^\star|-1)^2}{|I^\star|}\left(\sum_{j\in I^\star}\mathbb{P}_x(j)\right)^2.\end{aligned}$$

This implies that the term multiplying $(1-c)^{2k}$ is greater than

$$\begin{aligned}\left(\sum_{j\in I^\star}\mathbb{P}_x(i)\right)^2\left(\frac{|I^\star|-|I^\star|^2c+c(|I^\star|-1)^2}{|I^\star|(1-c)^2}-1\right) &\geq \left(\sum_{j\in I^\star}\mathbb{P}_x(i)\right)^2\left(\frac{(1-2c)|I^\star|+c}{|I^\star|(1-c)^2}-1\right), \\ &= \left(\sum_{j\in I^\star}\mathbb{P}_x(i)\right)^2\left(\frac{c(1-|I^\star|c)}{|I^\star|(1-c)^2}\right).\end{aligned}$$

Note that under condition (c), $|I^\star|c < 1$, establishing (3.13).

We now consider the asymptotic behavior of $\xi_{0,k}(k)$ as $k \to \infty$. Let

$$T := \arg\min_{n\geq 1}\{I_x \cap I_y \subset \{Y_1, \ldots, Y_n\}\}.$$

That is T is the first draw when the complete set $I_x \cap I_y$ has been observed in draws from urn-y. Let

$$c' := |I_x \cap I_y|.$$

That is c' is the number of elements in $I_x \cap I_y$. We may bound the probability of T being large by

$$\mathbb{P}(T > k) \leq c'(1-c)^k.$$

It follows that

$$\xi_{0,k}(k) \leq c'(1-c)^k + (\theta(k) - \theta(\infty))^2. \tag{3.17}$$

Here $c'(1-c)^k$ is a bound on the probability that $I_x \cap I_y \not\subset \{Y_1, \ldots, Y_k\}$, and $(\theta(k) - \theta(\infty))^2$ is the error resulting from replacing $\theta(k)$ by $\theta(\infty)$ when $I_x \cap I_y \subset \{Y_1, \ldots, Y_k\}$.

Note that from (1.4),

$$\begin{aligned}\theta(k) - \theta(\infty) &:= \sum_{i \in I_x} \mathbb{P}_x(i)(1 - \mathbb{P}_y(i))^k - \sum_{i \notin I_y} \mathbb{P}_x(i), \\ &= \sum_{i \in I_x \cap I_y} \mathbb{P}_x(i)(1 - \mathbb{P}_y(i))^k = (1-c)^k \sum_{i \in I^\star} \mathbb{P}_x(i) + o\left((1-c)^k\right).\end{aligned}$$

and (3.14) follows.

It follows from this and (3.17) that for some $c'' > 0$,

$$\xi_{0,k}(k) \leq c'(1-c)^k + c''(1-c)^{2k}.$$

This shows (3.15). To show (3.16) note that (3.10) implies

$$\begin{aligned}\xi_{1,k}(k) - \xi_{1,0}(k) &= \theta(k) - \theta(2k), \\ &= \sum_{i \in I_x} \mathbb{P}_x(i)(1 - \mathbb{P}_y(i))^k (1 - (1 - \mathbb{P}_y(i))^k), \\ &= \sum_{i \in I_x \cap I_y} \mathbb{P}_x(i)(1 - \mathbb{P}_y(i))^k (1 - (1 - \mathbb{P}_y(i))^k), \\ &= (1-c)^k \sum_{i \in I^\star} \mathbb{P}_x(i) + o\left((1-c)^k\right),\end{aligned}$$

which establishes (3.16). □

Under conditions (a)-(c), $\hat{\theta}_P(k)$ is the non-degenerate sum of two independent sums of i.i.d. random variables and satisfies the hypotheses of the LLN and CLT. When $R(k)$ is small relative to $\hat{\theta}_P(k)$, it follows that $\hat{\theta}(k)$ is well approximated by its projection, $\hat{\theta}_P(k)$. Further, the variance of the projection statistic is easier to analyze and estimate than the variance of $\hat{\theta}(k)$, which is relevant to address consistency for the jackknife estimation of variance. The next lemma summarizes results about the asymptotic properties of $R(k)$, particularly with relation to the scale of $\hat{\theta}_P(k)$ as given by its variance.

Lemma 3.9. *The variance of the projection, $\hat{\theta}_P(k)$ is given by*

$$\mathbb{V}(\hat{\theta}_P(k)) = n_x^{-1}\xi_{1,0}(k) + k^2 n_y^{-1}\xi_{0,1}(k). \tag{3.18}$$

The error in the approximation of $\hat{\theta}(k)$ by its projection, obeys the following bounds with constants that hold uniformly for $k = 1 : \lfloor \log(n_y) \rfloor$:

$$\mathbb{E}(R^2(k)) = O\left(\frac{\log^2(n_y)}{n_y n_x} + \frac{\log^4(n_y)}{n_y^2}\right). \tag{3.19}$$

Further, under assumptions (a)-(d), there exists $b > 1$ such that for any $\epsilon > 0$,

$$\lim_{n_x,n_y\to\infty} \max_{k=1:\lfloor \log_b(n_y)\rfloor} \left| \frac{\mathbb{V}(\hat{\theta}(k))}{\mathbb{V}(\hat{\theta}_P(k))} - 1 \right| = 0; \tag{3.20}$$

$$\lim_{n_x,n_y\to\infty} \max_{k=1:\lfloor \log_b(n_y)\rfloor} \mathbb{P}\left(\left| \frac{\hat{\theta}(k) - \theta(k)}{\sqrt{\mathbb{V}(\hat{\theta}_P(k))}} - \frac{\hat{\theta}_P(k) - \theta(k)}{\sqrt{\mathbb{V}(\hat{\theta}_P(k))}} \right| > \epsilon \right) = 0. \tag{3.21}$$

Also, under assumptions (a)-(f) we have that for any $\epsilon > 0$,

$$\lim_{n_x,n_y\to\infty} \max_{k=1:n_y} \left| \frac{\mathbb{V}(\hat{\theta}(k))}{\mathbb{V}(\hat{\theta}_P(k))} - 1 \right| = 0; \tag{3.22}$$

$$\lim_{n_x,n_y\to\infty} \max_{k=1:n_y} \mathbb{P}\left(\left| \frac{\hat{\theta}(k) - \theta(k)}{\sqrt{\mathbb{V}(\hat{\theta}_P(k))}} - \frac{\hat{\theta}_P(k) - \theta(k)}{\sqrt{\mathbb{V}(\hat{\theta}_P(k))}} \right| > \epsilon \right) = 0. \tag{3.23}$$

Proof. A direct calculation from the form given in (3.2) gives that

$$\mathbb{E}(\hat{\theta}(k)|X_i) = n_x^{-1}\mathbb{P}(X_i \notin \{Y_1, \ldots, Y_k\}|X_i) + \left(1 - n_x^{-1}\right)\theta(k); \tag{3.24}$$

$$\mathbb{V}(\mathbb{E}(\hat{\theta}(k)|X_i)) = \frac{\xi_{1,0}(k)}{n_x^2};$$

$$\mathbb{E}(\hat{\theta}(k)|Y_i) = k n_y^{-1}\mathbb{P}(X_i \notin \{Y_i, \ldots, Y_{k+i-1}\}|Y_i) + \left(1 - k n_y^{-1}\right)\theta(k); \tag{3.25}$$

$$\mathbb{V}(\mathbb{E}(\hat{\theta}(k)|Y_i)) = \frac{k^2\xi_{0,1}(k)}{n_y^2}.$$

By (3.5), $\mathbb{V}(\hat{\theta}_P(k)) = n_x\mathbb{V}(\mathbb{E}(\hat{\theta}(k)|X_i)) + n_y\mathbb{V}(\mathbb{E}(\hat{\theta}(k)|Y_i))$, and (3.18) follows.

Next, we establish (3.19). Indeed, decomposing $\hat{\theta}(k)$ into a sum which depends on X_i or Y_i

and one that does not,

$$\mathbb{E}(\hat{\theta}(k)\mathbb{E}(\hat{\theta}(k)|X_i)) - \theta^2(k) = \frac{\xi_{1,0}(k)}{n_x^2};$$
$$\mathbb{E}(\hat{\theta}(k)\mathbb{E}(\hat{\theta}(k)|Y_i)) - \theta^2(k) = \frac{k^2\xi_{0,1}(k)}{n_y^2};$$
$$\mathrm{Cov}(\hat{\theta}(k), \hat{\theta}_P(k)) = \mathbb{V}(\hat{\theta}_P(k)).$$

It follows from the last relation that $\hat{\theta}_P(k)$ and $R(k)$ are uncorrelated. Of use is the relation that

$$\mathbb{E}(R^2(k)) = \mathbb{V}(\hat{\theta}(k)) - \mathbb{V}(\hat{\theta}_P(k)), \tag{3.26}$$

and in particular, $\mathbb{V}(\hat{\theta}(k)) \geq \mathbb{V}(\hat{\theta}_P(k))$. Following Hoeffding [29], we square the U-statistic form in (3.2), and count the pairs of indicators that share a sample of size $i \in \{0, 1\}$ from urn-x and $j \in \{0, \ldots, k\}$ from urn-y, giving that

$$\mathbb{V}(\hat{\theta}(k)) = \frac{\sum_{j=0}^{k} \binom{k}{j}\binom{n_y-k}{k-j}((n_x-1)\xi_{0,j}(k) + \xi_{1,j}(k))}{n_x\binom{n_y}{k}}. \tag{3.27}$$

Further, $\xi_{0,0}(k) = 0$ and $\xi_{i,j}(k) < 1$. We have by Lemma 2.10 and the identity in (3.18) that

$$\mathbb{V}(\hat{\theta}(k)) = \mathbb{V}(\hat{\theta}_P(k)) + O\left(\frac{k^2}{n_x n_y} + \frac{k^4}{n_y^2}\right),$$

with constants that are bounded independently of k, n_y and n_x, establishing (3.19). Note that the order estimate in (3.19) holds for any choice of base for the logarithm.

We now assume conditions (a)-(d) to insure that $\xi_{0,1}(k), \xi_{1,0}(k) > 0$ for all k. From Lemma 3.8 we have that for some $0 < \delta < 1$,

$$(1-\delta)^{2k} = \Theta(\xi_{0,1}(k));$$
$$(1-\delta)^{2k} = O(\xi_{1,0}(k));$$
$$\xi_{1,k}(k) - \xi_{1,0}(k) = \Theta\left((1-\delta)^k\right).$$

Further, we have that $\xi_{0,j}(k)$ and $\xi_{1,j}(k)$ are increasing functions in j. Note that for any $b > 1$,

there exists a finite $C > 0$ such that

$$\min_{k=1:\lfloor \log_b(n_y)\rfloor} \xi_{0,1}(k) \geq C n_y^{2\log_b(1-\delta)};$$

$$\min_{k=1:\lfloor \log_b(n_y)\rfloor} \xi_{1,0}(k) \geq C n_y^{2\log_b(1-\delta)},$$

and we use that $\xi_{0,1}(k), \xi_{1,0}(k) \leq 1$. It follows from Lemma 2.10, (3.27), and (3.18) that for some finite $C > 0$

$$\max_{k=1:\lfloor \log_b(n_y)\rfloor} \left(\mathbb{V}(\hat{\theta}(k)) - \mathbb{V}(\hat{\theta}_P(k))\right) \leq C\left(\frac{\log_b^2(n_y)}{n_x n_y} + \frac{\log_b^4(n_y)}{n_y^2}\right);$$

$$\min_{k=1:\lfloor \log_b(n_y)\rfloor} \mathbb{V}(\hat{\theta}_P(k)) \geq C n_y^{2\log_b(1-\delta)}\left(n_x^{-1} + n_y^{-1}\right).$$

It follows that for b chosen large enough to satisfy $2\log_b(1-\delta) > -0.5$, and some finite $C_b > 0$,

$$\max_{k=1:\lfloor \log_b(n_y)\rfloor} \frac{\mathbb{V}(\hat{\theta}(k))}{\mathbb{V}(\hat{\theta}_P(k))} \leq 1 + C_b \frac{\log_b^2(n_y)(n_x n_y)^{-1} + \log_b^4(n_y) n_y^{-2}}{n_y^{2\log_b(1-\delta)}\left(n_x^{-1} + n_y^{-1}\right)} = 1 + O\left(\log_b^4(n_y) n_y^{-0.5}\right).$$

which establishes (3.20). Recalling that $\mathbb{E}(R^2(k)) = \mathbb{V}(\hat{\theta}(k)) - \mathbb{V}(\hat{\theta}_P(k))$, it also follows that

$$\max_{k=1:\lfloor \log_b(n_y)\rfloor} \frac{\mathbb{E}(R^2(k))}{\mathbb{V}(\hat{\theta}_P(k))} = O(\log_b^4(n_y) n_y^{-0.5}),$$

which establishes (3.21).

We now show (3.22) and (3.23) under conditions (a)-(f). We begin by showing that under conditions (a)-(f) the b in the above argument may be chosen arbitrarily small. After this result we show that

$$\max_{k=\lfloor \log_b(n_y)\rfloor : n_y} \frac{\mathbb{E}(R^2(k))}{\mathbb{V}(\hat{\theta}_P(k))} = o(1),$$

where b is chosen sufficiently small, and note that this separation of proof for small and large k does not have a gap.

Note that under condition (e) following (3.12), $\xi_{1,0}(k) \to \theta(\infty)(1-\theta(\infty)) > 0$. This implies that

$$1 \geq \sup_{k\geq 1} \xi_{1,0}(k) \geq \inf_{k \geq 1} \xi_{1,0}(k) > 0.$$

Hence $\xi_{1,0}(k)$ is bounded below by a positive constant, and in the above argument we may define C_b such that

$$\max_{k=1:\lfloor \log_b(n_y)\rfloor} \frac{\mathbb{V}(\hat{\theta}(k))}{\mathbb{V}(\hat{\theta}_P(k))} \leq 1 + C_b \frac{\log_b^2(n_y)(n_x n_y)^{-1} + \log_b^4(n_y) n_y^{-2}}{n_x^{-1} + n_y^{2\log_b(1-\epsilon)} n_y^{-1}},$$
$$= 1 + O\left(\log_b^2(n_y) n_y^{-1} + n_x \log_b^4(n_y) n_y^{-2}\right).$$

By condition (f), $n_x = \Theta(n_y)$, and convergence holds with an arbitrarily small logarithmic base $b > 1$. Similarly for any arbitrarily small choice of b,

$$\max_{k=1:\lfloor \log_b(n_y)\rfloor} \frac{\mathbb{E}(R^2(k))}{\mathbb{V}(\hat{\theta}_P(k))} = O\left(\log_b^2(n_y) n_y^{-1} + n_x \log_b^4(n_y) n_y^{-2}\right).$$

We now show uniform convergence for $k \geq \lfloor \log_b(n_y)\rfloor$ for some $b > 1$. By Lemma 3.8, we have that for some finite $C > 0$ and $0 < \delta < 1$:

$$\xi_{0,k}(k) < C(1-\delta)^k;$$
$$\xi_{1,k}(k) - \xi_{1,0}(k) < C(1-\delta)^k;$$
$$\theta(k) - \theta(\infty) < C(1-\delta)^k.$$

The first two inequalities imply that for large k, all $\xi_{0,j}(k)$ are small, and all $\xi_{1,j}(k)$ are nearly identical. Further, $\xi_{0,j}(k)$ and $\xi_{1,j}(k)$ are increasing functions of j. Let

$$\xi_1(\infty) = \theta(\infty) - \theta^2(\infty) > 0.$$

As

$$\xi_{1,k}(k) - \xi_1(\infty) = \theta(k) - \theta(\infty) + \theta^2(\infty) - \theta^2(k),$$

we may increase C if necessary so as to have that

$$|\xi_{1,k}(k) - \xi_1(\infty)| < C(1-\delta)^k.$$

Note that for each j and k, $\xi_{0,j}(k) < C(1-\delta)^k$ and $|\xi_{1,j}(k) - \xi_1(\infty)| < C(1-\delta)^k$. It follows by

(3.27) that for $k \geq k_n$,

$$\begin{aligned}
\max_{k_n \leq k \leq n_y} \mathbb{V}(\hat{\theta}(k)) &\leq C(1-\delta)^{k_n} + \frac{\xi_1(\infty) + C(1-\delta)^{k_n}}{n_x}; \\
\min_{k_n \leq k \leq n_y} \mathbb{V}(\hat{\theta}_P(k)) &\geq n_x^{-1} \left(\xi_1(\infty) - C(1-\delta)^{k_n} \right); \\
\max_{k_n \leq k \leq n_y} \frac{\mathbb{V}(\hat{\theta}(k))}{\mathbb{V}(\hat{\theta}_P(k))} &\leq \frac{C(1-\delta)^{k_n} + n_x^{-1}(\xi_1(\infty) + C(1-\delta)^{k_n})}{n_x^{-1}\left(\xi_1(\infty) - C(1-\delta)^{k_n}\right)}, \\
&\leq \frac{\xi_1(\infty) + C(1-\delta)^{k_n}}{\xi_1(\infty) - C(1-\delta)^{k_n}} + \frac{C(1-\delta)^{k_n}}{(n_x)^{-1}\left(\xi_1(\infty) - C(1-\delta)^{k_n}\right)}.
\end{aligned}$$

Exchanging k_n for $\lfloor \log_b(n_y) \rfloor$ implies that

$$\max_{\log_b(n_y) \leq k \leq n_y} \frac{\mathbb{V}(\hat{\theta}(k))}{\mathbb{V}(\hat{\theta}_P(k))} \leq \frac{\xi_1(\infty) + C n_y^{\log_b(1-\delta)}}{\xi_1(\infty) - C n_y^{\log_b(1-\delta)}} + \frac{C n_y^{\log_b(1-\delta)}}{(n_x)^{-1}\left(\xi_1(\infty) - C n_y^{\log_b(1-\delta)}\right)}.$$

We now choose the base, b, of our logarithm to be as small as necessary to insure that $\log_b(1-\delta) < -1$. Hence when $n_x = \Theta(n_y)$, we have that for some $b > 1$ and all $k \geq \lfloor \log_b(n_y) \rfloor$ that

$$\begin{aligned}
\max_{\lfloor \log_b(n_y) \rfloor \leq k \leq n_y} \frac{\mathbb{V}(\hat{\theta}(k))}{\mathbb{V}(\hat{\theta}_P(k))} &= 1 + o(1); \\
\max_{\lfloor \log_b(n_y) \rfloor \leq k \leq n_y} \frac{\mathbb{E}(R^2(k))}{\mathbb{V}(\hat{\theta}_P(k))} &= o(1).
\end{aligned}$$

Hence, we establish (3.22) and (3.23). □

This lemma allows us to prove the remaining theorems. We begin by showing that as a function, $\hat{\theta}(k)$ and similar U-statistics converge in probability, uniformly for $k = 1 : n_y$.

3.4.1 Consistency

We first present a result that is useful for proving Theorem 3.15, as well as showing that $\hat{\theta}(k)$ is a uniformly consistent estimator of $\theta(k)$ in the appropriate range for k.

Theorem 3.10. *Define $\lambda(k) := \mathbb{E}(h(X_1, Y_1, \ldots, Y_k))$, where $h(x_1, y_1, \ldots, y_k)$ is a bounded $(1,k)$-symmetric kernel statistic, and let $\hat{\lambda}(k) = \hat{\lambda}_{n_x,n_y}(k)$ be the associated U-statistic using n_x draws from urn-x and n_y draws from urn-y; in particular, $\mathbb{E}(\hat{\lambda}(k)) = \lambda(k)$. Furthermore, assume that*

(i) $h \in [0,1]$.

(ii) There is a function $f : I_x \to [0,1]$ *such that* $\lim_{k\to\infty} h(X_1, Y_1, \ldots, Y_k) \overset{a.s}{=} f(X_1)$.

(iii) $\hat{\lambda}(k) \geq \hat{\lambda}(k+1)$*; in particular,* $\lambda(k) \geq \lambda(k+1)$.

(iv) $\log(n_y) = o(n_x)$.

It follows that

$$\lim_{n_x, n_y \to \infty} \mathbb{E}\left(\max_{k=1:n_y} \left| \hat{\lambda}(k) - \lambda(k) \right|^2 \right) = 0,$$

that is, $\hat{\lambda}(k)$ *converges to its expected value uniformly in* $\mathcal{L}^2$.

Proof. We prove this theorem for large and small k seperately. Let $k_n := \log\lfloor n_y \rfloor$. We show that

$$\lim_{n_x, n_y \to \infty} \mathbb{E}\left(\max_{k=k_n:n_y} |\hat{\lambda}(k) - \lambda(k)|^2 \right) = 0; \tag{3.28}$$

$$\lim_{n_x, n_y \to \infty} \mathbb{E}\left(\max_{k=1:k_n} |\hat{\lambda}(k) - \lambda(k)|^2 \right) = 0. \tag{3.29}$$

Note that similarly to (3.27), because h is symmetric, we have for

$$\xi_{0,j}(k) := \mathbb{V}(\mathbb{E}(h(X_1, Y_1, \ldots, Y_k)|Y_1, \ldots, Y_j));$$

$$\xi_{1,j}(k) := \mathbb{V}(\mathbb{E}(h(X_1, Y_1, \ldots, Y_k)|X_1, Y_1, \ldots, Y_j)),$$

that

$$\mathbb{V}\left(\hat{\lambda}(k)\right) = \frac{\sum_{j=0}^{k} \binom{k}{j}\binom{n_y - k}{k-j}((n_x - 1)\xi_{0,j}(k) + \xi_{1,j}(k))}{n_x \binom{n_y}{k}}. \tag{3.30}$$

Note also that with assumptions (i)-(ii) and the Bounded Convergence Theorem, it follows that

$$\lim_{k\to\infty} \xi_{0,k}(k) = 0.$$

For each $j \leq k$, $\xi_{0,j}(k) \leq \xi_{0,k}(k)$ and $\xi_{1,j}(k) \leq 1$. As a result,

$$\mathbb{V}\left(\hat{\lambda}(k)\right) \leq \xi_{0,k}(k) + \frac{1}{n_x}.$$

Furthermore, due to assumption (iii),

$$\mathbb{E}\left(\max_{k=k_n:n_y} |\hat{\lambda}(k) - \lambda(k)|^2 \right) \leq 2|\lambda(k_n) - \lambda(n_y)|^2 + 2\mathbb{V}(\hat{\lambda}(k_n)) + 2\mathbb{V}(\hat{\lambda}(n_y)).$$

As each term on the right-hand side above tends to zero as $n_x, n_y \to \infty$, (3.28) follows.

We now show (3.29). As $\xi_{i,j}(k) \leq 1$ and $\xi_{0,0}(k) = 0$, it follows by (3.30) and Lemma 2.10 that

$$\mathbb{V}\left(\hat{\lambda}(k)\right) \leq 1 - \frac{\binom{n_y-k}{k}}{\binom{n_y}{k}} + \frac{1}{n_x} = \frac{1}{n_x} + \sum_{j=1}^{k} \frac{\binom{k}{j}\binom{n_y-k}{k-j}}{\binom{n_y}{k}} = \frac{1}{n_x} + \frac{k^2}{n_y} + O\left(\frac{k^4}{n_y^2}\right),$$

uniformly for $k = 1 : k_n$ as $n_x, n_y \to \infty$. In particular,

$$\mathbb{E}\left(\max_{k=1:k_n} |\hat{\lambda}(k) - \lambda(k)|^2\right) \leq \sum_{k=1}^{k_n} \mathbb{V}(\hat{\lambda}(k)) \leq \frac{k_n}{n_x} + \frac{k_n^3}{n_y} + O\left(\frac{k_n^5}{n_y^2}\right).$$

Under assumption (iv), the right-hand side above tends to zero, showing (3.29) and proving the theorem. □

A relatively direct consequence of the previous theorem is the following result concerning the consistency of $\hat{\theta}(k)$. Before stating the result, observe that $\theta(k) = \mathbb{E}(h(X_1, Y_1, \ldots, Y_k))$ with $h(x_1, y_1, \ldots, y_k) = [\![x_1 \notin \{y_1, \ldots, y_k\}]\!]$.

Theorem 3.11. *Under the assumption that* $\log(n_y) = o(n_x)$*, we have*

$$\lim_{n_x, n_y \to \infty} \mathbb{P}\left(\max_{k=1:n_y} |\hat{\theta}(k) - \theta(k)| > \epsilon\right) = 0. \tag{3.31}$$

Proof. We note that $\mathcal{L}^2$-convergence implies convergence in probability and show that $\hat{\theta}(k)$ satisfies the hypotheses of Theorem 3.10.

Assumptions (i) and (ii) are immediate from (3.1), while assumption (iv) is a hypothesis. To show assumption (iii), recall that $S_{k,n}$ is the set of one-to-one functions from $\{1, \ldots, k\}$ into $\{1, \ldots, n\}$; in particular, $|S_{k+1,n_y}| = (n_y - k)|S_{k,n_y}|$. Next, note that for each indicator of the form $[\![X_1 \notin \{Y_{\sigma(1)}, \ldots, Y_{\sigma(k)}\}]\!]$, with $\sigma \in S_{k+1,n_y}$, there are $n - k$ choices of $\sigma(k+1) \in 1 : n_y$ which are not present in $\{\sigma(1), \ldots, \sigma(k)\}$. As $[\![X_1 \notin \{Y_{\sigma(1)}, \ldots, Y_{\sigma(k)}\}]\!] \geq [\![X_1 \notin \{Y_{\sigma(1)}, \ldots, Y_{\sigma(k+1)}\}]\!]$, it follows that $\hat{\theta}(k) \geq \hat{\theta}(k+1)$ for all $k = 1 : n_y - 1$. □

3.4.2 Asymptotic Normality

We are able to show convergence of $\hat{\theta}(k)$ to a normal distribution, which under assumptions (a)-(f) is shown to be uniform for $k = 1 : n_y$. In particular, while the joint distribution of $\hat{\theta}(k)$

for various k's is complicated due to the intricate dependencies, the marginal distribution of the U-statistic for a fixed k is approximately normal, uniformly for $k = 1 : n_y$.

Theorem 3.12. *Let $Z \sim \mathcal{N}(0,1)$. If we assume conditions (a)-(f), then for any t,*

$$\lim_{n_x,n_y\to\infty} \max_{k=1:n_y} \left| \mathbb{P}\left(\frac{\hat{\theta}(k) - \theta(k)}{\sqrt{\mathbb{V}(\hat{\theta}(k))}} \leq t \right) - \mathbb{P}(Z \leq t) \right| = 0. \tag{3.32}$$

Hence, $\hat{\theta}(k)$ is approximately normally distributed for $k = 1 : n_y$.

Proof. For a fixed k, note that assuming (a)-(c) $\hat{\theta}_P(k)$ is the sum of two independent sums of non-degenerate i.i.d. random variables, and thus $(\hat{\theta}_P(k) - \theta(k))/\sqrt{\mathbb{V}(\hat{\theta}_P(k))}$ is asymptotically $\mathcal{N}(0,1)$ by the CLT. However, we wish to show convergence uniformly for $k = 1 : n_y$ as $n_x, n_y \to \infty$ of $(\hat{\theta}_P(k) - \theta(k))/\sqrt{\mathbb{V}(\hat{\theta}_P(k))}$ in distribution to a standard normal random variable by the Berry-Esseen inequality [59]. Motivated by this we define the random variables

$$X_i'(k) := \frac{\mathbb{E}(\hat{\theta}(k)|X_i) - \theta(k)}{\sqrt{\mathbb{V}(\hat{\theta}_P(k))}};$$
$$Y_j'(k) := \frac{\mathbb{E}(\hat{\theta}(k)|Y_j) - \theta(k)}{\sqrt{\mathbb{V}(\hat{\theta}_P(k))}}.$$

Note that $\mathbb{E}(X_i'(k)) = \mathbb{E}(Y_j'(k)) = 0$, and that

$$\sum_{i=1}^{n_x} \mathbb{E}(|X_i'(k)|^2) + \sum_{j=1}^{n_y} \mathbb{E}(|Y_j'(k)|^2) = 1.$$

Motivated by the Berry-Esseen inequality, we wish to show that uniformly in k,

$$\sum_{i=1}^{n_x} \mathbb{E}(|X_i'(k)|^3) + \sum_{j=1}^{n_y} \mathbb{E}(|Y_j'(k)|^3) = o(1).$$

Note that from (3.24) and (3.25)

$$\left|\mathbb{E}(\hat{\theta}(k)|X_i) - \theta(k)\right|^3 = \frac{|\mathbb{P}(X_i \notin \{Y_1, \ldots, Y_k\}|X_i) - \theta(k)|^3}{n_x^3};$$
$$\left|\mathbb{E}(\hat{\theta}(k)|Y_i) - \theta(k)\right|^3 = \frac{k^3|\mathbb{P}(X_1 \notin \{Y_i, \ldots, Y_{k+1-i}\}|Y_i) - \theta(k)|^3}{n_y^3}.$$

Let

$$\eta_{1,0}(k) := \mathbb{E}|\mathbb{P}(X_i \notin \{Y_1, \ldots, Y_k\}|X_i) - \theta(k)|^3;$$
$$\eta_{0,1}(k) := \mathbb{E}|\mathbb{P}(X_1 \notin \{Y_i, \ldots, Y_{k+1-i}\}|Y_i) - \theta(k)|^3.$$

It follows that

$$\sum_{i=1}^{n_x}\mathbb{E}(|X_i'(k)|^3)+\sum_{j=1}^{n_y}\mathbb{E}(|Y_j'(k)|^3)=\frac{\eta_{1,0}(k)/n_x^2+k^3\eta_{0,1}(k)/n_y^2}{\left(\mathbb{V}(\hat{\theta}_P(k))\right)^{3/2}}=\frac{\eta_{1,0}(k)/n_x^2+k^3\eta_{0,1}(k)/n_y^2}{\left(\xi_{1,0}(k)/n_x+k^2\xi_{0,1}(k)/n_y\right)^{3/2}}.$$

Note that $0\leq\eta_{0,1}(k)\leq\xi_{0,1}(k)$. Since, under assumption (d), Lemma 3.8 implies that $\xi_{0,1}(k)$ decreases exponentially fast, we obtain

$$\frac{k^3\eta_{0,1}(k)}{n_y^2}=O(n_y^{-2}),$$

uniformly for all $k=1:n_y$ as $n_y\to\infty$. On the other hand, $0\leq\eta_{1,0}(k)\leq\xi_{1,0}(k)\leq 1$. Furthermore, under assumptions (a) and (e), (3.12) implies that $\inf_{k\geq 1}\xi_{1,0}(k)>0$. Note that by (f), $n_x=\Theta(n_y)$ and hence for some finite $C>0$

$$\sum_{i=1}^{n_x}\mathbb{E}(|X_i'(k)|^3)+\sum_{j=1}^{n_y}\mathbb{E}(|Y_j'(k)|^3)\leq C\frac{1/n_x^2+1/n_y^2}{\left(\inf\limits_{k\geq 1}\xi_{1,0}(k)/n_x\right)^{3/2}}=O\left(\frac{1}{\sqrt{n_x}}\right).$$

The above establishes convergence in distribution of $(\hat{\theta}_P(k)-\theta(k))/\sqrt{\mathbb{V}(\hat{\theta}_P(k))}$ to a standard normal random variable, uniformly for $k=1:n_y$. The end of the proof is an adaptation of the proof of Slutsky's Theorem [60]. Indeed, note that

$$\mathbb{P}\left(\frac{\hat{\theta}(k)-\theta(k)}{\sqrt{\mathbb{V}(\hat{\theta}(k))}}\leq t\right)=\mathbb{P}\left(\frac{\hat{\theta}_P(k)-\theta(k)}{\sqrt{\mathbb{V}(\hat{\theta}_P(k))}}+\frac{\hat{\theta}(k)-\hat{\theta}_P(k)}{\sqrt{\mathbb{V}(\hat{\theta}_P(k))}}\leq t\sqrt{\frac{\mathbb{V}(\hat{\theta}_P(k))}{\mathbb{V}(\hat{\theta}(k))}}\right).\tag{3.33}$$

From this identity, it follows for any fixed $\epsilon>0$ that

$$\mathbb{P}\left(\frac{\hat{\theta}(k)-\theta(k)}{\sqrt{\mathbb{V}(\hat{\theta}(k))}}\leq t\right)\leq\mathbb{P}\left(\frac{\hat{\theta}_P(k)-\theta(k)}{\sqrt{\mathbb{V}(\hat{\theta}_P(k))}}\leq t\sqrt{\frac{\mathbb{V}(\hat{\theta}_P(k))}{\mathbb{V}(\hat{\theta}(k))}}+\epsilon\right)+\mathbb{P}\left(\left|\frac{\hat{\theta}(k)-\hat{\theta}_P(k)}{\sqrt{\mathbb{V}(\hat{\theta}_P(k))}}\right|\geq\epsilon\right).$$

Under assumptions (a)-(f), the first term on the right-hand side of the above inequality can be made as close to $\mathbb{P}[Z\leq t+\epsilon]$ as wanted, uniformly for $k=1:n_y$, as $n\to\infty$, because of (3.22). On the other hand, the second term tends to 0 uniformly for $k=1:n_y$ because of (3.23). Letting $\epsilon\to 0^+$, shows that

$$\limsup_{n_x,n_y\to\infty}\max_{k=1:n_y}\mathbb{P}\left(\frac{\hat{\theta}(k)-\theta(k)}{\sqrt{\mathbb{V}(\hat{\theta}(k))}}\leq t\right)\leq\mathbb{P}[Z\leq t].$$

Similarly, using (3.33), we have:

$$\mathbb{P}\left(\frac{\hat{\theta}(k)-\theta(k)}{\sqrt{\mathbb{V}(\hat{\theta}(k))}}\leq t\right)\geq\mathbb{P}\left(\frac{\hat{\theta}_P(k)-\theta(k)}{\sqrt{\mathbb{V}(\hat{\theta}_P(k))}}\leq t\sqrt{\frac{\mathbb{V}(\hat{\theta}_P(k))}{\mathbb{V}(\hat{\theta}(k))}}-\epsilon\right)-\mathbb{P}\left(\left|\frac{\hat{\theta}(k)-\hat{\theta}_P(k)}{\sqrt{\mathbb{V}(\hat{\theta}_P(k))}}\right|\geq\epsilon\right),$$

and a similar argument as before shows now that

$$\liminf_{n_x,n_y\to\infty}\max_{k=1:n_y}\mathbb{P}\left(\frac{\hat{\theta}(k)-\theta(k)}{\sqrt{\mathbb{V}(\hat{\theta}(k))}}\leq t\right)\geq\mathbb{P}[Z\leq t],$$

which completes the proof of the theorem. □

3.5 Variance Estimation

In this section we estimate the variance of $\hat{\theta}(k)$, present a computationally convenient form for this variance estimate, and show that this estimate is consistent. We use a jackknife approach to account for variance in $\hat{\theta}(k)$ as a result of uncertainty in the X data, and use a separate jackknife to account for variance in $\hat{\theta}(k)$ as a result of uncertainty in the Y data. The sum of these jackknifes gives our consistent estimate of the variance of $\hat{\theta}(k)$.

3.5.1 Jackknife Estimation

We assume that $n_x\geq 2$ and let

$$S_x^2(k)=\frac{n_x-1}{n_x}\sum_{j=1}^{n_x}\left(\frac{1}{(n_x-1)|S_{k,n_y}|}\sum_{i\neq j}\sum_{\sigma\in S_{k,n_y}}[\![X_i\notin\{Y_{\sigma(1)},\ldots,Y_{\sigma(k)}\}]\!]-\hat{\theta}(k)\right)^2. \tag{3.34}$$

This estimate accounts for variability of $\hat{\theta}(k)$ due to variance in the X data through a leave-one-out jackknife estimate. To account for variability in the Y data, let S_r be the set of one-to-one functions from $\{1,\ldots,k\}$ into $\{1,\ldots,n_y\}/\{r\}$, note that $|S_r|=\binom{n_y-1}{k}k!$, and let

$$S_y^2(k)=\frac{n_y-1}{n_y}\sum_{r=1}^{n_y}\left(\frac{1}{n_x|S_r|}\sum_{i=1}^{n_x}\sum_{\sigma\in S_r}[\![X_i\notin\{Y_{\sigma(1)},\ldots,Y_{\sigma(k)}\}]\!]-\hat{\theta}(k)\right)^2, \tag{3.35}$$

which as an estimate accounts for variability of $\hat{\theta}(k)$ due to variance in the Y data by a leave-one-out jackknife estimate. We use this definition for $k=1:n_y-1$ and set $S_y^2(n_y)=0$. We define our

jackknife estimate in terms of $S_x^2(k)$ and $S_y^2(k)$ by

$$S^2(k) := S_x^2(k) + S_y^2(k), \tag{3.36}$$

for $k = 1 : n_y$. The corresponding estimate of the standard error is obtained by taking a square root, which is given by $S(k)$. This form of the variance estimate is convenient for analysis, but is not useful for computation.

3.5.2 Computationally Convenient Jackknife Estimation

Recall the $M(i,j)$ from (3.3). Let

$$c_j(k) := \frac{\binom{n_y-j-1}{k}}{n_x\binom{n_y-1}{k}}; \tag{3.37}$$

$$\hat{\theta}_y(k) := \frac{1}{n_x}\sum_{j=0}^{n_y-k-1} c_j Q(j), \tag{3.38}$$

which we use to show the following theorem.

Theorem 3.13. *We may write (3.34) as*

$$S_x^2(k) := \frac{1}{n_x(n_x-1)}\sum_{j=0}^{n_y-k} Q(j)\left(\hat{\theta}(k) - \frac{\binom{n_y-j}{k}}{\binom{n_y}{k}}\right)^2. \tag{3.39}$$

Similarly, we may write (3.35) as

$$S_y^2(k) := \frac{n_y-1}{n_y}\sum_{i=1}^{n_x}\sum_{j=1}^{n_y-k} jM(i,j)\left(i(c_{j-1}(k)-c_j(k)) + \hat{\theta}_y(k) - \hat{\theta}(k)\right)^2. \tag{3.40}$$

Proof. Note that removing a color from the X data which would add to $Q(j)$, decrements $Q(j)$. As a result each inner sum varies from $\hat{\theta}(k)$ only through the (n_x-1) multiplying constant and a correction to each $Q(j)$. Let Q_i denote the Q-statistics for the sum on the left, when X_i is removed while Q denotes those for $\hat{\theta}(k)$. Note that as each draw from urn-x contributes to exactly one $Q_i(j)$,

$Q_i(j) = Q(j)$ for all j except for some $j_i^\star$ where $Q_i(j_i^\star) = Q(j_i^\star) - 1$. We have that

$$\begin{aligned}
S_x^2(k) &= \frac{n_x - 1}{n_x} \sum_{i=1}^{n_x} \left(\sum_{j=1}^{n_y - k} \left(\frac{\binom{n_y - j}{k} Q_i(j)}{(n_x - 1)|S|} - \frac{\binom{n_y - j}{k} Q(j)}{n_x |S|} \right) \right)^2, \\
&= \frac{n_x - 1}{n_x} \sum_{i=1}^{n_x} \left(\sum_{j \neq j_i^\star} \frac{\binom{n_y - j}{k} Q(j)}{n_x (n_x - 1)|S|} + \frac{\binom{n_y - j_i^\star}{k} ((Q(j_i^\star) - 1) n_x - Q(j_i^\star)(n_x - 1))}{n_x (n_x - 1)|S|} \right)^2, \\
&= \frac{n_x - 1}{n_x} \sum_{i=1}^{n_x} \left(\sum_{j=1}^{n_y - k} \frac{\binom{n_y - j}{k} Q(j)}{n_x (n_x - 1)|S|} - \frac{\binom{n_y - j_i^\star}{k}}{|S|(n_x - 1)} \right)^2, \\
&= \frac{1}{n_x (n_x - 1)} \sum_{i=1}^{n_x} \left(\hat{\theta}(k) - \frac{\binom{n_y - j_i^\star}{k}}{|S|} \right)^2.
\end{aligned}$$

Note that there are $Q(j)$ draws from urn-x which contribute to $Q(j)$, allowing us to transform the sum from $i = 1 : n_x$ to the sum given in (3.39).

On the other hand, $S_y^2(k)$ corresponds to the jackknife summed over each deletion of Y data. Recall that S_r is the set of one-to-one functions from $\{1, \ldots, k\}$ into $\{1, \ldots, r-1, r+1, \ldots, n_y\}$, that is we remove the point Y_r from consideration in the U-statistic. Recall that $M(i, j)$ is the number of colors seen i times in draws from urn-x and j times in draws from urn-y, giving that $\sum_i iM(i, j) = Q(j)$. Further the point Y_r is of a color that contributes to $M(i_r^\star, j_r^\star)$ for some $i_r^\star, j_r^\star$. Removing it decrements $M(i_r^\star, j_r^\star)$ and increments $M(i_r^\star, j_r^\star - 1)$. Proceeding similarly as in the case for $S_x^2(k)$ we have that if M_r denotes the M statistics when Y_r is removed from the sample, then

$$S_y^2(k) = \frac{n_y - 1}{n_y} \sum_{r=1}^{n_y} \left(\sum_{j=0}^{n_y - k} \left(\frac{\binom{n_y - j - 1}{k} j}{n_x |S_r|} \sum_{i=1}^{n_y} M_r(i, j) - \sum_{j=0}^{n_y - k} \frac{\binom{n_y - j}{k} j}{n_x |S|} \sum_{i=1}^{n_y} M(i, j) \right) \right)^2.$$

Recalling the definition of $\hat{\theta}_y(k)$ and $c_j(k)$ from (3.37) and (3.38), we have that

$$\begin{aligned}
S_y^2(k) &= \frac{n_y - 1}{n_y} \sum_{r=1}^{n_y} \left(i_r^\star \frac{\binom{n_y - j_r^\star}{k}}{n_x |S_r|} - i_r^\star \frac{\binom{n_y - j_r^\star - 1}{k}}{n_x |S_r|} + \hat{\theta}_y(k) - \hat{\theta}(k) \right)^2, \\
&= \frac{n_y - 1}{n_y} \sum_{r=1}^{n_y} \left(i_r^\star \left(c_{j_r^\star - 1}(k) - c_{j_r^\star}(k) \right) + \hat{\theta}_y(k) - \hat{\theta}(k) \right)^2.
\end{aligned}$$

Noting that for each $M(i, j)$, there are j draws from urn-y that contribute, summing over each possible draw removed from urn-y gives the form in (3.40). □

3.5.3 Consistency

The consistency of the variance estimation of $S^2(k)$ from (3.36) is again closely related to the projection method. In particular $S^2(k)$ is shown to be asymptotically consistent as an estimate of $\mathbb{V}(\hat{\theta}_P(k))$. When $\mathbb{V}(\hat{\theta}(k))$ converges to $\mathbb{V}(\hat{\theta}_P(k))$, we have that $S^2(k)$ is a consistent estimator of $\mathbb{V}(\hat{\theta}_P(k))$. As $S^2(k)$, $\mathbb{V}(\hat{\theta}_P(k))$, and $\mathbb{V}(\hat{\theta}(k))$ each tend to zero, the unnormalized consistency result is unsatisfactory. As an alternative, we show that $S_x^2(k)$ and $S_y^2(k)$ are a consistent estimator relative to the appropriate terms of $\mathbb{V}(\hat{\theta}_P(k))$. We first show the following technical lemma. To state the lemma, recall that $S_{k,n}$ is the set of one-to-one functions from $\{1,\ldots,k\}$ into $\{1,\ldots,n\}$.

Lemma 3.14. *Let $S_{k,n}^i$ be the set of one-to-one functions from $\{1,\ldots,k\}$ into $\{1,\ldots,n\}/\{i\}$. Consider the transformation $h(x_1,y_1,\ldots,y_k) := [\![x_1 \notin \{y_1,\ldots,y_k\}]\!]$ and define*

$$\hat{\theta}_x^i(k) := \frac{1}{|S_{k,n_y}|}\sum_{\sigma\in S_{k,n_y}}\frac{1}{n_x-1}\sum_{j=1,j\neq i}^{n_x} h(X_j,Y_{\sigma(1)},\ldots,Y_{\sigma(k)}); \tag{3.41}$$

$$\hat{\theta}_x^{i\prime}(k) := \frac{1}{|S_{k,n_y}|}\sum_{\sigma\in S_{k,n_y}} h(X_i,Y_{\sigma(1)},\ldots,Y_{\sigma(k)}); \tag{3.42}$$

$$\hat{\theta}_y^i(k) := \frac{1}{|S_{k,n_y}^i|}\sum_{\sigma\in S_{k,n_y}^i}\frac{1}{n_x}\sum_{j=1}^{n_x} h(X_j,Y_{\sigma(1)},\ldots,Y_{\sigma(k)}); \tag{3.43}$$

$$\hat{\theta}_y^{i\prime}(k) := \frac{1}{|S_{k-1,n_y}^i|}\sum_{\sigma\in S_{k-1,n_y}^i}\frac{1}{n_x}\sum_{j=1}^{n_x} h(X_j,Y_i,Y_{\sigma(1)},\ldots,Y_{\sigma(k-1)}). \tag{3.44}$$

If $\log(n_y) = o(n_x)$, then for each $\epsilon > 0$,

$$\lim_{n_x,n_y\to\infty}\max_{k=1:n_y}\mathbb{P}\left(\left|\sum_{i=1}^{n_x}\frac{\left(\hat{\theta}_x^i(k)-\hat{\theta}_x^{i\prime}(k)\right)^2}{n_x}-\xi_{1,0}(k)\right|>\epsilon\right)=0; \tag{3.45}$$

$$\lim_{n_x,n_y\to\infty}\max_{k=1:n_y-1}\mathbb{P}\left(\left|\sum_{i=1}^{n_y}\frac{\left(\hat{\theta}_y^i(k)-\hat{\theta}_y^{i\prime}(k)\right)^2}{n_y}-\xi_{0,1}(k)\right|>\epsilon\right)=0. \tag{3.46}$$

Proof. We first show that

$$\lim_{n_x,n_y\to\infty} \max_{k=1:n_y} \mathbb{E}\left(\hat{\theta}_x^1(k) - \theta(k)\right)^2 = 0; \tag{3.47}$$

$$\lim_{n_x,n_y\to\infty} \max_{k=1:n_y} \mathbb{E}\left(\hat{\theta}_x^{1\prime}(k) - \mathbb{E}(h(X_1, Y_1, \ldots, Y_k)|X_1)\right)^2 = 0; \tag{3.48}$$

$$\lim_{n_x,n_y\to\infty} \max_{k=1:n_y-1} \mathbb{E}\left(\hat{\theta}_y^1(k) - \theta(k)\right)^2 = 0; \tag{3.49}$$

$$\lim_{n_x,n_y\to\infty} \max_{k=1:n_y-1} \mathbb{E}\left(\hat{\theta}_y^{1\prime}(k) - \mathbb{E}(h(X_1, Y_1, \ldots, Y_k)|Y_1)\right)^2 = 0. \tag{3.50}$$

Indeed, note that $\hat{\theta}_x^1(k)$ and $\hat{\theta}_y^1(k)$, much like $\hat{\theta}(k)$, are U-statistics associated with the kernel h; in particular, due to Theorem 3.10, (3.47) and (3.49) follow. On the other hand, observe that

$$\mathbb{E}\big(\hat{\theta}_x^{1\prime}(k) - \mathbb{E}(h(X_1, Y_1, \ldots, Y_k)|X_1\big)^2 = \sum_{j\in I_x} \mathbb{P}[X_1 = j] \cdot \mathbb{E}\big(\hat{\lambda}_{x,j}(k) - h(j, Y_1, \ldots, Y_k)\big)^2; \tag{3.51}$$

$$\mathbb{E}\big(\hat{\theta}_y^{1\prime}(k) - \mathbb{E}(h(X_1, Y_1, \ldots, Y_k)|Y_1\big)^2 = \sum_{j\in I_y} \mathbb{P}[Y_1 = j] \cdot \mathbb{E}\big(\hat{\lambda}_{y,j}(k) - h(X_1, j, Y_2, \ldots, Y_k)\big)^2; \tag{3.52}$$

where

$$\begin{aligned}
\hat{\lambda}_{x,j}(k) &:= \frac{1}{|S_{k,n_y}|} \sum_{\sigma\in S_{k,n_y}} h(j, Y_{\sigma(1)}, \ldots, Y_{\sigma(k)}); \\
\hat{\lambda}_{y,j}(k) &:= \frac{1}{|S^1_{k-1,n_y}|} \sum_{\sigma\in S^1_{k-1,n_y}} \frac{1}{n_x} \sum_{i=1}^{n_x} h(X_i, j, Y_{\sigma(1)}, \ldots, Y_{\sigma(k-1)}).
\end{aligned}$$

Clearly, for each $j \in I_x$, $\hat{\lambda}_{x,j}(k)$ is the U-statistic associated with the kernel $h(j, Y_1, \ldots, Y_k)$. Alternatively, similarly to $\hat{\theta}(k)$ but when urn-x is only composed by balls of color j, $\hat{\lambda}_{x,j}(k)$ is the U-statistic associated with the kernel $h(X_1, Y_1, \ldots, Y_k)$. We may thus again appeal to Theorem 3.10 to conclude that

$$\lim_{n_x,n_y\to\infty} \max_{j\in I_x, j\le\ell} \max_{k=1:n_y} \mathbb{E}\left(\big(\hat{\lambda}_{x,j}(k) - h(j, Y_1, \ldots, Y_k)\big)^2\right) = 0,$$

for each $\ell < \infty$. Since $\left|\hat{\lambda}_{x,j}(k) - h(j, Y_1, \ldots, Y_k)\right| \le 1$, for all $j \in I_x$, and the summation in (3.51) may be approximated to any accuracy by truncating it to $j \le \ell$, with ℓ large enough, (3.48) follows. A similar argument shows (3.50).

Next, observe that

$$\left|(A-B)^2 - (C-D)^2\right|^2 \le 8\,|A-C|^2 + 8\,|B-D|^2\,,$$

for any real numbers A, B, C and D in the interval $[0,1]$. Noting that each of the terms involved in (3.47)-(3.50) are bounded between zero and one, it is immediate from the above inequality that

$$\lim_{n_x,n_y\to\infty} \max_{k=1:n_y} \mathbb{E}\left(\left(\hat{\theta}_x^1(k)-\hat{\theta}_x^{1\prime}(k)\right)^2-(\theta(k)-\mathbb{E}(h(X_1,Y_1,\ldots,Y_k)|X_1))^2\right)^2 = 0; \quad (3.53)$$

$$\lim_{n_x,n_y\to\infty} \max_{k=1:n_y-1} \mathbb{E}\left(\left(\hat{\theta}_y^1(k)-\hat{\theta}_y^{1\prime}(k)\right)^2-(\theta(k)-\mathbb{E}(h(X_1,Y_1,\ldots,Y_k)|Y_1))^2\right)^2 = 0. \quad (3.54)$$

Furthermore, because

$$\begin{aligned}\xi_{1,0}(k) &= \mathbb{E}\left(\theta(k)-\mathbb{E}(h(X_1,Y_1,\ldots,Y_k)|X_1)\right)^2;\\ \xi_{0,1}(k) &= \mathbb{E}\left(\theta(k)-\mathbb{E}(h(X_1,Y_1,\ldots,Y_k)|Y_1)\right)^2;\end{aligned}$$

we conclude that

$$\lim_{n_x,n_y\to\infty} \max_{k=1:n_y}\left|\mathbb{E}\left(\hat{\theta}_x^1(k)-\hat{\theta}_x^{1\prime}(k)\right)^2-\xi_{1,0}(k)\right| = 0; \quad (3.55)$$

$$\lim_{n_x,n_y\to\infty} \max_{k=1:n_y-1}\left|\mathbb{E}\left(\hat{\theta}_y^1(k)-\hat{\theta}_y^{1\prime}(k)\right)^2-\xi_{0,1}(k)\right| = 0. \quad (3.56)$$

We finally show (3.45) paying attention to the details of the proof of Theorem 1.5.4 in [17], for the convergence in probability of triangular arrays. The argument for (3.46) is analogous and therefore omitted.

Fix $\epsilon > 0$. Define

$$S_n(k) := \sum_{i=1}^{n_x} \frac{\left(\hat{\theta}_x^i(k)-\hat{\theta}_x^{i\prime}(k)\right)^2}{n_x}.$$

Due to the Cauchy-Schwarz inequality (expressed in terms of the $\mathcal{L}^2$-norm) and using that $2ab \le a^2+b^2$, for any real numbers a and b, we obtain

$$\begin{aligned}\mathbb{P}\big[|S_n(k)-\xi_{1,0}(k)|\ge\epsilon\big] &\le \frac{\|S_n(k)-\xi_{1,0}(k)\|_2^2}{\epsilon^2},\\ &\le \frac{\left(\sqrt{\mathbb{V}(S_n(k))}+|\mathbb{E}(S_n(k))-\xi_{1,0}(k)|\right)^2}{\epsilon^2},\\ &\le \frac{2}{\epsilon^2}\Big\{\mathbb{V}(S_n(k))+|\mathbb{E}(S_n(k))-\xi_{1,0}(k)|^2\Big\}.\end{aligned}$$

But note that $\mathbb{E}(S_n(k)) = \mathbb{E}\big(\hat{\theta}_x^1(k)-\hat{\theta}_x^{1\prime}(k)\big)^2$, hence

$$\max_{k=1:n_y}\mathbb{P}[|S_n(k)-\xi_{1,0}(k)|\ge\epsilon] \le \frac{2}{\epsilon^2}\left\{\max_{k=1:n_y}\mathbb{V}(S_n(k))+\max_{k=1:n_y}\left|\mathbb{E}\big(\hat{\theta}_x^1(k)-\hat{\theta}_x^{1\prime}(k)\big)^2-\xi_{1,0}(k)\right|\right\}.$$

In particular, due to (3.55), to complete the proof of (3.45) it suffices to show that $\mathbb{V}(S_n(k))$ converges to 0 uniformly for $k = 1 : n_y$. For this, observe first that

$$\left|\mathrm{Cov}(U', V')\right| \leq \|U - U'\|_2 \|V'\|_2 + \|V' - V\|_2 \|U\|_2 + \left|\mathrm{Cov}(U, V)\right|,$$

for any random variables U, V, U' and V' with finite square-mean. In particular, if we let $U'(k) := \left(\hat{\theta}_x^1(k) - \hat{\theta}_x^{1\prime}(k)\right)^2$, $U(k) := \left(\theta(k) - \mathbb{E}(h(X_1, Y_1, \ldots, Y_k)|X_1)\right)^2$, $V'(k) := \left(\hat{\theta}_x^2(k) - \hat{\theta}_x^{2\prime}(k)\right)^2$ and $V(k) := \left(\theta(k) - \mathbb{E}(h(X_1, Y_1, \ldots, Y_k)|X_2)\right)^2$ then

$$\begin{aligned}
\mathbb{V}(S_n(k)) &= \frac{\mathbb{V}\left(\hat{\theta}_x^1(k) - \hat{\theta}_x^{1\prime}(k)\right)^2}{n_x} + \frac{n_x - 1}{n_x} \mathrm{Cov}\left(\left(\hat{\theta}_x^1(k) - \hat{\theta}_x^{1\prime}(k)\right)^2, \left(\hat{\theta}_x^2(k) - \hat{\theta}_x^{2\prime}(k)\right)^2\right), \\
&\leq \frac{1}{n_x} + \|U(k) - U'(k)\|_2 + \|V(k) - V'(k)\|_2,
\end{aligned}$$

because $\mathrm{Cov}(U(k), V(k)) = 0$, and $\hat{\theta}_x^1(k), \hat{\theta}_x^{1\prime}(k), \hat{\theta}_x^2(k)$ and $\hat{\theta}_x^{2\prime}(k)$ are each bounded between zero and one. The identity in (3.45) is now a direct consequence of (3.53). □

Theorem 3.15. *Assume conditions (a)-(d) and that* $\log(n_y) = o(n_x)$. *Then for any* $\epsilon > 0$,

$$\lim_{n_x, n_y \to \infty} \max_{k=1:n_y} \mathbb{P}\left(\left|S_x^2(k) - \frac{\xi_{1,0}(k)}{n_x}\right| > \frac{\epsilon}{n_x}\right) = 0; \tag{3.57}$$

$$\lim_{n_x, n_y \to \infty} \max_{k=1:n_y-1} \mathbb{P}\left(\left|S_y^2(k) - \frac{k^2 \xi_{0,1}(k)}{n_y}\right| > \frac{k^2 \epsilon}{n_y}\right) = 0. \tag{3.58}$$

Furthermore, under the above conditions, for each fixed k,

$$\lim_{n_x, n_y \to \infty} \mathbb{P}\left(\left|\frac{S^2(k)}{\mathbb{V}(\hat{\theta}(k))} - 1\right| > \epsilon\right) = 0, \tag{3.59}$$

for all $\epsilon > 0$.

Proof. Using (3.2) we have that

$$\begin{aligned}
\hat{\theta}(k) &= \left(1 - \frac{1}{n_x}\right) \hat{\theta}_x^i(k) + \frac{1}{n_x} \hat{\theta}_x^{i\prime}(k); \\
\hat{\theta}_x^i(k) - \hat{\theta}(k) &= \frac{1}{n_x}\left(\hat{\theta}_x^i(k) - \hat{\theta}_x^{i\prime}(k)\right); \\
\hat{\theta}(k) &= \left(1 - \frac{k}{n_y}\right) \hat{\theta}_y^i(k) + \frac{k}{n_y} \hat{\theta}_y^{i\prime}(k); \\
\hat{\theta}_y^i(k) - \hat{\theta}(k) &= \frac{k}{n_y}\left(\hat{\theta}_y^i(k) - \hat{\theta}_y^{i\prime}(k)\right),
\end{aligned}$$

It follows by (3.34) and (3.35) that

$$S_x^2(k) = \frac{n_x - 1}{n_x} \cdot \frac{X_S(k)}{n_x}, \tag{3.60}$$

$$S_y^2(k) = \frac{n_y - 1}{n_y} \cdot \frac{k^2 Y_S(k)}{n_y}, \tag{3.61}$$

where

$$X_S(k) := \sum_{i=1}^{n_x} \frac{\left(\hat{\theta}_x^i(k) - \hat{\theta}_x^{i\prime}(k)\right)^2}{n_x};$$

$$Y_S(k) := \sum_{i=1}^{n_y} \frac{\left(\hat{\theta}_y^i(k) - \hat{\theta}_y^{i\prime}(k)\right)^2}{n_y}.$$

Furthermore, observe that

$$S^2(k) = S_x^2 + S_y^2 = \frac{n_x - 1}{n_x} \cdot \frac{X_S(k)}{n_x} + \frac{n_y - 1}{n_y} \cdot \frac{k^2 Y_S(k)}{n_y}.$$

By Lemma 3.14, $X_S(k)$ converges in probability to $\xi_{1,0}(k)$ uniformly for $k = 1 : n_y$, which shows (3.57), while similarly $Y_S(k)$ converges in probability to $\xi_{0,1}(k)$ uniformly for $k = 1 : n_y - 1$ which shows (3.58).

To show (3.59), recall $\mathbb{V}(\hat{\theta}_P(k))$ as given by (3.18). Note that by (3.20) of Lemma 3.9, for some $b > 0$ we may substitute $\mathbb{V}(\hat{\theta}_P(k))$ for $\mathbb{V}(\hat{\theta}(k))$ uniformly for $k = 1 : \log_b(\lfloor n_y \rfloor)$ under assumptions (a)-(d), which with (3.57), (3.58) and the Continuous Mapping Theorem, shows (3.59). □

3.6 Case Study: Human Microbiome Project

We use our estimators to analyze data from the Human Microbiome Project. In particular, our samples are V35 16S data, processed by the Quantitative Insights into Microbial Ecology (Qiime) software package [7] into a taxonomic unit count table format. We let environment-z fill the role of urn-z. Each of the 266 samples analyzed have more than 5000 bacteria successfully assigned to an Operational Taxonomic Unit (OTU). We sort these samples by the body location metadata describing the origin of the sample. This sorting yields the assignments displayed in Table 3.1.

Table 3.1: **Summary of V35 16S data from the Human Microbiome Project.**

Body Supersite	Body Subsite	Assigned Labels
Airways	Anterior Nares	1-5
	Throat	6-17
Gastrointestinal Tract	Stool	18-47
Oral	Attached/Keratinized Gingiva	48-59
	Buccal Mucosa	60-76
	Hard Palate	77-90
	Palatine Tonsils	91-112
	Saliva	113-122
	Subgingival Plaque	123-144
	Supragingival Plaque	145-167
	Tongue Dorsum	168-191
Skin	Left Antecubital Fossa	192-195
	Left Retroauricular Crease	196-217
	Right Antecubital Fossa	218-222
	Right Retroauricular Crease	223-242
Urogenital Tract	Mid Vagina	243-248
	Posterior Fornix	249-259
	Vaginal Introitus	260-266

We present our estimates of $\hat{\theta}(n_y)$ for all $266 \cdot 265$ sample comparisons in Figure 3.1. That is we estimate the average dissimilarity of the environment-x relative to the full sample from environment-y. At the given sample sizes we can differentiate four broad groups of environments. Namely, there are stool, vagina, oral/throat and skin/nostril environments. We also differentiate a larger proportion of oral/throat bacteria found in stool than stool bacteria found in the oral/throat environments. At these sample sizes we may differentiate the throat, gingival and saliva samples, but cannot reliably differentiate between tongue and throat or between the subgingival and supragingival plaques. At this level of sampling the stool samples have larger proportions of unique bacteria relative to other stool samples of the same type, as do vaginal samples. In contrast the skin/nostril samples have relatively few bacteria that are not identified in other skin/nostril samples.

These effects may be a property of the environments from which samples are taken, or an effect of noise from inaccurate estimates due to sampling. To rule out the latter interpretation

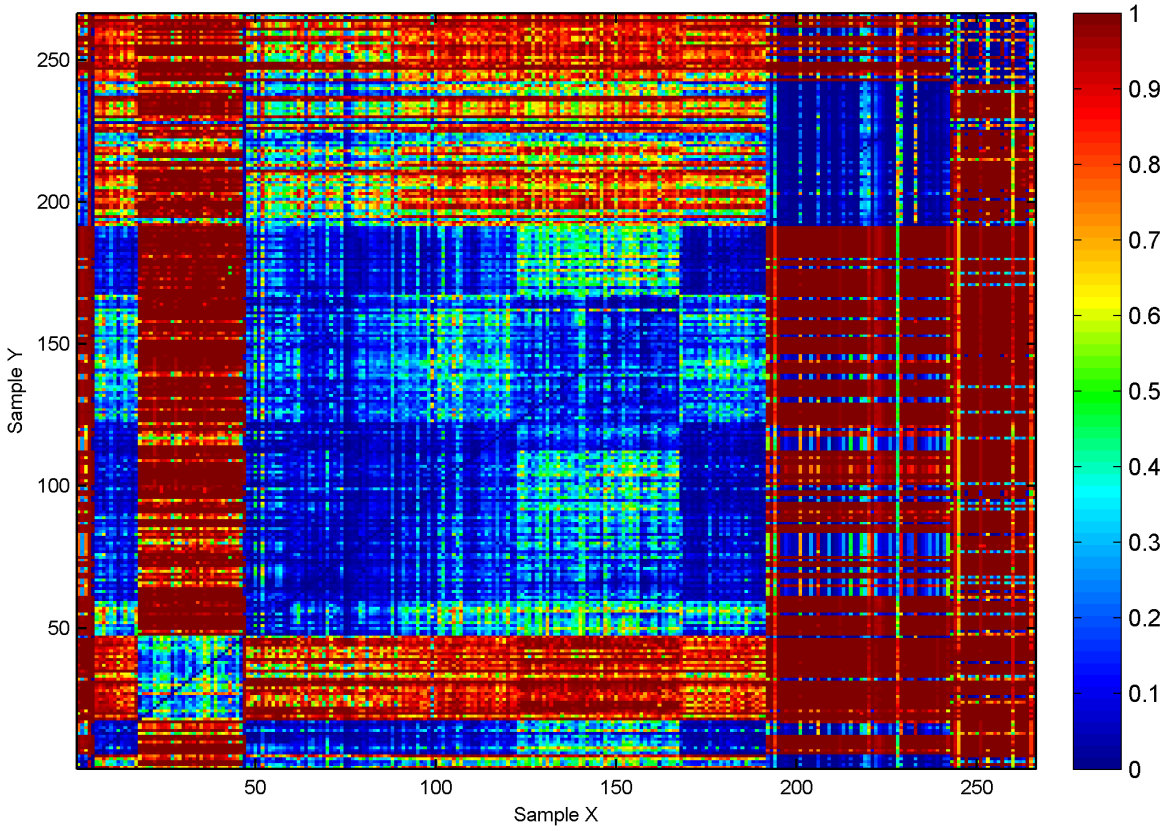

Figure 3.1: **Dissimilarity estimates for Human Microbiome Data.** A heat map of $\hat{\theta}(n_y)$ sorted by metadata. Here the x-axis gives the environment corresponding to urn-x, from which a sample was taken, and similarly the y-axis gives the environment corresponding to urn-y. The entries on the diagonal are set to zero.

we show estimates of the standard deviation of $\hat{\theta}(n_y)$ using the jackknife estimate from (3.36) in Figure 3.2. As $S_y^2(n_y)$ is zero (see 3.25), the error estimate is given by $S_x(n_y)$. We see from (3.39) with $k = n_y$ that this depends only on $Q(0)$ and n_x. As a result, larger $\hat{\theta}(n_y)$ correspond to smaller variance estimates. If we assume that the jackknife estimate of variance is accurate and that $\hat{\theta}(n_y)$ is normally distributed, then the 95% confidence interval $\theta(n_y)$ will be contained in the interval $(\hat{\theta}(n_y) - 0.01, \hat{\theta}(n_y) + 0.01)$ uniformly over any choice of sample comparisons; in particular, on a linear scale the estimates in Figure 3.1 should be accurate to at least 2 decimal places.

Figure 3.3 shows our estimate of the discrete derivative $|\theta(n_y) - \theta(n_y - 1)|$ for each pair of samples. These derivatives are uniformly small, suggesting that the distance between $\theta(n_y)$ and $\theta(\infty)$ is of order 10^{-5} for the majority of the comparisons, while this derivative spikes to

10^{-4} for particular environment-y of varied environment types when environment-x is associated with skin or vaginal samples. In particular, while many environments appear to be sufficiently sampled to saturation, further sampling effort from environments associated with certain vaginal, oral, and stool samples are likely to reveal bacteria associated with broadly defined skin or vaginal environments. Quantifying the gain in information from further sampling, and determining sample strategies based on the estimates of $\theta(k)$ is addressed in Chapter 4.

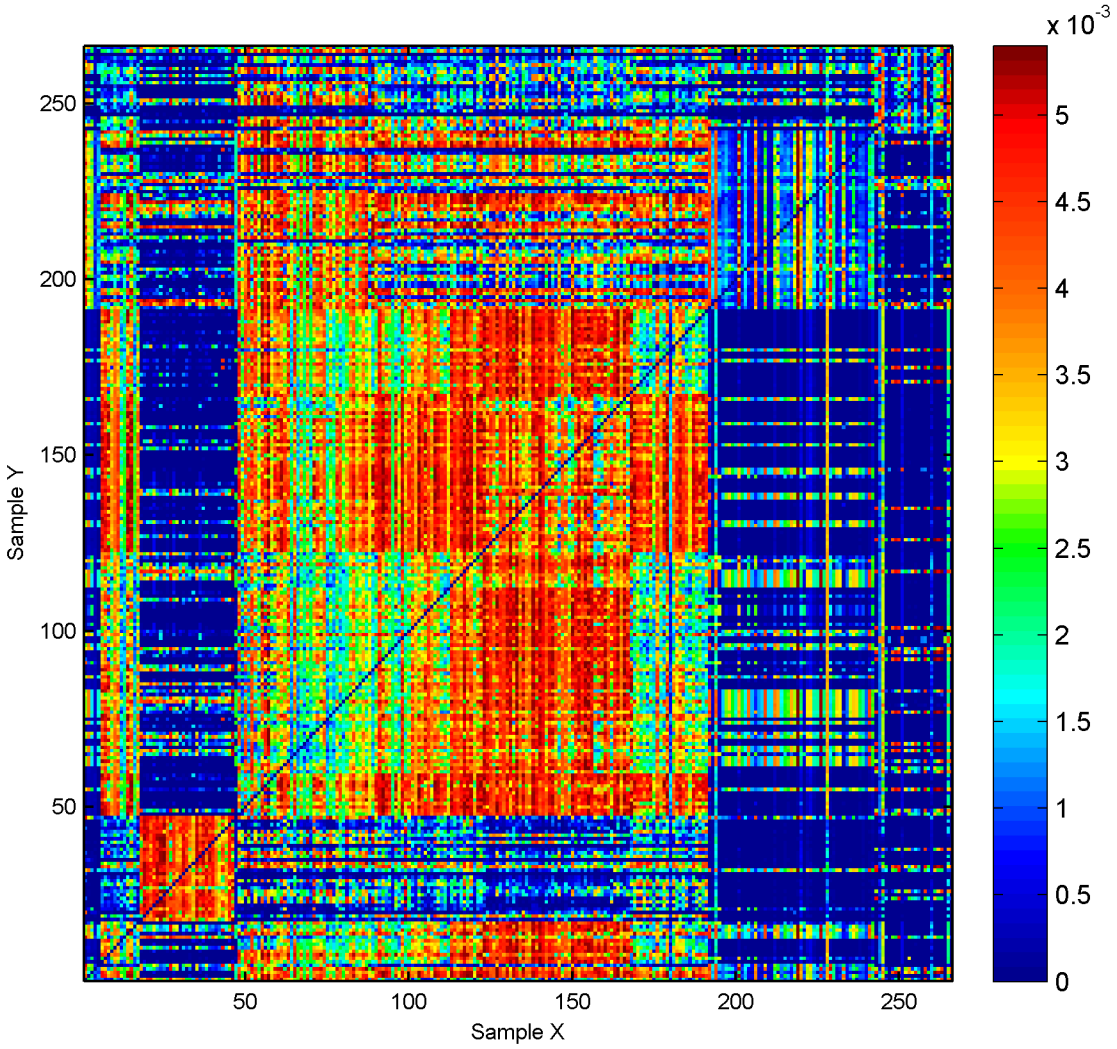

Figure 3.2: **Error estimates for Human Microbiome Data.** A heat map of $S(n_y)$ obtained from $S^2(n_y)$ given in (3.36), sorted by metadata. Here the x-axis gives the environment corresponding to urn-x, from which a sample was taken, and similarly the y-axis gives the environment corresponding to urn-y. The entries on the diagonal are set to zero.

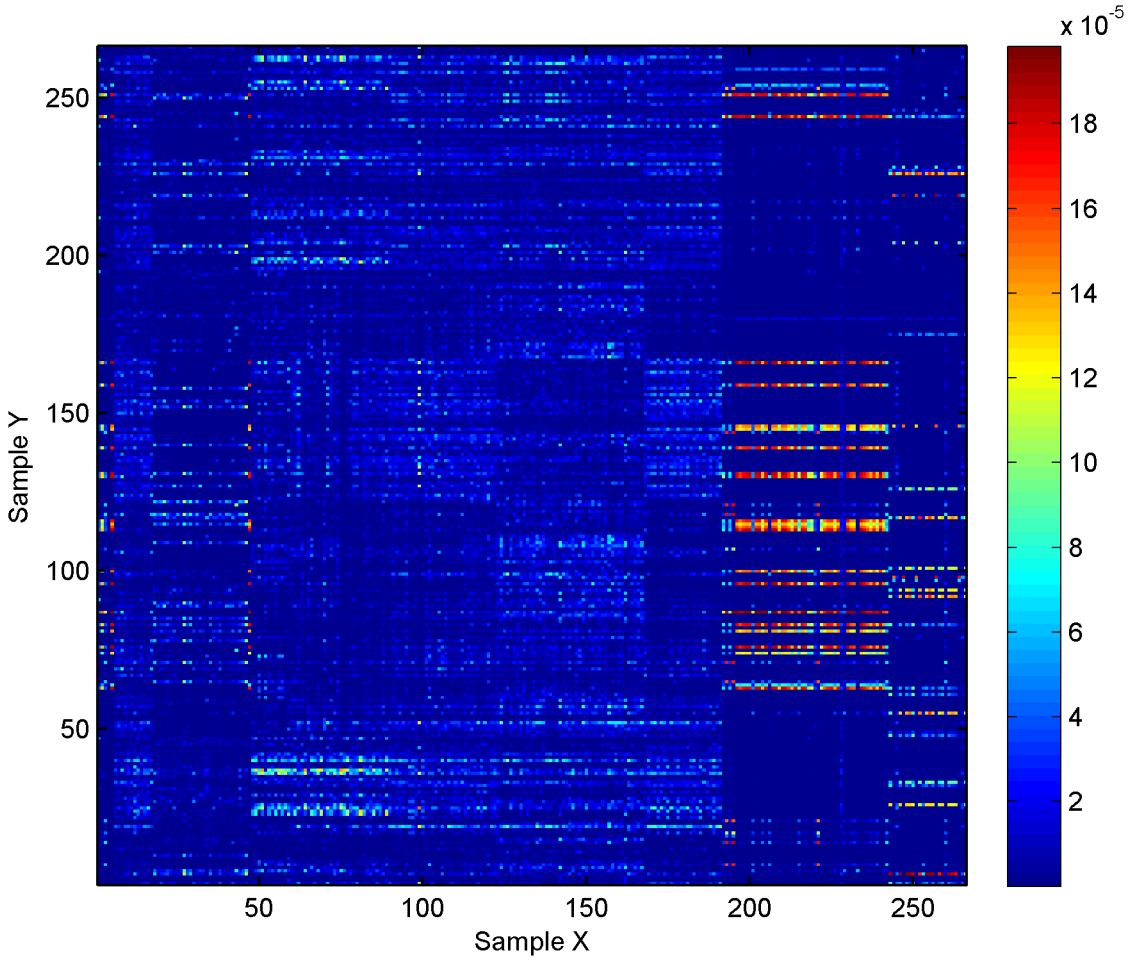

Figure 3.3: **Discrete derivative estimates for Human Microbiome Data.** A heat map of $|\hat{\theta}(n_y) - \hat{\theta}(n_y - 1)|$, sorted by metadata. Here the x-axis gives the environment corresponding to urn-x, from which a sample was taken, and similarly the y-axis gives the environment corresponding to urn-y. The entries on the diagonal are set to zero.

Chapter 4

Optimal Sampling

In this chapter we present new results to form regression functions from the U-statistics used to estimate $\psi(k), \phi(k)$, and $\theta(k)$ as studied in Chapters 2 and 3. In this chapter we assume that all urns are finite, and in particular, $\phi(\infty) < \infty$. For simplicity of presentation, algorithms are presented in terms of ϕ, but they may be used with only minor changes for ψ and θ.

In Section 4.1, we approximate estimates as a function of k with a weighted sum of exponential functions satisfying restrictions given in Table 4.1. In Theorem 4.1 we show that if exponents input into the regression algorithm are asymptotically close to theoretical exponents and the estimated parameter values converge to the theoretical values in a certain sense, then the output regression functions converge uniformly to the functions they approximate. Section 4.2 uses these regression functions as approximations to the theoretical functions to allocate additional sampling resources in a way which is optimal with respect to a given score function. In Section 4.2.1 we describe a method which allows us to reduce error in the allocation as arises from errors associated with estimation or regression. We demonstrate these methods on data from the Human Microbiome Project, as well as simulated data from theoretical urns in Section 4.3.

4.1 Regression

Here we present the methods through which we study our regression and optimization based on the parameters in (1.6)-(1.8). Further, we present a method for individual weighting of parameter functions to control error. For the sake of simplicity, we describe our methods only for the parameter

function ϕ, though analogous results hold by similar arguments for ψ and θ.

Let n be the initial sample size from an urn. We now address the regression of the function $\phi(k)$ in the range $k = 2 : n$ via the U-statistic values, $\hat{\phi}_n(k)$ from (2.4), on the same range. For convenience we add $\hat{\phi}_n(1) = 1$ to the set on which regression is performed. Motivated by the expression in (1.2), we seek a T, and a set of w_i and λ_i to give a function

$$\hat{\phi}_n^R(k) := \sum_{i=1}^{T} w_i e^{\lambda_i k}, \tag{4.1}$$

which is defined for all real $k \geq 0$. Restrictions on the weights w_i and exponents λ_i are suggested by (1.2) and summarized in Table 4.1 with the restrictions suggested by (1.1) and (1.4). We choose T weights, w_i, and exponents, λ_i, to approximately minimize the square of the $\mathcal{L}^2$-distance:

$$e_R := \sum_{k=1}^{n} \left(\hat{\phi}^R(k) - \hat{\phi}_n(k) \right)^2. \tag{4.2}$$

The heuristic here is that the estimated values, $\hat{\phi}_n(k)$, approach the theoretical values, $\phi(k)$, as shown in Theorem 2.12, and our regression function, $\hat{\phi}_n^R(k)$ approaches the estimated values, and as a result so too does our regression function approach the theoretical function.

Table 4.1: **Restrictions in exponential regression.** Given a parameter function to approximate and an exponent in a particular range, the range for the associated weights is given.

Parameter	If	then
$\phi(k)$	$\lambda_i < 0$	$w_i \leq 0$
	$\lambda_i = 0$	$w_i > 0$
	$\lambda_i > 0$	$w_i = 0$
$\psi(k)$	$\lambda_i < 0$	$w_i \geq 0$
	$\lambda_i \geq 0$	$w_i = 0$
$\theta(k)$	$\lambda_i \leq 0$	$w_i \geq 0$
	$\lambda_i > 0$	$w_i = 0$

To describe our regression suppose that we have an ordering for unique exponents, λ_i, and wish to identify weights, w_i, so as to fit $(\hat{\phi}_n(k))_{k=1}^n$. Let A be a matrix with a number of columns equal to the number of exponents, and the number of rows equal to the number of k where estimation of $\phi(k)$ has occurred. We define the entries of A by $A(k, i) = e^{-\lambda_i k}$. Let b be a vector with

$b(k) = \hat{\phi}_n(k)$. Note that we may remove those rows corresponding to k where estimation did not occur. The weights are determined by the vector $\vec{w}$, with $\vec{w}(i) = w_i$. Here $\vec{w}$ is chosen to satisfy

$$w = \arg\min_x \|Ax - b\|_2, \text{ subject to } l \leq x \leq u, \tag{4.3}$$

where $l \leq x \leq u$, means that each element of x is greater than the corresponding element in l, but less than the corresponding element in u. This computation may be completed using the method of Coleman and Li [14]. This constraint allows us to calculate optimal weights satisfying the restrictions on weights and exponents given in Table 4.1.

To state our next result let $\hat{\phi}_n(k)$, with $k = 1 : n$, denote the U-statistic associated with $\phi(k)$ when the number of draws from a certain urn-z is n. The following theorem shows that if the set of exponents used in the calculation of the regression $\hat{\phi}_n^R$ is a good representation of the true exponents implied by (1.2), then the regression function $\hat{\phi}_n^R$ is a consistent estimator of ϕ in the appropriate range. We remark that condition (4.5) in our next result is stronger than that provided by (2.31) in Theorem 2.12. We conjecture that the conclusion of Theorem 2.12 may be strengthened to have (4.5) though certain technical details remain still open to settle this conjecture.

Theorem 4.1. *Assume that I_z is a finite set. Let $\mathcal{E}_n$ be the random set of exponents used to determine regression weights in (4.3), when the number of draws from urn-z is n, and assume that $0 \in \mathcal{E}_n$. Furthermore, assume that for each exponent $\lambda_i := \log(1 - \mathbb{P}_z(i))$, with $i \in I_z$, as implied by (1.2),*

$$\lim_{n\to\infty} \mathbb{P}\left(\min_{s\in\mathcal{E}_n} |s - \lambda_i| > \epsilon\right) = 0, \tag{4.4}$$

and

$$\lim_{n\to\infty} \mathbb{P}\left(\sum_{k=1}^{n} \left|\hat{\phi}_n(k) - \phi(k)\right|^2 > \epsilon\right) = 0, \tag{4.5}$$

for all $\epsilon > 0$. Under these assumptions, it follows that

$$\lim_{n\to\infty} \mathbb{P}\left(\sum_{k=1}^{n} \left|\hat{\phi}_n^R(k) - \phi(k)\right|^2 > \epsilon\right) = 0. \tag{4.6}$$

If in addition to the above, $\left(\hat{\phi}_n^R(n) - \hat{\phi}_n^R(\infty)\right)$ *converges to* 0 *in probability, then*

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{k\geq 1}\left|\hat{\phi}_n^R(k) - \phi(k)\right| > \epsilon\right) = 0. \tag{4.7}$$

Proof. In what follows, we write $\|\cdot\|_2$ to refer to the $\mathcal{L}^2$-norm over an appropriate domain of the form $k = 1 : n$. To show (4.6), define

$$\check{\phi}_n^R(k) := |I_z| - \sum_{i\in I_z} e^{s_{i,n}k},$$

where

$$s_{i,n} := \underset{s\in\mathcal{E}_n, s\neq 0}{\arg\min} |e^s - e^{\lambda_i}|.$$

Due to (4.4), observe that $s_{i,n} \to \lambda_i$ in probability, as $n \to \infty$. Furthermore, because $0 \in \mathcal{E}_n$, the assigning of weights according to (4.3) in $\hat{\phi}_n^R$ leads to a fit of $\hat{\phi}_n$ that is at least as accurate in the $\mathcal{L}^2$-norm as the one given by $\check{\phi}_n^R$, i.e. $\|\hat{\phi}_n^R - \hat{\phi}_n\|_2 \leq \|\check{\phi}_n^R - \hat{\phi}_n\|_2$. In particular we obtain that

$$\begin{aligned}
\|\hat{\phi}_n^R - \phi\|_2 &\leq \|\hat{\phi}_n^R - \hat{\phi}_n\|_2 + \|\hat{\phi}_n - \phi\|_2, \\
&\leq \|\check{\phi}_n^R - \hat{\phi}_n\|_2 + \|\hat{\phi}_n - \phi\|_2, \\
&\leq \|\check{\phi}_n^R - \phi\|_2 + 2\|\hat{\phi}_n - \phi\|_2.
\end{aligned}$$

Due to (4.5), $\|\hat{\phi}_n - \phi\|_2 \to 0$ in probability as $n \to \infty$. To complete the proof of (4.6), it suffices therefore to show that $\|\check{\phi}_n^R - \phi\|_2 \to 0$ also in probability i.e. that

$$\lim_{n\to\infty} \mathbb{P}\left(\sum_{k=1}^{n}\left(\check{\phi}_n^R(k) - \phi(k)\right)^2 > \epsilon\right) = 0, \tag{4.8}$$

for all $\epsilon > 0$. Indeed, observe that

$$\sum_{k=1}^{n}(\check{\phi}_n^R(k) - \phi(k))^2 \leq \sum_{k=1}^{n}\left(\sum_{i\in I_z}\left|e^{s_{i,n}k} - e^{\lambda_i k}\right|\right)^2 \leq C_n^2 \cdot |I_z|^2 \cdot \sum_{k=1}^{n}\frac{1}{k^2} \leq \frac{C_n^2|I_z|^2\pi^2}{6},$$

where

$$C_n := \max_{i\in I_z,\, k\geq 1} k \cdot \left|e^{s_{i,n}k} - e^{\lambda_i k}\right|.$$

Since $|I_z| < \infty$, to show (4.8), it suffices therefore to show that C_n converges to 0 in probability. But recall that $s_{i,n} \to \lambda_i$ in probability, as $n \to \infty$. In particular, since $\lambda_i < 0$ for each i in the finite set I_z, the event "$3\lambda_i/2 \leq s_{i,n} \leq \lambda_i/2$, for all $i \in I_z$" has asymptotic probability one. From this, a short calculation shows that

$$\lim_{n\to\infty} \mathbb{P}\,(A_n) = 1, \tag{4.9}$$

where A_n is the event defined as "$k|e^{s_{i,n}k} - e^{\lambda_i k}| \leq 2k \exp\{k \min_{i\in I_z} \lambda_i/2\},$ for all $k \geq 1$ and $i \in I_z$". Given $\delta > 0$, choose an integer $K \geq 1$ such that $2k \exp\{k \min_{i\in I_z} \lambda_i/2\} < \delta$, for all $k > K$, which is possible because $\min_{i\in I_z} \lambda_i < 0$. It follows that

$$\mathbb{P}(C_n \geq \delta) \leq \mathbb{P}(A_n^c) + \sum_{i\in I_z} \sum_{k=1}^{K} \mathbb{P}(|k\, e^{s_{i,n}k} - k\, e^{\lambda_i k}| \geq \delta).$$

Due to (4.9), the first term on the right-hand side of the above inequality tends to 0 as $n \to \infty$. On the other hand, $k\, e^{s_{i,n}k} \to k\, e^{\lambda_i k}$ in probability, for all $i \in I_z$ and $k = 1 : K$, because of the Continuous Mapping Theorem. Since $|I_z| < \infty$, the second term on the right-hand side above also converges to 0. Hence $C_n \to 0$ in probability as $n \to \infty$, which shows (4.8) and hence also (4.6).

To complete the proof of the theorem, observe that as $\hat{\phi}_n^R$ and ϕ are bounded and increasing functions,

$$\max_{k>n} |\hat{\phi}_n^R(k) - \phi(k)| \leq |\hat{\phi}_n^R(n) - \phi(n)| + |\hat{\phi}_n^R(n) - \hat{\phi}_n^R(\infty)| + |\phi(n) - \phi(\infty)|.$$

Note that the first term on the right-hand side of the inequality above converges to 0 in probability because of (4.6). On the other hand, the second term converges to 0 by assumption. Since the third term converges to 0 as well, we conclude that $\max_{k>n} |\hat{\phi}_n^R(k) - \phi(k)| \to 0$ in probability as $n \to \infty$, which shows (4.7) and completes the proof of the theorem. □

We use three methods to identify useful exponents to be used in determining weights. We are willing to identify more exponents than necessary as extraneous exponents will receive a weight of 0. However, extra exponents do increase computational overhead for the determination of weights. The first method uses the $\hat{\phi}_n(k)$ to identify exponents which explain the data, and is useful in identifying exponents near zero. The second method uses exponents derived from an estimated distribution

of the urn, which is useful in identifying exponents further from 0, and provides the asymptotic convergence of estimated exponents to the actual exponents as required by (4.4) in Theorem 4.1. The third method maps the range over which estimation points are used for regression into usable exponents.

In more detail, the first method follows from Osborne and Smyth [49], and uses the $\hat{\phi}_n(k)$ to find exponents which explain data. We let this method be given by OsborneSmyth($\hat{\phi}$). One problem with this method is that returned exponents may not satisfy our restrictions placed on the exponents. Another is that the exponents may be useful in fitting the data only when weights do not satisfy the imposed restrictions. Further the algorithm may be susceptible to numerical instability when finding a large number of exponents. We utilize this method to identify the first seven exponents which best explain the data, but one may run this algorithm iteratively until an inappropriate exponent appears according to the restrictions in Table 4.1.

The second method estimates the distribution of the urn to infer exponents. In (1.2) each $\log(1 - \mathbb{P}(c))$ is an exponent in the exponential decomposition of $\phi(k)$. Using observed data we estimate $\mathbb{P}(c)$ for each c. For simplicity, we estimate $\mathbb{P}(c)$ by

$$\hat{\mathbb{P}}_n(c) := \frac{(1-p)}{n} \sum_{i=1}^{n} [\![X_i = c]\!],$$

where p is the estimate of unobserved probability in the sample as given by Good [23]. We define the estimates of the set of c by $\hat{\mathbb{P}}_n$, which estimates the proportion of every observed color, but is not a distribution as

$$\sum_{c} \hat{\mathbb{P}}_n(c) \leq 1.$$

The remaining probability may be assigned to one or several unobserved colors, but we do not do so. Given $\hat{\mathbb{P}}_n(c)$, the associated estimated exponent is given by $\log(1 - \hat{\mathbb{P}}_n(c))$. As $\hat{\mathbb{P}}_n(c)$ converges to $\mathbb{P}(c)$ in probability, uniformly in c, it follows that condition (4.4) of Theorem 4.1 is satisfied by the use of this method for including exponents. However this method fails to accurately identify colors with small proportions in the urn, and thus fails to accurately identify exponents near zero.

The third method uses the fact that in k draws from an urn a given color c is expected to be seen $k\mathbb{P}(c)$ times. In particular, if $\mathbb{P}(c) = (1 - e^{-1/k})$ then color c is expected to be seen approximately once at a sample depth of k. Equivalently, $\log(1 - \mathbb{P}(c))$ is approximately $-k^{-1}$. This relation, and the range of k for which we have data, suggest a set of exponents for regression, given by $-k^{-1}$, associated with colors that are not always observable. Another heuristic justifying this method is that it assists in keeping the mesh of exponents from becoming too coarse. In our algorithm, we choose twenty evenly spaced k between 1 and n, from which we form exponents, $-k^{-1}$. Although this method does not take into account any information contained in the data, and has the drawback that a fine mesh of exponents leads to a large number of exponents and therefore significant computational effort in the identification of weights, it appears to perform well in practice.

Recall that given the exponents from these three methods, we may determine weights by performing the optimization in (4.3) using the method of Coleman and Li [14]. We summarize this computation by the function lsqlin() which optimizes within provided constraints.

$$\text{lsqlin}(A, b, l, u, C, d) := \arg\min_x \|Ax - b\|_2, \text{ subject to } l \leq x \leq u \text{ and } Cx = d. \tag{4.10}$$

We often do not require the $Cx = d$ condition. In such cases the optimization is only a function of A, b, l and u. The method by which we identify the exponents and weights to form the function $\hat{\phi}_n^R$, is given by the algorithm in Table 4.2.

4.2 Optimal Allocation of Draws

We now analyze an ensemble of r urns, with data $X_j(1), \ldots, X_j(n_j)$ denoting the sample of size n_j from urn-j for $j = 1 : r$. Let $\hat{\phi}_j^R(k)$, be the regression function calculated by the Algorithm in Table 4.2 using data from urn-j. These functions are used to allocate m additional draws amongst each of the r urns. We may use the $\hat{\phi}_j^R$ to extrapolate estimates for each k satisfying $n_j \leq k \leq n_j + m$. We would like to choose m_j satisfying $\sum_{j=1}^{r} m_j = m$ to minimize the score function

Table 4.2: **Regression algorithm:** The regression function, $\hat{\phi}_n^R$, is determined by the exponents $\vec{\lambda}$, the weights $\vec{w}$ and (4.1).

Input:	$(\hat{\phi})$ is the set of estimates for $\phi(k)$ for $k = 1, \cdots, n$. $(\hat{\mathbb{P}})$ is the estimated proportions of colors in the urn.
Output:	$\vec{w}$ is the set of weights for the regression function $\hat{\phi}_n^R$. $\vec{\lambda}$ is the set of exponents for the regression function $\hat{\phi}_n^R$.
1	$\mathcal{S} = \{0\}$ % 0 is a necessary exponent, used to approximate $\lvert I_j \rvert$ as in (1.2).
2	$\mathcal{S} = \mathcal{S} \cup$OsborneSmyth$(\hat{\phi})$ % Add negative exponents returned from Osborne-Smyth().
3	**For all** c **such that** $\hat{\mathbb{P}}(c) > 0$
4	$\mathcal{S} = \mathcal{S} \cup \log(1 - \hat{\mathbb{P}}(c))$.
5	**End For loop**
6	$\mathcal{S} = \mathcal{S} \cup \{-\frac{n}{20}, -\frac{2n}{20}, \ldots, -\frac{20n}{20}\}$.
7	Let $\vec{\lambda}$ order $\mathcal{S}$ in decreasing order. % Note that the first exponent in $\vec{\lambda}$ is 0.
8	$l = (0, -\infty, \cdots, -\infty)$ % $\lambda = 0$, has lower bound of 0, other exponents have no lower bound.
9	$u = (\infty, 0, \cdots, 0)$ % $\lambda = 0$, has no upper bound, other exponents have no upper bound.
10	$A(k, i) = e^{k\lambda_i}$ % A is a matrix with n rows, and a number of columns equal to the size of $\vec{\lambda}$.
11	$\vec{w} =$lsqlin$(A, \hat{\phi}, l, u)$ % Weights are determined from the $\hat{\phi}(k)$, and the exponents chosen.

given by

$$S_\phi(\vec{m}) := \sum_{j=1}^{r} \hat{\phi}_j^R(n_j + m_j). \tag{4.11}$$

Our optimization algorithm identifies $\vec{m}$ using the algorithm given in Table 4.3

Table 4.3: **Optimization algorithm:** The optimal allocation of draws for a subsequent sample of m draws from the ensemble of urns are determined one at a time.

Input:	$(\hat{\phi}_j^R)$ contains the regression functions which approximate each ϕ_j.
Output:	$\vec{m}$ contains the allocated draws m_j for each urn-j.
1	$\vec{m} = \vec{0}$ % We begin with no draws assigned and assign them sequentially.
2	**While** $\sum m_j < m$ **do**
3	$j = \arg\max_{j=1:r} \left(\hat{\phi}_j^R(n_j + m_j + 1) - \hat{\phi}_j^R(n_j + m_j) \right)$.
4	$m_j = m_j + 1$ % The urn associated with the maximal discrete derivative will be incremented.
5	**End While loop**

Note that our optimization may be applied with few changes to several other functions.

As an example we may replace ϕ_j with $\log(\phi_j)$ in (4.11) which maximizes a multiplication of the $\phi_j(n_j + m_j)$ in place of a sum. The algorithm in Table 4.3 may be altered to produce the optimal allocation under any S satisfying three sufficient conditions. First, S is monotonic in each m_j, insuring that the effect on S of incrementing m_j does not change sign. Second, $|\partial S/\partial m_j|$ is decreasing for each j, so that each increment of m_j leads to a smaller change in the magnitude of S than the previous increment to m_j. Finally, $\partial^2 S/\partial m_i \partial m_j = 0$ for $i \neq j$, so that changes to S from incrementing m_i do not affect changes to S in regards to m_j.

4.2.1 Parameter Weighting

For given data, estimation error may be unacceptably large. Suppose that we have estimates for the variance of $\hat{\phi}_j(k)$ in estimating $\phi_j(k)$ given by $V_j(k)$. Suppose that we wish to weight regression functions such that those with large error are weighted less heavily than those with small error. Choose a $0 < \delta < 1$. Let $\vec{p}$ be such that p_j corresponds to the weights of each estimated function. We choose $\vec{p}$ to minimize

$$S_\delta := \sum_{j=1}^{r} p_j \sum_{k=1}^{n_j} V_j(k), \tag{4.12}$$

subject to the constraints

$$\sum_{j=1}^{r} p_j = 1;$$

$$r^{-1} - \delta \leq p_j \leq r^{-1} + \delta \text{ for all } j,$$

The first constraint normalizes the p_j, while the second constrains deviation from uniform weighting. This calculation minimizes the average variance across all estimates that are used in variance calculation, subject to any two arbitrary weights not differing by more than 2δ. Let

$$v_j := \sum_{k=1}^{n_j} V_j(k).$$

It follows from (4.12) that we seek to minimize the inner product between $\vec{p}$ and $\vec{v}$ with $\vec{v}(j) = v_j$. Let $\mathbf{1}$ be an $r \times 1$ vector with 1 as every entry. Finding $\vec{p}$ is equivalent to solving

$$\vec{p} = \arg\min_x \vec{v}^T x, \text{ subject to } (r^{-1} - \delta)\mathbf{1} \leq x \leq (r^{-1} + \delta)\mathbf{1} \text{ and } \mathbf{1}^T\vec{p} = 1.$$

This optimization may be accomplished as follows. Without loss of generality we may order the urns such that $v_1, \leq, \ldots, \leq v_r$. Let $p_1, \ldots, p_{\lfloor r/2 \rfloor}$ each equal $(r^{-1} + \delta)$, and let $p_{\lceil (r+2)/2 \rceil}, \ldots, p_r$ each equal $(r^{-1} - \delta)$. In the case that r is odd, define $p_{(r+1)/2}$ to be r^{-1}.

4.3 Sample Allocation

4.3.1 Case Study: Human Microbiome Project

We apply our methods to skin, stool, oral, and vaginal samples from V13 16S data associated with the Human Microbiome Project [51, 50, 63, 15, 16]. Our ensemble consists of one sample of each type, giving $r = 4$ urns representing environments from which each sample was taken.

We selected the above environments to represent broad microbiomes, with samples chosen to be those with the highest number of successfully identified organisms (OTUs) within that environment type. There are 42999 identified organisms in the skin sample, 20165 in the stool, 37081 in the oral, and 20702 in the vagina sample. These sample sizes correspond to the n_j which determine the size of our initial samples.

In Figure 4.1, we see estimates, regressions and extrapolations for ϕ_j, ψ_j, and θ_j with $j = 1 : 4$. Here, the estimation of $\theta_j(k)$ is accomplished by averaging over each estimate of $\theta_{i,j}(k)$ as studied in Chapter 3. The regression is performed using 20% of potential estimated values, evenly spaced. We see the estimated relative error of parameter estimates in Figure 4.2 calculated using the square-root of a delete-1 jackknife [18, 57, 58] estimation of the variance presented in Sections 2.5.1 and 3.5 at each point, divided by the appropriate parameter estimate. That is for each choice of λ, we plot $S_\lambda(k)/\hat{\lambda}(k)$. In the case of θ, this variance estimation assumes that the $\theta_{i,j}(k)$ are not correlated with respect to differing i. Note the different curve orderings and shapes, demonstrating that the choice of parameter has a significant effect on the measure of sample quality in each urn. Further

these estimates suggest that $\psi_j(k)$ has the least accurate estimations. Note that the relative error in both of $\phi_j(k)$ and $\psi_j(k)$ increases with k.

We consider allocations of draws for a subsequent sample of up to one million draws ($m = 10^6$) from these four environments, with m_j draws allocated to urn-j. The allocation is chosen to maximize $\sum \hat{\phi}_j^R(n_j + m_j)$, or minimize one of $\sum \hat{\psi}_j^R(n_j + m_j)$ or $\sum \hat{\theta}_j^R(n_j + m_j)$, subject to the constraint that $\sum m_j = m$. These allocations are displayed in Figure 4.3. We see that stool and vagina environments are the two environments with appreciable sampling for smaller m. Depending on the measure, oral or skin environments may receive a significant portion of samples for larger m. If we assume that the error in extrapolation along a regression curve increases with k, then the reliability of draw allocation decreases as m increases, in particular when m is large compared to each n_j. Still, we may conclude that of these four environments the stool and vagina environments are the best candidates to receive further sampling, where exact proportions depend on the measure of interest. If m is large then further sampling from the oral or skin environment may be considered, depending on which measure is used, though it is unclear how close the computed allocations are relative to the optimal allocations.

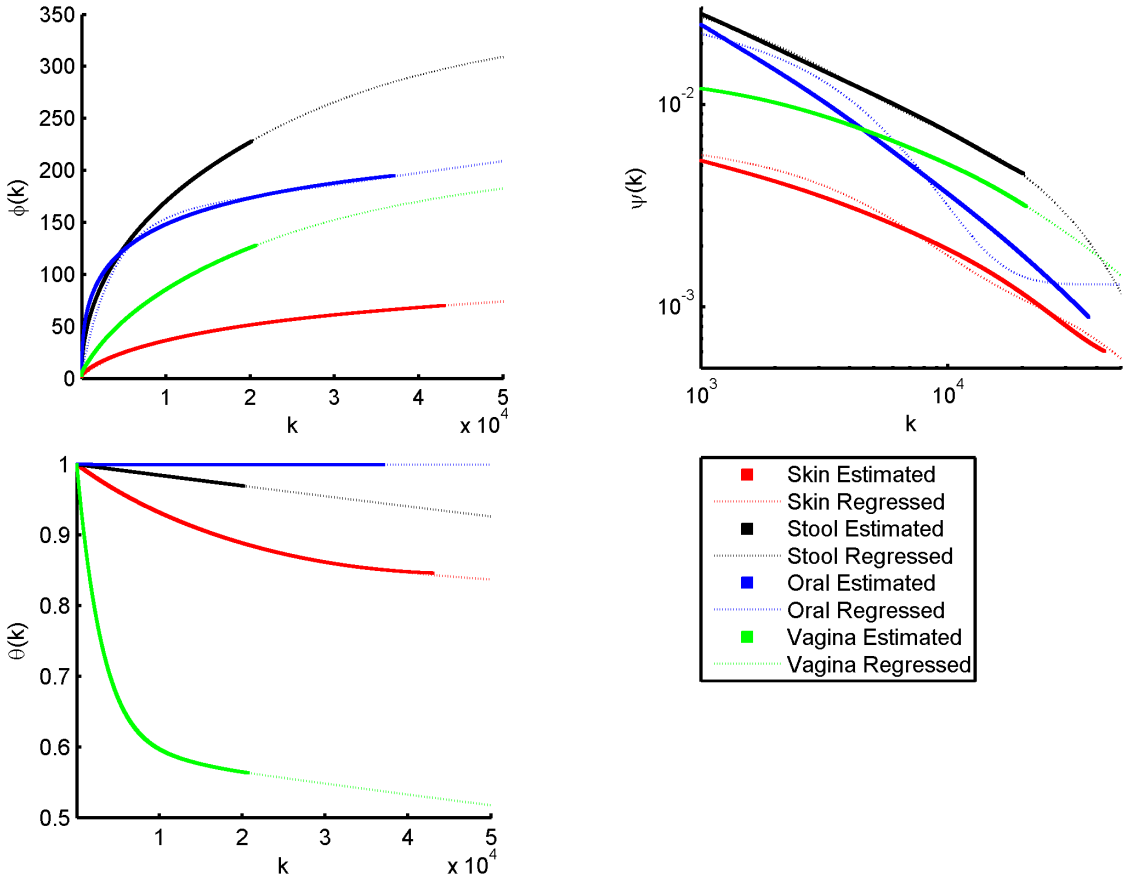

Figure 4.1: **Estimates, regressions and extrapolations for Human Microbiome Data.** The top left shows output of our methods for approximating $\phi(k)$, the top right for $\psi(k)$, and the bottom left for $\theta(k)$. The U-statistic estimates are calculated for k in increments of 5 starting from 1. The plots for $\psi(k)$ are in a log-log scale to better appreciate differences between the environments.

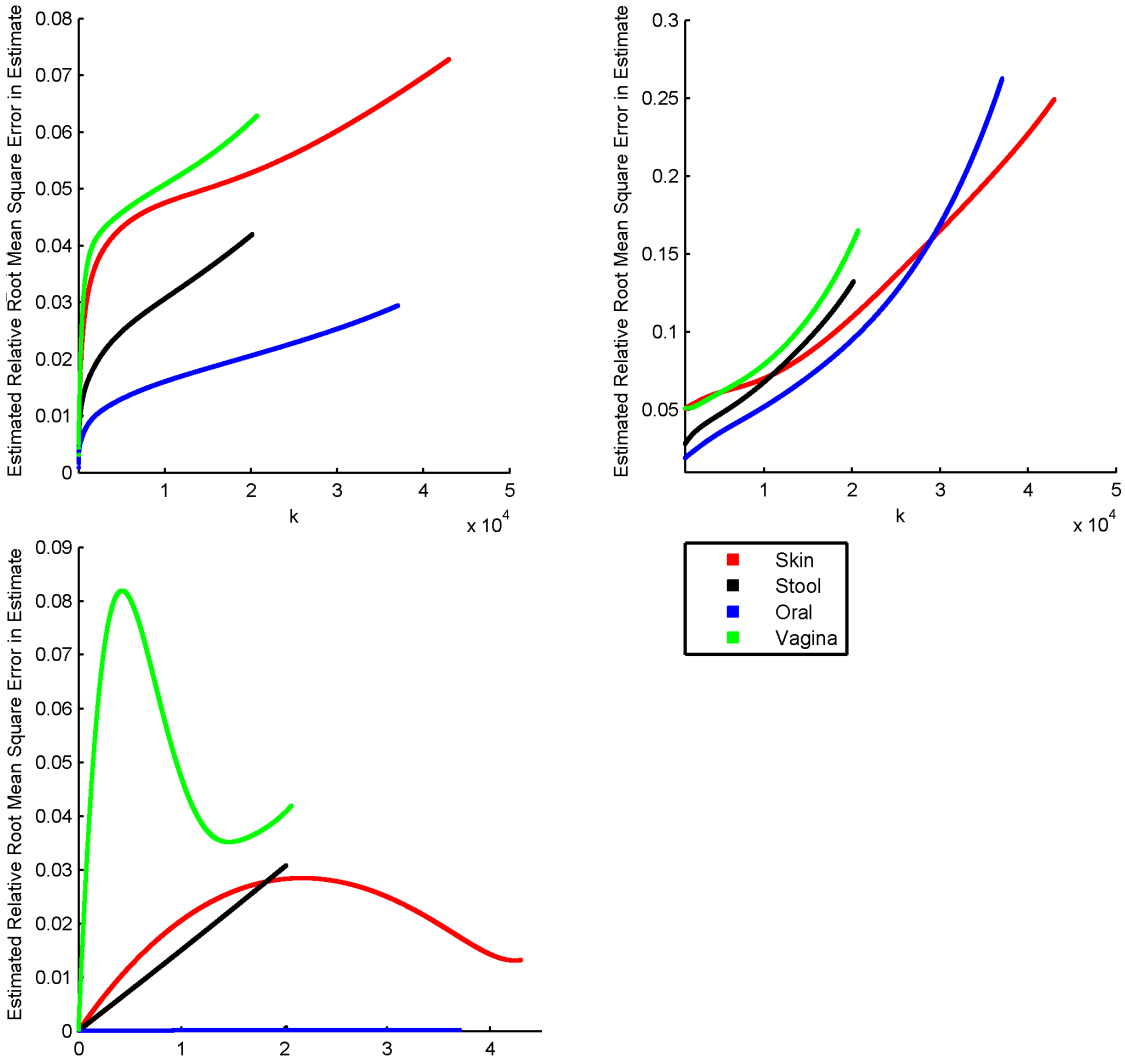

Figure 4.2: **Relative Error Estimates for Human Microbiome Data.** The top left shows root mean square error estimates relative to the U-statistic estimate for $\phi(k)$, the top right for $\psi(k)$, and the bottom left for $\theta(k)$, calculated by delete-1 jackknifes. The error estimates are calculated for k in increments of 5 starting at $k = 1$.

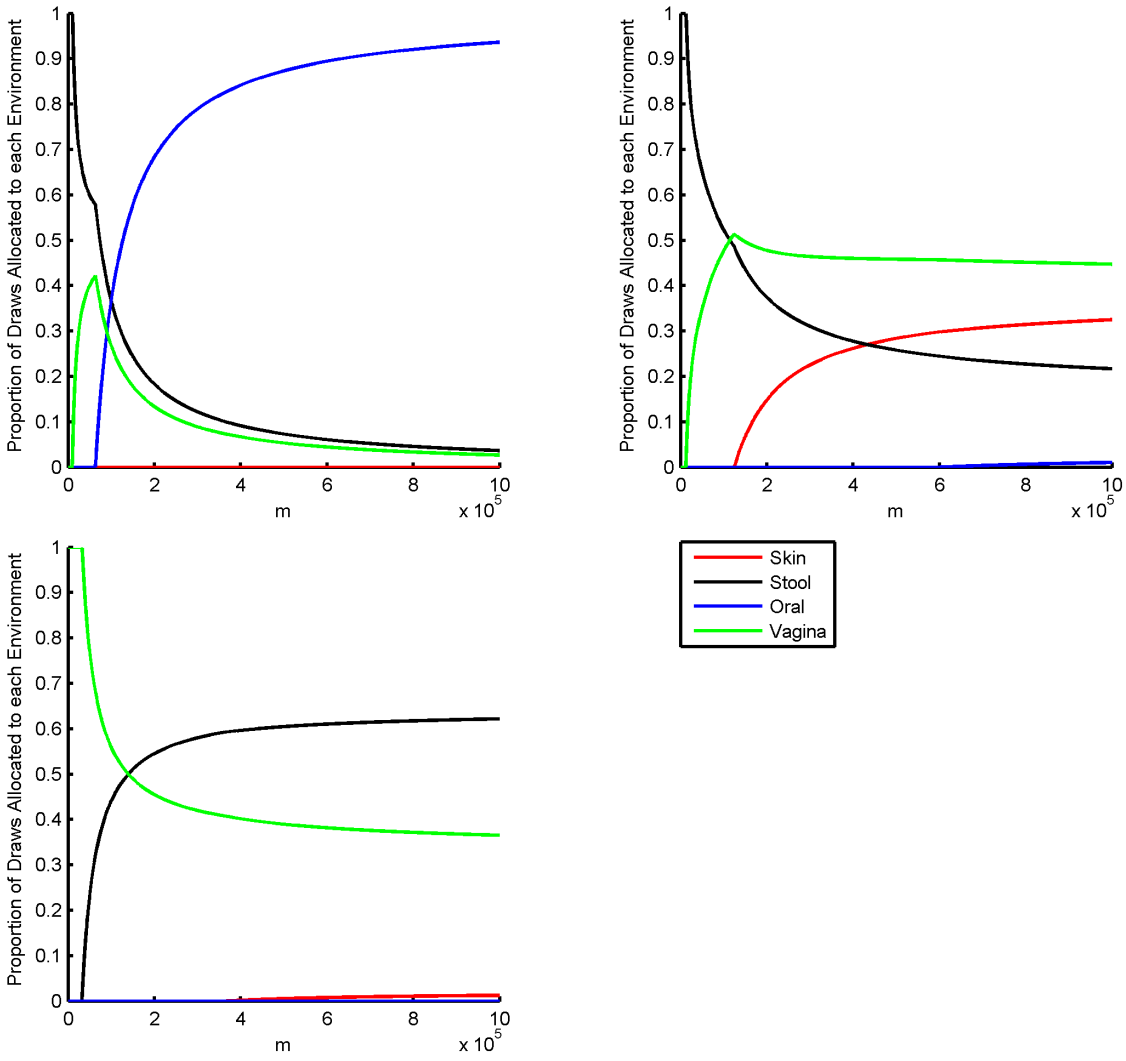

Figure 4.3: **Estimated Sample Allocations for Human Microbiome Data.** Here we see estimated optimal sample allocations for a subsequent sample. The top left shows allocations to maximize $\sum \hat{\phi}_j^R(n_j+m_j)$, the top right to minimize $\sum \hat{\psi}_j^R(n_j+m_j)$, and the bottom left to minimize $\sum \hat{\theta}_j^R(n_j+m_j)$.

4.3.2 Case Study: Theoretical Urns

Now we explore the results of our algorithms on a set of three theoretical urns, chosen to demonstrate the differences among measures. First we discuss optimal urn allocations using perfect information about the urns. Then we discuss estimation, regression, and draw allocation using draws with replacement from the urns, comparing results to those which would be obtained with perfect information.

As a theoretical example on which to test our methods we have a set of three urns, that is $r = 3$. The supports and distributions of these urns are given in Figure 4.4.

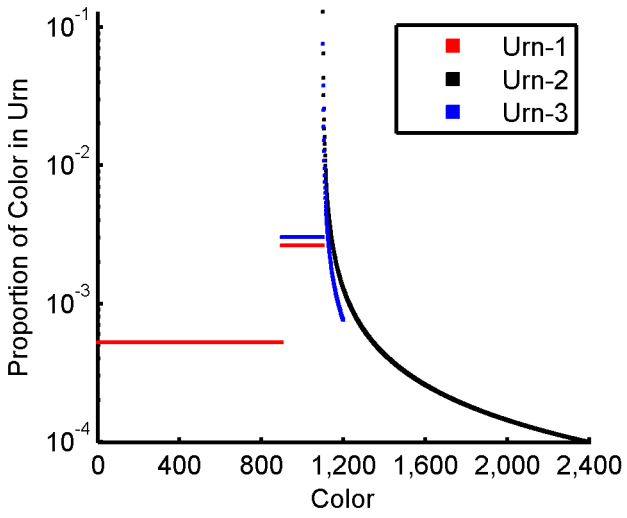

Figure 4.4: **The theoretical urn distribution for three urns.** Points of the appropriate color are plotted at those colors and proportions where the corresponding urn is represented with positive probability. The y-axis is displayed in log-scale to better display differences in the urn distributions.

Notice how Urn-1 and Urn-2 have no colors in common, and how the composition of all three urns differs, both in shape of the distribution and number of colors contained. These urns are designed to give qualitatively different results for draws assigned according to the parameters in (1.6), (1.7), and (1.8). We begin our discussion without reference to estimation. Instead we calculate each of these parameters exactly, as displayed in Figure 4.5.

Further, we may calculate the limits of each parameter as a function of k. Recall that

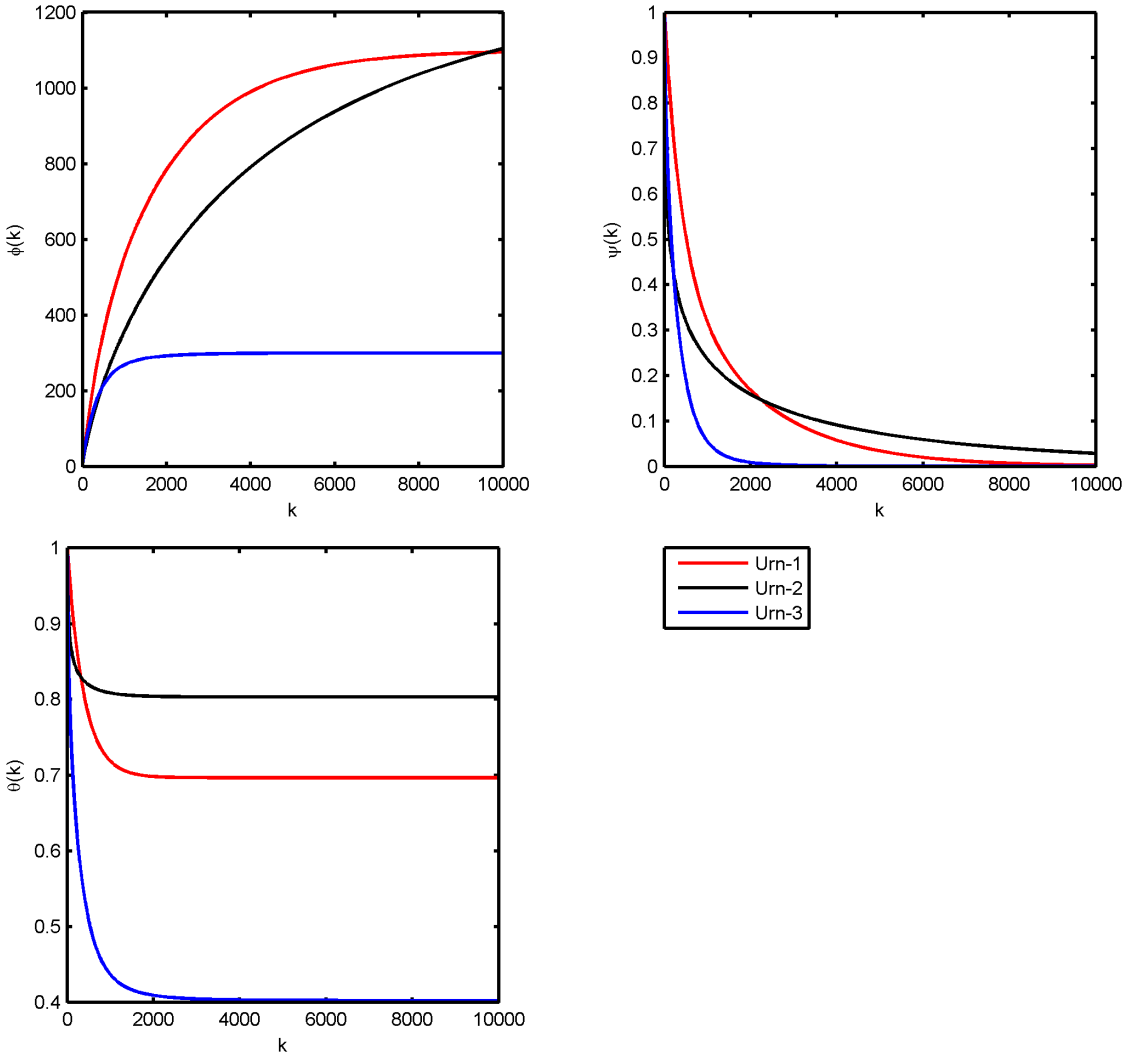

Figure 4.5: **The theoretical functions associated with three urns.** Here we see theoretical parameter values for each urn. The top left shows $\phi(k)$, the top right shows $\psi(k)$, and the bottom left shows $\theta(k)$.

$\psi(\infty) = 0$ for any urn, and the other limits are given by

$$\phi_1(\infty) = 1100, \qquad \phi_2(\infty) = 1300, \qquad \phi_3(\infty) = 300,$$

$$\theta_1(\infty) \approx 0.6967, \qquad \theta_2(\infty) \approx 0.8033, \qquad \theta_3(\infty) \approx 0.4021.$$

For our three urns suppose that we have an initial sample of n_j draws from each urn-j, and m draws to assign between the three urns in a second sample. Let $\vec{m} = (m_1, m_2, m_3)$ count the draws assigned to each urn, constrained such that $m_1 + m_2 + m_3 = m$, and $m_j \geq 0$. We define score

functions for the assignments by

$$S_\phi(\vec{m}) := \sum_{j=1}^{3} \phi_j(n_j + m_j); \qquad S_\psi(\vec{m}) := \sum_{j=1}^{3} \psi_j(n_j + m_j); \qquad S_\theta(\vec{m}) := \sum_{j=1}^{3} \theta_j(n_j + m_j).$$

Suppose first that we wish to allocate draws optimally without any initial sample, that is when each $n_j = 0$. Assume that we have m draws to assign and are to use one of these score functions to determine a sampling allocation by maximizing $S_\phi(\vec{m})$ or minimizing one of $S_\psi(\vec{m})$ or $S_\theta(\vec{m})$ over the set of admissible $\vec{m}$. Using the theoretical curves shown in Figure 4.5 leads to allocation schemes as shown in Figure 4.6.

We see that different score functions lead to different decisions concerning sample allocation. In mathematical terms, these decisions are based on discrete derivatives. For example, if we seek to allocate a sample of one draw which will maximize S_ϕ we need only identify the j which maximizes the difference between $\phi_j(n_j+1)$ and $\phi_j(n_j)$. The derivatives for each score are given in Figure 4.7. Notice the exponential decay of these derivatives, implying that for larger $n_j + m_j$, more accurate derivative estimates are necessary for accuracy in optimal draw allocation.

4.3.3 Regression and Optimal Allocation from Data

The previous discussion involves decisions based on perfect knowledge of the urns in question. In applications we use data to estimate these curves, and as such the error in our estimations will affect our decisions. The monotonicity and smoothness of the parameters and their derivatives imply that given an accurate estimation over a certain range, we may extrapolate in k along the constrained regression which approximates estimated parameter values. The accuracy of these regression extrapolations for large k may not be necessary for practical applications as all samples, initial and subsequent, are finite. Finally, we optimize our score functions through the use of these regression functions.

The error between the regression functions and the theoretical functions they estimate depends on the data, urns, the statistics used in estimation, and the methods used for regression. We assume an initial sample of $n_j = 1,000$ draws from each urn. Using this data we estimate

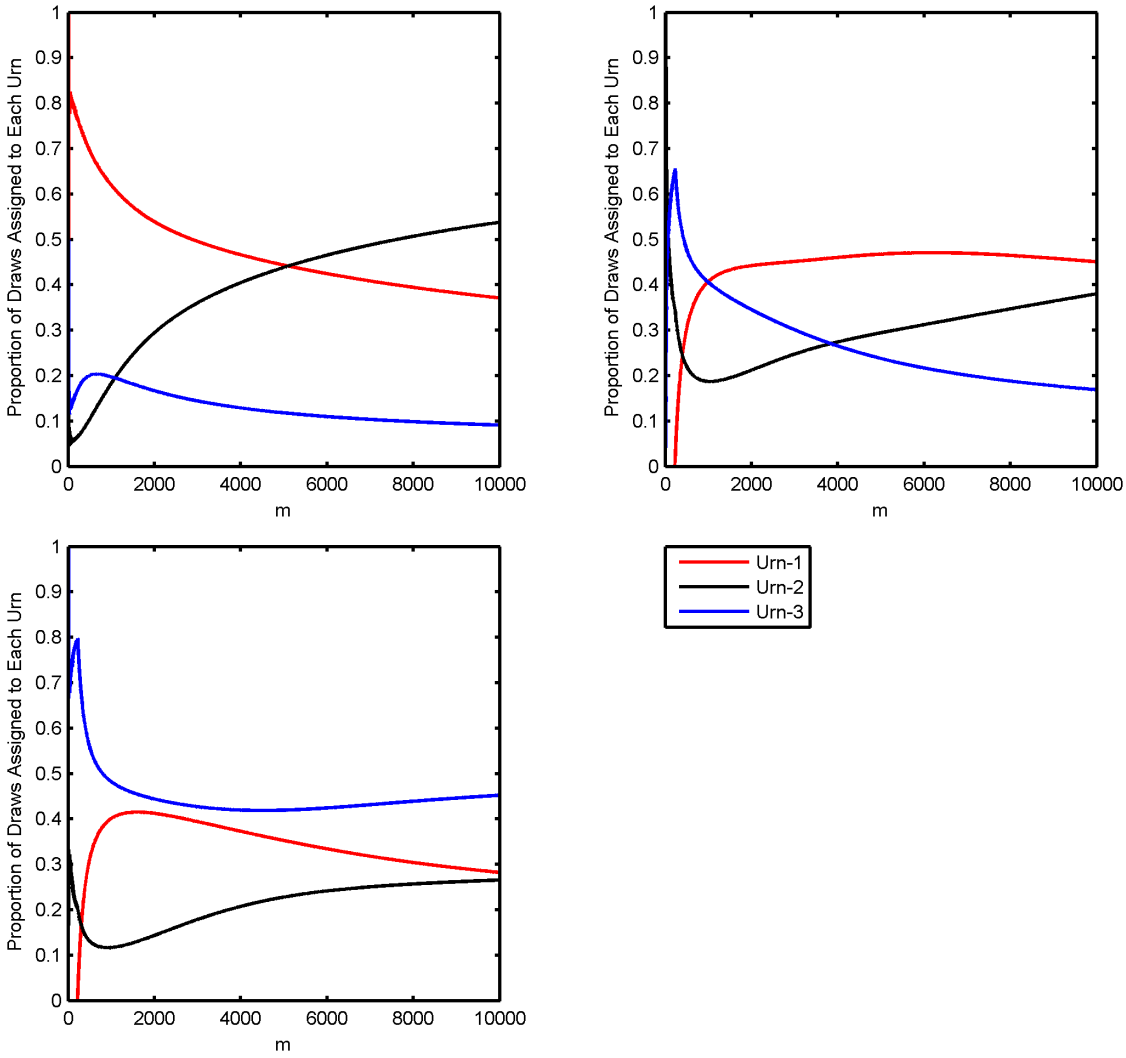

Figure 4.6: **The optimal urn allocations when each** $n_j = 0$. The top left shows allocations which maximize $S_\phi(\vec{m})$, the top right those which minimize $S_\psi(\vec{m})$, and the bottom left those which minimize $S_\theta(\vec{m})$.

parameter functions by the appropriate U-statistic where the estimates are defined. Specifically we form $\hat{\phi}_j(k)$ and $\hat{\theta}_j(k)$ for $k = 1 : n_j$ and $\hat{\psi}_j(k)$ for $k = 1 : n_j - 1$. We perform regressions on these estimates to form functions $\hat{\phi}_j^R(k)$, $\hat{\psi}_j^R(k)$, and $\hat{\theta}_j^R(k)$. We use these regressed functions to identify a sampling allocation. In Figure 4.8 we show for each measure, the bias between the regression function and the true parameter value, relative to the true parameter value. As the estimates we use are unbiased, the regression has small bias in the range where estimates are available. The bias grows further from this range, with rates that are linear in the case of regressions of ϕ and θ. The

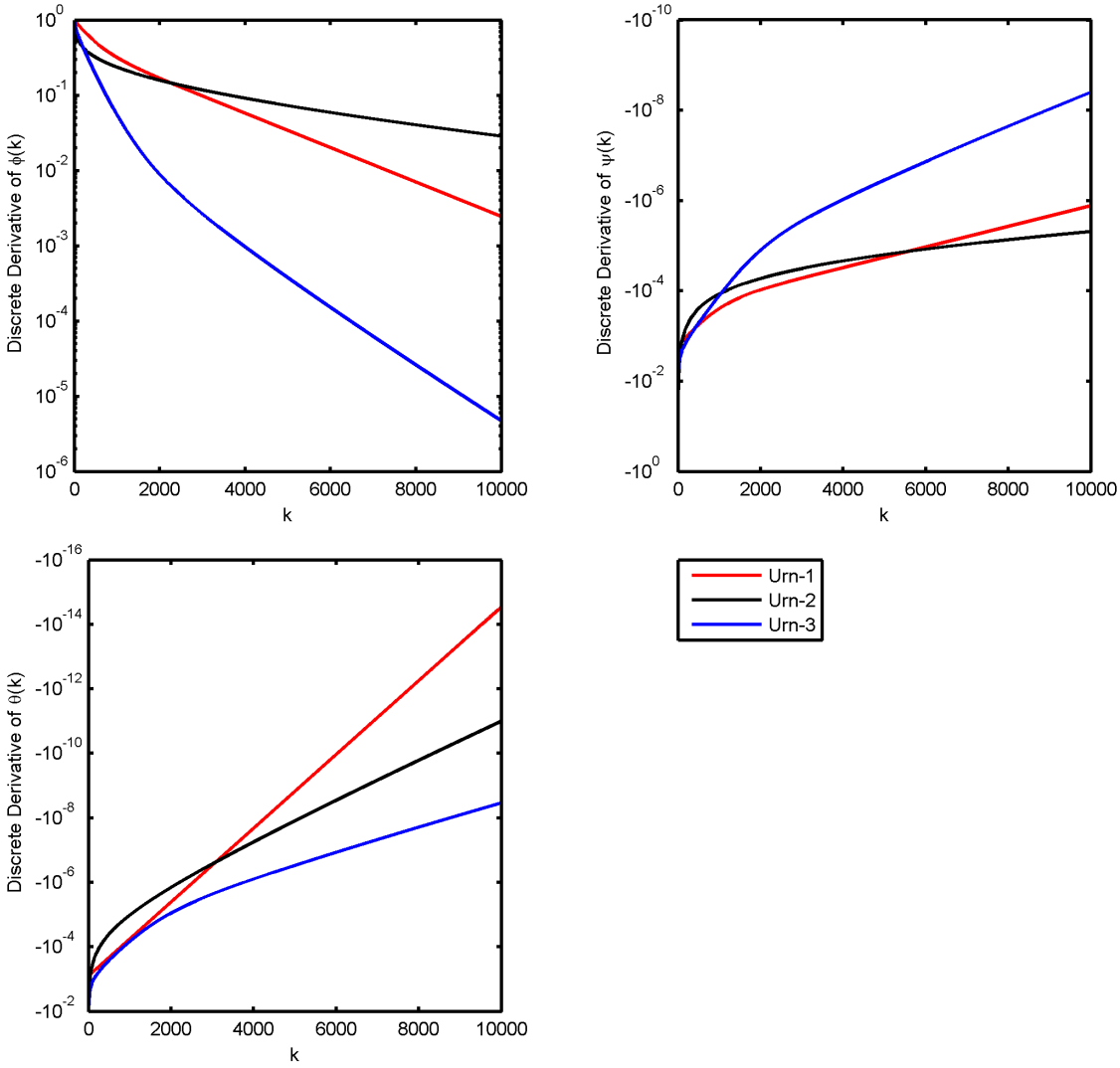

Figure 4.7: **Theoretical discrete derivatives**. The top left shows the discrete derivatives for $\phi(k)$, the top right for $\psi(k)$, and the bottom left for $\theta(k)$.

regressions of θ have a consistently negative bias, and for large m, the relative bias is of smallest magnitude for θ and ϕ, and can grow large for ψ. The standard error in parameter regression relative to the expected regression is shown in Figure 4.9. We see that in relative terms regressions of $\psi(k)$ may be extremely inaccurate.

With each $n_j = 1000$ as an initial sample, we can analyze the distribution of draw allocation. In Figure 4.10, the optimal draw allocation is shown, as well as the expected draw allocation. While the allocations may have significant bias for large m, the allocations are typically biased

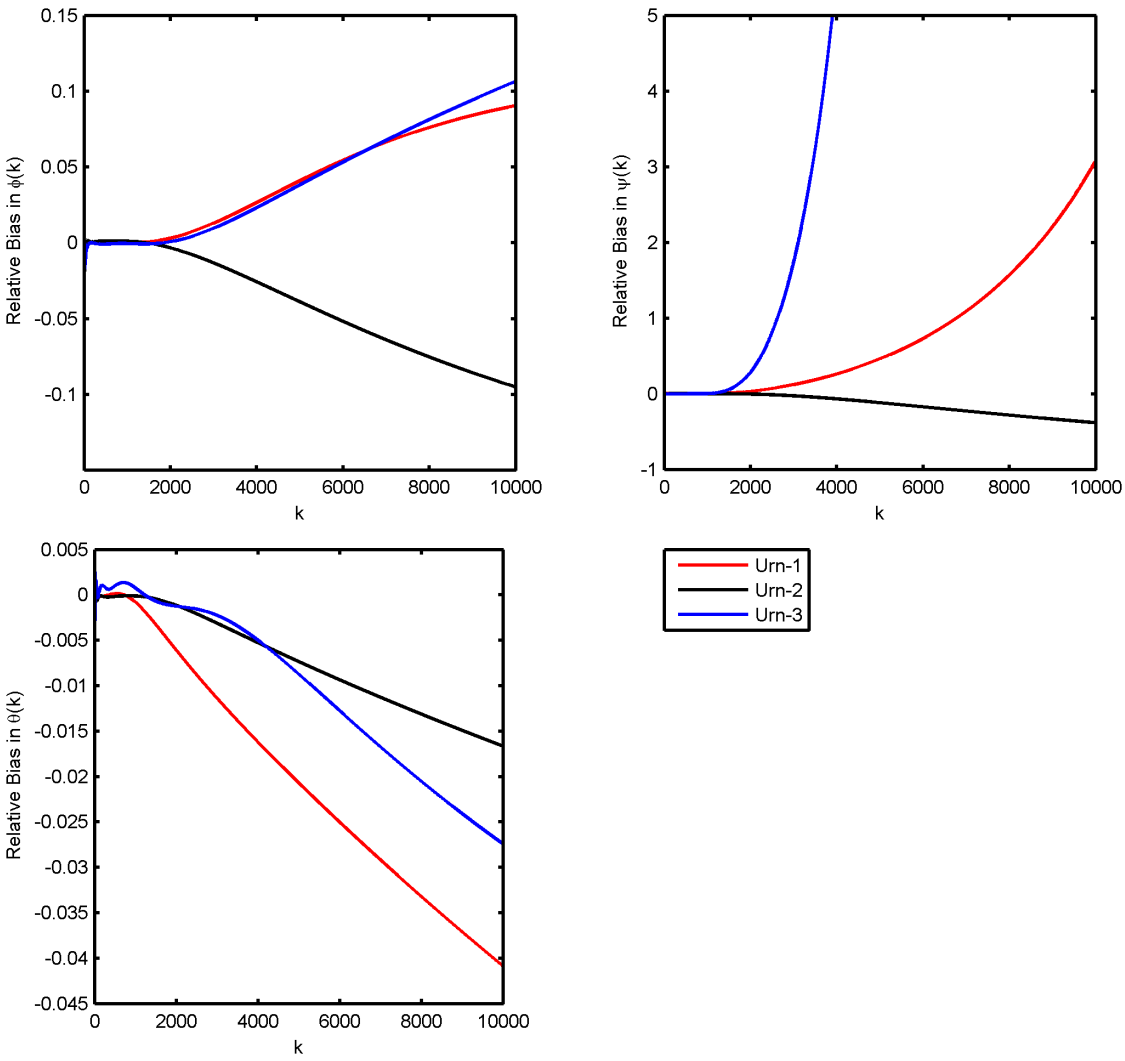

Figure 4.8: **Relative bias in regression curves**. The top left shows the bias relative to the mean in approximating $\phi_j(k)$ with $\hat{\phi}_j^R(k)$, the top right for $\hat{\psi}_j^R(k)$, and the bottom left for $\hat{\theta}_j^R(k)$. All biases are normalized by the theoretical functions to provide scale. The curve for Urn-3 in the $\psi(k)$ graph is truncated to differentiate curves, and continues to grow rapidly, reaching 887 at $k = 10000$.

towards more uniform sampling amongst urns. We view this as a more conservative sampling scheme as uniform sampling is a reasonable sampling scheme in the absence of information about urns. The effect of uniform, optimal, and estimated allocations on the score function is given in Figure 4.11. Notice that in the allocation based on ψ, the score function between the estimated and the optimal are indistinguishable, while the expected does perform better for small m in all situations, and over the entire range of m in the case of $\psi(k)$ and $\phi(k)$. The standard deviation in

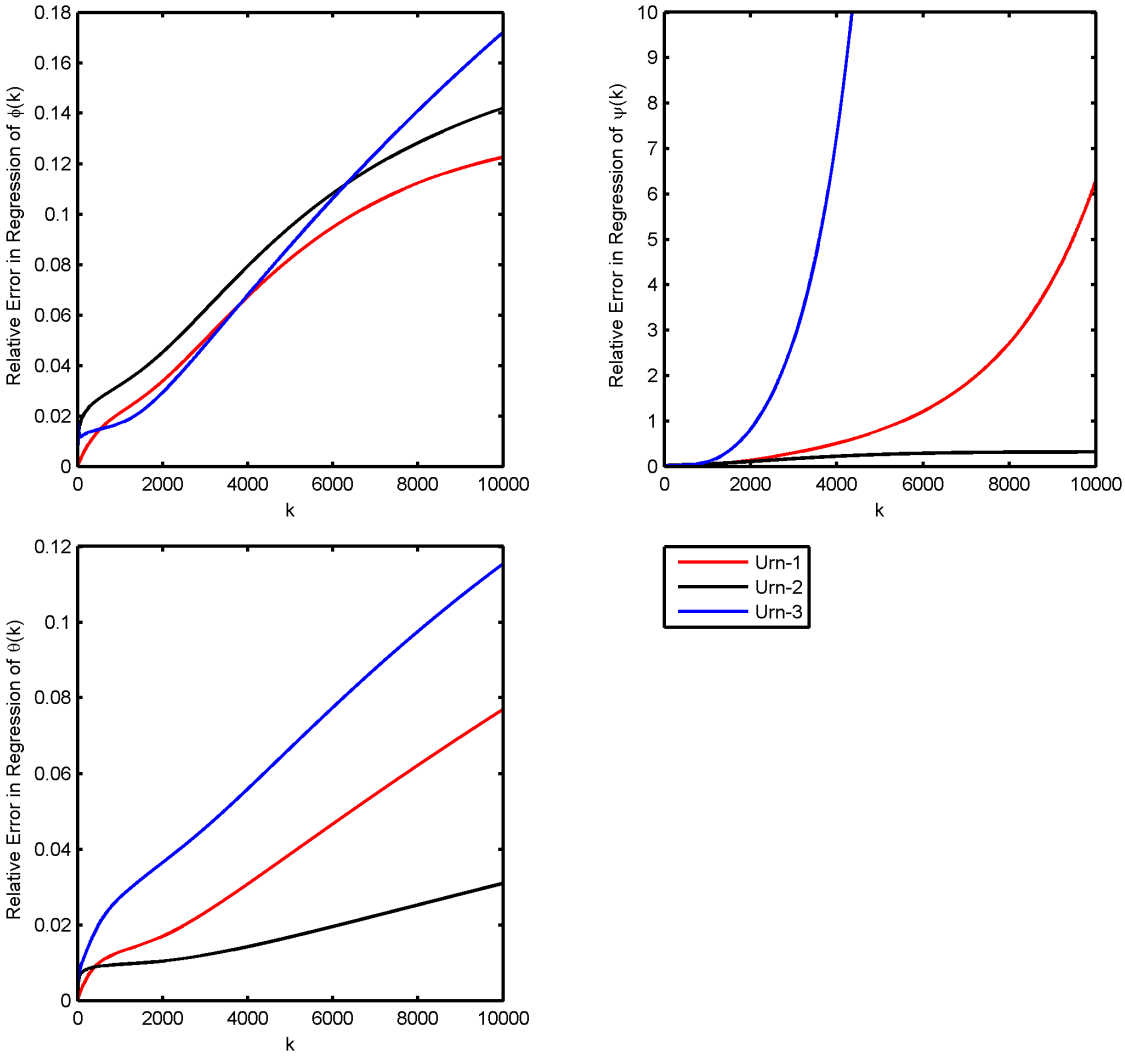

Figure 4.9: **Relative error in regression curves**. The top left shows the standard error in $\hat{\phi}_j^R(k)$, the top right in $\hat{\psi}_j^R(k)$, and the bottom left in $\hat{\theta}_j^R(k)$. All errors are normalized by the theoretical functions to provide scale. The curve for Urn-3 in the $\psi(k)$ graph is truncated to differentiate curves, and continues to grow rapidly, reaching 1052 at $k = 10000$.

draw allocation proportions is given in Figure 4.12. We see that there is significant variability in the draw allocations and that the allocations with regard to the θ measure are most volatile.

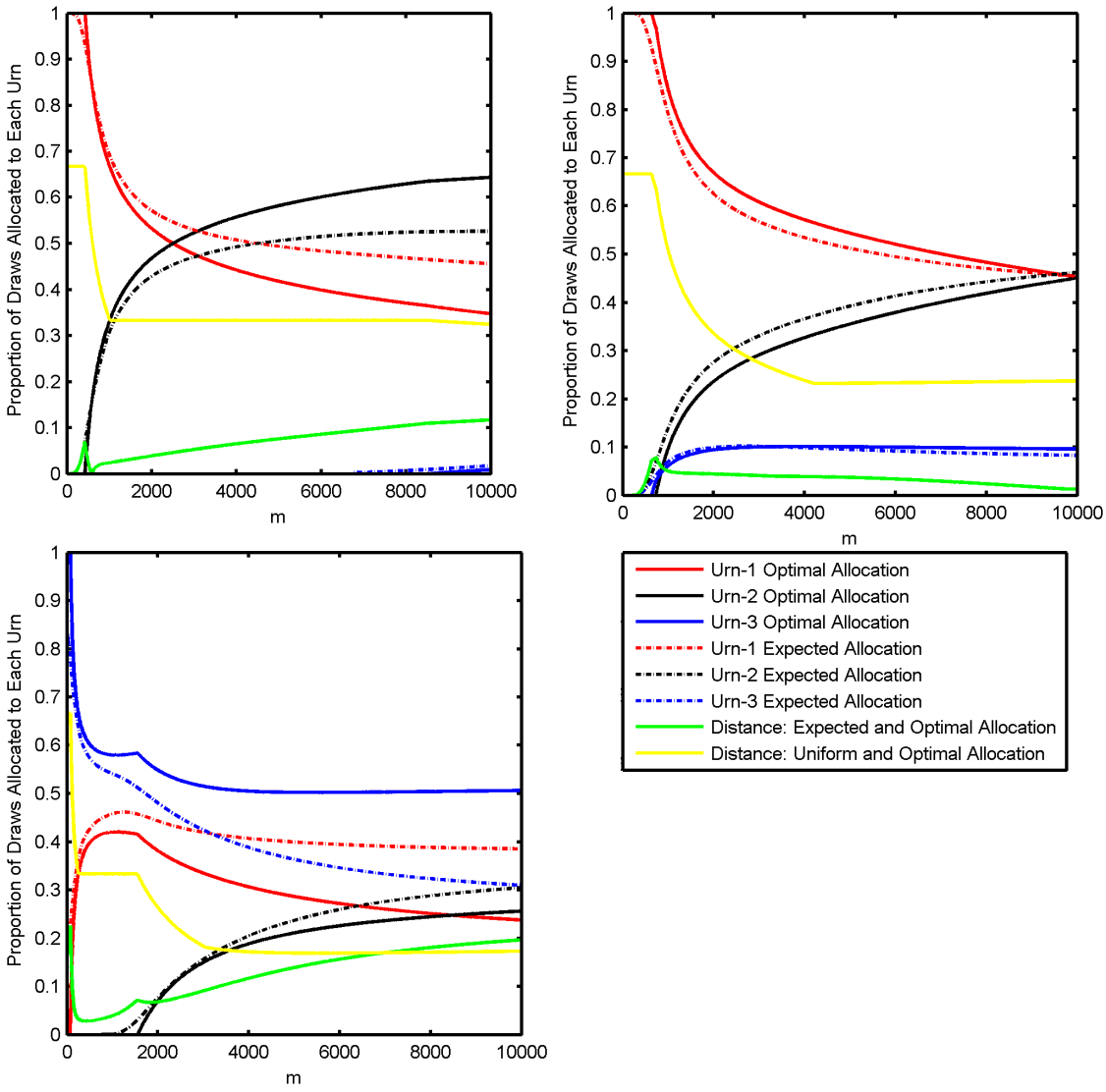

Figure 4.10: **Bias in draw allocation**. These graphs show the allocation to exactly maximize or minimize our measure of interest, as well as the expected allocation using the data. Total variation distance between the optimal allocation and the expected as well as uniform allocations are also displayed. The measure for the top left is $S_\phi(\vec{m})$, for the top right is $S_\psi(\vec{m})$, and for the bottom left is $S_\theta(\vec{m})$.

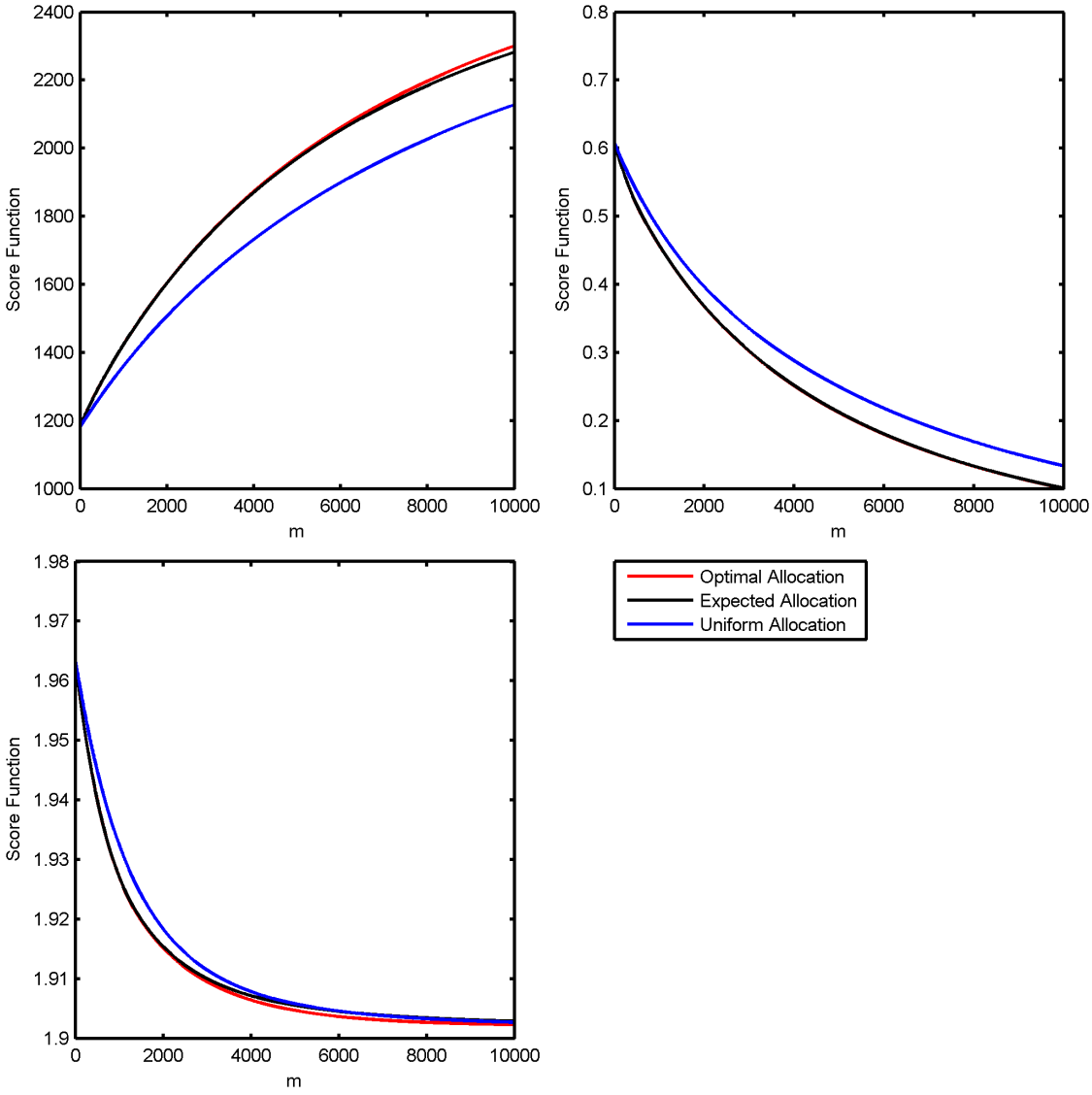

Figure 4.11: **Scores for draw allocations**. These graphs show the score functions for the optimal allocation of m draws, the expected allocation of m draws using data, and a uniform allocation of m draws. The score function for the top left is $S_\phi(\vec{m})$, for the top right is $S_\phi(\vec{m})$, and for the bottom left is $S_\phi(\vec{m})$. In the top right, the expected and optimal allocations give indistinguishable score curves.

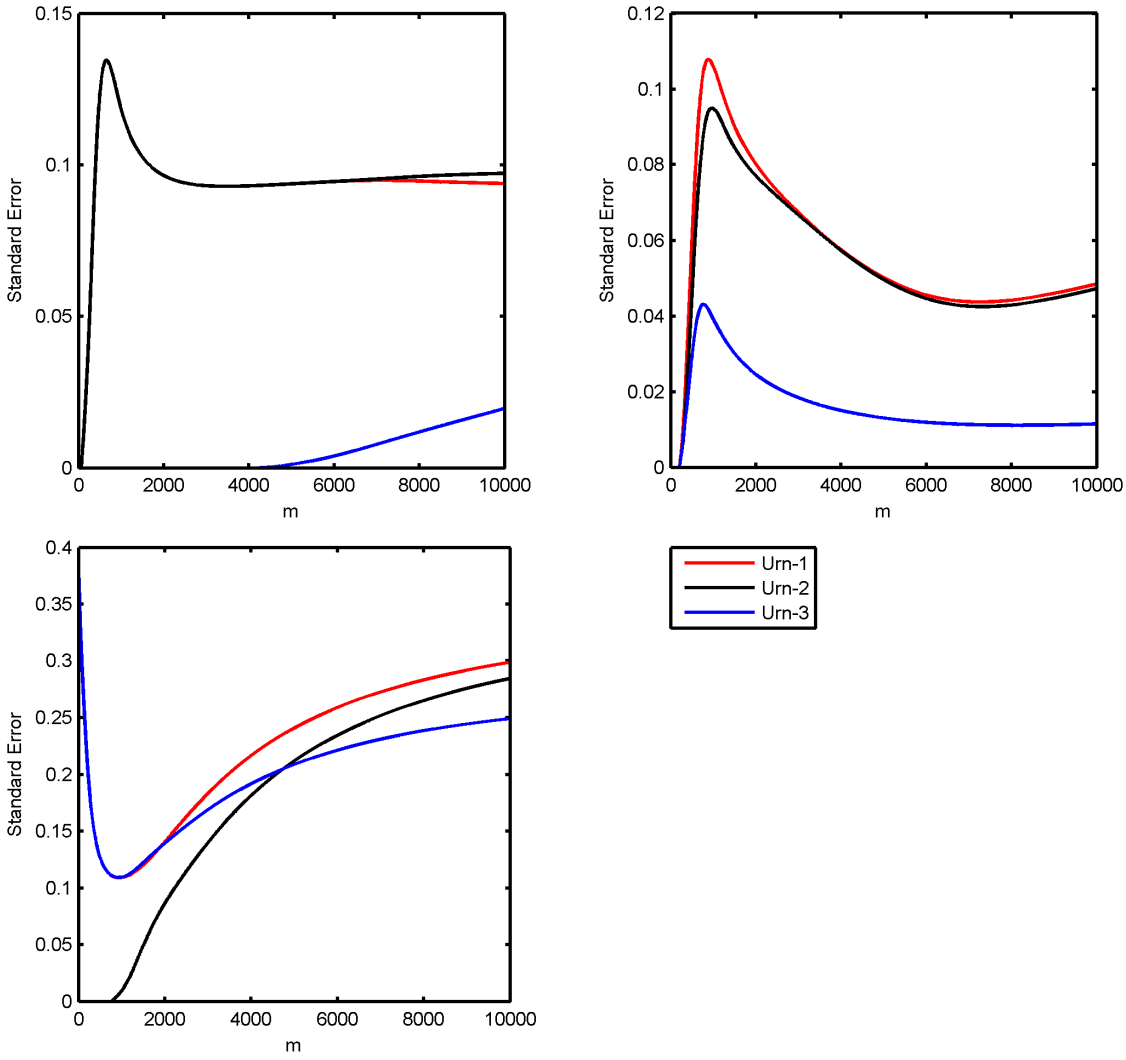

Figure 4.12: **Error in draw allocation**. These graphs show the standard deviation in the returned allocation of draws around the expected allocation of draws. The measure for the top left is $S_\phi(\vec{m})$, for the top right is $S_\psi(\vec{m})$, and for the bottom left is $S_\theta(\vec{m})$.

Chapter 5

Conclusions

Here we summarize the conclusions of this work, particularly as they apply to the statistical theory and data from the Human Microbiome Project. In particular, we summarize the found results, and present future problems that are likely to yield useful advancements with further work.

5.1 Results

The statistical results focus on pointwise convergence that is strengthened to uniform convergence over a range of k, and the estimation of $\psi(k), \phi(k)$ and $\theta(k)$ as well as their variances. The results applying to the Human Microbiome Project concern the identification of environment types having interesting properties, and others which may benefit from further sampling.

5.1.1 Statistical Methods

In Chapter 2 we have explored the consistent estimation of $\phi(k)$ as in (1.1) by the UMVUE in (2.4). We have a form that is amenable to computation over a range of k as in (2.7). This estimator has been shown to have similar properties to $\hat{\psi}(k)$ given by (2.3), in particular with regards to the asymptotic normality as shown in Section 2.4.2. We have also been able to show the convergence, uniformly in k, of $\hat{\psi}(k)$ and $\hat{\phi}(k)$ to their means in Section 2.4.1. In Section 2.5 we have explored the jackknife estimation of the variance of $\hat{\psi}(k)$ and $\hat{\phi}(k)$.

In Chapter 3 we presented the estimation of $\theta(k)$ as in (1.4) by the UMVUE, $\hat{\theta}(k)$, in (3.2).

In Section 3.4.1, we have shown the convergence, uniformly in k, of $\hat{\theta}(k)$ in probability. We have uniform convergence results of marginal distributions as shown in Section 3.4.2, and we have addressed the estimation of the variance, showing a uniform consistency of the estimates $S_x^2(k)$ and $S_y^2(k)$ in Section 3.5.

In Chapter 4, we addressed a regression of our estimates and allocation of sampling resources using those regressions. Specifically, in Section 4.1 we demonstrated an algorithm which gives asymptotic uniform convergence of our regression function to the function it approximates. In Section 4.2, we used these regression functions to extrapolate estimates and allocate sampling resources. In Section 4.2.1, we addressed the minimization of noise in our estimates subject to a constraint on the weights from deviating too far from uniform weighting.

5.1.2 Human Microbiome Project

In Section 3.6 we have seen estimation of $\theta(k)$ applied to samples from various environments. We saw that there is a rich structure of overlap between samples taken from various environmental groups. We may compare average dissimilarity between samples from differing environment types, as well as samples from the same environment type. Further, we saw that we are able to differentiate between environments based on these estimates.

The sample allocation methods were applied to Human Microbiome Project Data in Section 4.3.3, and we were able to identify interesting relations in sampling allocations and estimated parameters for representative stool, vagina, oral, and skin samples. Particularly, we see that stool and vagina environments are likely to benefit from deeper sampling in terms of revealing a number of bacteria with new taxonomic labels, uncovering significant proportions of the environments, and in terms of unveiling taxonomic labels that are present in other environments.

5.2 Future Problems

The problems we have addressed in this thesis suggest U-statistics problems worthy of further study, and other parameters than those presented in (1.1), (1.2), and (1.4) that may be applicable

to robust and reliable estimation of the UniFrac distance.

5.2.1 U-Statistics Theory

The U-statistics results presented here allow the kernel statistic to vary over the range of k where unbiased estimation is possible, and have depended on a careful analysis of conditional variances related to the kernel statistic. In particular, the method of the projection statistic as presented in Sections 2.4 and 3.4 depends on conditional variances as given in (2.14), (3.7), and (3.8). Letting ξ denote a generic conditional variance, we are interested in properties of ξ related to those given in Lemma 3.8. A study in a more general setting depending only on the properties of the kernel statistic and associated ξ may allow the methods used here to prove theorems with a more precise order of convergence, or uniform convergence over a larger range of k. As we see in Lemma 3.9, the convergence results presented here are stronger when $\xi_{1,0}(k)$ does not converge to 0. Intuitively this condition implies that dependence on $\hat{\theta}(k)$ for large k depends mostly on draws from urn-x. Given n_x and n_y growing at certain relative rates we may expect that even if $\xi_{1,0}(k)$ goes to zero, it may be sufficient to maintain a uniform convergence in k, particularly if $\xi_{1,0}(k)$ tends to zero more slowly than $\xi_{0,1}(k)$.

We may also consider the study of the U-statistics built from functions like

$$h_1(k_x, k_y) := |\{X_1, \ldots, X_{k_x}\} \cap \{Y_1, \ldots, Y_{k_y}\}|;$$
$$h_2(k_x, k_y) := \mathbb{P}(\{X_1, \ldots, X_{k_x}\} \cap \{Y_1, \ldots, Y_{k_y}\} = \emptyset),$$

which depend on k_x and k_y, both of which may be variable, and approach interesting quantities. In particular as $k_x, k_y \to \infty, \mathbb{E}(h_1(k_x, k_y)) \to |I_x \cap I_y|$, and $\mathbb{E}(h_2(k_x, k_y)) \to [\![I_x \cap I_y = \emptyset]\!]$. These quantities measure qualities of the sampling of intersections of two urns, and may be useful in estimation of the UniFrac distance.

5.2.2 Expanded Mathematical Model

In Section 1.3.3 we discussed the calculation of the UniFrac distance. Of particular notice is that the UniFrac distance includes a metric on the colors in the urn. Contrasting this, the methods presented in this study assume that all distinct colors are equidistant. An extended model would measure the effect of identifying a new color, not only in its novelty, but in how varied that color is from other colors in the urns of interest. As an example in the language of UniFrac distance, assume that the evolutionary distance between a reddish color and a greenish color is large, while the evolutionary distance between two reddish or two greenish colors is small. Further, consider an urn of primarily reddish colors and an urn of primarily greenish colors. If we wish to measure the UniFrac distance between these urns, identifying a greenish color in the reddish urn will more dramatically effect the distance estimate than identifying another reddish color in the reddish urn. We may desire that our model account for this phenomena by including the appropriate color information. This expansion would require the facilitation of new parameters more sophisticated than those presented in Section 1.3, particularly, a dependence on colors would be necessary. For example, $\psi(k)$ could measure the probability of observing a new color after k-samples multiplied by the expected distance of that color to the average observed color, or $\phi(k)$ could count the sum of distances between colors observed in k-samples.

Similarly, to give more sophistication to optimal allocation as discussed in Section 4.2, we may adjust the measure in (1.8) so as to not weight each $\theta_{i,j}$ uniformly, but with regards to estimated UniFrac distances between pairs of environments/urns. Often, for a set of environments there are two groups of closely related environments which differ from each other as groups. Uniform weighting may suggest significantly more sampling weight to the larger group of similar environments, and relatively little to the smaller group, when perhaps we would prefer an even sampling between environments in the two groups. An approach to such a normalization may include the incorporation of the estimated UniFrac distance or θ information covering all possible environmental comparisons.

Bibliography

[1] A. A. Ahmad. On the Berry-Esseen Theorem for Random U-Statistics. The Annals of Statistics, 8(6):1395–1398, 1980.

[2] J. N. Arvesen. Jackknifing U-Statistics. The Annals of Mathematical Statistics, 40(6):2076–2100, 1969.

[3] G. Beylkin and L. Monzón. On Approximation of Functions by Exponential Sums. Applied and Computational Harmonic Analysis, 19:17–48, 2005.

[4] G. Beylkin and L. Monzón. Approximation by Exponential Sums Revisited. Applied and Computational Harmonic Analysis, 28:131–149, 2010.

[5] J. Bunge and M. Fitzpatrick. Estimating the Number of Species: A Review. Journal of fthe American Statistical Association, 88(421):364–373, 1993.

[6] H. Callaert and P. Janssen. The Berry-Esseen Theorem for U-Statistics. The Annals of Statistics, 6(2):417–421, 1978.

[7] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight. QIIME allows analysis of high-throughput community sequencing data. Nature Methods, 7(5):335–336, Apr. 2010.

[8] G. Casella. Statistical Inference, page 369. Duxbury Press, 2001.

[9] Y.-K. Chan and J. Wierman. On the Berry Esseen Theorem for U-Statistics. The Annals of Probability, 5(1):136–139, 1977.

[10] A. Chao, R. L. Chazdon, R. K. Colwell, and T.-J. Shen. A New Statistical Approach for Assessing Similarity of Species Composition with Incidence and Abundance Data. Ecology Letters, 8:148–159, 2005.

[11] A. Chao, R. L. Chazdon, R. K. Colwell, and T.-J. Shen. Abundance-Based Similarity Indices and Their Estimation When There Are Unseen Species in Samples. Biometrics, 62:361–371, 2006.

[12] A. Chao and S.-M. Lee. Estimating the Number of Classes via Sample Coverage. Journal of the American Statistical Association, 87(417):210–217, 1992.

[13] M. K. Clayton and E. M. Frees. Linear Estimation of Discovering a New Species. Journal of the American Statistical Association, 82(397):305–311, 1987.

[14] T. F. Coleman and Y. Li. On the Convergence of Interior-Reflective Newton Methods for Nonlinear Minimization Subject to Bounds. Mathematical Programming, 67:189–224, 1994.

[15] H. Consortium. Mapping the human microbiota: resources from the human microbiome project. Nature, Submitted.

[16] H. Consortium. Structure, function and diversity of the human microbiome in an adult reference population. Nature, Submitted.

[17] R. Durrett. Probability: Theory and Examples. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010.

[18] B. Efron and C. Stein. The Jackknife Estimate of Variance. The Annals of Statistics, 9(3):586–596, 1981.

[19] W. W. Esty. A Normal Limit Law for a Nonparametric Estimator of the Coverage of a Sample. The Annals of Statistics, 11(3):905–912, 1983.

[20] S. J. Finch, N. R. Mendell, and J. Henry C. Thode. Probabilistic Measures of Adequacy of a Numerical Search for a Global Maximum. Journal of the American Statistical Association, 84(408):1020–1023, 1953.

[21] P. Flajolet, D. Gardy, and L. Thimonier. Birthday Paradox, Coupon Collectors, Caching Algorithms, and Self-Organizing Search. Discrete Applied Mathematics, 39:207–229, 1992.

[22] J. A. Gilbert, F. Meyer, D. Antonopoulos, P. Balaji, C. T. Brown, C. T. Brown, N. Desai, J. A. Eisen, D. Evers, D. Field, W. Feng, D. Huson, J. Jansson, R. Knight, J. Knight, E. Kolker, K. Konstantindis, J. Kostka, N. Kyrpides, R. Mackelprang, A. McHardy, C. Quince, J. Raes, A. Sczyrba, A. Shade, and R. Stevens. Meeting report: the terabase metagenomics workshop and the vision of an Earth microbiome project. Stand Genomic Sci, 3:243–248, 2010.

[23] I. J. Good. The Population Frequencies of Species and the Estimation of Population Parameters. Biometrika, 40(3-4):237–264, 1953.

[24] W. F. Grams and R. J. Serfling. Convergence Rate for U-Statistics and Related Statistics. The Annals of Statistics, 1(1):153–160, 1973.

[25] J. Hajek. Asymptotic Normality of Simple Linear Rank Statistics Under Alternatives. The Annals of Mathematical Statistics, 39(2):325–346, 1968.

[26] P. R. Halmos. The Theory of Unbiased Estimation. The Annals of Mathematical Statistics, 17(1):34–43, 1946.

[27] J. Hampton and M. E. Lladser. Allocation of New Draws for Optimal Sampling of Urn Ensembles with Application to the Human Microbiome Project. Submitted.

[28] J. Hampton and M. E. Lladser. Estimation of Distribution Overlap of Urn Models. Submitted.

[29] W. Hoeffding. A Class of Statistics with Asymptotically Normal Distribution. The Annals of Mathematical Statistics, 19(3):293,325, 1948.

[30] O. Hölder. Ueber einen Mittelwertsatz. Göttingen Nachr., pages 38–47, 1889.

[31] Hwang, S. Janson, and A. Sinica. Local Limit Theorems for Finite and Infinite Urn Models. Ann. Probab, 2007.

[32] J. Jensen. Sur les fonctions convexes et les ingalits entre les valeurs moyennes. Acta Mathematica, 30:175–193, 1906. 10.1007/BF02418571.

[33] D. W. Kammler. L_1-Approximation of Completely Monotonic Functions by Sums of Exponentials. SIAM Journal on Numerical Analysis, 16(1):30–45, 1979.

[34] E. Lehmann and H. Scheffé. Completeness, Similar Regions, and Unbiased Estimation: Part I. Sankhyā: The Indian Journal of Statistics, 10(4):305–340, 1950.

[35] E. Lehmann and H. Scheffé. Completeness, Similar Regions, and Unbiased Estimation: Part II. Sankhyā: The Indian Journal of Statistics, 15(3):219–236, 1955.

[36] E. L. Lehmann. Consistency and Unbiasedness of Certain Nonparametric Tests. The Annals of Mathematical Statistics, 22(2):165–179, 1951.

[37] H. Li, C. Kuo, and M. G. Rusell. The impact of perceived channel utilities, shopping orientations, and demographics on the consumer's online buying behavior. Journal of Computer-Mediated Communication, 5(2):0–0, 1999.

[38] A. Lijoi, R. H. Mena, and I. Prünster. Bayesian Nonparametric Estimation of the Probability of Discovering New Species. Biometrika, 94(4):769–786, 2007.

[39] M. E. Lladser. Prediction of Unseen Proportions in Urn Models with Restricted Sampling. In ANALCO2009, 2009.

[40] M. E. Lladser, R. Goeuet, and J. Reeder. Extrapolation of Urn Models via Poissonization: Accurate Measurements of the Microbial Unknown. PLoS One, 6(6), 2011.

[41] C. Lozupone and R. Knight. UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. Applied and Environmental Microbiology, 71(12):8228–8235, 2005.

[42] C. Lozupone, M. E. Lladser, D. Knights, J. Stombaugh, and R. Knight. UniFrac: An Effective Distance Metric for Microbial Community Comparison. The ISME Journal, 5:169–172, 2011.

[43] H. B. Mann and A. Wald. On stochastic limit and order relationships. The Annals of Mathematical Statistics, 14(3):pp. 217–226, 1943.

[44] C. X. Mao. Predicting the Conditional Probability of Finding a New Class. Journal of the American Statistical Association, 99(468):1108–1118, 2004.

[45] C. X. Mao and B. G. Lindsay. A Poisson Model for the Coverage Problem with a Genomic Application. Biometrika, 89(3):669–681, 2002.

[46] C. X. Mao and B. G. Lindsay. Estimating the Number of Classes. The Annals of Statistics, 35(2):917–930, 2007.

[47] J. F. Nash. Equilibrium points in n-person games. Proceedings of the National Academy of Sciences, 36(1):48–49, 1950.

[48] A. Orlitsky, N. P. Santhanam, and J. Zhang. Always Good Turing: Asymptotically Optimal Probability Estimation. Science, 302:427–431, 2003.

[49] M. R. Osborne and G. K. Smyth. An Algorithm for Exponential Fitting Revisited. Journal of Applied Probability, 23:419–430, 1986.

[50] J. Peterson, S. Garges, M. Giovanni, P. McInnes, L. Wang, J. A. Schloss, V. Bonazzi, J. E. McEwen, K. A. Wetterstrand, C. Deal, C. C. Baker, V. Di Francesco, T. K. Howcroft, R. W. Karp, R. D. Lunsford, C. R. Wellington, T. Belachew, M. Wright, C. Giblin, H. David, M. Mills, R. Salomon, C. Mullins, B. Akolkar, L. Begg, C. Davis, L. Grandison, M. Humble, J. Khalsa, A. R. Little, H. Peavy, C. Pontzer, M. Portnoy, M. H. Sayre, P. Starke-Reed, S. Zakhari, J. Read, B. Watson, and M. Guyer. The NIH Human Microbiome Project. Genome Res., 19:2317–2323, Dec 2009.

[51] L. M. Proctor. The Human Microbiome Project in 2011 and beyond. Cell Host Microbe, 10:287–291, Oct 2011.

[52] H. E. Robbins. Estimating the Total Probability of the Unobserved Outcomes of an Experiment. The Annals of Mathematical Statistics, 39(1):256–257, 1968.

[53] R. Schaback. Suboptimal Exponential Approximations. SIAM Journal on Numerical Analysis, 16(6):1007–1018, 1979.

[54] P. D. Schloss and J. Handelsman. Introducing SONS, a tool for operational taxonomic unit-based comparisons of microbial community memberships and structures. Appl. Environ. Microbiol., 72:6773–6779, Oct 2006.

[55] P. K. Sen. Almost Sure Convergence of Generalized U-Statistics. The Annals of Probability, 5(2):pp. 287–290, 1977.

[56] R. J. Serfling. Approximation Theorems of Mathematical Statistics, chapter 5. John Wiley & Sons, 1980.

[57] J. Shao. The Efficiency and Conistency of Approximations to the Jackknife Variance Estimators. Journal of the American Statistical Association, 84(405):114–119, 1989.

[58] J. Shao and C. Wu. A General Theory for Jackknife Variance Estimation. The Annals of Statistics, 17(3):1176–1197, 1989.

[59] I. G. Shevtsova. An Improvement of Convergence Rate Estimates in the Lyapunov Theorem. Daklady Mathematics, 82(3):862–864, 2010.

[60] E. Slutsky. Über stochastische Asymptoten und Grenzwerte. Metron, 5(3):3–89, 1925.

[61] N. Starr. Linear Estimation of Discovering a New Species. The Annals of Statistics, 7(3):644–652, 1979.

[62] J. Tang and R. R. Breaker. Structural diversity of self-cleaving ribozymes. Proceedings of the National Academy of Sciences, 97(11):5784–5789, 2000.

[63] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon. The human microbiome project. Nature, 449:804–810, Oct 2007.

[64] H. Washington. Diversity, Biotic and Similarity Indices: A Review with Special Relevance to Aquatic Ecosystems. Water Research, 18(6):653 – 694, 1984.

[65] R. H. Whittaker. Vegetation of the siskiyou mountains, oregon and california. Ecological Monographs, 30(3):pp. 279–338, 1960.

[66] H. Wolda. Similarity Indices, Sample Size and Diversity. Oecologia, 50:296–302, 1981.

Appendix A

Matlab Code: psiEst.m

```
function [err,psis] = psiEst(skipDepth,table)
%%%%%%%%%%%%%%%%%%%%%%%%
%%% psiEst.m       %%%
%%% Jerrad Hampton %%%
%%% © 2011/2012    %%%
%%%%%%%%%%%%%%%%%%%%%%%%

% Calculates ψ̂(k) and S_ψ(k), for k = 1 : n − 1.

%%%%%%%%%%%%%%%
%%% Inputs %%%
%%%%%%%%%%%%%%%

% skipDepth: (Positive Integer)
%   This controls the number of points estimated. Every 1 in skipDepth points
%   are estimated.

% table: (Matrix)
%   Data where each row corresponds to an color, each column corresponds to
%   an urn, and each entry is a number of observations.

%%%%%%%%%%%%%%%%
%%% Outputs %%%
%%%%%%%%%%%%%%%%

% err: (1 Dimensional Array of Vectors)
%   S_ψ(k), for k = 1 : n − 2.

% psis: (1 Dimensional Array of Vectors)
%   ψ̂(k), for k = 1 : n − 1.

%%%%%%%%%%%%
%%% Code %%%
%%%%%%%%%%%%
numRows = size(table,1); % Upper bound on number of colors in Urn
numCols = size(table,2); % Number of Urns
psis = cell(numCols,1);
err = cell(numCols,1);
```

```
sSize = sum(table,1); % Holds sample sizes from each Urn
parfor col = 1:numCols
    n = sSize(col);
    tab = table(:,col);
    ind = unique(tab); % Determines what Q statistics may arise
    ind = ind(ind>0); % 0 is not useful for these statistics.
    iSize = size(ind,1);
    kSize = floor((n-1)/skipDepth);
    cEst = zeros(kSize,iSize);
    cEstPrime = zeros(kSize,iSize);
    cEstCor = zeros(kSize,iSize);
    cEstGam = gammaln(n); %(n-1)!
    cEstPrimeGam = gammaln(n-1);
    for kk = 1:kSize
        k = kk*skipDepth;
        cEstGamK = gammaln(n-k)-cEstGam;
        cEstPrimeGamK = gammaln(n-k-1)-cEstPrimeGam;
        for jj = 1:iSize
            j = ind(jj);
            if(jj<n-k)
                temp = gammaln(n-j+1)-gammaln(n-j-k+1);
                cEstGamJ = cEstGamK+temp;
                cEstPrimeGamJ = cEstPrimeGamK+gammaln(n-j)-gammaln(n-j-k);
                cEstCorGamJ = cEstPrimeGamK+temp;
                cEst(kk,jj) = exp(cEstGamJ)/n;
                cEstPrime(kk,jj) = exp(cEstPrimeGamJ)/(n-1);
                cEstCor(kk,jj) = (j-1)*exp(cEstCorGamJ)/(n-1)...
                    -j*cEstPrime(kk,jj);
            end
            if(jj==n-k)
                temp = gammaln(k+1)+cEstGamK;
                cEst(kk,jj) = exp(temp)/n;
            end
        end
    end
    Q = zeros(iSize,1);
    for m = 1:numRows   %First element to Q(1), etc.
        if(tab(m)>0)
            i = find(ind==tab(m),1);
            Q(i) = Q(i)+tab(m);
        end
    end
    psis{col} = cEst*Q;
    cEstCor = (cEstCor+(cEstPrime*Q-psis{col})*ones(1,iSize)).^2;
    err{col} = sqrt(((n-1)/n)*cEstCor*Q);
    err{col} = err{col}(~isnan(err{col}));
end
end
```

Appendix B

Matlab Code: phiEst.m

```
function [err,phis] = phiEst(skipDepth,table)
%%%%%%%%%%%%%%%%%%%%%%%%
%%% phiEst.m       %%%
%%% Jerrad Hampton %%%
%%% © 2011/2012    %%%
%%%%%%%%%%%%%%%%%%%%%%%%

% Calculates φ̂(k) and S_φ(k), for k = 1:n.

%%%%%%%%%%%%%%%
%%% Inputs %%%
%%%%%%%%%%%%%%%

% skipDepth: (Positive Integer)
%   This controls the number of points estimated. Every 1 in skipDepth points
%   are estimated.

% table: (Matrix)
%   Data where each row corresponds to an color, each column corresponds to
%   an urn, and each entry is a number of observations.

%%%%%%%%%%%%%%%%
%%% Outputs %%%
%%%%%%%%%%%%%%%%

% err: (1 Dimensional Array of Vectors)
%   S_φ(k), for k = 1:n-1

% psis: (1 Dimensional Array of Vectors)
%   φ̂(k), for k = 1:n.

%%%%%%%%%%%%
%%% Code %%%
%%%%%%%%%%%%
numRows = size(table,1); % Upper bound on number of colors in Urn
numCols = size(table,2); % Number of Urns
phis = cell(numCols,1);
err = cell(numCols,1);
```

```
sSize = sum(table,1); % Holds sample sizes from each Urn
parfor col = 1:numCols
    n = sSize(col);
    tab = table(:,col);
    ind = unique(tab); % Determines what Q statistics may arise
    ind = ind(ind>0); % 0 is not useful for these statistics.
    iSize = size(ind,1);
    kSize = floor(n/skipDepth);
    cEst = zeros(kSize,iSize);
    cEstPrime = zeros(kSize,iSize);
    cEstCor = zeros(kSize,iSize);
    cEstGam = gammaln(n+1); %n!
    cEstPrimeGam = gammaln(n);
    for kk = 1:kSize
        k = kk*skipDepth;
        cEstGamK = gammaln(k+1)+gammaln(n-k+1)-cEstGam;
        cEstPrimeGamK = gammaln(k+1) + gammaln(n-k)-cEstPrimeGam;
        for jj = 1:iSize
            j = ind(jj);
            cEstGamJ = cEstGamK+gammaln(n-j+1)+gammaln(j+1);
            cEstPrimeGamJ = cEstPrimeGamK+gammaln(n-j)+gammaln(j+1);
            cEstCorGamJ = cEstPrimeGamK+gammaln(n-j+1)+gammaln(j);
            if(j<n-k)
                for i = 1:j
                    temp = -gammaln(k-i+1)-gammaln(j-i+1)-gammaln(i+1);
                    cEstGamI = cEstGamJ-gammaln(n-j-k+i+1)+temp;
                    cEstPrimeGamI = cEstPrimeGamJ-gammaln(n-j-k+i)+temp;
                    cEstCorGamI = cEstCorGamJ-gammaln(n-j-k+i+1)...
                        -gammaln(k-i+1)-gammaln(j-i)-gammaln(i+1);
                    cEst(kk,jj) = cEst(kk,jj)+exp(cEstGamI);
                    cEstPrime(kk,jj) = cEstPrime(kk,jj)+exp(cEstPrimeGamI);
                    cEstCor(kk,jj) = cEstCor(kk,jj)+exp(cEstCorGamI);
                end
                cEst(kk,jj) = cEst(kk,jj)/j;
                cEstCor(kk,jj) = cEstCor(kk,jj)-cEstPrime(kk,jj);
                cEstPrime(kk,jj) = cEstPrime(kk,jj)/j;
            end
            if(j==n-k)
               for i = 1:j
                    cEstGamI = cEstGamJ-gammaln(n-j-k+i+1)...
                        -gammaln(k-i+1)-gammaln(j-i+1)-gammaln(i+1);
                    cEstCorGamI = cEstCorGamJ-gammaln(n-j-k+i+1)...
                        -gammaln(k-i+1)-gammaln(j-i)-gammaln(i+1);
                    cEst(kk,jj) = cEst(kk,jj)+exp(cEstGamI);
                    cEstCor(kk,jj) = cEstCor(kk,jj)+exp(cEstCorGamI);
               end
               cEst(kk,jj) = cEst(kk,jj)/j;
               cEstCor(kk,jj) = cEstCor(kk,jj)-1;
               cEstPrime(kk,jj) = 1/j;
            end
            if(j>n-k)
                cEst(kk,jj) = 1/j;
                cEstPrime(kk,jj) = 1/j;
```

```
                cEstCor(kk,jj) = 0;
            end
        end
    end
    Q = zeros(iSize,1);
    for m = 1:numRows   %First element to Q(1), etc.
        if(tab(m)>0)
            i = find(ind==tab(m),1);
            Q(i) = Q(i) + tab(m);
        end
    end
    phis{col} = cEst*Q;
    cEstCor = (cEstCor+(cEstPrime*Q-phis{col})*ones(1,iSize)).^2;
    err{col} = sqrt(((n-1)/n)*cEstCor*Q);
    err{col} = err{col}(~isnan(err{col}));
end
end
```

Appendix C

Matlab Code: thetaEst.m

```
function [err,thetas] = thetaEst(skipDepth,table)
%%%%%%%%%%%%%%%%%%%%%%%%
%%% thetaEst.m      %%%
%%% Jerrad Hampton %%%
%%% © 2011/2012    %%%
%%%%%%%%%%%%%%%%%%%%%%%%

% Calculates θ̂(k) and S_θ(k), for k = 1 : n_y.

%%%%%%%%%%%%%%%
%%% Inputs %%%
%%%%%%%%%%%%%%%

% skipDepth: (Positive Integer)
%   This controls the number of points estimated. Every 1 in skipDepth points
%   are estimated.

% table: (Matrix)
%   Data where each row corresponds to an color, each column corresponds to
%   an urn, and each entry is a number of observations.

%%%%%%%%%%%%%%%%
%%% Outputs %%%
%%%%%%%%%%%%%%%%

% err: (1 Dimensional Array of Vectors)
%   S_θ(k), for k = 1 : n_y

% psis: (1 Dimensional Array of Vectors)
%   θ̂(k), for k = 1 : n_y.

%%%%%%%%%%%%
%%% Code %%%
%%%%%%%%%%%%
numRows = size(table,1); % Upper bound on number of colors in Urn
numCols = size(table,2); % Number of Urns
thetas = cell(numCols,1);
err = cell(numCols,1);
```

```
sSize = sum(table,1); % Holds sample sizes from each Urn
parfor iii = 1:numCols
    n = sSize(iii);
    tab = table(:,iii);
    ind = unique(tab); % Determines what Q statistics may arise
    ind = ind(ind>0); % 0 is not useful for these statistics.
    iSize = size(ind,1);
    kSize = floor(n/skipDepth);
    cEst = zeros(kSize,iSize);
    cEstErr = zeros(kSize,iSize);
    cEstErrTwo = zeros(kSize,iSize);
    cEstGam = gammaln(n+1); %n!
    cEstErrGam = gammaln(n);
    for kk = 1:kSize
        k = skipDepth*kk;
        cEstGamK = gammaln(n-k+1)-cEstGam;
        cEstErrGamK = gammaln(n-k)-cEstErrGam;
        for jj = 1:iSize
            j = ind(jj);
            if(j<=n-k-1)
                temp = gammaln(n-j+1)-gammaln(n-k-j+1);
                tempErr = gammaln(n-j)-gammaln(n-k-j);
                cEst(kk,jj) = exp(cEstGamK+temp);
                cEstErr(kk,jj) = exp(cEstErrGamK+tempErr);
                cEstErrTwo(kk,jj) = exp(cEstErrGamK+temp);
            end
            if(j==n-k)
                cEst(kk,jj) = exp(cEstGamK+gammaln(k+1));
                cEstErrTwo(kk,jj)=exp(cEstErrGamK+gammaln(k+1));
            end
        end
    end
    Q = zeros(iSize,numCols);
    Q0 = zeros(1,numCols);
    for m = 1:numRows   %First element to Q(1), etc.
        if(tab(m)>0)
            i = find(ind==tab(m),1);
            Q(i,:) = Q(i,:)+table(m,:);
        else
            Q0 = Q0 + table(m,:);
        end
    end
    thetaEst = cEst*Q+ones(kSize,1)*Q0; %θ̂n_x
    thetaErrEst =cEstErr*Q+ones(kSize,1)*Q0; %For S_y later
    for i = [1:iii-1 iii+1:numCols]
        thetaEst(:,i) = thetaEst(:,i)/sSize(i);
        thetaErrEst(:,i) = thetaErrEst(:,i)/sSize(i);
    end
    err{iii} = zeros(size(thetaEst)); %For S_x later
    for i = [1:iii-1 iii+1:numCols]
        cEstT = zeros(size(cEst));
        for k = 1:kSize
            for jj = 1:iSize;
```

```
                ii = ind(jj);
                if(ii<=n-k)
                    cEstT(k,jj) = (cEst(k,jj)-thetaEst(k,i)).^2;
                end
            end
        end
        err{iii}(:,i) = cEstT*Q(:,i)+(1-thetaEst(:,i)).^2*Q0(i);
        err{iii}(:,i) = err{iii}(:,i)/((sSize(i)-1)*sSize(i));
    end
    % For Y Variance
    indy = ind;
    ySize = iSize;
    for i = [1:iii-1 iii+1:numCols]
        indx = unique(table(:,i));
        indx = indx(indx>0);
        xSize = size(indx,1);
        M = zeros(xSize,ySize);
        for m = 1:numRows   %For M Statistics
            if(tab(m)>0)
                yy = find(indy==tab(m),1);
                if(table(m,i)>0)
                    xx = find(indx==table(m,i),1); %#ok<PFBNS>
                    M(xx,yy) = M(xx,yy)+1;
                end
            end
        end
        for k = 1:kSize
            tempErr = 0;
            for x = 1:xSize
                xx = indx(x);
                for y = 1:ySize
                    yy = indy(y);
                    tempErr = tempErr+yy*M(x,y)*((xx*(cEstErrTwo(k,y)...
                        -cEstErr(k,y)))/sSize(i)+thetaErrEst(k,i)...
                        -thetaEst(k,i)).^2; %#ok<PFBNS>
                end
            end
            err{iii}(k,i) = err{iii}(k,i)+(n-1)/n*tempErr;
        end
    end
    thetas{iii} = thetaEst(:,[1:iii-1 iii+1:numCols]);
    err{iii} = sqrt(err{iii}(:,[1:iii-1 iii+1:numCols]));
end
end
```

Appendix D

Matlab Code: psiReg.m

```
function [regExps,regWeights,outError] =...
                                        psiReg(iData,iExps,iLim,skipDepth,tol)
%%%%%%%%%%%%%%%%%%%%%%%
%%% psiReg.m       %%%
%%% Jerrad Hampton %%%
%%% © 2011/2012    %%%
%%%%%%%%%%%%%%%%%%%%%%%

% This method computes exponents and weights for $\hat{\psi}^R$

%%%%%%%%%%%%%%%
%%% Inputs %%%
%%%%%%%%%%%%%%%

% iData: (Vector)
%   Data to be approximated.

% iExps: (Vector)
%   Exponents to be included.

% iLim: (Positive Integer)
%   An upper bound on number of iterations when determining exponents and
%   weights.

% skipDepth: (Positive Integer)
%   This is the skipDepth used in psiEst.m

% tol: (Positive Real)
%   Iterative steps will stop if error is within tolerance.

%%%%%%%%%%%%%%%%
%%% Outputs %%%
%%%%%%%%%%%%%%%%

% regExps: (1 Dimensional Array of Vectors)
%   Exponents for each average of comparisons, couples with regWeights for
%   function approximation.
```

```
% regWeights: (1 Dimensional Array of Vectors)
%   Weights for each average of comparisons, couples with regExps
%   for function approximation.

% outError: (Vector)
%   Contains error between input and regression.

%%%%%%%%%%%%
%%% Code %%%
%%%%%%%%%%%%
osbExps = 7; % Exponents for OsborneSmyth
dSize = size(iData,1);
evnExps = 20; % Exponents evenly distributed
skipLength = floor(dSize/evnExps);
Y = hankel(iData(1:dSize-osbExps),iData(dSize-osbExps:dSize));
[c,lam] = eig(Y'*Y); % c an initial guess at an eigenvector.
lam = diag(lam);
c = c(:,lam>0);
[~, pos] = min(lam(lam>0));
c = c(:,pos);
X = fliplr(hankel([zeros(dSize-osbExps-1,1); c]));
X = X(:,(osbExps+1):dSize);
X = (X'*X)\Y;
X = Y'*X - c'*(X'*X)*c;
[c, lam] = eig(X);
lam = diag(lam);
c = c(:,lam>0);
[lam, pos] = min(lam(lam>0));
c = c(:,pos);
c = c/norm(c,2);
i = 1;
while(any(lam > tol) && any(i <= iLim))
    i = i + 1;
    X = fliplr(hankel([zeros(dSize-osbExps-1,1); c]));
    X = X(:,(osbExps+1):dSize);
    X = (X'*X)\Y;
    X = Y'*X - c'*(X'*X)*c;
    [c, lam] = eig(X);
    lam = diag(lam);
    c = c(:,lam>0);
    [lam, pos] = min(lam(lam>0));
    c = c(:,pos);
    c = c/norm(c,2);
end
exps = [iExps; real(log(roots(flipud(c))))];
exps = exps(exps<0);
exps = [exps; -(1:skipLength:dSize).^(-1)'];
exps = unique(exps);
X = exp((1:dSize)'*exps');
[weights,outError,~] = lsqnonneg(X,iData);
regExps = exps(weights > 0)/skipDepth;
regWeights = weights(weights > 0);
end
```

Appendix E

Matlab Code: phiReg.m

```
function [regExps,regWeights,outError] =...
                                        phiReg(iData,iExps,iLim,skipDepth,tol)
%%%%%%%%%%%%%%%%%%%%%%%
%%% phiReg.m       %%%
%%% Jerrad Hampton %%%
%%% © 2011/2012    %%%
%%%%%%%%%%%%%%%%%%%%%%%

% This method computes exponents and weights for $\hat{\phi}^R$

%%%%%%%%%%%%%%%
%%% Inputs %%%
%%%%%%%%%%%%%%%

% iData: (Vector)
%   Data to be approximated.

% iExps: (Vector)
%   Exponents to be included.

% iLim: (Positive Integer)
%   An upper bound on number of iterations when determining exponents and
%   weights.

% skipDepth: (Positive Integer)
%   This is the skipDepth used in phiEst.m

% tol: (Positive Real)
%   Iterative steps will stop if error is within tolerance.

%%%%%%%%%%%%%%%%
%%% Outputs %%%
%%%%%%%%%%%%%%%%

% regExps: (1 Dimensional Array of Vectors)
%   Exponents for each average of comparisons, couples with regWeights for
%   function approximation.
```

```
% regWeights: (1 Dimensional Array of Vectors)
%   Weights for each average of comparisons, couples with regExps
%   for function approximation.

% outError: (Vector)
%   Contains error between input and regression.

%%%%%%%%%%%%
%%% Code %%%
%%%%%%%%%%%%
osbExps = 7; % Exponents for OsborneSmyth
dSize = size(iData,1);
evnExps = 20; % Exponents evenly distributed
skipLength = floor(dSize/evnExps);
Y = hankel(iData(1:dSize-osbExps),iData(dSize-osbExps:dSize));
[c,lam] = eig(Y'*Y); % c an initial guess at an eigenvector.
lam = diag(lam);
c = c(:,lam>0);
[~, pos] = min(lam(lam>0));
c = c(:,pos);
X = fliplr(hankel([zeros(dSize-osbExps-1,1); c]));
X = X(:,(osbExps+1):dSize);
X = (X'*X)\Y;
X = Y'*X - c'*(X'*X)*c;
[c, lam] = eig(X);
lam = diag(lam);
c = c(:,lam>0);
[lam, pos] = min(lam(lam>0));
c = c(:,pos);
c = c/norm(c,2);
i = 1;
while(any(lam > tol) && any(i <= iLim))
    i = i + 1;
    X = fliplr(hankel([zeros(dSize-osbExps-1,1); c]));
    X = X(:,(osbExps+1):dSize);
    X = (X'*X)\Y;
    X = Y'*X - c'*(X'*X)*c;
    [c, lam] = eig(X);
    lam = diag(lam);
    c = c(:,lam>0);
    [lam, pos] = min(lam(lam>0));
    c = c(:,pos);
    c = c/norm(c,2);
end
exps = [iExps; real(log(roots(flipud(c))))];
exps = [0; exps(exps<0)];
exps = [exps; -(1:skipLength:dSize).^(-1)'];
exps = unique(exps);
X = exp((1:dSize)'*exps');
[weights,outError,~] = lsqlin(X,iData,-ones(1,size(exps,1)),0, ...
    exp(exps(1:size(exps)))',1, ...
    [-Inf*ones(size(exps,1)-1,1); 0],[zeros(size(exps,1)-1,1); Inf],...
    [],optimset('Display','off','LargeScale','off'));
```

```
regExps = exps(weights~=0)/skipDepth;
regWeights = weights(weights~=0);
end
```

Appendix F

Matlab Code: thetaReg.m

```
function [regExps,regWeights,outError] =...
                                    thetaReg(iData,iExps,iLim,skipDepth,tol)
%%%%%%%%%%%%%%%%%%%%%%%
%%% thetaReg.m     %%%
%%% Jerrad Hampton %%%
%%% © 2011/2012    %%%
%%%%%%%%%%%%%%%%%%%%%%%

% This method computes exponents and weights for $\hat{\theta}^R$

%%%%%%%%%%%%%%%
%%% Inputs %%%
%%%%%%%%%%%%%%%

% iData: (Vector)
%   Data to be approximated.

% iExps: (Vector)
%   Exponents to be included.

% iLim: (Positive Integer)
%   An upper bound on number of iterations when determining exponents and
%   weights.

% skipDepth: (Positive Integer)
%   This is the skipDepth used in thetaEst.m

% tol: (Positive Real)
%   Iterative steps will stop if error is within tolerance.

%%%%%%%%%%%%%%%%
%%% Outputs %%%
%%%%%%%%%%%%%%%%

% regExps: (1 Dimensional Array of Vectors)
%   Exponents for each average of comparisons, couples with regWeights for
%   function approximation.
```

```
% regWeights: (1 Dimensional Array of Vectors)
%   Weights for each average of comparisons, couples with regExps
%   for function approximation.

% outError: (Vector)
%   Contains error between input and regression.

%%%%%%%%%%%%
%%% Code %%%
%%%%%%%%%%%%
osbExps = 7; % Exponents for OsborneSmyth
dSize = size(iData,1);
evnExps = 20; % Exponents evenly distributed
skipLength = floor(dSize/evnExps);
Y = hankel(iData(1:dSize-osbExps),iData(dSize-osbExps:dSize));
[c,lam] = eig(Y'*Y); % c an initial guess at an eigenvector.
lam = diag(lam);
c = c(:,lam>0);
[~, pos] = min(lam(lam>0));
c = c(:,pos);
X = fliplr(hankel([zeros(dSize-osbExps-1,1); c]));
X = X(:,(osbExps+1):dSize);
X = (X'*X)\Y;
X = Y'*X - c'*(X'*X)*c;
[c, lam] = eig(X);
lam = diag(lam);
c = c(:,lam>0);
[lam, pos] = min(lam(lam>0));
c = c(:,pos);
c = c/norm(c,2);
i = 1;
while(any(lam > tol) && any(i <= iLim))
    i = i + 1;
    X = fliplr(hankel([zeros(dSize-osbExps-1,1); c]));
    X = X(:,(osbExps+1):dSize);
    X = (X'*X)\Y;
    X = Y'*X - c'*(X'*X)*c;
    [c, lam] = eig(X);
    lam = diag(lam);
    c = c(:,lam>0);
    [lam, pos] = min(lam(lam>0));
    c = c(:,pos);
    c = c/norm(c,2);
end
exps = [iExps; real(log(roots(flipud(c))))];
exps = [0;exps(exps<0)];
exps = [exps; -(1:skipLength:dSize).^(-1)'];
exps = unique(exps);
X = exp((1:dSize)'*exps');
[weights,outError,~] = lsqnonneg(X,iData);
regExps = exps(weights > 0)/skipDepth;
regWeights = weights(weights > 0);
end
```

Appendix G

Matlab Code: allocation.m

```
function [draws,score] = allocation(iDepth,maxDepth,exps,weights,stepSize)
%%%%%%%%%%%%%%%%%%%%%%%%
%%% allocation.m   %%%
%%% Jerrad Hampton %%%
%%% © 2011/2012    %%%
%%%%%%%%%%%%%%%%%%%%%%%%

% This method calculates optimal draw allocation.

%%%%%%%%%%%%%%%
%%% Inputs %%%
%%%%%%%%%%%%%%%

% iDepth: (Vector)
%   Depth of initial sampling, a vector of n_j.

% maxDepth: (Positive Integer)
%   Maximum number of draws to allocate.

% exps: (Array of Vectors)
%   Each vector corresponds to a set of exponents.

% weights: (Array of Vectors)
%   Each vector corresponds to a set of weights.

% stepSize: (Integer)
%   Number of draws to allocate in each step. 1 provides exact solution.

%%%%%%%%%%%%%%%%
%%% Outputs %%%
%%%%%%%%%%%%%%%%

% draws: (Matrix)
%   Number of draws, m_j, assigned to each urn for each calculated m.

% score: (vector)
%   Score at each optimal allocation and initial allocation.

%%%%%%%%%%%%
```

```
%%% Code %%%
%%%%%%%%%%%%
numUrns = size(iDepth,1); % Number of Urns;
steps = ceil(maxDepth/stepSize);
fp = zeros(numUrns,1);
score = zeros(steps,1);
scTemp = zeros(numUrns,1);
draws = zeros(numUrns,steps);
for j = 1:numUrns
    for k = 1:size(weights{j},1)
        temp = weights{j}(k)*exp(exps{j}(k)*iDepth(j));
        scTemp(j) = scTemp(j) + temp;
        fp(j) = fp(j) + exps{j}(k)*temp;
    end
end
score(1) = sum(scTemp);
[~,j] = max(abs(fp));
draws(j,1) = stepSize;
iDepth(j) = iDepth(j)+stepSize;
for i = 2:steps
    scTemp(j) = 0;
    fp(j) = 0;
    for k = 1:size(weights{j},1)
        temp = weights{j}(k)*exp(exps{j}(k)*iDepth(j));
        scTemp(j) = scTemp(j) + temp;
        fp(j) = fp(j) + exps{j}(k)*temp;
    end
    [~,j] = max(abs(fp));
    iDepth(j) = iDepth(j)+stepSize;
    draws(:,i) = draws(:,i-1);
    draws(j,i) = draws(j,i) + stepSize;
    score(i) = sum(scTemp);
end
end
```