CONTRIBUTIONS TO ANCESTRAL INFERENCE FOR SUPERCRITICAL BRANCHING PROCESSES AND HIGH-DIMENSIONAL DATA ANALYSIS

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by
Bret Michael Hanlon
August 2009

CONTRIBUTIONS TO ANCESTRAL INFERENCE FOR SUPERCRITICAL BRANCHING PROCESSES AND HIGH-DIMENSIONAL DATA ANALYSIS

Bret Michael Hanlon, Ph.D.

Cornell University 2009

This thesis is concerned with statistical methods that are relevant in the scientific study of gene expression data. It is customary in these areas to use microarray technology as a first step in identifying the genes that are differentially expressed followed by using quantitative polymerase chain reaction (qPCR) as a confirmatory tool. The first part of thesis addresses statistical analysis for qPCR data, while the second part of the thesis addresses the so-called large p, small n problem, using microarray gene expression data as the motivating example.

Description of the gene expression profiles from PCR can be cast within the more general framework of ancestral inference for branching processes. Accordingly, part one of the thesis is devoted to the study of branching processes initiated by a random number of ancestors. We address issues concerning modeling, inference, and asymptotic justification of the proposed methodologies.

The second part of the thesis focuses on microarray data, specifically developing multivariate techniques for identifying differentially expressed genes. The results can be viewed in the more general context of multiple hypothesis testing or the multivariate testing problem.

BIOGRAPHICAL SKETCH

Bret Hanlon graduated from Jackson Memorial High School in Jackson, New Jersey in 1997. He then attended the University of North Carolina where he earned a bachelor's degree (mathematics and economics) in 2001 and a master's degree (statistics) in 2003. In 2005, Bret completed a master's degree in mathematics from Texas Tech University.

ACKNOWLEDGEMENTS

I would like to thank the Department of Statistical Science for the opportunity to study for my PhD. I would also like to thank my committee members, Professors Jim Booth, Tom DiCiccio, Rob Strawderman, and Anand Vidyashankar, for their time and advice. Finally, I thank my thesis advisor Anand Vidyashankar for his guidance, patience, and enthusiasm. I have thoroughly enjoyed working with him and learning from him.

I would also like to acknowledge the financial support that I received during my graduate studies. First of all I want to thank Professor John Guckenheimer, the director of Cornell's NSF-IGERT in nonlinear systems. In addition to supporting my studies from September 2005 through August 2007, the IGERT program provided me with a unique interdisciplinary educational opportunity. Again, I want to thank my advisor Anand Vidyashankar who supported me as a research assistant from September 2007 through August 2009.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND ORGANIZATION OF THE THESIS

This thesis is concerned with statistical methods that are relevant in the scientific study of gene expression data. It is customary in these areas to use microarray technology as a first step in identifying the genes that are differentially expressed followed by using quantitative polymerase chain reaction (qPCR) as a confirmatory tool (Ferré, 1998). While several papers have dealt with the statistical issues for the microarray problem (Dudoit et al., 2002b), statistical analysis of qPCR data has received considerably less attention.

This thesis contains two parts: Part 1 (Chapters 2, 3, and 4) addresses statistical analysis of qPCR data, while Part 2 (Chapters 5 and 6) address statistical issues related to gene expression microarray data. Statistical analysis of qPCR data can be cast within a more general framework of ancestral inference for branching processes. Accordingly, Part 1 of the thesis is devoted to the study of branching processes initiated by a random number of ancestors. We address issues concerning modeling, inference, and asymptotic justification of the proposed methodologies.

We begin with a brief description of qPCR. The polymerase chain reaction (PCR) is a biochemical technique used to amplify the number of copies of a specific DNA fragment. For qPCR, the scientific goal is estimation of the initial number of molecules present in a genetic material. qPCR is an important and widely used tool for gene expression experiments (Ferré, 1998; Kubista et al., 2006; Nolan et al., 2006). A typical PCR experiment is run for 40 cycles; theoretically, the number of molecules doubles in every cycle. In practice, only some fraction of the molecules actually replicate in a given cycle. Hence, a supercritical Galton-Watson branching process with a Bernoulli offspring distribution provides a natural model to describe the dynamics of PCR. Under a branching process model, the question of quantitation is tantamount to estimating the initial number of ancestors of the process. There are several papers which model PCR as a branching process and then estimate the initial number of ancestors (Nedelman et al., 1992; Jacob and Peccoud, 1998; Lalam, 2007; Lalam and Jacob, 2007). This work is based on observing a single realization of a branching process. In fact, a typical qPCR experiment produces data from 96 or 384 separate reactions.

Ancestral inference for branching processes is characterized by four parameters: the ancestor mean, the offspring mean, ancestor variance, and offspring variance. In Chapter 2, we focus on the binary branching process model suggested by the PCR dynamics. We propose a new design, namely replicated PCR experiments, and develop the associated branching process model with random effects to account for variability between replicates. We develop generalized method of moments estimators for the model parameters and establish their consistency and asymptotic normality. The advantage of our method compared to traditional (non-branching process) methods, such as the comparative C_T and the standard curve method (Livak, 2001), is that our estimates have smaller bias and tighter confidence intervals. Furthermore, because our method incorporates sources of variability into the model in a systematic way it yields confidence intervals that attain the nominal coverage probability. These advantages are illustrated through both simulation and experimental data.

Chapter 3 develops analysis of variance (ANOVA) models for supercritical branching processes. These models are useful to compare the expression profiles of multiple genes or more generally the mean number of ancestors across various "groups." The proposed methodology is dependent on the equality of the offspring means. For this reason, we undertake a detailed description of the methodology for comparison of the ancestral means and offspring means. This paper does not incorporate a random effect for replicates; however, it treats branching processes with unbounded support. The methods are again illustrated through both simulation and qPCR data.

Chapter 4 focuses on limit theorems for branching processes under minimal moment conditions. These results, besides being useful for inferential purposes, develop new aspects of some of the classical martingales studied in the branching process literature. These results are based on new ideas concerning the harmonic moments of random variables. We establish the joint asymptotic normality of the appropriately centered and scaled versions of the estimators of the ancestor mean, offspring mean, and offspring variance. We show that the asymptotic limit distributions of the offspring mean and offspring variance are independent while that of the ancestor mean and ancestor variance are not independent. Furthermore, the asymptotic limit distributions of appropriately centered and scaled ancestor mean and offspring mean are independent.

Part 2 of the thesis explores the so-called large p, small n problem, using microarray gene expression data as the motivating example. Chapter 5 develops a multivariate technique for identifying differentially expressed genes using the sup-norm statistic recently studied in Kuelbs and Vidyashankar (2009). Chapter 6 addresses the question of robust and adaptive multivariate methods.

The results presented in Chapters 5 and 6 can be viewed in the more general context of multiple hypothesis testing or the multivariate testing problem. The results presented in the thesis are strong competitors for some of the traditionally available methods in the multivariate literature.

CHAPTER 2

INFERENCE FOR QUANTITATION PARAMETERS IN POLYMERASE CHAIN REACTIONS VIA BRANCHING PROCESSES WITH RANDOM EFFECTS

2.1 Introduction

The polymerase chain reaction (PCR) is a biochemical technique used to amplify the number of copies of a specific DNA fragment. It is one of the most utilized scientific experiments (Mullis et al., 1994; Kubista et al., 2006; Nolan et al., 2006) with applications in diverse areas such as forensic science, virology, and gene expression (Ferré, 1998). We are specifically concerned with quantitative PCR (qPCR), where the goal is inference concerning the initial number of molecules present in a genetic material. Informally, our goal is to statistically identify, using Figure 2.2, the ratio of mean intensities of blue to mean intensities of red in cycle 0 using the data from cycles 21 through 40.

In this paper, we propose a new replicated design for PCR experiments. We then develop a novel generalized method of moments approach for inference concerning the initial number of molecules. We establish the asymptotic validity of our approach and show, using simulations, that our methodology yields results that are close to the nominal values in small samples. Furthermore, in the analysis of experimental data, our method yields estimates with a relative bias of approximately 4.4% compared to traditional methods which yield point estimates with at least 15% relative bias.

2.1.1 PCR concepts

We begin with a summary of critical details concerning PCR; see Mullis et al. (1994) and Ferré (1998) for references and more detailed descriptions. As described above, PCR is a biochemical technique for replicating DNA. A PCR vial contains the following components for DNA duplication: a piece of DNA, large quantities of nucleotides, large quantities of primer sequence, and the DNA polymerase. A single cycle of PCR consists of three steps: denaturing, annealing, and extension. During the denaturing step, the vial is heated thereby separating the two DNA chains in the helix. In the annealing step the vial is cooled causing the primers to anneal to the ends of DNA strands. Finally, in the extension step the vial is heated again and the polymerase begins adding nucleotides to the primer and eventually makes a complementary copy of the DNA template. This completes the first cycle. At the end of the first cycle, it is believed that each piece of DNA in the vial has been duplicated. The cycles are repeated several times. An animation of this process can be found on the website `http://www.sumanasinc.com/webcontent/animations/content/pcr.html`.

It is not always the case that at the end of each cycle the number of DNA molecules have doubled. By the nature of the experiment, it follows that denaturing, annealing, and extension of each template are independent of one another and hence each template either duplicates or does not duplicate independently of the other. Since the amount of resources available to all the templates are the same within a cycle, the probability of duplication, denoted by p, is the same for all the templates within a cycle. If enough resources are available, then one can assume that this probability remains constant between cycles. The parameter p is called the efficiency of the PCR. Thus, under the assumption that enough resources are available during all cycles, the above biochemical process has been modeled as a single type branching process with offspring distribution $P(N_1 = 1) = 1 - p$ and $P(N_1 = 2) = p$, where N_1 is the number of templates at the end of the first cycle starting with a single template. Early probabilistic results modeling PCR as a Galton-Watson branching process have been discussed in Krawczak et al. (1989), Reiss et al. (1990), Hayashi (1990), and Maruyama (1990).

The amount of genetic material from each reaction is measured via the intensity of the fluorescence signal and not the number of molecules; these fluorescence measurements are available for every cycle of the reaction. The cycles of a PCR can be classified into three phases depending on the amount and rate of accumulated genetic material. These are (i) the initial phase, (ii) the exponential phase, and (iii) the plateau phase. Figure 2.1, which plots fluorescence data from three reactions, illustrates these three phases. An important feature of qPCR is that fluorescence can only be detected above a given threshold; it is this detection limitation that separates the initial phase from the exponential phase. More precisely, during the initial phase the product accumulates exponentially but sufficient genetic material is not present for an accurate fluorescence reading. The exponential phase corresponds to cycles at which sufficient genetic material is available yielding detectable fluorescence levels. During this phase, the plot of the fluorescence intensity against the cycle number displays an exponential curve. After a few cycles of exponential growth, the rate of fluorescence accumulation begins to slow down resulting in linear growth. Beyond this linear phase, the accumulation stops and intensity plateaus yielding the plateau phase.

The relationship between the number of molecules and the fluorescence intensity is a subject of intense research in biochemistry and related scientific areas.

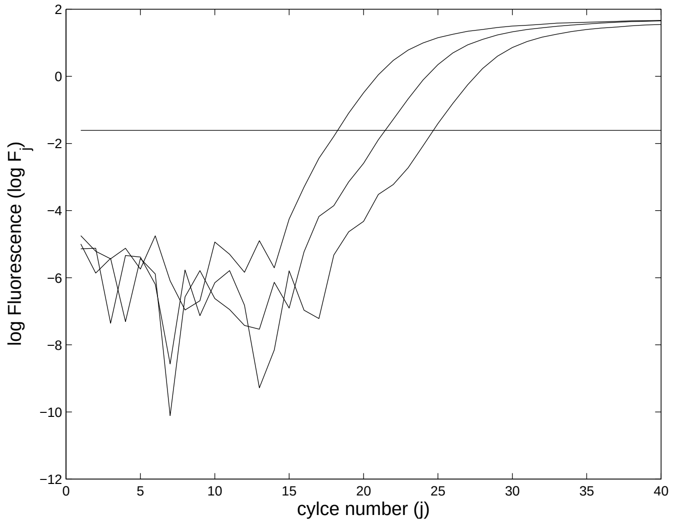


Figure 2.1: This figures illustrates the three phases of PCR, plotting log fluorescence versus cycle number ($\log F_j$ vs. j) for three separate reactions.

The number of molecules N corresponding to a fluorescence intensity F depends on factors such as amplicon size, polymerase, and the type of dye used (Rutledge, 2004; Alvarez et al., 2007). Under optimal experimental conditions, the number of molecules is a constant multiple of the intensity and this constant depends on the PCR system used for the amplification. In fact, Rutledge (2004) established that

$$N = \left(\frac{CF \times 9.1 \times 10^{11}}{A_S} \right) F \tag{2.1}$$

where A_S is the amplicon size and CF is the calibration factor, which represents the number of nanograms of double-stranded DNA per fluorescence unit (ng/FU). For statistical purposes it is convenient to view CF as a random variable with mean $c^\star$ and variance σ^2_{CF}. Thus, the precision of our inference for quantitation parameters is dependent on the assumptions concerning $c^\star$ and σ^2_{CF}. We examine this question in the simulation section.

2.1.2 PCR Quantitation

Quantitation, or absolute quantitation, refers to estimation of the copy number, i.e. the number of DNA molecules in the initial amount of genetic material. One way to estimate the copy number is to first estimate the fluorescence intensity and convert it to the number of molecules using (2.1). However, the fluorescence intensity corresponding to the target is noisy and hence the amplification process is needed to accumulate sufficient product and use this to estimate the fluorescence intensity corresponding to the target material. Thus, an understanding of the accumulation process, its relationship to the target material, and an accurate knowledge of the parameter $c^\star$ are needed for absolute quantitation. qPCR experiments have been performed for estimating the parameter $c^\star$ (Goll et al., 2006; Ming and Kwok, 2003; Monis et al., 2005).

Relative quantitation is concerned with the estimation of the ratio of the copy numbers of a target genetic material to that of a reference genetic material. As in absolute quantitation, this is achieved using the fluorescence intensities of the accumulated products of the target and reference genetic materials. Under the assumption that $c^\star$ is the same for both the amplifications, the estimate of the ratio does not involve this parameter.

Traditional techniques for quantitation involve a linear or non-linear regression model for the fluorescence data. These methods are based on either the log-linear relationship between the initial number of molecules and the threshold cycle C_T or the sigmoidal nature of the cycle number versus the fluorescence intensity curve. The log-linear relationship is developed using a deterministic model for the amplification of the DNA molecules. Furthermore, this deterministic model can be obtained as the expectation of the n^{th} generation of the branching process.

Other contemporary efforts in quantitation involve non-homogeneous models for the amplification processes. The primary drawback of these approaches is that all available data are not utilized for statistical analyses. Furthermore, the estimates of the standard error provided by these techniques are usually incorrect because they do not account for all sources of variation.

The branching process model described above is more suited for the underlying count of DNA molecules rather than for the fluorescence data. However, the state space of the branching process, which describes the number of DNA molecules, can be estimated using (2.1) or other methods.

2.1.3 Outline

The primary purpose of this paper is to clarify, improve, and address some of the limitations of the existing statistical methods and models concerning inference for the quantitation parameters of PCR. Motivated by this, we introduce a mixture model, namely a collection of conditionally independent branching processes each initiated by a random number of initial molecules, as a statistical model for data from PCR experiments. We undertake a systematic asymptotic approach towards inference. We emphasize here that the asymptotic framework used in this paper is not intended to understand the behavior of the reactions when infinitely many cycles are run; rather, we seek to explain several features that arise in a typical PCR from the behavior of the limits of various statistical quantities encountered in the data analyses. This is facilitated by the exponential rate of convergence of various statistics involved in our study. Thus the asymptotic framework serves to set up standards for comparing various results.

In Section 2 we describe the data structures encountered in a typical PCR experiment and the sources of variability associated with them. Additionally, we develop a hierarchical, non-homogenous binary branching process model for describing the dynamics of PCR experiments. In Section 3 we state our main results concerning the quantitation parameters of PCR experiments. We apply our methodology to simulated data in Section 4 and experimental data in Section 5. Section 6 contains concluding remarks. The proofs are given in Section 7.

2.2 Data Structures and Statistical Models

We begin with a description of the data for a single real-time qPCR. Let $\{F_j : j = 0, 1, 2, \cdots\}$ denote the fluorescence intensities at various cycles and $\{N_j : j = 0, 1, 2, \cdots\}$ denote the underlying unobservable branching process initiated by N_0 molecules. Let $m_a = E(N_0)$ and $\sigma_a^2 = Var(N_0)$. Let m_e denote the efficiency parameter describing the mean splitting rate of the process and σ_e^2 denote the variance of the splitting rate. Under the binary splitting model, the efficiency of the PCR is given by $p = m_e - 1$ and $\sigma_e^2 = (m_e - 1)(2 - m_e)$. Our focus is on the vector of parameters (m_a, m_e), which we call the quantitation parameters. However, estimates of the variance parameters σ_a^2 and σ_e^2 are needed for standard error calculations and for other inferential purposes. It is assumed, and experimentally verified (Goll et al., 2006), that the amount of fluorescence is proportional to the number of DNA molecules; that is $N_j = cF_j$, where, using (2.1), $c = \frac{c^{\star} \times 9.1 \times 10^{11}}{A_S}$. We will assume that c is known even though in several PCR assays c is estimated. When c is estimated from data, our results will hold conditionally on c, even though marginal results depend on the variability introduced due to the estimation of c. We will return to this issue in the simulation section.

2.2.1 Replicated Experiments

A typical qPCR experiment produces data from either 96 or 384 separate reactions. Replication of the PCR experiment under identical conditions yields data that can be represented as $\{F_{k,j} : k \geq 1, j \geq 1\}$; that is, $F_{k,j}$ represents the fluorescence intensity from the j^{th} cycle of the k^{th} replicate (or well). Let the underlying unobservable branching process of DNA molecules associated with the fluorescence intensities be denoted by $\{N_{k,j}, j \geq 0\}$. We will assume that the processes $\{N_{k,j}, \quad j \geq 0\}$ are independent and identically distributed (i.i.d.) for all $k \geq 1$, resulting in $F_{k,j}$ being i.i.d. random variables in k. Let $r(n)$ denote the number of replicates. Then the data for the problem are $\{F_{k,j}, 1 \leq k \leq r(n), j = 1, 2, \cdots n\}$. To ensure that these assumptions are satisfied, it is important that replicate experiments be carried out by splitting a master-mix containing all components of the PCR. This minimizes the unwarranted variability introduced by non-uniform presence of amplifying substance between replicates.

2.2.2 Dilution Experiments

Another set of experimental data usually collected, is the so-called dilution data. In such data, the mean initial number of molecules is $E(N_{k,0}) = m_a d_k$, where d_k are called the dilution constants. The d_k's are controlled by the scientist, modulo experimental error, and are considered known constants. Contrary to the name, these constants can take values larger than one. Thus, PCR experiments with dilution consists of fluorescence data as described in the previous section, but the initial number of molecules are assumed to be independent but not identically distributed. Thus the statistical model used for dilution experiments are i.i.d.

branching processes generated by independent initial random variables rather than i.i.d. initial random variables. For $d_k = 1$, the dilution data reduce to replicated data.

2.2.3 Sources of variability and non-homogeneous models

Accuracy of inference concerning the quantitation parameters depends on identifying different sources of variability and controlling for them. In a PCR experiment, variability is introduced at various stages of the experiment. First, it is impossible to exactly identify the number of DNA molecules even for a known genetic material. Hence, from a practical perspective, it is convenient to view the initial number as a random variable and the variance of this random variable determines the precision of the estimator of m_a. Since N_0 is unobservable, and the results are based on accumulated product, the variability in the rate of accumulation plays an important role. Experimental evidence shows that the rate of accumulation changes within a reaction and between reactions. The changes within a reaction are essentially due to various biochemical reasons while the variability between reactions can be caused due to many factors one of which is the so called pipetting error.

Variants of simple branching processes have been suggested to model the evolution of the process across cycles. For example, Schnell and Mendoza (1997) suggest a kinetic model to describe the PCR dynamics and Jagers and Klebaner (2003a) use a size-dependent branching process model as a discrete time approximation to the dynamic model that allows for stochasticity. Lalam et al. (2004a) suggest a size dependent model for the non-exponential phase and a simple branching process model for the exponential phase. Since we focus mainly on the exponential phase for quantitation, we deal with the branching process model.

The variability in efficiency between reactions is important and needs to be addressed. From a statistical modeling perspective, not much information concerning variability within a reaction is available. Hence while non-homogeneous models can be proposed, it is difficult to ascertain the practical effect of such modeling. However, the non-homogeneity can be addressed empirically by using reaction and cycle dependent estimators of efficiency.

On the other hand, information concerning between reaction variability can be extracted by using either replicated or dilution data. Thus, we account for between reaction variability, by modeling the efficiency of the exponential phase of the k^{th} reaction as a random variable p_k with some distribution G on (0,1). During the exponential phase, experimental evidence suggests that the support of G lies in the interval $(1-\epsilon, 1)$ for some "small" ϵ. Thus, the model proposed for describing PCR dynamics is,

$$p_k \overset{i.i.d.}{\sim} G(.), \quad 1 \leq k \leq r(n), \tag{2.2}$$

$$N_{k,0} \overset{independent}{\sim} H_k(.), \quad 1 \leq k \leq r(n), \tag{2.3}$$

and, given $N_{k,0},\ p_k,\ \{N_{k,j} : j \geq 1\}$ is a binary branching process initiated by $N_{k,0}$ ancestors with splitting probability p_k. That is,

$$N_{k,j+1}|N_{k,j}, p_k \sim N_{k,j} + Bin(N_{k,j}, p_k),\ \text{for all } j \geq 0, \tag{2.4}$$

where $Bin(N_{k,j}, p_k)$ is a binomial random variable with parameters $N_{k,j}$ and p_k. The sequence $\{p_k : k \geq 1\}$ representing the splitting rate is assumed to be independent of the sequence $\{N_{k,0} : k \geq 1\}$ of initial amounts of genetic materials. We call this model a branching process model incorporating random effects or more generally a mixture model using branching processes. Recently, similar ideas have been used to model between subject variability. For instance, Altman (2007) studies hidden Markov models with random effects, and Hyrien and Zand (2008) use mixture models in multi-type age-dependent branching processes in the context of CFSE-labeling experiments.

We will use the notation $E(N_{k,0}) = m_a d_k$, $Var(N_{k,0}) = \sigma_a^2 d_k^2$, $E(N_{k,0}^j) = m_{j,0} d_k^j$, $j = 3, 4$, for all $k \geq 1$. The behavior of the constants d_k will be critical in the small and large sample study of our methodology. In particular, we will encounter the behavior of $D_j(n) = (r(n))^{-1} \sum_{k=1}^{r(n)} d_k^j$ for $j = 1, 2, 3, 4$. The following condition concerns the stability of the sequence $D_j(n)$ as n increases and is analogous to conditions of similar type adopted when studying regression problems.

Condition 1. *Assume that $D_j(n) \to D_j > 0$ as $n \to \infty$ for all $j = 1, 2, 3, 4$. Furthermore, assume that $\sum_{k \geq 1} k^{-2} d_k^{2j} < \infty$ for j=1, 2.*

Condition 1 will be called the regularity of the dilution constants and we will assume that this condition holds throughout the paper. The condition on $D_4(n)$ and the convergence of $\sum_{k \geq 1} k^{-2} d_k^4$ are needed only for studying the consistency of the variance estimate.

2.3 Inference for Copy Number

2.3.1 Moments and Martingales

It is instructive to write down the model in a more transparent form as follows:

$$N_{k,j+1} = \sum_{l=1}^{N_{k,j}} \xi_{k,j,l}, \tag{2.5}$$

where the random variable $\xi_{k,j,l}$ represents whether the l^{th} DNA strand in the j^{th} cycle of the k^{th} reaction splits or not. In terms of the random variables $\xi_{k,j,l}$, our assumption states that for every fixed k and p_k, the random variables are i.i.d. with distribution

$$P(\xi_{k,j,l} = 2|p_k) = p_k \quad P(\xi_{k,j,l} = 1|p_k) = 1 - p_k. \tag{2.6}$$

Thus, $E(\xi_{k,j,l}|p_k) = 1 + p_k \equiv m_k$ and $Var(\xi_{k,j,l}|p_k) = p_k(1 - p_k) \equiv \sigma_k^2$. We will use the notation $E_k(.)$ to denote the conditional expectation $E(.|p_k)$ and $Var_k(.)$ to denote the conditional variance $Var(.|p_k)$. Hence, it follows that

$$E_k(N_{k,j+1}) \quad = \quad E_k(E_k(N_{k,j+1}|N_{k,j})) = m_k E_k(N_{k,j}). \tag{2.7}$$

Iterating the above identity, it follows that

$$E_k(N_{k,j+1}) = m_k^{j+1} m_a d_k. \tag{2.8}$$

In the case of fluorescence data, (2.8) and (2.1) imply that $E_k(F_{k,j+1}) = c^{-1} m_k^{j+1} m_a d_k$. Thus, conditioned on the random effect p_k, $V_{k,j} \equiv m_k^{-j} N_{k,j}$ is a positive martingale sequence with respect to the sigma field containing information up to $(j-1)$ cycles and the value of the random effect. A consequence of this observation is that as $j \to \infty$, $V_{k,j}$ converges to $V_k^\star$, where $V_k^\star > 0$ (Athreya and Ney, 1972). Furthermore, since $V_{k,j}$ has uniformly bounded marginal and conditional moments of at least order four (Lemma 6), the sequence $\{V_{k,n}^2 : n \geq 1\}$ is uniformly integrable and hence $V_{k,n}$ converges in L_2 to $V_k^\star$.

The moments of the limit random variable $V_k^\star$ will play an important role in our analyses. It is easy to see that the marginal and the conditional means of $V_k^\star$

coincide and are given by $m_a d_k$. Next,

$$\begin{aligned} Var_k(V_k^\star) &= E_k(Var_k(V_k^\star|N_{k,0})) + Var_k(E_k(V_k^\star|N_{k,0})) && (2.9) \\ &= E_k(N_{k,0}\frac{\sigma_k^2}{m_k(m_k-1)}) + Var_k(N_{k,0}) && (2.10) \\ &= m_a d_k \frac{\sigma_k^2}{m_k(m_k-1)} + \sigma_a^2 d_k^2, && (2.11) \end{aligned}$$

where the last equality follows from the assumed independence of $N_{k,0}$ and p_k. Now, since the conditional expectation of $V_k^\star$ is $m_a d_k$, the variance of the conditional expectation is zero. Hence, using $\sigma_k^2 = p_k(1-p_k)$ and $m_k = 1 + p_k$, the marginal variance of $V_k^\star$ is given by

$$\omega_k^2 = m_a d_k E(\frac{1-p_1}{1+p_1}) + \sigma_a^2 d_k^2, \qquad (2.12)$$

where we used that p_k's have the same distribution. The marginal variance of the limiting random variable $V_k^\star$ depends on the reaction only via the dilution factor used in that reaction.

2.3.2 Absolute Quantitation

Information about m_a is contained both in $N_{k,0}$ and in $V_k^\star$. Let us assume, for the moment that $d_k = 1$. If one can obtain a random sample of size r from $N_{k,0}$, then the resulting sample mean is an unbiased estimator of m_a. The variance of this estimator is then $r^{-1}\sigma_a^2$ and the problem would be completely resolved. However, since it is not possible to obtain observable samples from $N_{k,0}$, one could use instead the sample mean of a random sample from $V_k^\star$ to estimate m_a. The variance of this estimator would be $r^{-1}(\sigma_a^2 + m_a E(\frac{1-p_1}{1+p_1}))$. Since $V_k^\star$ are unobservable, once again this recipe is not feasible. The discussion however suggests that if one were to average observable data over replicates then it is plausible that consistent estimators of m_a may exist. Since, one can obtain fluorescence data at every cycle, it is natural to use data from the cycles in the exponential phase to obtain estimators of m_a. Thus, the first step is to identify cycles belonging to the exponential phase.

It is common to associate cycles in the exponential phase with a parameter called C_T. C_T is defined to be that cycle at which the accumulated product crosses a specified threshold. This threshold is a user defined quantity. In Section 5 we propose an alternative method to identify the cycles in the exponential phase. Let τ_k and n_k denote the first and last cycles of the exponential phase, respectively, in the k^{th} reaction. Then, the cycles in the exponential phase of that reaction can be denoted by $\tau_k, \tau_{k+1}, \cdots n_k$. To make the conditions more transparent when studying asymptotics, we will take $n_k = n$ and $\tau_k = \tau$. This does not entail any loss of generality and also minimizes cumbersome notation. Alternate conditions involving $\wedge_{k=1}^{r(n)} n_k$ and $\wedge_{k=1}^{r(n)} \tau_k$ can be written down for large sample analysis. In our data analysis, we do not make this assumption.

Since more than one cycle is involved during the exponential phase, we consider the total accumulated fluorescence during the exponential phase, namely

$$Y_{k,n} = \sum_{j=\tau}^{n} F_{k,j}. \tag{2.13}$$

Our formulation of the inference problem in terms of the generalized method of moments technique will involve the behavior of $Y_{k,n}$ and not $F_{k,n}$. The proposition below describes the asymptotic behavior of $Y_{k,n}$ for every reaction k.

Proposition 1. *Under the assumptions of our model, conditioned on the random effect p_k, with probability one*

$$\lim_{n\to\infty} \frac{Y_{k,n}}{m_k^n} = c^{-1}(\frac{m_k}{m_k - 1})V_k^{\star}. \tag{2.14}$$

Motivated by the above proposition, we consider the following generalized method of moments estimator of m_a, given by

$$\tilde{m}_{a,n} = \frac{c}{r(n)D_1(n)} \sum_{k=1}^{r(n)} \frac{\tilde{p}_{k,n}}{\tilde{m}_{k,n}^{n+1}} Y_{k,n}, \tag{2.15}$$

where $\tilde{p}_{k,n}$ is an estimator of the efficiency for the k^{th} reaction and $\tilde{m}_{k,n} = 1 + \tilde{p}_{k,n}$. For replicated data, $D_j(n) \equiv 1$. The estimator $\tilde{m}_{a,n}$ takes into account the variability in amplification rates between cycles and scales the product from the k^{th} reaction by the amplification rate of that reaction. The factor c is needed to convert the fluorescence information into number of molecules. As one would expect, the asymptotic properties of $\tilde{m}_{a,n}$ depend on the properties of the estimator of efficiency. While several estimators for efficiency are available, we use the weighted conditional least squares estimator of the reaction efficiency since it is based on the total accumulated fluorescence during the exponential phase. The estimator is given by,

$$\tilde{p}_{k,n} = \frac{\sum_{j=\tau}^{n-1}(F_{k,j+1} - F_{k,j})}{\sum_{j=\tau}^{n} F_{k,j}} = \frac{Y_{k,n} - F_{k,\tau} - Y_{k,n-1}}{Y_{k,n-1}}. \tag{2.16}$$

The challenge to studying the asymptotic properties of our estimators is related to the phenomenon called the propagation of variability. The power of n in the estimators increase the variability, and the rate of convergence of $\tilde{p}_{k,n}$ to p_k when raised to the n^{th} power becomes a critical issue. We develop a novel method to address this issue using harmonic moments (see Lemmas 1, 3, and 4). Our first result describes the consistency and asymptotic normality properties of our estimator of m_a.

Theorem 1. *Assume that the dilution constants are regular and define $D_L = D_1 D_2^{-1}$. Let the number of replicates $r(n)$ be such that $r(n) \to \infty$ with $r(n)n^{-1} \to 0$ as $n \to \infty$. Then, $\tilde{m}_{a,n}$ is a strongly consistent estimator of m_a. Furthermore,*

$$\sqrt{r(n)D_1(n)}(\tilde{m}_{a,n} - m_a) \xrightarrow{d} H, \tag{2.17}$$

where $H \sim N(0, \sigma_L^2)$, *where* $\sigma_L^2 = m_a E(\frac{1-p_1}{1+p_1}) + D_L \sigma_a^2$.

Thus, it follows from the theorem that

$$\tilde{m}_{a,n} \stackrel{\cdot}{\sim} N(\tilde{m}_{a,n}, \frac{\tilde{\sigma}_{L,n}^2}{r(n)D_1(n)}), \tag{2.18}$$

where $\tilde{\sigma}_{L,n}^2$ is an estimate of the variance based on the data. When $p_1 \equiv 1$, then the genetic material exactly doubles in each cycle, and the only variation in the estimation comes from the variability in the initial amount of genetic material. If $\sigma_a^2 = 0$, then one can quantitate exactly and the results reduce to classical results from the PCR literature. Of course, neither of these are feasible and the above theorem shows the precise nature of the variability in the quantitation process, providing a decomposition along the lines of classical analysis of variance. Finally, notice that for the replicated data structure $(D_1(n), D_L) = (1, 1)$ and the limiting variance does not involve the dilution parameters.

2.3.3 Relative Quantitation

We now turn to relative quantitation. In this case, we have two sets of genetic materials, the calibrator and the target. We will add a subscript C and T to our notation to distinguish between data collected from calibrator and target materials. Hence $F_{k,j,C}$ and $F_{k,j,T}$ will represent the fluorescence from the j^{th} cycle of the k^{th} reaction from the calibrator and target materials, respectively. The unobservable branching process associated with these fluorescence data will be denoted by $N_{k,j,C}$ and $N_{k,j,T}$ respectively. Let $E(N_{k,0,C}) = m_{a,C}d_k$ and $E(N_{k,0,T}) = m_{a,T}d_k$. Let $\sigma_{a,C}^2 d_k^2$ and $\sigma_{a,T}^2 d_k^2$ denote the variance of $N_{0,C}$ and $N_{0,T}$, respectively. To complete the description of the model, we will assume that $\{p_{k,C}, p_{k,T} : k \geq 1\}$ to be a collection of independent random variables with distribution $G_C(.)$ and $G_T(.)$ respectively and that the support of G_C and G_T are $(1-\epsilon, 1)$.

In relative quantitation, the object of interest is R, where

$$R = \frac{m_{a,T}}{m_{a,C}}. \tag{2.19}$$

Analogous to the absolute quantitation case, we can define the non-parametric maximum likelihood estimator of the reaction efficiency using the accumulated fluorescence as follows:

$$\tilde{p}_{k,n,C} = \frac{Y_{k,n,C} - F_{k,\tau,C} - Y_{k,n-1,C}}{Y_{k,n-1,C}}, \quad \tilde{p}_{k,n,T} = \frac{Y_{k,n,T} - F_{k,\tau,T} - Y_{k,n-1,T}}{Y_{k,n-1,T}}. \tag{2.20}$$

This yields, $\tilde{m}_{k,n,I} = 1 + \tilde{p}_{k,n,I}$ and $\hat{m}_{k,n,I} = 1 + \hat{p}_{k,n,I}$ for $I = C, T$. Hence, one can now estimate the ratio R using

$$\tilde{R}_n = \frac{\tilde{m}_{A,n,T}}{\tilde{m}_{A,n,C}}, \tag{2.21}$$

where

$$\tilde{m}_{a,n,T} = \frac{1}{r(n)} \sum_{k=1}^{r(n)} \frac{\tilde{p}_{k,n,T}}{\tilde{m}_{k,n,T}^{n+1}} Y_{k,n,T}, \quad \tilde{m}_{a,n,C} = \frac{1}{r(n)} \sum_{k=1}^{r(n)} \frac{\tilde{p}_{k,n,C}}{\tilde{m}_{k,n,C}^{n+1}} Y_{k,n,C}. \tag{2.22}$$

Our main result concerning the relative quantitation is given below.

Theorem 2. *(Relative Quantitation) Under the assumptions of Theorem 1, $\tilde{R}_n$ is a strongly consistent estimator of R. Furthermore,*

$$\sqrt{r(n) D_1(n)} (\tilde{R}_n - R) \xrightarrow{d} G_2, \tag{2.23}$$

where $G_2 \sim N(0, \sigma_R^2)$. The limiting variance σ_R^2 is given by

$$\sigma_R^2 = R^2 (\sigma_{L,T}^2 + \frac{\sigma_{L,C}^2}{m_{a,C}^2}), \tag{2.24}$$

where

$$\sigma_{L,I}^2 = m_{a,I} E(\frac{1 - p_{1,I}}{1 + p_{1,I}}) + D_L \sigma_{a,I}^2, \quad I = C, T, \tag{2.25}$$

and D_L is as in Theorem 1.

2.3.4 Bias Correction

The estimator of m_a proposed in the previous section can be improved by accounting for the cycles during the initial noisy phase. To address this issue, we observe that the mean fluorescence during the exponential phase of the k^{th} reaction is given by $c^{-1}m_k^{n+1}(m_k-1)^{-1}(1-m_k^{\tau-(n+1)})$. Since, $(1-\tilde{m}_{k,n}^{\tau-(n+1)})$ converges to one exponentially fast, one can show that the bias corrected estimator

$$\tilde{m}_{a,n}^{(b)} = \frac{c}{r(n)D(n)} \sum_{k=1}^{r(n)} \frac{\tilde{p}_{k,n}}{\tilde{m}_{k,n}^{n+1}}(1-\tilde{m}_{k,n}^{\tau-(n+1)})^{-1}Y_{k,n}, \tag{2.26}$$

inherits the asymptotic properties of $\tilde{m}_{a,n}$. For this reason, we use and recommend this estimator for data analysis.

2.3.5 Inference for PCR Efficiency

As seen in the previous section, inference for quantitation depends critically on the estimator of the efficiency of the PCR experiment. As described so far in the text, there are two notions of efficiency; the reaction efficiency (conditional efficiency) and the marginal efficiency. The reaction efficiency is useful for quantitation purposes and is estimated as the conditional weighted least squares estimator and is given by the (2.16). The following proposition describes the asymptotic limit distribution of $\tilde{p}_{k,n}$.

Proposition 2. *Under the assumptions of our model, for every fixed k,*

$$\sqrt{Y_{k,n-1}}(\tilde{p}_{k,n}-p_k) \xrightarrow{d} H_2, \tag{2.27}$$

where $P(H_2 \le x) = \int_{1-\epsilon}^{1} \Phi(\frac{x}{t(1-t)})dG(t)$.

The marginal efficiency, which is helpful in determining the efficiency of the PCR equipment, is defined to be Ep_1. The estimator of marginal efficiency, is obtained by averaging the reaction efficiencies and is given by

$$\tilde{p}_{n,pool} = \frac{1}{r(n)} \sum_{k=1}^{r(n)} \tilde{p}_{k,n}. \tag{2.28}$$

Our next result is concerned with strong consistency and asymptotic normality of the estimator of marginal efficiency.

Theorem 3. *Under the assumptions of Theorem 1, $\tilde{p}_{n,pool}$ is a strongly consistent estimator of the overall efficiency of the PCR, namely $E(p_1)$. Furthermore,*

$$\sqrt{r(n)}(\tilde{p}_{n,pool} - E(p_1)) \xrightarrow{d} H_1, \tag{2.29}$$

where $H_1 \sim N(0, \sigma_G^2)$, where σ_G^2 is the variance of the random variable p_1.

We notice that the rate of convergence is only $\sqrt{r(n)}$ for marginal efficiency where as it is "exponential" for the reaction efficiency.

2.3.6 Estimation of Variability

Estimation of the variability is important for performing inference concerning the quantitation parameters. In this section we provide consistent estimators of the limiting variance σ_L^2 and that of σ_a^2. We begin with σ_L^2. Define

$$\tilde{\sigma}_{L,n}^2 = \frac{c^2}{r(n)D_1(n)} \sum_{k=1}^{r(n)} (\frac{\tilde{p}_{k,n}}{\tilde{m}_{k,n}^{n+1}} Y_{k,n} - \tilde{m}_{a,n} d_k)^2, \tag{2.30}$$

and

$$\tilde{\theta}_{1,n} = \frac{1}{r(n)} \sum_{k=1}^{r(n)} \frac{1-\tilde{p}_{k,n}}{1+\tilde{p}_{k,n}} \quad \tilde{\theta}_{2,n} = \frac{1}{r(n)} \sum_{k=1}^{r(n)} \frac{\tilde{p}_{k,n}}{1+\tilde{p}_{k,n}}. \tag{2.31}$$

Theorem 4. *Under the assumptions of Theorem 1, $\sigma^2_{L,n}$ is a consistent estimator of σ^2_L. Furthermore, $\tilde{\theta}_{1,n}$ and $\tilde{\theta}_{2,n}$ are consistent estimators of $E(\frac{1-p_1}{1+p_1})$ and $E(\frac{p_1}{1+p_1})$ respectively.*

An immediate consequence of Theorem 4 is the following corollary concerning consistent estimation of σ^2_a.

Corollary 1. *Define*

$$\tilde{\sigma}^2_{a,n} = \frac{\tilde{\sigma^2}_{L,n} - \tilde{m}_{a,n}\tilde{\theta}_{1,n}}{D_{L,n}}, \tag{2.32}$$

where $D_{L,n} = D_2(n)D_1^{-1}(n)$. Then, $\sigma^2_{a,n}$ is a consistent estimator of σ^2_a.

2.4 Simulation Experiment

In this section we describe our simulation results to evaluate the performance of the proposed methodology and compare them with other procedures studied in the literature. The results in this section are based on 5000 simulations.

We compare our branching process estimator to two common techniques for relative quantitation, the comparative C_T method and the standard curve method. Both methods summarize the PCR with the single value C_T. The comparative C_T method assumes perfect doubling for both the target and calibrator and estimates R using the formula $\hat{R} = 2^{C_{T,C}-C_{T,T}}$, where $C_{T,C}$ is the C_T value for the calibrator and $C_{T,T}$ is the C_T value for the target. In the presence of replicates, we use the averaged values of $C_{T,C}$ and $C_{T,T}$ in the above formula for $\hat{R}$. The standard curve method requires a third collection of reactions, in addition to the target and calibrator, referred to as the standard data; the standard data is dilution data. The technique is to fit a simple linear regression of the C_T values, corresponding to different dilutions of the standard, to the log of the dilution constants. Now, using the estimated regression line together with the $C_{T,T}$ and $C_{T,C}$ values one can determine the log of the dilution constants corresponding to $C_{T,T}$ and $C_{T,C}$. The ratio of these dilution constants are then used to obtain an estimate of R. In the presence of replicates, we repeat the estimation of the log dilution constants for each replicate and then average these for the target and the calibrator separately. Now, the ratio of these averages is used as an estimate of R. The ABI User's Manual (Livak, 2001) provides a complete description of both of these methods.

We generate data from three different models. We use the notation $X \sim Bern(p)$ to mean that $P(X = 2) = 1 - P(X = 1) = p$. The first model we study is an example of the random effect model proposed in the paper; specifically we use a beta distribution to describe the random effect.

Model 1. *(Random effects). For $I = C, T$, let $F_{k,j,I} = N_{k,j,I}$, where $N_{k,j,I}$ has offspring distribution $Bern(p_{k,I})$, with $p_{k,I} \sim^{iid} Beta(90, 10)$.*

We also wish to address the robustness of our procedure to certain model assumptions. First we consider the impact of random environments (Smith and Wilkinson, 1969). In the current paper, we assume that the splitting probability remains constant across cycles, for a given replicate. However, it is frequently argued in the PCR literature that the splitting probability varies across cycles as a function of the remaining product. To address this issue, we study the following random environment model.

Model 2. *(Random environments). For $I = C, T$. $F_{k,j,I} = N_{k,j,I}$, where $N_{k,j,I}$ has offspring distribution $Bern(p_{k,j,I})$, with $p_{k,j,I} \sim^{iid} Beta(90, 10)$.*

As discussed in Section 2, the fluorescence constant c varies within and between reactions. It is difficult to quantify the magnitude of this variability. To identify how this variability can affect our results, we consider the following model involving random fluorescence coefficients.

Model 3. *(Random fluorescence coefficient). For* $I = C, T$*.* $F_{k,j,I} = c_{k,j,I} N_{k,j,I}$*, where* $N_{k,j,I}$ *has offspring distribution* $Bern(p_{k,I})$*, with* $p_{k,I} \sim^{iid} Beta(90, 10)$*. And* $c_{k,j,I} \sim gamma(1, 10^{-3})$*, i.e.* $Ec_{k,j,I} = 1$ *and* $var(c_{k,j,I}) = 10^{-3}$*.*

In all three models, $N_{k,0,T} \sim Poiss(10^3)$ and $N_{k,0,C} \sim Poiss(10^2)$; hence the true value for relative quantitation is 10.

We compare three inferential methods: the branching process method, which is developed in this paper (see results in Table 2.1), the comparative C_T method (Table 2.2) and the standard curve method (Table 2.3). All of the results are based on $n = 20$ cycles and $r(n) = 20$ replicates. For the branching process estimator, generations 15 to 20 are used. For calculating C_T the threshold which marks the beginning of the exponential phase is set at $F^\star = 10^6$, i.e. $C_T \equiv \inf \left\{ j : F_j > F^\star \right\}$.

The standard curve method additionally requires the use of standards. In each simulation three replicates of a five fold dilution series were used to form the standard curve. More specifically, the initial number for the dilution series is a Poisson random variable with means $80, 400, 2000, 10,000$ and $50,000$.

Since the asymptotic behavior of the estimators is unknown for the C_T method and the standard curve method we use the bootstrap method (by resampling reactions) to construct confidence intervals for the quantitation parameters. All confidence intervals presented are 95% confidence intervals; all bootstrap confidence intervals are based on 2000 bootstrap samples. The bootstrap sample size was taken to be $r(n)$.

The comparative C_T method does not perform well under any of the three models. In contrast, the branching process estimator performs well under all three models. For the branching process estimator the increased variability present in Model 2 and Model 3 is reflected in increased variance of the point estimate and increased length of the confidence intervals.

Finally, we comment on the coverage of the confidence intervals for the branching process method (see Table 2.1). The confidence intervals based on the t distribution have the best coverage (closest to the nominal 95%), whereas the bootstrap confidence intervals have the worst coverage. One possible cause for the suboptimal bootstrap coverage is the random effect. Marginally the data is i.i.d. but conditionally (because of the random effect) it is not. It is possible that a weighted resampling scheme would improve the bootstrap coverage. We are currently investigating this issue.

2.5 Analysis of Experimental Data

The PCR data analyzed in this section were collected from a ABI Prism 7700 Sequence Detection System. We collected data on four replicates from an eight point dilution series. Let LH_1 denote the first term in the dilution series and LH_8 denote the eighth term. The master mix for LH_i is obtained from the master mix of LH_{i-1} using a dilution factor of 2.9505. Thus, if $m_{a,LH_1} = m_a$, then $m_{a,LH_i} = \frac{m_a}{2.9505^{i-1}}$ $i = 2, ..., 8$. Additionally, we collected 12 more replicates for LH_1 and LH_2 yielding a total of sixteen replicates from each these two groups. In this data analysis we proceed as if LH_1 is the target group and LH_2 is the calibrator group. Thus, the desired answer for relative quantitation is 2.9505.

Table 2.1: Branching process estimator. Point mean and Point var give the mean and variance of the point estimates, over the 5000 simulations. For the confidence intervals, Cov gives the simulated coverage and mean gives the mean length of the confidence interval over the 5000 simulations. B is for the bootstrap confidence interval; G is for the confidence interval based on asymptotic normality; t is for the confidence interval based on asymptotic normality, using the t distribution.

	Model 1	Model 2	Model 3
Point mean	10.0014	10.0310	10.0831
Point var	0.0581	0.3759	1.1992
B Cov	0.9292	0.9278	0.9286
B Mean	0.9098	2.2771	4.1704
G Cov	0.9374	0.9374	0.9416
G Mean	0.9331	2.3357	4.2485
t Cov	0.9528	0.9536	0.9580
t Mean	0.9964	2.4943	4.5369

Table 2.2: Comparative C_T estimator. Point mean and Point var give the mean and variance of the point estimates, over the 5000 simulations. B Cov gives the simulated coverage of the bootstrap confidence interval, and B mean gives the mean length of the confidence interval over the 5000 simulations.

	Model 1	Model 2	Model 3
Point mean	12.1523	12.0897	12.1486
Point var	0.8215	0.1488	0.8298
B Cov	0.3664	0	0.3782
B Mean	3.4412	1.4760	3.4678

Table 2.3: Standard curve estimator. Point mean and Point var give the mean and variance of the point estimates, over the 5000 simulations. B Cov gives the simulated coverage of the bootstrap confidence interval, and B mean gives the mean length of the confidence interval over the 5000 simulations.

	Model 1	Model 2	Model 3
Point mean	9.9212	9.9776	9.8900
Point var	0.6624	0.1428	0.6591
B Cov	0.8504	0.8472	0.8406
B Mean	2.4744	1.1253	2.4842

Traditionally biological labs use the so-called standard curve method for quantitation. To compare our methods to the standard curve method, we use two eight point dilution series to form the standard curve; this yields 14 replicates of LH_1 and LH_2 to compute the estimator of R. The data for the LH_1 and LH_2 reactions are displayed in Figure 2.2. In our analysis we excluded data from one of the reactions for LH_2, since it did not reach the appropriate C_T level. Similarly, we excluded data from an LH_1 replicate since its C_T value was much larger than those of other replicates.

2.5.1 Branching Process Method

Our methodology requires identification of the exponential phase. The strategy is to choose those cycles which yield a fluorescence of at least $F^\star$ and per-cycle amplification of at least m_c. The following algorithm identifies the cycles of data belonging to the exponential phase. In our analyses we chose $F^\star = 0.2$ and $m_c = 1.5$. The choice of $F^\star$ was suggested by the manufacturers of the equipment.

2.5.2 Results

Replicate Data. In this section we provide the conclusions of our analyses using the replicate data. The point estimate of R was 2.8221 and the 95% confidence interval using the asymptotic Gaussian limit was $(1.8624, 3.7817)$. The confidence interval based on the asymptotic t distribution was determined to be $(1.7719, 3.8722)$. Even though this analysis suffices, we developed bootstrap intervals so as to compare them with the other methods. The 95% bootstrap confidence interval using 2000 resampling of the reactions was determined to be $(1.6870, 3.6013)$.

Dilution Data. The dilution data yields $\tilde{R}_n = 2.4185$ and the 95% confidence interval based on the asymptotic Gaussian distribution to be $(1.6153, 3.2216)$ while that based on the t distribution to be $(1.5450, 3.2919)$. The 95% bootstrap confidence interval based on 2000 resampling of reactions was determined to be $(0.9829, 3.3616)$.

Other Methods. As mentioned previously, two of the methods used in biological labs are the so-called C_T method and the standard curve method. Using the C_T method $\tilde{R}_n = 3.3108$. The 95% bootstrap confidence interval using 2000 resampling of the reactions was determined to be $(2.7935, 3.7477)$. The corresponding values for the standard curve method were determined to be 3.5558 and $(2.8646, 4.1355)$.

From the data analysis, it is clear that the proposed branching process method using replicate data yields point estimates with smaller bias than the other methods. Furthermore, in contrast to the traditional methods, the confidence intervals based on our branching process method are supported by asymptotic theory.

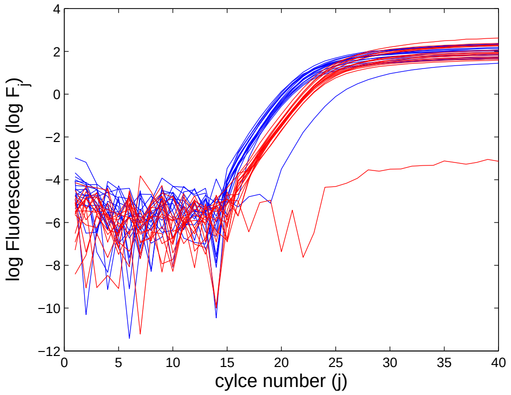


Figure 2.2: Plot of cycle number versus log fluorescence (j vs. $\log F_j$) for all 16 replicates of LH_1 (in blue) and all 16 replicates of LH_2 (in red).

2.6 Discussion and Concluding Remarks

In this paper we suggested a new design namely, replicated PCR experiments for quantitation. We developed branching process models with random effects to account for various sources of variability present in the PCR. Next, we developed a novel generalized method of moments approach for inference concerning the quantitation parameters and established consistency and asymptotic normality of these estimators. In our simulations we evaluated the behavior of our methodology under scenarios that are considerably different from the assumed model and illustrated the robust behavior of our proposed methodology. Data analysis reveals these aspects discovered in our theory and simulations.

Even though we do not advocate the use of dilution data, the methods of the paper show how to use such information if one has access only to dilution data. Frequently, this is encountered in various biological labs and the methods of this paper show one can get some quantitative information concerning the parameters of interest.

The advantage of the methods proposed in this paper compared to other existing methods like comparative C_T and the standard curve method is that our estimates have smaller bias and the confidence intervals supported by asymptotic theory. Additionally, analysis of calibrated experimental data using our method yields point estimates with smaller relative error (4.5%) compared to traditional methods (15%).

2.7 Proofs

In this section we present the proofs of our main theorems. Without loss of generality we will assume that $c = 1$ since otherwise all our estimates hold with a factor of c. Under this simplification, $Y_{k,n}$ represents total number of molecules in the reaction during the exponential phase, namely $Y_{k,n} = \sum_{j=\tau}^{n} N_{k,j}$. In the following C (or C_ϵ) denotes a generic constant that could change between successive lines and between successive inequalities.

Proof of Proposition 1. Conditioned on the random effect p_k, $N_{k,n}$ is an Galton-Watson process with finite conditional and marginal second moments. The proof then follows using the Toeplitz lemma and Theorem 8.1 of Harris (2002).

The proof of Theorem 1 involves several steps and hence we proceed by proving several lemmas. Our first lemma is concerned with the behavior of the harmonic moment of $N_{k,n}$.

Lemma 1. *Let $r \geq 1$. Under the assumptions of our model,*

$$E(\frac{1}{N_{k,n}^r}) \leq (1 - \frac{1}{2}E(p_1))^n. \tag{2.33}$$

Proof: It is sufficient to consider the case $N_{k,0} = 1$ and $r = 1$, since $N_{k,n}^r \geq N_{k,n}$ for all $r \geq 1$ and $N_{k,n} = \sum_{l=1}^{N_{k,0}} N_{k,n,l} \geq N_{k,n,1}$, where $N_{k,n,j}$ is the number of DNA templates in the n^{th} cycle initiated by the j^{th} template in the 0^{th} cycle of the k^{th} reaction. Now,

$$\begin{aligned} E(\frac{1}{N_{k,n}}|N_{k,0} = 1) &= E(\frac{1}{N_{k,n-1}}E(\frac{1}{N_{k,n-1}^{-1}N_{k,n}}|N_{k,n-1})|N_{k,0} = 1) && (2.34) \\ &\leq E(\frac{1}{N_{k,n-1}}|N_{k,0} = 1)E(\frac{1}{N_{k,1}}|N_{k,0} = 1), && (2.35) \end{aligned}$$

where the last step follows using the inequality concerning the arithmetic mean and harmonic mean. Now iterating the above inequality, it follows that

$$E(\frac{1}{N_{k,n}}|N_{k,0} = 1) \leq (E(\frac{1}{N_{k,1}}|N_{k,0} = 1))^n. \tag{2.36}$$

Now, observe that

$$\begin{aligned} E(\frac{1}{N_{k,1}}|N_{k,0} = 1) &= E(E(\frac{1}{N_{k,1}}|N_{k,0} = 1, p_k)) && (2.37) \\ &= E(1 - p_k + \frac{1}{2}p_k)) = (1 - \frac{1}{2}E(p_1)) < 1, && (2.38) \end{aligned}$$

where the last inequality follows from $E(p_1) > 0$. This completes the proof of Lemma 1.

Our next lemma is concerned with the bound on the $E(\sqrt{Y_{k,n-1}}(\tilde{m}_{k,n} - m_k))^{2r}$.

Lemma 2. *Under the assumptions of our model, there exists a universal constant C such that*

$$E(\sqrt{Y_{k,n-1}}(\tilde{m}_{k,n} - m_k))^4 \leq Cn^4. \tag{2.39}$$

Proof: We note that

$$\begin{aligned}\sqrt{Y_{k,n-1}}(\tilde{m}_{k,n} - m_k) &= \sqrt{Y_{k,n-1}}(\frac{Y_{k,n} - F_{k,\tau}}{Y_{k,n-1}} - m_k) \qquad (2.40)\\ &= \sum_{j=\tau}^{n} \frac{N_{k,j+1} - m_k N_{k,j}}{\sqrt{N_k, j}} w_{k,n,j}, \qquad (2.41)\end{aligned}$$

where

$$w_{k,n,j}^2 = \frac{N_{k,j}}{Y_{k,n-1}}. \qquad (2.42)$$

Thus, setting $X_{k,j} = \frac{N_{k,j+1} - m_k N_{k,j}}{\sqrt{N_k, j}}$, we have that

$$\begin{aligned}(\sqrt{Y_{k,n-1}}(\tilde{m}_{k,n} - m_k))^4 &\leq n^4(\frac{1}{n-\tau}\sum_{j=\tau}^{n-1} |X_{k,j}|)^4 \qquad (2.43)\\ &\leq n^4\left(\frac{1}{n-\tau}\sum_{j=\tau}^{n-1} X_{k,j}^4\right), \qquad (2.44)\end{aligned}$$

where the last inequality follows from Jensen's inequality for convex functions. Now, conditioned on the random effect p_k and $N_{k,j-1}$, the numerator of $X_{k,j}$ is $Bin(N_{k,j-1}, p_k) - E_k(Bin(Z_{k,j-1}, p_k)|Z_{k,j-1})$. Now, using the formula for the fourth central moment of a binomial random variable, it follows that

$$E_k(X_{k,j}^4|N_{k,j-1})^4 = N_{k,j-1}^{-1} p_k q_k (3N_{j,k-1} p_k q_k - 6p_k q_k + 1), \qquad (2.45)$$

where $q_k = 1 - p_k$. Hence, since $(1-\epsilon) \leq p_k \leq 1$, it follows that

$$\begin{aligned}E_k(X_{k,j}^4|N_{k,j-1}, p_k) &\leq 3(p_k(1-p_k))^2 + 1 \qquad (2.46)\\ &\leq 3\epsilon^2 + 1. \qquad (2.47)\end{aligned}$$

Now taking expectation with respect to $N_{k,j-1}$ and with respect to the distribution of the random effect, it follows that

$$E(X_{k,j}^4) \leq 3\epsilon^2 + 1. \qquad (2.48)$$

Finally, taking expectation in (3.4), and using (2.48) it follows that

$$E(\sqrt{Y_{k,n-1}}(\tilde{m}_{k,n} - m_k))^4 \leq n^4(1 + 3\epsilon^2). \qquad (2.49)$$

This completes the proof of the lemma.

Our next lemma is concerned with the almost sure behavior of $\vee_{k=1}^{r(n)}(\frac{\tilde{m}_{k,n}-1}{m_{k-1}-1}-1)$.

Lemma 3. *Under the conditions of Theorem 1, it happens with probability one that*

$$\lim_{n\to\infty}\sqrt{r(n)D_1(n)}\max_{1\le k\le r(n)}|\frac{\tilde{m}_{k,n}-1}{m_k-1}-1|=0. \tag{2.50}$$

Proof: It is sufficient to show that for all $\eta > 0$

$$\sum_{n\ge1} r(n)max_{1\le k\le r(n)}P(|\frac{\tilde{m}_{k,n}-m_k}{p_k}|>\frac{\eta}{\sqrt{r(n)D_1(n)}})<\infty. \tag{2.51}$$

By Markov's inequality,

$$\begin{aligned} P(|\frac{\tilde{m}_{k,n}-m_k}{p_k}|>\frac{\eta}{\sqrt{r(n)D_1(n)}}) &\le (\frac{\sqrt{r(n)D_1(n)}}{\eta})^2E|\frac{\tilde{m}_{k,n}-m_k}{p_k}|^2 & (2.52)\\ &\le (\frac{\sqrt{r(n)D_1(n)}}{\eta(1-\epsilon)})^2E|\tilde{m}_{k,n}-m_k|^2 & (2.53)\\ &\le (\frac{\sqrt{r(n)D_1(n)}}{\eta(1-\epsilon)})^2d_n(1)d_n(2), & (2.54)\end{aligned}$$

where

$$d_n(1)=(E(|\sqrt{Y_{k,n-1}}(\tilde{m}_{k,n}-m_k))^4|)^{1/2},\ \text{ and }\ d_n(2)=E(\frac{1}{Y_{k,n-1}^2})^{1/2}, \tag{2.55}$$

and the last inequality follows by first multiplying and dividing by $\sqrt{Y_{k,n-1}}$ inside the expectation in (3.4) and then applying the Cauchy-Schwarz inequality. Now by Lemma 2, $d_n(1)\le Cn^2$, where C is a deterministic constant. By Lemma 1, it follows that $E(d_n(2))\le C\gamma^n$ where $0<\gamma<1$. Thus,

$$P(|\frac{\tilde{m}_{k,n}-m_k}{p_k}|>\frac{\eta}{\sqrt{r(n)}})\ \le\ C(\frac{\sqrt{r(n)D_1(n)}}{\eta(1-\epsilon)})^2n^2\gamma^n.$$

Thus, it follows from the regularity of the dilution constants and that $r(n)n^{-1}\to 0$

that

$$\begin{aligned}\sum_{n\geq 1} r(n) max_{1\leq k\leq r(n)} P(|\frac{\tilde{m}_{k,n}-m_k}{p_k}| > \frac{\eta}{\sqrt{r(n)}}) &\leq C\sum_{n\geq 1} r^2(n)D_1(n)n^2\gamma^n \\ &\leq C\sum_{n\geq 1} n^4\gamma^n < \infty,\end{aligned}$$

where the finiteness is established using the ratio test.

Lemma 4. *Under the conditions of Theorem 1, with probability one*

$$\lim_{n\to\infty} \sqrt{r(n)D_1(n)} \max_{1\leq k\leq r(n)} |(\frac{m_k^n}{\tilde{m}_k^n} - 1)| = 0. \tag{2.56}$$

Proof: It is sufficient, using Borel-Cantelli, to show that for any $\eta > 0$,

$$\sum_{n\geq 1} r(n) \max_{1\leq k\leq r(n)} P(|(\frac{\tilde{m}_{k,n}}{m_k})^n - 1)| > \frac{\eta}{\sqrt{r(n)D_1(n)}}) < \infty. \tag{2.57}$$

We will now obtain estimates on $P(|(\frac{\tilde{m}_{k,n}}{m_k})^n - 1)| > \frac{\eta}{\sqrt{r(n)D_1(n)}})$. To this end, it is easy to see that

$$P(|(\frac{\tilde{m}_{k,n}}{m_k})^n - 1)| > \frac{\eta}{\sqrt{r(n)D_1(n)}}) \quad = \quad J_n(1) + J_n(2), \tag{2.58}$$

where

$$J_n(1) = P(\tilde{m}_{k,n} - m_k > m_k a_1(n)) \tag{2.59}$$

$$J_n(2) = P(\tilde{m}_{k,n} - m_k < m_k a_2(n)), \tag{2.60}$$

$a_1(n) = (1 + \frac{\eta}{\sqrt{r(n)D_1(n)}})^{\frac{1}{n}} - 1$ and $a_2(n) = (1 - \frac{\eta}{\sqrt{r(n)D_1(n)}})^{\frac{1}{n}} - 1$. We will deal with $J_n(1)$ as the proof of the other term is similar. By Markov's inequality,

$$\begin{aligned}J_n(1) &\leq E(\frac{E_k(|\tilde{m}_{k,n} - m_k|)}{m_k a_1(n)}) \qquad (2.61)\\ &\leq \frac{1}{(2-\epsilon)a_1(n)} E|m_{k,n} - m_k| \qquad (2.62)\\ &\leq \frac{(E(|\sqrt{Y_{k,n-1}}|\tilde{m}_{k,n} - m_k|)^2)^{\frac{1}{2}}}{(2-\epsilon)a_1(n)}(E(Y_{k,n-1}^{-1}))^{\frac{1}{2}} \qquad (2.63)\\ &\leq \frac{C}{(2-\epsilon)a_1(n)} n^2\gamma^n \qquad (2.64)\end{aligned}$$

Using the mean value theorem and $r(n) \leq n$, one can show that $a_1^{-1}(n) \leq Cn^2$. Using this estimate and the ratio test it follows that $\sum_{n\geq 1} J_n(1) < \infty$. A similar calculation for $J_n(2)$ then yields the lemma.

Lemma 5. *Under the conditions of Theorem 1, for l=1, 2, with probability one,*

$$\lim_{n\to\infty} \frac{1}{r(n)D_1(n)} \sum_{k=1}^{r(n)} |\frac{Y_{k,n}}{(1+p_k)^n} - V_k)(\frac{p_k}{1+p_k})|^l = 0, \tag{2.65}$$

where $V_k = V_k^\star(\frac{m_k}{m_k-1})$.

Proof: Let $\theta_k = \frac{p_k}{1+p_k}$. We begin by developing an estimate of $Var[(\frac{Y_{k,n}}{(1+p_k)^n} - V_k)\theta_k]$. Using $V_k = V_k^\star \sum_{j\geq 0} m_k^{-j}$ and a change of variables, it follows that

$$\begin{aligned} \frac{Y_{k,n}}{(1+p_k)^n} - V_k &= \sum_{j=0}^{n-\tau}(V_{k,n-j} - V_k^\star)m_k^{-j} - V_k^\star \sum_{j\geq n+1-\tau} m_k^{-j} && (2.66) \\ &= J_n(1,k) - J_n(2,k) && (2.67) \end{aligned}$$

Thus,

$$Var[(\frac{Y_{k,n}}{m^n} - V_k)\theta_k] = Var(J_n(1,k)\theta_k) + Var(J_n(2,k)\theta_k) - 2Cov(J_n(1,k)\theta_k, J_n(2,k)\theta_k). \tag{2.68}$$

Now, setting $S(k,n,j) = \theta_k \sum_{j\geq n+1-\tau} m_k^{-j}$

$$\begin{aligned} Var(J_n(2,k)\theta_k) &= Var(E(V_k^\star S(k,n,j)|p_k)) + E(Var(V_k^\star S(k,n,j)|p_k)) \\ &\leq E(S^2(k,n,j)(m_a^2 d_k^2 + Var_k(V_k^\star))). && (2.69) \end{aligned}$$

Now, using $m_k \geq (2-\epsilon)$ and $\theta_k \leq 1$, it follows that $S^2(n,k,j) \leq ((1-\epsilon)(2-\epsilon)^n)^{-1}$. Using this estimate in (2.69) it follows that

$$Var(J_n(2,k)\theta_k) \leq ((1-\epsilon)(2-\epsilon)^n)^{-1}(m_a^2 d_k^2 + \omega_k^2). \tag{2.70}$$

We next study the behavior of $Var(J_n(1,k)\theta_k)$. Now, using conditioning it follows that

$$Var(J_n(1,k)\theta_k) = E(Var(\sum_{j=0}^{n-\tau}(V_{k,n-j} - V_k^\star)m_k^{-j}\theta_k)|p_k)). \tag{2.71}$$

Now,

$$Var(\sum_{j=0}^{n-\tau}(V_{k,n-j} - V_k^\star)m_k^{-j}\theta_k|p_k) = J_n(1,1,k) + J_n(1,2,k), \tag{2.72}$$

where

$$J_n(1,1,k) = \sum_{j=0}^{n-\tau} Var(V_{k,n-j} - V_k^\star|p_k)m_k^{-2j}\theta_k^2, \tag{2.73}$$

and

$$J_n(1,2,k) = \sum_{j=0}^{n-\tau}\sum_{j\neq l=0}^{n-\tau} \frac{\theta_k^2}{m_k^{j+l}} Cov(V_{k,n-j} - V_k^\star, V_{k,n-l} - V_k^\star|p_k). \tag{2.74}$$

Using the branching property it follows that,

$$Var(V_{k,n-j} - V_k^\star|p_k) \leq C_\epsilon m_a d_k(2-\epsilon)^{n-j}, \tag{2.75}$$

where C_ϵ is a finite positive constant independent of k. Now, using this estimate and that $\theta_k \leq 1$ it follows that

$$J_n(1,1,k) \leq C_\epsilon m_a d_k(2-\epsilon)^{-n}. \tag{2.76}$$

Now, we deal with $J_n(1,2,k)$. Using the Cauchy-Schwarz inequality and (2.75) it follows that

$$|Cov(V_{k,n-j} - V_k^\star, V_{k,n-l} - V_k^\star|p_k)| \leq C_\epsilon m_a d_k(2-\epsilon)^{n-(j+l)/2}. \tag{2.77}$$

Using this estimate and $\theta_k \leq 1$ it follows that

$$J_n(1,2,k) \leq C_\epsilon m_a d_k(2-\epsilon)^{-n}. \tag{2.78}$$

Now, combining the estimates for $J_n(1,1,k)$ and $J_n(1,2,k)$ we get

$$Var(J_n(1,k)\theta_k) \leq C_\epsilon d_k(2-\epsilon)^{-n}. \tag{2.79}$$

Again using the Cauchy-Schwarz inequality and $\theta_k \leq 1$, it follows that

$$Cov(J_n(1,k)\theta_k, J_n(2,k)\theta_k|p_k) \leq C_\epsilon(2-\epsilon)^{-n}(C_{1,\epsilon}d_k^2 + C_{2,\epsilon}d_k^3)^{1/2}. \tag{2.80}$$

Thus combining all the estimates, taking expectation with respect to the distribution of p_k, summing over k and using the Cauchy-Schwarz inequality, one can show, using the regularity of the dilution constants, that

$$\sum_{k=1}^{r(n)} Var[(\frac{Y_{k,n}}{m_k^n} - V_k)\theta_k] \leq C_{3,\epsilon} r(n)(2-\epsilon)^{-n}. \qquad (2.81)$$

Next, we obtain an estimate of $|E[(\frac{Y_{k,n}}{m_k^n} - V_k)\theta_k]|$. Again, using the decomposition (2.67) and $E(J_n(1,k)\theta_k) = 0$, it follows that

$$|E[(\frac{Y_{k,n}}{m_k^n} - V_k)\theta_k]| \leq |E(\theta_k m_a d_k \sum_{j \geq n+1} m_k^{-j})| \leq C_{4,\epsilon}(2-\epsilon)^{-n} d_k. \qquad (2.82)$$

Now, using (2.81), (2.82), and the regularity of the dilution constants it follows that

$$\begin{aligned} E(\sum_{k=1}^{r(n)} \theta_k(\frac{Y_{k,n}}{m_k^n} - V_k))^2 &= \sum_{k=1}^{r(n)} Var[(\frac{Y_{k,n}}{m_k^n} - V_k)\theta_k] + (\sum_{k=1}^{r(n)} E[(\frac{Y_{k,n}}{m_k^n} - V_k)\theta_k])^2 \\ &\leq C_{5,\epsilon} r(n)(2-\epsilon)^{-n}, \qquad (2.83) \end{aligned}$$

where $0 < C_{5,\epsilon} < \infty$ is some constant depending on ϵ. Finally, using Markov's inequality and (2.83) it follows that for $l = 1, 2$,

$$\begin{aligned} P(\frac{1}{r(n)}|\sum_{k=1}^{r(n)} \theta_k(\frac{Y_{k,n}}{(1+p_k)^n} - V_k)|^l > \eta) &\leq \frac{1}{\eta^2 r(n)} E|\sum_{k=1}^{r(n)} \theta_k(\frac{Y_{k,n}}{m_k^n} - V_k)|^2 \\ &\leq C_{5,\epsilon}(2-\epsilon)^{-n}. \qquad (2.84) \end{aligned}$$

Since the RHS of (2.84) is summable, (2.65) follows using the Borel-Cantelli lemma.

Our next lemma is concerned with the moment behavior of the limit random variable $V_k^\star$ when the process is initiated by a single ancestor.

Lemma 6. *Let $N_{k,0} = 1$ for all $k \geq 1$. Then there exists a finite positive constant C such that $E(V_1^{\star 4}) \leq C$.*

Proof: First note that for all k and j $E(V_{k,j}) = 1$. Also using the representation $N_{k,j+1} = N_{k,j} + Bin(N_{k,j}, p_k)$, where $Bin(N_{k,j}, p_k)$ is a binomial random variable (given $N_{k,j}$ and p), one can show that

$$E(V_{k,j}^2) \leq E(V_{k,j-1}^2) + E(\frac{1}{m_k^j}). \tag{2.85}$$

Now, iterating the above and using Tonelli's theorem, it follows that

$$E(V_{k,j}^2) \leq \sum_{l \geq 0} E(\frac{1}{m_k^l}) = E(\frac{1 + p_k}{p_k}) \equiv C < \infty. \tag{2.86}$$

We next show that $E(V_{k,j}^3)$ is uniformly bounded. Using the representation of $V_{k,j}$ alluded to above and the uniform boundedness of the first and second moments it follows that

$$E(V_{k,j}^3) \leq E(V_{k,j-1}^3) + E(\frac{C}{m_k^j}). \tag{2.87}$$

The uniform boundedness follows by iteration and summing as before. Now, using the uniform boundedness of $V_{k,j}^3$ and using the fourth moment of a binomial random variable one can show that

$$E(V_{k,j}^4) \leq E(V_{k,j-1}^4) + E(\frac{C}{m_k^j}). \tag{2.88}$$

Iterating and summing, it follows that $E(V_{k,j}^4)$ is uniformly bounded. Now it follows using Jensen's inequality, uniform boundedness of the fourth moment of $V_{k,n}$, and that $V_k^\star - V_{k,n}$ are identically distributed in k that

$$E(V_k^{\star 4}) \leq 4(\sup_{n \geq 1} E(V_{k,n}^4) + E|V_1^\star - V_{1,n}|^4) \leq C + E|V_1^\star - V_{1,n}|^4, \tag{2.89}$$

where C is some finite positive constant. Thus, to complete the proof of the lemma, it is sufficient to show that the second term of the RHS of (2.89) is bounded in n. We will actually show that $E|V_1^\star - V_{1,n}|^4 \to 0$ as $n \to \infty$. To this end, it is sufficient to show that $\{V_{1,n} : n \geq 1\}$ is a Cauchy sequence in L_4 space. Now, using conditioning, the Marcinkiewicz-Zygmund inequality for independent random variables (Chow and Teicher, 1997) and the branching property, it can be seen that

$$E(|V_{1,k+n} - V_{1,n}|^4|p_1) \leq (2\sqrt{2})^4 E(N_{1,n}^{1/2}) m_1^{-4n} E|V_{1,k} - 1|^4 \tag{2.90}$$

$$\leq (2\sqrt{2})^4 E|V_{1,k} - 1|^4 (2-\epsilon)^{-7n/2}. \tag{2.91}$$

Now, using the uniform boundedness of the fourth moments of $V_{1,k}$ and that $0 < \epsilon < 1$, it follows first by taking expectations with respect to the distribution of p_1 and then taking the supremum over k that

$$\sup_{k\geq 1} E|V_{1,k+n} - V_{1,n}|^4 \leq C(2-\epsilon)^{-7n/2}, \tag{2.92}$$

establishing the L_4 convergence of $V_{k,n}$ to $V_k^\star$.

Lemma 7. *Under the conditions of Theorem 1, with probability one,*

$$\lim_{n\to\infty} \frac{1}{r(n)D_1(n)} \sum_{k=1}^{r(n)} V_k^\star = m_a, \tag{2.93}$$

and

$$\frac{1}{\sqrt{r(n)D_1(n)}} \sum_{k=1}^{r(n)} (V_k^\star - m_a d_k) \xrightarrow{d} G_1, \tag{2.94}$$

where $G_1 \sim N(0, \sigma_L^2)$ and σ_L^2 is defined in Theorem 1.

Proof. Note that the random variables $V_k^\star$ are independent with mean $m_a d_k$ and variance ω_k^2. Thus, by regularity of the dilution constants, it follows that

$$\sum_{k\geq 1} \frac{E(V_k - m_a d_k)^2}{k^2} = \sum_{k\geq 1} \frac{\omega_k^2}{k^2} < \infty. \tag{2.95}$$

Hence, by Loeve's generalization of Kolmogorov's laws of large numbers (Chow and Teicher, 1997), it follows that $\frac{1}{r(n)} \sum_{k=1}^{r(n)} V_k^\star$ converges almost surely to m_a. To establish the asymptotic normality, we will verify the Liapounov condition for independent random variables. To this end, we consider $E|V_k^\star - m_a d_k|^3$. By the branching property and using $E(N_{k,0}) = m_a d_k$, it follows that

$$\begin{aligned} E|V_k^\star - m_a d_k|^3 &= E|(\sum_{j=1}^{N_{k,0}} (V_{k,j}^\star - 1) + (N_{k,0} - E(N_{k,0})|^3 & (2.96) \\ &\leq 4(E|\sum_{j=1}^{N_{k,0}} (V_{k,j}^\star - 1)|^3 + E|(N_{k,0} - E(N_{k,0})|^3), & (2.97) \end{aligned}$$

where $V_{k,j}^\star$ are independent random variables (and independent of $N_{k,0}$) with $E(V_{k,j}^\star) = 1$. Now, by first conditioning on $N_{k,0}$ and then using conditional Jensen's inequality it follows that

$$E|\sum_{j=1}^{N_{k,0}} (V_{k,j}^\star - 1)|^3 \leq E(N_{k,0}^3 E(|\frac{1}{N_{k,0}}|\sum_{j=1}^{N_{k,0}} (V_{k,j}^\star - 1)|^3 | N_{k,0})). \qquad (2.98)$$

Now, using the independence of $V_{k,j}^\star$ and $N_{k,0}$ and that for each fixed k, $EV_{k,j}^{\star 3} = EV_{k,1}^{\star 3}$, it follows that

$$\begin{aligned} E|\sum_{j=1}^{N_{k,0}} (V_{k,j}^\star - 1)|^3 &\leq E(N_{k,0}^3) E(V_{k,1}^{\star 3}) & (2.99) \\ &\leq CE(N_{k,0}^3) = C m_{3,0} d_k^3, & (2.100) \end{aligned}$$

where the last inequality follows from Lemma 6 and the parametrization for the third moment. Hence,

$$(\frac{1}{r(n)})^{\frac{3}{2}} \sum_{k=1}^{r(n)} E|\sum_{j=1}^{N_{k,0}} (V_{k,j}^\star - 1)|^3 \leq C(\frac{1}{r(n)})^{\frac{1}{2}} D_3(n). \qquad (2.101)$$

Now, by the regularity of the dilution constants, $\{D_3(n) : n \geq 1\}$ is a bounded sequence. This implies that

$$\lim_{n\to\infty} (\frac{1}{r(n)})^{\frac{3}{2}} \sum_{k=1}^{r(n)} E|\sum_{j=1}^{N_{k,0}} (V_{k,j}^\star - 1)|^3 = 0. \qquad (2.102)$$

Now, using the fact that

$$\lim_{n\to\infty} \left(\frac{1}{r(n) D_1(n)} \sum_{k=1}^{r(n)} \omega_k^2 \right) = \sigma_L^2, \qquad (2.103)$$

the lemma follows.

Proof of Theorem 1. First we express $\tilde{m}_{a,n}$ as

$$\tilde{m}_{a,n} - m_a = T_n(1) + (T_n(2) - m_a), \tag{2.104}$$

where

$$T_n(1) = \frac{1}{r(n)D_1(n)} \sum_{k=1}^{r(n)} \frac{Y_{k,n}}{(1+p_k)^n}(\frac{p_k}{1+p_k})((\frac{\tilde{p}_{k,n}}{p_k})(\frac{(1+p_k)^{n+1}}{(1+\tilde{p}_{k,n})^{n+1}}) - 1), \quad \text{and}$$

$$T_n(2) = \frac{1}{r(n)D_1(n)} \sum_{k=1}^{r(n)} \frac{Y_{k,n}}{(1+p_k)^n}(\frac{p_k}{1+p_k}).$$

We begin with a decomposition for $T_n(2)$ to obtain an expression for $T_n(2) - m_a$.

$$T_n(2) - m_a = T_n(3) + T_n(4), \tag{2.105}$$

where

$$T_n(3) = \frac{1}{r(n)D_1(n)} \sum_{k=1}^{r(n)} (\frac{Y_{k,n}}{m_k^n} - V_k)(\frac{p_k}{m_k}), \text{and} \tag{2.106}$$

$$T_n(4) = \frac{1}{r(n)D_1(n)} \sum_{k=1}^{r(n)} (V_k^\star - m_a d_k). \tag{2.107}$$

Returning to $T_n(1)$ we have

$$|T_n(1)| \leq \max_{1\leq k\leq r(n)} |(\frac{\tilde{p}_{k,n}}{p_k})(\frac{(1+p_k)^{n+1}}{(1+\tilde{p}_{k,n})^{n+1}} - 1)|T_n(2). \tag{2.108}$$

Now by Lemma 5, $T_n(3)$ converges to zero with probability one and by Lemma 7, $T_n(4)$ converges to 0 with probability one. Combining the results we get that $|T_n(2) - m_a|$ converges to zero with probability one. Also, we obtain the convergence to zero of $|T_n(1)|$ using Lemma 3 and Lemma 4. This yields the strong consistency of $\tilde{m}_{a,n}$. To establish the asymptotic normality, first note that by Lemma 7,

$$(r(n)D_1(n))^{1/2} T_n(4) \xrightarrow{d} N(0, \sigma_L^2). \tag{2.109}$$

Define $\theta_k \equiv \frac{p_k}{1+p_k}$. For any $\eta > 0$, using Chebychev's inequality

$$\begin{aligned} P(|(r(n)D_1(n))^{1/2}T_n(3)| > \eta) &\leq \frac{1}{\eta^2 r(n)D_1(n)}(E\sum_{k=1}^{r(n)}\theta_k(\frac{Y_{k,n}}{m_k^n} - V_k))^2 \\ &\to 0, \end{aligned}$$

where the last convergence follows form (2.83). Finally, using Lemma 3 and Lemma 4, it follows that $(r(n)D_1(n))^{1/2}T_n(1)$ converges to zero in probability. Combining the convergences, asymptotic normality follows using Slutsky's lemma.

Proof of Theorem 2. Theorem 2 follows by an application of delta method to the function $f(x,y) = \frac{x}{y}$.

Proof of Proposition 2. Conditioned on the random effect, the process $\{N_{k,n} : n \geq 1\}$ is a branching process with offspring distribution $1 + X$, where $X \sim Ber(p_k)$ denotes a Bernoulli random variable with $P(X = 1|p_k) = p_k$. Hence, it follows that

$$\lim_{n\to\infty} P(\sqrt{Y_{k,n-1}}(\tilde{m}_{k,n} - m_k) \leq x|p_k) = P(N(0, p_k(1-p_k)) \leq x|p_k). \qquad (2.110)$$

Thus by the bounded convergence theorem, it follows that

$$\lim_{n\to\infty} E(P(\sqrt{Y_{k,n-1}}(\tilde{m}_{k,n} - m_k) \leq x|p_k)) = \int_{1-\epsilon}^{1} \Phi(\frac{x}{t(1-t)})dG(t). \qquad (2.111)$$

The proposition follows since the random variables p_k are identically distributed.

Proof of Theorem 3. First we rewrite

$$\begin{aligned} \frac{1}{r(n)}\sum_{k=1}^{r(n)}(\tilde{p}_{k,n} - E(p_1)) &= \frac{1}{r(n)}\sum_{k=1}^{r(n)}(\tilde{p}_{k,n} - p_k) + \frac{1}{r(n)}\sum_{k=1}^{r(n)}(p_k - E(p_1)) \\ &= T_n(1) + T_n(2) \end{aligned}$$

and verify that $T_n(1) \to 0$ with probability 1. Now by Chebychev's inequality and the independence of $(\tilde{p}_{k,n} - p_k)$ in k,

$$\begin{aligned} P(|T_n(1)| > \eta) &\leq \eta^{-2} E(T_n^2(1)) = \eta^{-2}(Var(T_n(1)) + (E(T_n(1)))^2) & (2.112) \\ &\leq \frac{C}{r^2(n)} (\sum_{k=1}^{r(n)} (E(\tilde{p}_{k,n} - p_k)^2 + (E(\tilde{p}_{k,n} - p_k)^2)^{1/2}), & (2.113) \end{aligned}$$

where the last inequality follows by bounding the variance term by the second moment and using the Cauchy-Schwarz inequality on the expectation term. Now,

$$\begin{aligned} E(\tilde{p}_{k,n} - p_k)^2 &= E(\tilde{m}_{k,n} - m_k)^2 & (2.114) \\ &= E((\tilde{m}_{k,n} - m_k)^2 Y_{k,n} Y_{k,n}^{-1}) & (2.115) \\ &\leq (E(\tilde{m}_{k,n} - m_k)^4 Y_{k,n}^2))^{1/2} (E(Y_{k,n}^{-2}))^{1/2}, & (2.116) \end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality. Now applying Lemma 1 and Lemma 2 it follows that for some $0 < C < \infty$ and $0 < \gamma < 1$

$$E(\tilde{p}_{k,n} - p_k)^2 \leq C n^2 \gamma^{n/2}. \tag{2.117}$$

Now using this estimate in (2.113) it follows that $P(|T_n(1)| > \eta)$ is bounded above by $Cn^2\gamma^{n/4}$. By ratio test, the above probability sums there by yielding the almost sure convergence to 0 of $T_n(1)$. Since $T_n(2)$ is a sum of i.i.d. random variables with finite second moments, the theorem follows via the law of large numbers and central limit theorem for i.i.d. random variables.

Proof of Theorem 4. The estimator of variance can be expressed as

$$\tilde{\sigma}_{L,n}^2 = \frac{1}{r(n) D_1(n)} \sum_{k=1}^{r(n)} (T_n(1,k) + T_n(2,k) + T_n(3,k))^2, \tag{2.118}$$

where

$$T_n(1,k) = (\frac{Y_{k,n} \tilde{p}_{k,n}}{\tilde{m}_{k,n}^{n+1}} - V_k^\star), \tag{2.119}$$

$$T_n(2,k) = V_k^\star - m_a d_k \text{ and } T_n(3,k) = (m_a - \tilde{m}_{a,n} d_k). \tag{2.120}$$

One can show using the Cauchy-Schwarz inequality and Lemma 5 that the cross-product terms in the expansion of (3.21) converge to zero with probability one. Furthermore, normalized sums of squares of $T_n(3,k)$ converges to zero with probability one by regularity of the dilution constants and strong consistency of $\tilde{m}_{a,n}$. Also the normalized sums of squares of $T_n(1,k)$ converges to zero by Lemma 5. Finally, by using the arguments in Lemma 7 and the regularity of the dilution constants it follows that normalized sums of squares $T_n(2,k)$ converges to σ_L^2. This yields the strong consistency of $\tilde{\sigma}_{L,n}^2$. Strong consistency of $\tilde{\theta}_{1,n}$ and $\tilde{\theta}_{2,n}$ follow from Lemma 5 and the strong law of large numbers for i.i.d. random variables $(1-p_k)^{-1}(1+p_k)$.

Proof of Corollary 1. The proof follows from the strong consistency of $\sigma_{L,n}^2$ $\tilde{m}_{a,n}$, $\tilde{\theta}_{1,n}$, the regularity of the dilution constants , and the definition of σ_L^2.

CHAPTER 3

ANALYSIS OF VARIANCE MODELS RELATED TO ANCESTRAL INFERENCE FOR SUPERCRITICAL BRANCHING PROCESSES

3.1 Introduction

This paper is concerned with analysis of variance (ANOVA) models for comparing the means of the ancestor distributions in supercritical branching processes initiated by a random number of ancestors. We present applications of our methodology to analysis of data from quantitative polymerase chain reaction (qPCR) experiments. We begin with a brief description of the PCR experiment. PCR is a biochemical technique used to amplify the number of copies of a specific DNA fragment. This paper specifically investigates qPCR, where the scientific goal concerns the estimation of the initial number of molecules present in a genetic material. qPCR is an important tool for gene expression experiments (Ferré, 1998; Kubista et al., 2006; Nolan et al., 2006). A typical qPCR experiment is run for 40 cycles; theoretically, the number of molecules doubles in every cycle. In practice, only some fraction of the molecules actually replicate in a given cycle. Hence, a supercritical Galton-Watson branching process with a Bernoulli offspring distribution provides a natural model to describe the dynamics of PCR. Under a branching process model, the question of quantitation is tantamount to estimating the initial number of ancestors of the process. There are several papers which model PCR as a branching process and then estimate the initial number of ancestors (Nedelman et al., 1992; Jacob and Peccoud, 1998; Lalam, 2007; Lalam and Jacob, 2007). These works are based on observing a single realization of a branching process. In fact, a typical qPCR experiment produces data from 384 separate reactions. In the present paper, we work with replicated qPCR experiments, where replicates (or i.i.d. branching processes) are observed for several experimental groups.

Specifically, we address the a sample problem, where a denotes the number of groups. Multiple-sample problems are important from a scientific perspective, allowing scientists to address canonical questions, such as "is gene X expressed significantly more in males than females," or an extension of such questions to a finite number of experimental groups. Assuming a branching process model, PCR is governed by two distributions: (i) the ancestor distribution, which characterizes the initial number of particles in each process, and (ii) the offspring distribution, which characterizes the dynamics of the reaction. In this paper, we address testing for the equality of the offspring mean and the ancestor mean across experimental groups.

We begin with a brief review of one-way ANOVA problems for independent data. In a classical one-way ANOVA model with homoscedastic normal errors, the F statistic for testing the equality of means is defined as the ratio of mean regression sum of squares to mean error sum of squares. In this case, the statistic has an exact null distribution, namely the F distribution with degrees of freedom determined by the number of treatments and the sample size. The F test is robust to the normality assumption if the number of observations in each group is large. Specifically, with a fixed number of groups and each group size going to infinity, under the null hypothesis, the properly scaled F statistic has a limiting χ^2 distribution (Arnold, 1980; Ito, 1980). Under heteroscedasticity, this result breaks down. A consequence of heteroscedasticity when using the F statistic for testing the equality of means is the loss of power (Krutchkoff, 1988, 1989). Argaç (2004) describes several corrections for the classical F test under heteroscedasticity and performs extensive simulations to test their performance. The conclusion of the analysis is that there does not exist a 'best' method for addressing heteroscedasticity. Other works concerning heteroscedastic one-way models include Lee and Ahn (2003), Kulinskaya et al. (2003), and Krishnamoorthy et al. (2007). As discussed below, the issue of unequal variance is of primary importance in the present paper; the ANOVA problem we consider inherently involves heteroscedasticity.

Finally, we briefly mention some work on ANOVA models for dependent data. Bedall (1978) considers simulation experiments for Markov chain data. Brillinger (1980) considers various ANOVA problems for stationary time series data. More recently, De Iorio et al. (2004) developed ANOVA models for dependent random measures, while Mykland and Zhang (2006) consider ANOVA for diffusions Itô processes.

The rest of this paper is organized as follows. Section 2 defines the notation and assumptions used throughout the paper. Section 3 contains our result concerning the behavior of the F statistic for testing the equality of the offspring means. Section 4 is devoted to the study of the ANOVA model for ancestor means. In Section 5 we present algorithms for implementing our methodology. The methodology is then implemented in Section 6 on simulated data and Section 7 on experimental data. The proofs are contained in Sections 8 and 9.

3.2 Notation and assumptions

3.2.1 Notation for the branching process data

We observe data from a independent groups of branching processes; each group is defined by an ancestor distribution, which describes the initial number of particles of the branching process, and an offspring distribution, which describes the dynamics of how the branching process grows over time. To be precise, fix i, $1 \leq i \leq a$, and define two independent collections of random variables both distributed on the positive integers $\mathbb{N}$, the ancestor collection $\{Z_{i,j}(0) : 1 \leq j \leq r_i(n)\}$ and the offspring collection $\{\xi_{i,j,k}(n) : n \geq 1, 1 \leq i \leq r_i(n), k \geq 1\}$. For each i, $\{Z_{i,j}(0) : 1 \leq j \leq r_i(n)\}$ is a collection of independent and identically distributed (i.i.d.) random variables with mean $m_{A,i} \equiv E\, Z_{i,j}(0)$ and variance $\sigma^2_{A,i} \equiv var(Z_{i,j}(0))$; similarly, for each i, $\{\xi_{i,j,k}(n) : n \geq 1, 1 \leq j \leq r_i(n), k \geq 1\}$ is a collection of i.i.d. random variables with representation ξ_i, mean $m_{o,i} \equiv E\xi_i$ and variance $\sigma^2_{o,i} \equiv var(\xi_i)$. The branching process is defined recursively as

$$Z_{i,j}(n+1) = \sum_{k=1}^{Z_{i,j}(n)} \xi_{i,j,k}(n),$$

where $\xi_{i,j,k}(n)$ is interpreted as the number of children produced by the k^{th} parent in the n^{th} generation of the j^{th} replicate from group i.

We assume that the data from each group is independent. That is, we assume that the ancestor collections, $\{Z_{i,j}(0) : 1 \leq j \leq r_i(n)\}$, are independent random variables across i and the offspring collections, $\{\xi_{i,j,k}(n) : n \geq 1, 1 \leq i \leq r_i(n), k \geq 1\}$, are independent across i. Summarizing, $\{Z_{i,j}(n) : 1 \leq j \leq r_i(n)\}$ denotes a collection of i.i.d. branching processes initiated by a random number of ancestors $Z_{i,j}(0)$, and the data from different groups are independent.

The data from each replicate is observed starting at some (non-random) generation $\tau_{i,j}$ until generation $n_{i,j}$. To make the conditions more transparent when studying asymptotics, we will assume $n_{i,j} = n$ and $\tau_{i,j} = \tau$. This assumption does not entail any loss of generality and also minimizes cumbersome notation. Alternate conditions involving $\wedge_{i,j} n_{i,j}$ can be written down for large sample analysis. However, in our data analysis, we do not make this assumption; instead, we allow each replicate to have its own starting and ending observation time. To summarize, the data for the problem are the generation sizes starting at some generation τ going to n generations, namely $\{Z_{i,j}(k) : \tau \leq k \leq n, 1 \leq j \leq r_i(n)\}$, $1 \leq i \leq a$. For a given replicate, we define the sum of the observed generation sizes by $Y_{i,j}(n) \equiv \sum_{k=\tau}^{n} Z_{i,j}(k)$.

For asymptotic analysis, we need assumptions concerning the moments of the ancestor and offspring distributions. We assume that both ancestor and offspring distributions (of all of the groups) have finite second moments.

Assumption 1. *For each group i, $EZ_{i,j}^2(0) < \infty$ and $E\xi_i^2 < \infty$.*

Our results hold for both balanced and unbalanced data, but we do need to control the relative growth of the replicates in the unbalanced case. Additionally, we assume that the number of replicates goes to infinity slower than n. The following assumption makes these comments explicit.

Assumption 2. *As $n \to \infty$,*

1. $\min_{1 \leq i \leq a} r_i(n) \to \infty$,
2. *For each i, $1 \leq i \leq a$, $\frac{r_i(n)}{n} \to 0$,*
3. $\frac{r_i(n)}{r_j(n)} \to 1$.

As is discussed below, the limiting distribution of the F statistic is a linear combination of χ_1^2 random variables, where the constants are the relative sizes of the group variances. To make this idea precise, we use the notation of Marden (1995) (p.58) to define the following distribution. For any (possibly singular) $k \times k$ matrix Λ, let $\chi^2[\Lambda]$ denote the distribution of $\sum_{i=1}^{k} \lambda_i X_i^2$, where λ_i's are the eigenvalues of Λ, and X_i are i.i.d. standard Gaussian random variables. The *centering matrix* also plays a role in the limiting distribution of the F statistic. Let $\mathbf{I}_a$ be the $a \times a$ identity matrix, let $\mathbf{J}_a$ be the $a \times a$ matrix whose entries are all unity, and define $\mathbf{C}_a \equiv \mathbf{I}_a - \frac{1}{a}\mathbf{J}_a$. We also use the notation $diag\,(x_1, ..., x_a)$ to represent an $a \times a$ diagonal matrix with diagonal elements $x_1, ..., x_a$.

3.3 ANOVA for the offspring mean

In this section, we develop a test statistic for testing the null hypothesis

$$H_0 : m_{0,1} = m_{0,2} = \cdots m_{0,a}.$$

This problem, besides being of inherent interest, plays an important role in testing the equality of ancestor means.

Following the ideas from the independent data, we begin by developing the analogous quantities for testing H_0. First consider the case of estimating the offspring mean from the observation of a single branching process $\{Z_k : \tau \leq k \leq n\}$. For this data the non-parametric maximum likelihood estimator for the offspring mean is $\frac{\sum_{k=\tau+1}^{n} Z_k}{\sum_{k=\tau}^{n-1} Z_k}$ (Guttorp, 1991). The (random) sample size for this problem is thus $\sum_{k=\tau}^{n-1} Z_k$.

Under the present data structure, the non-parametric maximum likelihood estimator for the offspring mean of group i is

$$\hat{m}_{o,n,i} \equiv \frac{\sum_{j=1}^{r_i(n)} \big(Y_{i,j}(n) - Z_{i,j}(\tau)\big)}{\sum_{j=1}^{r_i(n)} Y_{i,j}(n-1)},$$

with sample size $w_{n,i} \equiv \sum_{j=1}^{r_i(n)} Y_{i,j}(n-1)$. Hence, under H_0, a natural estimator of the overall mean across all of the groups is

$$\hat{m}_{o,n} \equiv \frac{1}{w_n} \sum_{i=1}^{a} w_{n,i}\, \hat{m}_{o,n,i},$$

where $w_n \equiv \sum_{i=1}^{a} w_{n,i}$. We now define analogs of the standard ANOVA quantities in this context: treatment sum of squares (SST), mean treatment sum of squares (MST), error sum of squares (SSE), and the mean squared error (MSE). First, define the SST as the variation between the individual group means and the overall mean,

$$SST_{o,n} \equiv \sum_{i=1}^{a} w_{n,i} \left(\hat{m}_{o,n,i} - \hat{m}_{o,n}\right)^2.$$

Define the SSE, which measures the within group variation, as the weighted sum of the group variances, namely

$$SSE_{o,n} \equiv \sum_{i=1}^{a} (r_i(n) - 1)\, \hat{\sigma}^2_{o,n,i},$$

where the group variance estimator is given by

$$\hat{\sigma}^2_{o,n,i} \equiv \frac{1}{r_i(n) - 1} \sum_{j=1}^{r_i(n)} \frac{1}{n-\tau} \sum_{k=\tau}^{n-1} Z_{i,j}(k) \left(\frac{Z_{i,j}(k+1)}{Z_{i,j}(k)} - \hat{m}_{o,n,i} \right)^2.$$

Now define the appropriate averaged quantities for MST and MSE,

$$MST_{o,n} \equiv \frac{1}{a-1} SST_{o,n}, \quad MSE_{o,n} = \frac{1}{R(n) - a} SSE_{o,n},$$

where $R(n) = \sum_{i=1}^{a} r_i(n)$. Finally define the F statistic

$$F_{o,n} = \frac{MST_{o,n}}{MSE_{o,n}}.$$

Additionally, it is useful to define the average (over the a groups) of the ancestor mean and offspring variance,

$$\bar{m}_A \equiv \frac{1}{a}\sum_{i=1}^{a} m_{A,i}, \qquad \bar{\sigma}_o^2 \equiv \frac{1}{a}\sum_{i=1}^{a} \sigma_{o,i}^2 \,.$$

We now state the result which describes the limiting behavior of $F_{o,n}$ under the null hypothesis.

Theorem 5. *Let Assumptions 1 and 3 hold. Then under* H_0 *,*

$$(a-1)\, F_{o,n} \xrightarrow{d} \chi^2[\mathbf{L}_a \mathbf{\Sigma}_o \mathbf{L}_a^t] \text{ as } n \to \infty,$$

with $\mathbf{\Sigma}_o \equiv \frac{1}{\bar{\sigma}_o^2} diag\left(\sigma_{o,1}^2, ..., \sigma_{o,a}^2\right)$ *and* $\mathbf{L}_a \equiv \mathbf{I}_a - \frac{1}{a\,\bar{m}_A}\mathbf{s}_a\mathbf{s}_a^t$*, where*

$$\mathbf{s}_a \equiv \left(\sqrt{m_{A,1}}, ..., \sqrt{m_{A,a}}\right)^t .$$

Remark 1. *The results of the paper can be proved under more general conditions, but Assumption 2 helps to simplify the exposition. For instance, we could instead assume,* $\frac{r_i(n)}{r_j(n)} \to c_{ij}$*, for some constant* $c_{ij} > 0$*. These constants would show up in the limiting distribution, and thus our results would only be useful in the case where the constants are known (for instance if a scientist decided to collect twice as many replicates on a certain group, it would be possible to modify the results to account for this).*

We now develop the analogues of the ANOVA test for equality of ancestor means.

3.4 ANOVA for the ancestor mean

In this section we develop a test statistic test for the null hypothesis

$$H_{0,A} : m_{A,1} = m_{A,2,} = \cdots m_{A,a}.$$

The development relies on the classical martingale limit associated with a supercritical branching process. We review the key ideas here. It is well-known (Athreya and Ney, 1972) that there is a limiting random variable W, obtained by scaling the size of the n^{th} generation by its mean. Information concerning the ancestor distribution is contained in W, and hence W is a critical object of study. To be specific, we define $W_{i,j}(n) \equiv \frac{Z_{i,j}(n)}{m_{o,i}^n}$. For each fixed i and j, $W_{i,j}(n)$ is a positive martingale and hence has a limit $W_{i,j}$ which is non-degenerate under a finite second moment assumption. In fact, for each fixed (n, i), $W_{i,j}(n)$ are i.i.d. (across j), and hence, the limit random variables $W_{i,j}$ are i.i.d (across j).

The key result that will be useful for our purpose is that the moments of $W_{i,j}$ are functions of the moments of the ancestor and offspring distributions, namely,

$$E\,W_{i,j} = m_{A,i}, \qquad \sigma_{W,i}^2 = \frac{m_{A,i}\,\sigma_{o,i}^2}{m_{o,i}(m_{o,i}-1)} + \sigma_{A,i}^2, \tag{3.1}$$

where $\sigma_{W,i}^2 \equiv var(W_{i,j})$. Therefore, if $W_{i,j}$ were observable the ANOVA problem for the mean of the ancestor distributions would reduce to the case of (non-Gaussian, heterogenous) independent data. As mentioned above, the observed data for each group i is in fact $\{Z_{i,j}(k) : \tau \leq k \leq n, 1 \leq j \leq r_i(n)\}$. The following result is well-known (Guttorp, 1991)

$$\frac{Y_{i,j}(n)}{m_{o,i}^n} \xrightarrow{a.s.} \frac{m_{o,i}}{m_{o,i}-1}\, W_{i,j}\,. \tag{3.2}$$

In light of (4.1) and (3.2), it is instructive to define $\hat{V}_{i,j}(n) \equiv \frac{\hat{m}_{o,n,i}-1}{\hat{m}_{o,n,i}^{n+1}}\, Y_{i,j}(n)$.

Now, under the assumption of the equality of the offspring means across the groups, the estimate of $m_{o,i}$ can be improved by "borrowing strength" across the groups. Namely, the idea is to replace the estimator $\hat{m}_{o,n,i}$ with the pooled estimator of the offspring mean $\hat{m}_{o,n}$, which was defined in Section 3.3. Specifically, define $\tilde{V}_{i,j}(n) \equiv \frac{\hat{m}_{o,n}-1}{\hat{m}_{o,n}^{n+1}}\, Y_{i,j}(n)$.

For the question of ancestor inference, the effective sample size for each group is simply the number of replicates in that group. Borrowing ideas from classical ANOVA theory we define the relevant quantities that facilitate construction of the F statistic for testing $H_{0,A}$, namely the group means, the overall mean, the error sum of squares, the treatment sum of squares, and the F statistic. Let

$$\bar{V}_{i\cdot}(n) \equiv \frac{1}{r_i(n)} \sum_{j=1}^{r_i(n)} \hat{V}_{i,j}(n), \quad \bar{V}_{\cdot\cdot}(n) \equiv \frac{1}{R(n)} \sum_{i=1}^{a} \sum_{j=1}^{r_i(n)} \hat{V}_{i,j}(n),$$

$$SSE_{A,n} = \sum_{i=1}^{a} \sum_{j=1}^{r_i(n)} \left(\hat{V}_{i,j}(n) - \bar{V}_{i\cdot}(n)\right)^2, \quad SST_{A,n} \equiv \sum_{i=1}^{a} r_i(n) \left(\bar{V}_{i\cdot}(n) - \bar{V}_{\cdot\cdot}(n)\right)^2,$$

and

$$F_{A,n} = \frac{MST_{A,n}}{MSE_{A,n}},$$

where $MSE_{A,n} \equiv \frac{1}{R(n)-a} SSE_{A,n}$ and $MST_{A,n} \equiv \frac{1}{a-1} SST_{A,n}$. Finally, we define a variance estimator for each group and the average variance over all a groups,

$$\hat{\sigma}^2_{W,n,i} \equiv \frac{1}{r_i(n)} \sum_{j=1}^{r_i(n)} \left(\hat{V}_{i,j}(n) - \bar{V}_{i\cdot}(n)\right)^2, \quad \bar{\sigma}^2_W \equiv \frac{1}{a} \sum_{i=1}^{a} \sigma^2_{W,i}\,.$$

And for the case of $\tilde{V}_{i,j}(n)$ we define all of the analogous quantities, denoting them with a tilde.

$$\tilde{V}_{i\cdot}(n) \equiv \frac{1}{r_i(n)} \sum_{j=1}^{r_i(n)} \tilde{V}_{i,j}(n), \quad \tilde{V}_{\cdot\cdot}(n) \equiv \frac{1}{R(n)} \sum_{i=1}^{a} \sum_{j=1}^{r_i(n)} \tilde{V}_{i,j}(n),$$

$$S\tilde{S}E_A = \sum_{i=1}^{a} \sum_{j=1}^{r_i(n)} \left(\tilde{V}_{i,j}(n) - \tilde{V}_{i\cdot}(n)\right)^2, \quad S\tilde{S}T_A \equiv \sum_{i=1}^{a} r_i(n) \left(\tilde{V}_{i\cdot}(n) - \tilde{V}_{\cdot\cdot}(n)\right)^2,$$

$$M\tilde{S}E_A \equiv \frac{1}{R(n)-a} S\tilde{S}E_A, \quad M\tilde{S}T_A \equiv \frac{1}{a-1} S\tilde{S}T_A, \quad \tilde{F}_{A,n} = \frac{M\tilde{S}T_A}{M\tilde{S}E_A},$$

and

$$\tilde{\sigma}^2_{W,n,i} \equiv \frac{1}{r_i(n)} \sum_{j=1}^{r_i(n)} \left(\tilde{V}_{i,j}(n) - \tilde{V}_{i\cdot}(n)\right)^2.$$

We now state our main results concerning the hypothesis test for $H_{0,A}$.

Theorem 6. *Let Assumptions 1 and 3 hold. Then, under* $H_{0,A}$

$$(a-1)\,F_{A,n} \xrightarrow{d} \chi^2[\mathbf{C}_a \mathbf{\Sigma} \mathbf{C}_a^t], \quad as \quad n \to \infty,$$

where $\mathbf{\Sigma} \equiv \frac{1}{\sigma_W^2} diag\left(\sigma_{W,1}^2, ..., \sigma_{W,a}^2\right)$.

Theorem 7. *Let Assumptions 1 and 3 hold. Then under* H_0 *and* $H_{0,A}$

$$(a-1)\,\tilde{F}_{A,n} \xrightarrow{d} \chi^2[\mathbf{C}_a \mathbf{\Sigma} \mathbf{C}_a^t], \quad as \quad n \to \infty$$

where $\mathbf{\Sigma} \equiv \frac{1}{\sigma_W^2} diag\left(\sigma_{W,1}^2, ..., \sigma_{W,a}^2\right)$.

Remark 2. *It is worthwhile to consider when the homogeneity assumption is reasonable. First remember, that homogeneity only needs to hold under the null hypothesis. Also, using* (4.1)*, recall that* $\sigma_{W,i}^2$ *is a function of the ancestor mean, offspring mean, and offspring variance. If the ancestor distribution is a parametric family characterized by its mean, and additionally* $(m_{o,i}, \sigma_{o,i}^2) = (m_o, \sigma_0^2)$*, for all* i*, then under the null hypothesis,* $\sigma_{W,i}^2 = \sigma_W^2$ *(for all* i*). This situation is approximately true for PCR data. In that case, the offspring distribution (for all groups) is Bernoulli, so that if efficiency of the replication is equal across groups, then* $(m_{o,i}, \sigma_{o,i}^2) = (m_o, \sigma_0^2)$*, for all* i.

3.5 Implementation of the test procedures

As mentioned in the introduction and discussed in several references the problem with ANOVA models under heterogenous variances is power (Krutchkoff, 1988; Argaç, 2004). As will be seen in the simulation section, the procedures obtain close to the nominal size. The difficulty is to chose the most powerful procedure. For concreteness, we focus our comments on testing the ancestral mean equality with the statistic $F_{A,n}$. The algorithms we discuss are naturally extended for the statistics $\tilde{F}_{A,n}$ and $F_{o,n}$.

Argaç (2004) contains a nice summary of many of the commonly used procedures for correcting for heterogeneity. In the discussion, Argaç concludes that there is no 'best' way to account for heterogeneity. In fact, he writes "due to the non-standard assumptions...we make, there does not appear to be any systematic pattern in the simulations, and thus we cannot provide the reader with a general recommendation or a simple take-home message." In his simulations, Cochran's test is often the most powerful procedure; however, the results are not size-adjusted. This difficulty in evaluation is not surprising, since, as many authors have pointed out, the ANOVA model under heterogeneity is a Behrens-Fisher type problem. Interestingly, Argaç does not mention a procedure based on some analog of Theorem 6. Lee et al. (2007) prove an analog of Theorem 6 for heterogenous Gaussian data and provide a numerical method for computing the associated critical values.

We consider two basic procedures: the classical F-test and a procedure based on Theorem 6, which we refer to as the Monte-Carlo G-test. For the classical F-test we proceed ignoring the issue of variance heterogeneity; H_0 is rejected (at level α) if $F_{A,n} > F_{a-1,R(n)-a,1-\alpha}$, where $F_{\nu_1,\nu_2,1-\alpha}$ represents the $1-\alpha$ percentile of the F_{ν_1,ν_2} distribution. Theorem 6 gives an asymptotic result which we can use directly to test the null hypothesis of equal ancestor means. Of course, the limiting distribution depends on the eigenvalues of $\mathbf{C}_a\mathbf{\Sigma}\mathbf{C}_a^t$, which involves the unknown group variances. We will thus focus on computing the eigenvalues of $\mathbf{C}_a\hat{\mathbf{\Sigma}}_n\mathbf{C}_a^t$, where $\hat{\mathbf{\Sigma}}_n \equiv \frac{1}{\hat{\sigma}^2_{W,n}} diag\left(\hat{\sigma}^2_{W,n,1}, ..., \hat{\sigma}^2_{W,n,a}\right)$. Finally, the test is performed by rejecting (at level α) if $(a-1)\,F_{A,n} > \chi^2[\mathbf{C}_a\hat{\mathbf{\Sigma}}_n\mathbf{C}_a^t, 1-\alpha]$, where $\chi^2[\Lambda, 1-\alpha]$

represents $1-\alpha$ percentile of the $\chi^2[\Lambda]$ distribution. The percentiles of the $\chi^2[\Lambda]$ distribution are determined using Monte-Carlo methods; recalling the definition of $\chi^2[\Lambda]$ this involves determining the eigenvalues of Λ and simulating independent standard Gaussian pseudo-random variates. When the test is performed in this manner, we refer to it as the *Monte Carlo G-test.* Alternative numerical procedures are available. Lee et al. (2007) provide a non-random numerical algorithm for computing the critical values of a related test procedure. Strawderman (2004) studies the more general problem of computing tail probabilities for absolutely continuous distributions. We do not consider these methods in our analysis.

We make a brief comment on the eigenvalues of $\mathbf{C}_a\mathbf{\Sigma}\mathbf{C}_a^t$. In the equal variance case, $\mathbf{\Sigma} = \mathbf{I}_a$, hence $\mathbf{C}_a\mathbf{\Sigma}\mathbf{C}_a^t = \mathbf{C}_a\mathbf{C}_a^t = \mathbf{C}_a$. It is well-known (Moser, 1996) that $\mathbf{C}_a$ has one zero eigenvalue and $a-1$ eigenvalues equal to unity. It is easy to show that $rank\left(\mathbf{C}_a\mathbf{\Sigma}\mathbf{C}_a^t\right) = a-1$. And thus we need to approximate the $a-1$ non-zero eigenvalues of $\mathbf{C}_a\hat{\mathbf{\Sigma}}_n\mathbf{C}_a^t$.

3.6 Simulations

In this section we explore the size and power of ANOVA tests for differences in the ancestor means. We compare the performance of the classical F-test with the ANOVA G-test (based on Theorem 6).

All of the results in this section are based on 5000 simulations; additionally, the Monte Carlo G-test results are based on $M = 2000$ Monte Carlo samples. Throughout this section the size level is fixed at $\alpha = .05$. Also, the branching processes are observed for generations $\tau = 10$ to $n = 20$.

We begin by assessing size and power under asymptotic homogeneity with balanced data for three groups. The offspring distribution is Bernoulli (on $\{1, 2\}$) and the ancestor distribution is Poisson. The results of the simulation assessing size for varying numbers of replicates, offspring means, and ancestor means are shown in Tables 3.1 and 3.2. For tests of power, all three groups have the same Bernoulli offspring distribution, and a Poisson ancestor distribution with varying means. We vary the means in the following way. First fix the mean of one group at $m_{A,1}$, now let the means of group two and three be given by $m_{A,1} - \delta$ and $m_{A,1} + \delta$, respectively, for varying values of δ. The results of these simulations are shown in Figure 3.1.

Next we consider the case of (asymptotic) heterogeneity. Again, we consider three groups all with Poisson ancestor distributions and Bernoulli offspring distributions. The difference here is the three groups have different offspring means, $1.5, 1.8$ and 1.95. The size of these simulations, for various parameter values, are given in Table 3.3.

In the simulations considered, both tests achieve close to the nominal size; however, in every comparison the classical F-test is closer to the nominal value of .05. On the other hand, the Monte Carlo G-test is always more powerful than the classical F-test.

3.7 Data analysis

The PCR data analyzed in this section were collected from a ABI Prism 7700 Sequence Detection System. The data comes from the Luteinizing hormone taken from a mouse pituitary gland. We collected data on 16 replicates from three

Table 3.1: F test for m_A.

$m_A = 10$

replicates	$p = .98$	$p = .95$	$p = .5$	$p = .1$
10	0.0472	0.0536	0.0504	0.0481
20	0.05	0.0518	0.0524	0.0463
30	0.0476	0.053	0.0488	0.0457

$m_A = 50$

replicates	$p = .98$	$p = .95$	$p = .5$	$p = .1$
10	0.0516	0.052	0.0512	0.0524
20	0.0518	0.0454	0.0552	0.051
30	0.0474	0.0496	0.054	0.0482

$m_A = 100$

replicates	$p = .98$	$p = .95$	$p = .5$	$p = .1$
10	0.0496	0.0528	0.048	0.0494
20	0.0526	0.0494	0.0544	0.0508
30	0.0502	0.048	0.0494	0.055

different dilutions. Let LH_1, LH_2 and LH_3 denote the three dilutions. The master mix for LH_i is obtained from the master mix of LH_{i-1} using a dilution factor of 2.9505. Thus, if $m_{a,LH_1} = m_a$, then $m_{a,LH_i} = \frac{m_a}{2.9505^{i-1}}$ $i = 2, 3$.

The data are plotted in Figure 3.2; on the log scale, the three groups are visually separated in the exponential phase. In our analysis we excluded data from one of the reactions for LH_2, since it did not reach the appropriate C_T level. Similarly, we excluded data from one LH_1 replicate since its C_T value was much larger than those of the other LH_1 replicates.

Table 3.2: Monte Carlo G test for m_A.

$m_A = 10$

replicates	$p = .98$	$p = .95$	$p = .5$	$p = .1$
10	0.0648	0.0638	0.0576	0.0597
20	0.0586	0.055	0.054	0.0606
30	0.053	0.0528	0.0534	0.0544

$m_A = 50$

replicates	$p = .98$	$p = .95$	$p = .5$	$p = .1$
10	0.0588	0.0684	0.0614	0.0626
20	0.0588	0.0516	0.061	0.056
30	0.0532	0.0592	0.0604	0.0486

$m_A = 100$

replicates	$p = .98$	$p = .95$	$p = .5$	$p = .1$
10	0.0632	0.0654	0.0586	0.0634
20	0.0596	0.0568	0.0568	0.0604
30	0.0554	0.0542	0.0516	0.0562

Our methodology requires identification of the exponential phase. The strategy is to choose those cycles which yield a fluorescence of at least $F^\star$ and per-cycle amplification of at least m_c. The following algorithm identifies the cycles of data belonging to the exponential phase. Recall that $C_T \equiv \inf\left\{j : F_j > F^\star\right\}$. In our analyses we chose $F^\star = 0.2$ and $m_c = 1.5$. The choice of $F^\star$ was suggested by the manufacturers of the equipment.

First we consider pairwise comparisons. We quantitate LH_1 relative to LH_2; and then quantitate LH_3 relative to LH_2. Using the standard PCR terminology,

Table 3.3: F and Monte-Carlo G-test for m_A. The ancestor distribution is Poisson and the offspring distribution is Bernoulli. But the three groups have different offspring means: $1.5, 1.8$ and 1.95.

F-test

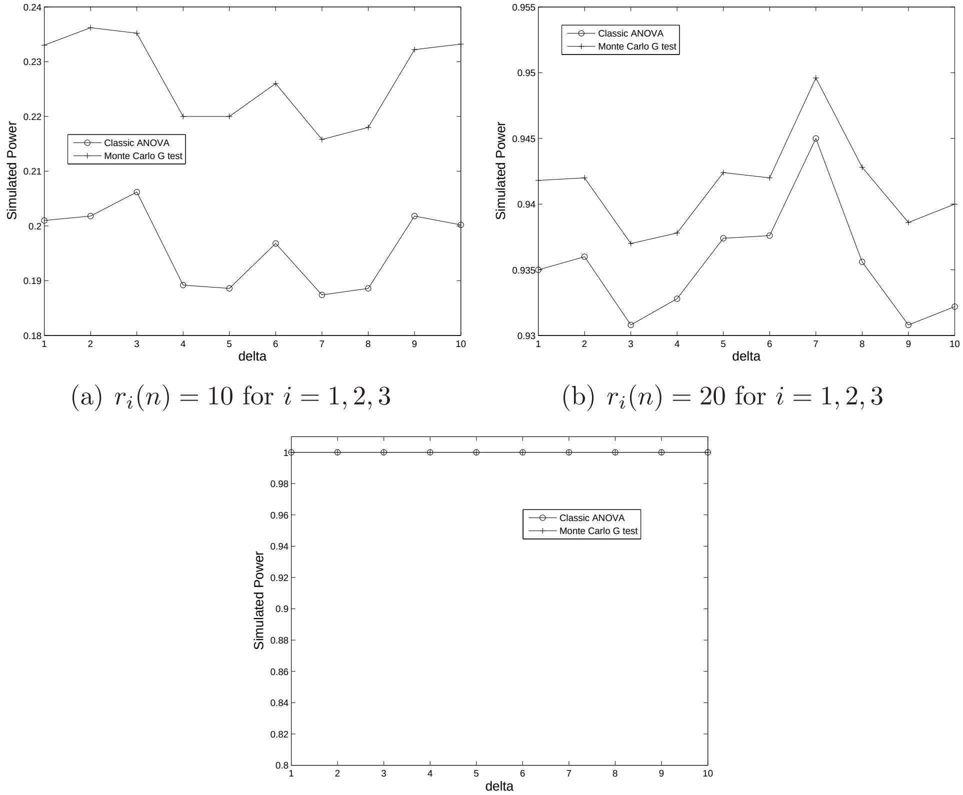
replicates	$m_A = 10$	$m_A = 20$	$m_A = 30$
10	0.0508	0.053	0.0532
20	0.0482	0.0556	0.052
30	0.0472	0.0514	0.0524

Monte Carlo G-test

replicates	$m_A = 10$	$m_A = 20$	$m_A = 30$
10	0.0672	0.0678	0.0618
20	0.058	0.0572	0.058
30	0.0554	0.0536	0.054

LH_1 and LH_3 are target groups, and LH_2 is the calibrator. Note that the optimal point estimates are: 2.9505 for LH_1 versus LH_2 and $\frac{1}{2.9505} = 0.3389$ for LH_3 versus LH_1. For LH_1 versus LH_2, the point estimate is 2.8221 with a 95% bootstrap confidence interval of $(1.6870, 3.6013)$. For LH_3 versus LH_2, the values are 0.4136 and $(0.2061, 0.5417)$. The bootstrap confidence intervals were based on 2000 bootstrap resamples of the reactions. Notice, that the two confidence intervals do not overlap.

Next we consider comparing the three groups using the methods presented in this paper. The F-statistic takes the value $F = 49.9477$. Both the classical F-test ($p < 10^{-6}$) and the Monte Carlo G-test ($p = 0$) reject the null hypothesis of equality of the ancestor means across the three groups. The Monte Carlo procedure is based on $M = 2000$ Monte-Carlo samples.



(a) $r_i(n) = 10$ for $i = 1, 2, 3$

(b) $r_i(n) = 20$ for $i = 1, 2, 3$

(c) $r_i(n) = 30$ for $i = 1, 2, 3$

Figure 3.1: Simulated power versus δ. All three groups have the same Bernoulli offspring distribution with success probability $p = .9$. The three groups each have an ancestor distribution given by $\max(1, Poiss(\lambda_i))$, for different values of λ_i. The first group has $\lambda_1 = 10$; the rates for the other two groups are given by $\lambda_1 - \delta$ and $\lambda_1 + \delta$.

3.8 Proofs of initial results

Before proving Theorems 5, 6, and 7, we need to establish some preliminary results.

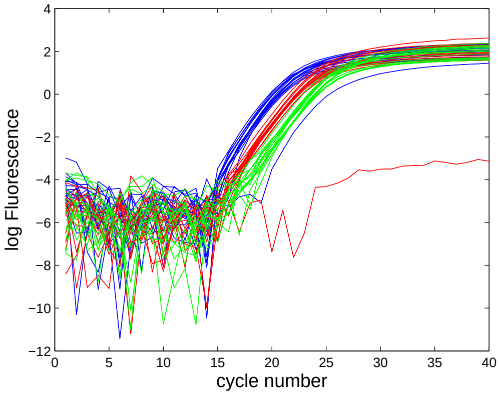


Figure 3.2: Plot of cycle number versus log fluorescence (j vs. $\log F_j$) for all 16 replicates of LH_1 (in blue), LH_2 (in red), and LH_3 (in green).

3.8.1 Initial Propositions

Before proving the key ANOVA theorems (Theorems 5, 6, and 7) we need to establish several intermediate results; first proving limit results for the offspring distribution estimators and then for the ancestor distribution estimators.

We begin by stating three propositions. First we state a harmonic moment result which is an immediate corollary of Theorem 1 in Ney and Vidyashankar (2003).

Proposition 3. *Let Assumption 1 hold. Fix $r \geq 1$. For each group i, there exists a $C_i > 0$ and a $\gamma \in (0,1)$ such that*

$$E\left(Y_{i,j}^{-r}(n)\right) \leq C_i \gamma^n.$$

The following result central limit theorem result is proved using a standard characteristic function argument.

Proposition 4. *Let $\{X_{n,i}\}$ be a collection random variables with $EX_{n,i} = 0$. Additionally, for each fixed n, $\left\{X_{n,i} : i \geq 1\right\}$ are i.i.d. with finite variance var $\left(X_{n,i}\right) =$*

σ_n^2. *If* $\sigma_n^2 \to \sigma^2$ *and* $k_n \to \infty$ *as* $n \to \infty$, *then*

$$\frac{1}{\sqrt{k_n}} \sum_{i=1}^{k_n} X_{n,i} \xrightarrow{d} N(0, \sigma^2).$$

Finally, we state a result which summarizes some important estimates concerning the covariance of the martingale sequence.

Proposition 5. *Let Assumption 1 hold. Fix group* i, *let* $C_{i,k}$ $k = 1, 2, 3$, *be positive constants. We have,*

$$var\left(W_{i,j}(n) - W_{i,j}\right) = \frac{C_{i,1}}{m_{o,i}^n},$$

$$\left|cov\left(W_{i,j}(n) - W_{i,j}, W_{i,j}(\ell) - W_{i,j}\right)\right| \leq \frac{C_{i,2}}{m_{o,i}^{(n+\ell)/2}},$$

$$\left|cov\left(W_{i,j}(n) - W_{i,j}, W_{i,j}\right)\right| \leq \frac{C_{i,3}}{m_{o,i}^{n/2}}.$$

3.8.2 Limit Results Related to the Offspring Distribution Estimators

We begin by proving estimates for the non-parametric maximum likelihood estimator of the offspring mean based on a single replicate (Guttorp, 1991). To be precise define

$$\hat{m}_{o,n,i,j} \equiv \frac{Y_{i,j}(n) - Z_{i,j}(\tau)}{Y_{i,j}(n-1)}.$$

Lemma 8. *Let Assumption 1 hold. For each* $\epsilon > 0$, *there exists a* $C > 0$ *and a* $\gamma \in (0, 1)$ *such that*

$$P\left(\left|\hat{m}_{o,n,i,j} - m_{o,i}\right| > \epsilon\right) \leq Cn\gamma^n.$$

Proof. We begin by proving the following estimate. There exists a $C > 0$, such that

$$E\left[\sqrt{Y_{i,j}(n-1)}\left(\hat{m}_{o,n,i,j} - m_{o,i}\right)\right]^2 \leq Cn^2. \tag{3.3}$$

To this end, we note that

$$\begin{aligned}\sqrt{Y_{i,j}(n-1)}\left(\hat{m}_{o,n,i,j} - m_{o,i}\right) &= \sqrt{Y_{i,j}(n-1)}\left(\frac{Y_{i,j}(n) - Z_{i,j}(\tau)}{Y_{i,j}(n-1)} - m_{o,i}\right) \\ &= \sum_{k=\tau}^{n} \frac{Z_{i,j}(k+1) - m_{o,i}\, Z_{i,j}(k)}{\sqrt{Z_{i,j}(k)}}\, w_{i,j}(n,k),\end{aligned}$$

where

$$w_{i,j}^2(n,k) \equiv \frac{Z_{i,j}(k)}{Y_{i,j}(n-1)}.$$

Thus, setting $X_{i,j}(k) = \frac{Z_{i,j}(k+1) - m_{o,i}\, Z_{i,j}(k)}{\sqrt{Z_{i,j}(k)}}$, we have that

$$\begin{aligned}\left(\sqrt{Y_{i,j}(n-1)}\left(\hat{m}_{o,n,i,j} - m_{o,i}\right)\right)^2 &\leq n^2\left(\frac{1}{n-\tau}\sum_{k=\tau}^{n-1}\left|X_{i,j}(k)\right|\right)^2 \\ &\leq n^2\left(\frac{1}{n-\tau}\sum_{k=\tau}^{n-1} X_{i,j}^2(k)\right),\end{aligned}$$

where the last inequality follows from Jensen's inequality for convex functions. Taking expectations gives (3.3).

By Markov's inequality,

$$\begin{aligned}P\left(\left|\hat{m}_{o,n,i,j} - m_{o,i}\right| > \epsilon\right) &\leq \frac{1}{\epsilon}E\left|\hat{m}_{o,n,i,j} - m_{o,i}\right| \\ &\leq \epsilon^{-2}\sqrt{d_n\left(1\right)d_n\left(2\right)}, \end{aligned} \tag{3.4}$$

where

$$d_n\left(1\right) = \left[E\left[\sqrt{Y_{i,j}(n-1)}\left(\hat{m}_{o,n,i,j} - m_{o,i}\right)\right]^2\right]^{1/2} \quad \text{and}$$

$$d_n\left(2\right) = \left[E\left(Y_{i,j}^{-1}(n-1)\right)\right]^{1/2}.$$

The last inequality follows by first multiplying and dividing by $\sqrt{Y_{i,j}(n-1)}$ inside the expectation in (3.4) and then applying the Cauchy-Schwarz inequality. The result now follows from Proposition 3 and (3.3). □

Next we consider the maximum deviation, across all replicates, of $\hat{m}_{o,n,i,j}$ from $m_{o,i}$. Namely, define

$$M^{\star}_{n,i} \equiv \max_{1\leq j\leq r_i(n)} \left|\hat{m}_{o,n,i,j} - m_{o,i}\right|.$$

Lemma 9. *Let Assumptions 1 and 3 hold. Then, with probability one,*

$$\lim_{n\to\infty} M^{\star}_{n,i} = 0.$$

Proof. Fix $\epsilon > 0$. Let $\alpha(n) = P\left(\left|\hat{m}_{o,n,i,j} - m_{o,i}\right| > \epsilon\right)$. Then, using the fact $a^n - b^n = (a-b)\sum_{k=0}^{n-1} a^k b^{(n-1)-k}$,

$$\begin{aligned}
\sum_{n\geq 1} P\left(M^{\star}_{n,i} > \epsilon\right) &= \sum_{n\geq 1} 1 - \left[1-\alpha(n)\right]^{r_i(n)} \\
&= \sum_{n\geq 1} \alpha(n) \sum_{\ell=0}^{r_i(n)-1} \left(1-\alpha(n)\right)^{\ell} \\
&\leq \sum_{n\geq 1} C\, r_i(n)\, n\gamma^n,
\end{aligned}$$

for some $\gamma \in (0,1)$, using Lemma 8. But, using the ratio test, $\sum_{n\geq 1} r_i(n)\, n\gamma^n < \infty$, hence the desired result follows from the Borel-Cantelli lemma. □

Next we prove analogous results for the variance estimator. First define a variance estimator, in the case the mean is known,

$$\sigma^2_{o,n,i} \equiv \frac{1}{r_i(n)} \sum_{j=1}^{r_i(n)} \sigma^2_{o,n,i,j},$$

where

$$\sigma^2_{o,n,i,j} \equiv \frac{1}{n-\tau} \sum_{k=\tau}^{n-1} Z_{i,j}(k) \left(\frac{Z_{i,j}(k+1)}{Z_{i,j}(k)} - m_{o,i}\right)^2.$$

We first state a result which considers the moments of this estimator.

Proposition 6. *Assume $EZ_0^2 < \infty$ and $E\xi^2 < \infty$.*

$$E\left(\sigma^2_{o,n,i,j}\right) = \sigma^2_{o,i}, \qquad var\left(\sigma^2_{o,n,i,j}\right) = \frac{a_n}{(n+1)^2},$$

where

$$a_n \equiv 2(n+1)\,\sigma^2_{o,i} + C\sum_{k=\tau}^{n} E\left(Z^{-1}_{i,j}(n)\right),$$

and $C = var\left(\xi_i - m_{o,i}\right)^2 - 2\,\sigma^2_{o,i}$.

Proof. The result follows from a minor modification of the moment arguments given in the proof of Theorem 1 in Dion (1975). □

To prove a result for the variance estimator, which is analogous to Lemma 9, define the following quantity,

$$V^{\star}_{n,i} \equiv \max_{1\le j\le r_i(n)} \left|\sigma^2_{o,n,i,j} - \sigma^2_{o,i}\right|.$$

Lemma 10. *Let Assumptions 1 and 3 hold. Then, as $n \to \infty$,*

$$V^{\star}_{n,i} \xrightarrow{P} 0. \tag{3.5}$$

It immediately follows, that

$$\sigma^2_{o,n,i} \xrightarrow{P} \sigma^2_{o,i}\,. \tag{3.6}$$

Proof. Fix $\epsilon > 0$. Let $\alpha(n) = P\left(\left|\sigma^2_{o,n,i,j} - \sigma^2_{o,i}\right| > \epsilon\right)$. Then, using the fact $a^n - b^n = (a-b)\sum_{k=0}^{n-1} a^k b^{(n-1)-k}$,

$$\begin{aligned} P\left(V^{\star}_{n,i} > \epsilon\right) &= 1 - \left[1-\alpha(n)\right]^{r_i(n)} \\ &= \alpha(n)\sum_{k=0}^{r_i(n)-1}\left(1-\alpha(n)\right)^k \\ &\le r_i(n)\,\alpha(n). \end{aligned}$$

But, using Chebyshev's inequality and Proposition 6

$$\begin{aligned} \alpha(n) &= P\left(\left|\sigma^2_{o,n,i,j} - \sigma^2_{o,i}\right| > \epsilon\right) \\ &\leq Cvar(\sigma^2_{o,n,i,j}). \end{aligned}$$

But using Assumption 3, $r_i(n)\, var(\sigma^2_{o,n,i,j}) \to 0$.

Now $\sigma^2_{o,n,i} = \frac{1}{r_i(n)} \sum_{j=1}^{r_i(n)} \sigma^2_{o,n,i,j}$, so (3.6) follows immediately from (3.5). □

We now state the key limit result needed to prove Theorem 5.

Lemma 11. *Let Assumptions 1 and 3 hold. For each i, as $n \to \infty$,*

$$\left(\hat{m}_{o,n,i}, \hat{\sigma}^2_{o,n,i}\right) \xrightarrow{P} (m_{o,i}, \sigma^2_{o,i})$$

and

$$\sqrt{A_{n,i}}\left(\hat{m}_{o,n,i} - m_{o,i}\right) \xrightarrow{d} N(0, \sigma^2_{o,i}),$$

where

$$A_{n,i} \equiv \sum_{j=1}^{r_i(n)} Y_{i,j}(n-1)\,.$$

Proof. We begin by proving consistency of $\hat{m}_{o,n,i}$. We have that

$$\begin{aligned} 0 &\leq \left|\hat{m}_{o,n,i} - m_{o,i}\right| = \left|\sum_{j=1}^{r_i(n)} \frac{Y_{i,j}(n-1)}{\sum_{\ell=1}^{r_i(n)} Y_{i,\ell}(n-1)} \left(\hat{m}_{o,n,i,j} - m_{o,i}\right)\right| \\ &\leq M^\star_{n,i}\,. \end{aligned}$$

The result now follows from Lemma 9.

Next consider the consistency of $\hat{\sigma}^2_{o,n,i}$. Basic algebra yields,

$$\hat{\sigma}^2_{o,n,i} = \sigma^2_{o,n,i} + J_{n,1} - J_{n,2},$$

where,

$$J_{n,1} \equiv (\hat{m}_{o,n,i} - m_{o,i})^2 \quad \frac{1}{r_i(n)(n-\tau)} \sum_{j=1}^{r_i(n)} \sum_{k=\tau}^{n-1} Z_{i,j}(k),$$

and

$$J_{n,2} \equiv 2(\hat{m}_{o,n,i} - m_{o,i}) \quad \frac{1}{r_i(n)(n-\tau)} \sum_{j=1}^{r_i(n)} \sum_{k=\tau}^{n-1} Z_{i,j}(k) \left(\frac{Z_{i,j}(k+1)}{Z_{i,j}(k)} - m_{o,i} \right).$$

Now Lemma 10 gives that $\sigma^2_{o,n,i} \xrightarrow{P} 0$; similar arguments show $J_{n,i} \xrightarrow{P} 0$, for $i = 1, 2$.

Finally, we prove the asymptotic normality result. Basic algebra yields,

$$\sqrt{A_{n,i}} \left(\hat{m}_{o,n,i} - m_{o,i} \right) = \frac{1}{\sqrt{m_{A,n-1,i}}} T_n,$$

where

$$T_n = \left\{ \frac{1}{\sqrt{r_i(n)}} \sum_{j=1}^{r_i(n)} \sqrt{V_{i,j}(n-1)} \sqrt{Y_{i,j}(n-1)} \left(\hat{m}_{o,n,i,j} - m_{o,i} \right) \right\}.$$

But from Corollary 2 $m_{A,n,i} \xrightarrow{a.s.} m_{A,i}$. Thus by Slutsky's Theorem, it is sufficient to prove

$$\frac{1}{\sqrt{r_i(n)}} \sum_{j=1}^{r_i(n)} \sqrt{V_{i,j}(n-1)} \sqrt{Y_{i,j}(n-1)} \left(\hat{m}_{o,n,i,j} - m_{o,i} \right) \xrightarrow{d} N(0, m_{A,i}\, \sigma^2_{o,i}). \tag{3.7}$$

But (3.7) follows immediately from Proposition 4 and standard moment calculations. □

3.8.3 Limit Results Related to the Ancestor Distribution Estimators

Asymptotic theory for the estimator of the ancestor mean relies critically on the behavior of the following term

$$\theta_{n,i} \equiv \left(\frac{\hat{m}_{o,n,i} - 1}{m_{o,i} - 1} \right) \left(\frac{m_{o,i}}{\hat{m}_{o,n,i}} \right)^{n+1}.$$

Before we address this term, we need an estimate for $E\left|\hat{m}_{o,n,i}-m_{o,i}\right|$.

Lemma 12. *Let Assumptions 1 and 3. Fix a group i. There exists a $C_i > 0$ and a $\gamma_i \in (0,1)$, such that*

$$E\left|\hat{m}_{o,n,i}-m_{o,i}\right| \leq C_i n \gamma_i^n .$$

Proof. Basic algebra gives,

$$\left|\hat{m}_{o,n,i}-m_{o,i}\right| \leq \sum_{j=1}^{r_i(n)} \frac{1}{\sqrt{Y_{i,j}(n-1)}} \cdot \sqrt{Y_{i,j}(n-1)}\left|\hat{m}_{o,n,i,j}-m_{o,i}\right| \tag{3.8}$$

The result now follows from (3.3) (see the proof of Lemma 8), (3.8), the Cauchy-Schwartz inequality, and Proposition 3. □

The next result proves that $\theta_{n,i}$ converges to unity, with probability one.

Lemma 13. *Let Assumptions 1 and 3. Then as $n \to \infty$,*

$$\sqrt{r_i(n)}\left|\left(\frac{m_{o,i}}{\hat{m}_{o,n,i}}\right)^n - 1\right| \xrightarrow{a.s.} 0, \tag{3.9}$$

and

$$\sqrt{r_i(n)}\left|\theta_{n,i}-1\right| \xrightarrow{a.s.} 0. \tag{3.10}$$

Proof. We begin by proving (3.9). Fix $\epsilon > 0$ and define $\psi_{n,i} \equiv \frac{m_{o,i}}{\hat{m}_{o,n,i}}$ Notice that

$$\begin{aligned} P\left(\sqrt{r_i(n)}\left|\psi_{n,i}^n - 1\right| > \epsilon\right) &= P\left(\psi_{n,i}^n > 1 + \frac{\epsilon}{\sqrt{r_i(n)}}\right) + P\left(\psi_{n,i}^n < 1 - \frac{\epsilon}{\sqrt{r_i(n)}}\right) \\ &\equiv a_{n,1} + a_{n,2} . \end{aligned}$$

From the Borel-Cantelli lemma it is sufficient to prove $\sum_{n\geq 1} a_{n,\ell} < \infty$, for $\ell = 1, 2$.

To this end, define $b_n = (1 + \frac{\epsilon}{\sqrt{r_i(n)}})^{-1/n}$ and use Markov's inequality to obtain

$$\begin{aligned} P\left(\psi_{n,i}^n > 1 + \frac{\epsilon}{\sqrt{r_i(n)}}\right) &= P\left(\psi_{n,i}^{-1} < b_n\right) \\ &\leq P\left(\left|\hat{m}_{o,n,i} - m_{o,i}\right| > m_{o,i}(1-b_n)\right) \\ &\leq \frac{1}{m_{o,i}(1-b_n)} E\left|\hat{m}_{o,n,i} - m_{o,i}\right| \\ &\leq \frac{Cn\gamma^n}{m_{o,i}(1-b_n)}, \end{aligned}$$

for some $\gamma \in (0,1)$, using Lemma 12. But, using the ratio test, $\sum_{n\geq 1} \frac{n\gamma^n}{1-b_n} < \infty$. A similar argument gives, $\sum_{n\geq 1} a_{n,2} < \infty$.

Notice (3.10) follows immediately from (3.9) and the consistency of $\hat{m}_{o,n,i}$ (see Lemma 11). □

Lemma 14. *Let Assumptions 1 and 3. Then*

$$\sum_{n\geq 1} var\left(V_{i,j}(n) - W_{i,j}\right) < \infty \tag{3.11}$$

and as $n \to \infty$, for $\ell = 1, 2$,

$$\frac{1}{r_i(n)} \sum_{j=1}^{r_i(n)} \left(V_{i,j}(n) - W_{i,j}\right)^\ell \xrightarrow{a.s.} 0. \tag{3.12}$$

Proof. In what follows, it is helpful to define $V_{i,j} \equiv W_{i,j} \frac{m_{o,i}}{m_{o,i}-1}$ and $\theta \equiv \left(\frac{m_{o,i}-1}{m_{o,i}}\right)$. We begin by developing an estimate for $var\left(V_{i,j}(n) - W_{i,j}\right)$. To this end, note that

$$V_{i,j}(n) - W_{i,j} = \left(\frac{Y_{i,j}(n)}{m_{o,i}^n} - V_{i,j}\right)\left(\frac{m_{o,i}-1}{m_{o,i}}\right), \tag{3.13}$$

and

$$\begin{aligned} \frac{Y_{i,j}(n)}{m_{o,i}^n} - V_{i,j} &= \sum_{k=0}^{n-\tau} \left(W_{i,j}(n-k) - W_{i,j}\right) m_{o,i}^{-k} \quad - \quad W_{i,j} \sum_{k>n-\tau} m_{o,i}^{-k} \\ &\equiv J_{n,j}(1) - J_{n,j}(2). \end{aligned} \tag{3.14}$$

Now, using (3.13), (3.14), and the standard formula for $var(X-Y)$, yields

$$\begin{aligned} var\left(V_{i,j}(n) - W_{i,j}\right) &= \theta^2 \Big(var\left(J_{n,j}(1)\right) + var\left(J_{n,j}(2)\right) - 2Cov\left(J_{n,j}(1), J_{n,j}(2)\right)\Big) \\ &\equiv \theta^2 \left(a_{n,1} + a_{n,2} - 2\, a_{n,3}\right). \qquad (3.15) \end{aligned}$$

We proceed to prove (3.11), by showing that $\sum_{n\geq 1} a_{n,\ell} < \infty$, for $\ell = 1, 2, 3$.

First consider $a_{n,1}$. It is helpful to define $X_{n,k} \equiv \left(W_{i,j}(n-k) - W_{i,j}\right)$. Now,

$$var(J_{n,j}(1)) = var\left(\sum_{k=0}^{n-\tau} X_{n,k}\, m_{o,i}^{-k}\right) = J_{n,j}(1,1) + J_{n,j}(1,2) \qquad (3.16)$$

where

$$J_{n,j}(1,1) \equiv \sum_{k=0}^{n-\tau} var\left(X_{n,k}\, m_{o,i}^{-k}\right) = \sum_{k=0}^{n-\tau} m_{o,i}^{-2k}\, var\left(X_{n,k}\right)$$

and

$$J_{n,j}(1,2) \equiv \sum_{k=0}^{n-\tau} \sum_{\ell\neq k} Cov\left(X_{n,k}\, m_{o,i}^{-k}, X_{n,\ell}\, m_{o,i}^{-\ell}\right) = \sum_{k=0}^{n-\tau} \sum_{\ell\neq k} m_{o,i}^{-(k+\ell)}\, Cov\left(X_{n,k}, X_{n,\ell}\right).$$

But using Proposition 5,

$$J_{n,j}(1,1) = \sum_{k=0}^{n-\tau} m_{o,i}^{-2k}\, var\left(X_{n,k}\right) = C\, m_{o,i}^{-n} \sum_{k=0}^{n-\tau} m_{o,i}^{-k}, \qquad (3.17)$$

and

$$\left|Cov\left(X_{n,k}, X_{n,\ell}\right)\right| \leq C m_{o,i}^{-(n-(j+\ell)/2)}. \qquad (3.18)$$

Hence, using (3.18),

$$\begin{aligned} \left|J_{n,j}(1,2)\right| &= \sum_{k=0}^{n-\tau} \sum_{\ell\neq k} m_{o,i}^{-(k+\ell)} \left|Cov\left(X_{n,k}, X_{n,\ell}\right)\right| \\ &\leq C\, m_{o,i}^{-n} \cdot \sum_{k=0}^{n-\tau} \sum_{\ell\neq k} m_{o,i}^{-(j+\ell)/2}. \end{aligned}$$

Combining the above estimates yield

$$\begin{aligned} \left|var(J_{n,j}(1))\right| &\leq \left|J_{n,j}(1,1)\right| + \left|J_{n,j}(1,2)\right| \\ &\leq C_1\, m_{o,i}^{-n} \sum_{k=0}^{n-\tau} m_{o,i}^{-k} \quad + \quad C_2\, m_{o,i}^{-n} \cdot \sum_{k=0}^{n-\tau} \sum_{\ell\neq k} m_{o,i}^{-(j+\ell)/2}, \qquad (3.19) \end{aligned}$$

and the two sums in (3.19) are finite.

The calculation for $a_{n,2}$ is trivial; the calculation for $a_{n,3}$ follows using Proposition 5 and similar arguments given for $a_{n,1}$. Hence, (3.11) follows.

We now proceed to prove (3.12). First consider $\ell = 1$. Using similar argument which were given to prove (3.11) and the ratio test, we have

$$\sum_{n \geq 1} \frac{EX_n^2}{r_n^2} < \infty, \tag{3.20}$$

where

$$X_n \equiv \sum_{j=1}^{r_i(n)} \left(V_{i,j}(n) - W_{i,j}\right).$$

Fix $\epsilon_1 > 0$. Using Markov's inequality,

$$P\left(\left|\frac{1}{r_i(n)} \sum_{j=1}^{r_i(n)} \left(V_{i,j}(n) - W_{i,j}\right)\right| > \epsilon_1\right) = P\left(X_n^2 > \epsilon_1^2 r_n^2\right) \leq \frac{EX_n^2}{\epsilon_1^2 r_n^2}.$$

The result for $\ell = 1$ now follows using (3.20) and the Borel-Cantelli lemma. The result for $\ell = 2$ follows similarly. □

Define $m_{A,n,i} \equiv \frac{1}{r_i(n)} \sum_{j=1}^{r_i(n)} V_{i,j}(n)$, then as an immediate corollary of Lemma 14 we have.

Corollary 2. *Let Assumption 3 hold and assume $EZ_0^2 < \infty$ and $E\xi^2 < \infty$. Then as $n \to \infty$,*

$$m_{A,n,i} \xrightarrow{a.s.} m_{A,i}\,.$$

Lemma 15. *Let Assumptions 1 and 3. Then as $n \to \infty$,*

$$\frac{1}{r_i(n)} \sum_{j=1}^{r_i(n)} \left(V_{i,j}^2(n) - W_{i,j}^2\right) \xrightarrow{a.s.} 0.$$

Proof. To this end, fix $\epsilon > 0$ and note that

$$\begin{aligned} P\left(\left|\frac{1}{r_i(n)}\sum_{j=1}^{r_i(n)}\left(V_{i,j}^2(n) - W_{i,j}^2\right)\right| > \epsilon\right) &\leq P\left(\frac{1}{r_i(n)}\sum_{j=1}^{r_i(n)}\left|V_{i,j}^2(n) - W_{i,j}^2\right| > \epsilon\right) \\ &\leq \frac{1}{\epsilon}E\left|V_{i,j}^2(n) - W_{i,j}^2\right| \\ &\equiv a_n. \end{aligned}$$

But using arguments similar to those given in the proof of Lemma 15, yields $\sum_{n\geq 1} a_n < \infty$. Thus, the result follows from the Borel-Cantelli lemma. □

Lemma 16. *Let Assumptions 1 and 3 hold. For each i, as $n \to \infty$,*

$$\left(\bar{V}_{i\cdot}(n), \hat{\sigma}_{W,n,i}^2\right) \xrightarrow{P} \left(m_{A,i}, \sigma_{W,i}^2\right)$$

and

$$\sqrt{r_i(n)}\left(\bar{V}_{i\cdot}(n) - m_{A,i}\right) \xrightarrow{d} N(0, \sigma_{W,i}^2).$$

Proof. The consistency of $\bar{V}_{i\cdot}(n)$ follows immediately from Lemmas 13 and 14. In fact, asymptotic normality follows along the same lines because Lemma 13 shows that θ_n is $\sqrt{r_i(n)}$ consistent.

Next consider the consistency of $\hat{\sigma}_{W,n,i}^2$, which can be expressed as

$$\hat{\sigma}_{W,n,i}^2 = \frac{1}{r_i(n)}\sum_{j=1}^{r_i(n)}\left(T_{n,j,1} + T_{n,j,2} + T_{n,j,3}\right)^2, \tag{3.21}$$

where

$$T_{n,j,1} \equiv \hat{V}_{i,j}(n) - W_{i,j}, \quad T_{n,j,2} \equiv W_{i,j} - m_{A,i}, \quad T_{n,j,3} \equiv m_{A,i} - \bar{V}_{i\cdot}(n)\,.$$

The strong law of large numbers yields,

$$\frac{1}{r_i(n)}\sum_{j=1}^{r_i(n)} T_{n,j,2}^2 \xrightarrow{a.s.} \sigma_{W,i}^2\,.$$

Hence, we proceed to showing the remaining terms in the expansion of (3.21) converge to zero with probability one. Using Lemmas 13 and 14,

$$\frac{1}{r_i(n)} \sum_{j=1}^{r_i(n)} T_{n,j,1}^2 \xrightarrow{a.s.} 0.$$

Finally, we get

$$\frac{1}{r_i(n)} \sum_{j=1}^{r_i(n)} T_{n,j,3}^2 \xrightarrow{a.s.} 0,$$

immediately from the strong consistency of $\bar{V}_{i\cdot}(n)$. Finally, using the Cauchy-Schwarz inequality combined with similar arguments given above, the cross-product terms in the expansion (3.21) converge to zero with probability one. □

Lemma 17. *Let Assumptions 1 and 3 hold. Additionally, assume $m_{o,i} = m_o$. For each i, as $n \to \infty$,*

$$\left(\tilde{V}_{i\cdot}(n), \tilde{\sigma}_{W,n,i}^2\right) \xrightarrow{P} \left(m_{A,i}, \sigma_{W,i}^2\right)$$

and

$$\sqrt{r_i(n)}\left(\tilde{V}_{i\cdot}(n) - m_{A,i}\right) \xrightarrow{d} N(0, \sigma_{W,i}^2).$$

Proof. This follows from similar arguments used to prove Lemma 16. □

3.9 Proofs of main results

In this section we prove Theorems 5, 6, and 7.

3.9.1 Proofs For $m_{o,i}$ ANOVA

In this section we prove Theorem 5. We begin by proving two key lemmas which give the limit behavior for the MSE and MST terms.

Lemma 18. *Let Assumptions 1 and 3 hold. As $n \to \infty$,*

$$MSE_{o,n} \xrightarrow{P} \bar{\sigma}_o^2 .$$

Proof. This follows immediately from Lemma 11 and the fact that $\frac{r_i(n)}{r_j(n)} \to 1$ (see Assumption 3). □

Lemma 19. *Let Assumptions 1 and 3 hold. Additionally, assume $m_{o,i} = m_o$ for all i, then, as $n \to \infty$,*

$$\sum_{i=1}^{a} \frac{w_{n,i}}{\bar{\sigma}_o^2} \left(\hat{m}_{o,n,i} - \hat{m}_{o,n}\right)^2 \xrightarrow{d} \mathbf{T}^t\mathbf{T},$$

$\mathbf{T} \sim N_a(\mathbf{0}, \mathbf{L}_a\mathbf{\Sigma}_o\mathbf{L}_a^t)$*, where $\mathbf{\Sigma}_o$ and $\mathbf{L}_a$ are defined as in Theorem 5.*

Proof. Define, $\mathbf{X}_n \equiv (X_{n,1}, ..., X_{n,a})^t$ and $\mathbf{T}_n \equiv (T_{n,1}, ..., T_{n,a})^t$, where

$$X_{n,i} \equiv \frac{\sqrt{w_{n,i}}}{\bar{\sigma}_o} \left(\hat{m}_{o,n,i} - m_o\right),$$

and

$$T_{n,i} \equiv \frac{\sqrt{r_i(n)}}{\bar{\sigma}_o} \left(\hat{m}_{o,n,i} - \hat{m}_{o,n}\right).$$

Basic algebra, yields $\mathbf{T}_n = \mathbf{L}_{a,n}\mathbf{X}_n$, where

$$(\mathbf{L}_{a,n})_{i,j} = \begin{cases} \frac{w_n - w_{n,i}}{w_n} & \text{if } i = j, \\ -\frac{w_{nj}}{w_n}\sqrt{\frac{w_{ni}}{w_{nj}}} & \text{if } i \neq j. \end{cases}$$

As $n \to \infty$, using the fact that $\frac{r_i(n)}{r_j(n)} \to 1$, $\mathbf{L}_{a,n} \xrightarrow{P} \mathbf{L}_a$ and, using Lemma 16, $\mathbf{X}_n \xrightarrow{d} N(0, \mathbf{\Sigma}_o)$. Now, using Slutsky's theorem and the continuous mapping theorem,

$$\sum_{i=1}^{a} \frac{w_{n,i}}{\bar{\sigma}_o^2} \left(\hat{m}_{o,n,i} - \hat{m}_{o,n}\right)^2 = \mathbf{T}_n^t\mathbf{T}_n \xrightarrow{d} \mathbf{T}^t\mathbf{T}.$$

□

Proof of Theorem 6. First notice that

$$(a-1)\,F_{o,n} = \frac{T_n}{MSE_{o,n}\,/\,\bar{\sigma}_o^2},$$

where

$$T_n = \sum_{i=1}^{a} \frac{w_{n,i}}{\bar{\sigma}_o^2}\left(\hat{m}_{o,n,i} - \hat{m}_{o,n}\right)^2.$$

Thus, using Slutsky's theorem and Lemma 18, it is sufficient to prove $T_n \xrightarrow{d} \chi^2[\mathbf{L}_a \mathbf{\Sigma}_o \mathbf{L}_a^t]$. But this follows immediately from Lemma 19 and the following fact (see Marden (1995), p. 300 Lemma 12.6). Let $\mathbf{X} \sim N_p(\mathbf{0}, \mathbf{V})$ and let $\mathbf{A}$ be a symmetric $p \times p$ (possibly singular) matrix, then $\mathbf{X}^t\mathbf{A}\mathbf{X} \sim \chi^2[\mathbf{A}\mathbf{V}]$. □

Remark 3. *We finish the above proof by quoting Lemma 12.6 in Marden (1995). Using modified arguments we could instead have utilized results for quadratic forms of Gaussian random vectors with singular covariance matrices (Styan, 1970; Mathai and Provost, 1992).*

3.9.2 Proofs For $m_{A,i}$ ANOVA

In this section we prove Theorems 6 and 7. We again proceed by first proving limit results for the MSE and MST terms.

Lemma 20. *Let Assumptions 1 and 3 hold. As $n \to \infty$,*

$$MSE_{A,n} \xrightarrow{P} \bar{\sigma}_W^2.$$

Proof. This follows immediately from the fact that, for each i, $\hat{\sigma}_{W,n,i}^2 \xrightarrow{P} \sigma_{W,i}^2$ (see Lemma 16) and $\frac{r_i(n)}{r_j(n)} \to 1$ (see Assumption 3). □

Lemma 21. *Let Assumptions 1 and 3 hold. Additionally, assume $m_{A,i} = m_A$ for all i, then, as $n \to \infty$,*

$$\sum_{i=1}^{a} \frac{r_i(n)}{\bar{\sigma}_W^2} \left(\bar{V}_{i\cdot}(n) - \bar{V}_{\cdot\cdot}(n)\right)^2 \xrightarrow{d} \mathbf{T}^t\mathbf{T},$$

with $\mathbf{T} \sim N_a(\mathbf{0}, \mathbf{C}_a\mathbf{\Sigma}\mathbf{C}_a^t)$, where $\mathbf{\Sigma} \equiv \frac{1}{\bar{\sigma}_W^2} diag\left(\sigma_{W,1}^2, ..., \sigma_{W,a}^2\right)$.

Proof. Define, $\mathbf{X}_n \equiv (X_{n,1}, ..., X_{n,a})^t$ and $\mathbf{T}_n \equiv (T_{n,1}, ..., T_{n,a})^t$, where

$$X_{n,i} \equiv \frac{\sqrt{r_i(n)}}{\bar{\sigma}_W} \left(\bar{V}_{i\cdot}(n) - m_A\right),$$

and

$$T_{n,i} \equiv \frac{\sqrt{r_i(n)}}{\bar{\sigma}_W} \left(\bar{V}_{i\cdot}(n) - \bar{V}_{\cdot\cdot}(n)\right).$$

Basic algebra, yields $\mathbf{T}_n = \mathbf{C}_{a,n}\mathbf{X}_n$, where

$$(\mathbf{C}_{a,n})_{i,j} = \begin{cases} \frac{R(n)-r_i(n)}{R(n)} & \text{if } i = j, \\ -\frac{r_j(n)}{R(n)}\sqrt{\frac{r_i(n)}{r_j(n)}} & \text{if } i \neq j. \end{cases}$$

As $n \to \infty$, using the fact that $\frac{r_i(n)}{r_j(n)} \to 1$, $\mathbf{C}_{a,n} \to \mathbf{C}_a$ and, using Lemma 16, $\mathbf{X}_n \xrightarrow{d} N(0, \mathbf{\Sigma})$. Now, using Slutsky's theorem and the continuous mapping theorem,

$$\sum_{i=1}^{a} \frac{r_i(n)}{\bar{\sigma}_W^2} \left(\bar{V}_{i\cdot}(n) - \bar{V}_{\cdot\cdot}(n)\right)^2 = \mathbf{T}_n^t\mathbf{T}_n \xrightarrow{d} \mathbf{T}^t\mathbf{T}.$$

□

We now provide the proofs of the main results.

Proof of Theorem 6. This result follows immediately from Lemmas 20 and 21 and the argument given in the proof Theorem 5. □

Proof of Theorem 7. We omit the proof of Theorem 7. This proof follows using a standard modification of the arguments used to prove Theorem 6. Namely, one first proves analogs of Lemmas 20 and 21; then the result follows using Lemma 17 and similar arguments given in the proof for Theorem 5 □

CHAPTER 4

LIMIT THEOREMS FOR A SUPERCRITICAL BRANCHING PROCESS INITIATED BY A RANDOM NUMBER OF ANCESTORS

4.1 Introduction

We study supercritical branching processes initiated by a random number of ancestors. Given data from branching process replicates, each governed by the same offspring distribution and initiated by a random number of ancestors, we seek to estimate the moments of the ancestor and offspring distributions.

This problem is motivated by quantitative polymerase chain reaction (qPCR), an important and widely used tool for gene expression experiments (Ferré, 1998; Kubista et al., 2006; Nolan et al., 2006). The polymerase chain reaction (PCR) is a biochemical technique used to amplify the number of copies of a specific DNA fragment; the goal of qPCR is estimation of the initial number of molecules present in a genetic material. A typical PCR is run for 40 cycles; theoretically, the number of molecules doubles in every cycle. In practice, only some fraction of the molecules actually replicate in a given cycle. Hence, a supercritical Galton-Watson branching process with a Bernoulli offspring distribution provides a natural model to describe the dynamics of PCR. Early probabilistic results modeling PCR as a Galton-Watson branching process have been discussed in Krawczak et al. (1989), Reiss et al. (1990), Hayashi (1990), Maruyama (1990), and Sun (1995), among others. More recent works, such as Jagers and Klebaner (2003b), Lalam et al. (2004b), Piau (2002, 2004, 2005), have considered generalizations of the Galton-Watson process to model PCR.

Aside from the connection to PCR, the study of ancestral inference is important in the general context of inference for branching processes. Denote the mean and variance of the offspring distribution as (m_o, σ_o^2), similarly denote the mean and variance of the ancestor distribution as (m_A, σ_A^2) (our notation is made precise in Section 4.2.1 below). Inference for (m_o, σ_o^2) when $Z_0 \equiv 1$ (Harris, 1948; Nagaev, 1967; Heyde, 1974; Dion, 1975; Guttorp, 1991), $Z_0 \equiv r(n)$, $r(n) \to \infty$ as $n \to \infty$, (Duby and Rouault, 1982; Yanev, 1975, 1985), and when Z_0 is random (Dion and Yanev, 1994, 1995, 1997; Stoimenova, 2005) is well-studied in the literature. Dion and Yanev (1994) contains an excellent review of this work. However, relatively little is known concerning inference for m_A and σ_A^2, under general conditions on the offspring distribution. We fill this gap in the supercritical context by studying the inference for the parameter vector $\left(m_o, \sigma_o^2, m_A, \sigma_A^2\right)$.

Ancestral inference has not been studied in the subcritical or critical cases. In the subcritical case, Rubin and Vere-Jones (1968) and Hoppe (1980) discuss the impact of the ancestor distribution on the Yaglom conditional limits. But the work does not discuss estimation of the initial number of particles.

Because of the connection to PCR, recently there has been considerable work for estimating the initial number of particles in a supercritical Galton-Watson process (with a Bernoulli offspring distribution), including Jacob and Peccoud (1998), Lalam (2007), Lalam and Jacob (2007), and Piau (2008). Jacob and Peccoud (1998) consider a single realization of a modified branching process. In their model, there is an unobserved supercritical branching process that experiences binomial emigration. The observation process is the number of emigrants from each generation, as opposed to the actual population size. Both Lalam and Jacob (2007) and Piau (2008) consider a Bayesian solution to the problem based on observing a single realization of a branching process with a Bernoulli offspring distribution. Finally, Lalam (2007) studies a hidden Markov model formulation of the problem incorporating the fact that in qPCR experiments only the fluorescence is observed not the actual number of molecules in each cycle.

The above work is based on observing a single realization of the branching process. Also, with the exception of Jacob and Peccoud (1998), it is based on the Bernoulli offspring distribution. Our paper is based on observing branching process replicates and our results are true under more general conditions for the offspring and ancestor distribution.

Our work also highlights inference for the distribution of the martingale limit W (see Section 4.2.2 for more details). As is discussed below, inference for the ancestor distribution is achieved via inference for the limiting random variable W. Now, $EW = m_A$, additionally, let σ_W^2 denote the variance of W. We prove consistency and joint asymptotic normality (after appropriate scaling and centering) of an estimator for the vector $\left(m_o, \sigma_o^2, m_A, \sigma_W^2\right)$.

The remainder of this paper is organized as follows. Section 2 develops the notation and states the main results. The results are then proved in Sections 3 through 7.

4.2 Definitions and main results

4.2.1 Notation

Define two independent collections of random variables both distributed on the positive integers $\mathbb{N}$, the ancestor collection $\{Z_{0,i} : 1 \leq i \leq r(n)\}$ and the offspring collection $\{\xi_{n,i,k} : n \geq 1, 1 \leq i \leq r(n), k \geq 1\}$. $\{Z_{0,i} : 1 \leq i \leq r(n)\}$ is a collection of independent and identically distributed (i.i.d.) random variables with representative Z_0; unless otherwise stated we assume $EZ_0^2 < \infty$ and define $m_A \equiv EZ_0$ and $\sigma_A^2 \equiv var(Z_0)$. $\{\xi_{n,i,k} : n \geq 1, 1 \leq i \leq r(n), k \geq 1\}$ is a collection of i.i.d. random variables with representative ξ; again unless otherwise stated we assume $E\xi^2 < \infty$ and define $m_o \equiv E\xi$ and $\sigma_o^2 \equiv var(\xi)$. As is standard we define each collection of branching processes recursively as

$$Z_{n+1,i} = \sum_{k=1}^{Z_{n,i}} \xi_{n,i,k},$$

where $\xi_{n,j,k}$ is interpreted as the number of children produced by the k^{th} parent in the n^{th} generation of the i^{th} branching process. Summarizing, $\{Z_{n,i} : 1 \leq i \leq r(n)\}$ denotes a collection of i.i.d. branching processes initiated by a random number of ancestors $Z_{0,i}$. At times we will refer to a generic branching process $\{Z_n : n \geq 1\}$ which is initiated by Z_0 ancestors and whose offspring distribution is described by ξ. We emphasize that our assumption $P(\xi \in \mathbb{N}) = 1$, implies that we are exclusively studying supercritical branching processes which diverge to infinity with probability one.

The data for the problem are the generation sizes from the $(n-1)^{th}$ and n^{th} generations, namely $\{\left(Z_{n-1,i}, Z_{n,i}\right) : 1 \leq i \leq r(n)\}$; the vector notation $\mathbf{Z_n} = \left(Z_{n,1}, ..., Z_{n,r(n)}\right)$ denotes all of the n^{th} generation data.

4.2.2 Motivating the estimators

It is well-known (Athreya and Ney, 1972) that there is a limiting random variable W obtained by scaling a supercritical branching process by its mean. Information concerning the ancestor distribution is contained in W, and hence W is a critical object of study. To be specific, we define, for $1 \leq i \leq n$, $W_{n,i} \equiv \frac{Z_{n,i}}{m_o^n}$. For i, $W_{n,i}$ is a positive martingale and thus there exists a random variable $W_{(i)}$ for which $W_{n,i} \xrightarrow{a.s.} W_{(i)}$. In fact, because $W_{n,i}$ are i.i.d. (across i), the limits $W_{(i)}$ are i.i.d. Fortunately, results exist which connect the moments of W_1 to the moments of ξ (and Z_0). From the Kesten-Stigum Theorem (Kesten and Stigum, 1966), $EW_1 < \infty$ if and only if $E\xi \log \xi < \infty$ (and $EZ_0 < \infty$). Athreya (1971) extended this result to show that for $k > 1$, $EW_1^k < \infty$ if and only if $E\xi^k < \infty$ (and $EZ_0^k < \infty$). Using the L_p Martingale Convergence Theorem, it is easily shown that the the moments of Z_0 are functions of the moments of W (and the offspring parameters), namely

$$m_A = EW, \qquad \sigma_A^2 = \sigma_W^2 - \frac{m_A \, \sigma_o^2}{m_o(m_o - 1)}, \tag{4.1}$$

where $\sigma_W^2 \equiv var(W)$. Hence we proceed by estimating $m_A, \sigma_W^2, m_o, \sigma_o^2$. As is standard for supercritical branching processes (Guttorp, 1991) we employ a weighted least squares estimator for m_o; we use (asymptotically) unbiased estimating equations for $m_A, \sigma_W^2, \sigma_o^2$. Define $\tilde{m}_{A,n} \equiv \frac{1}{r_n} \sum_{i=1}^{r_n} W_{n,i}$, the resulting estimating equations are

$$\begin{aligned}
\frac{1}{r_n}\sum_{i=1}^{r_n}\left(W_{n,i}-m_A\right) &= 0\\
\frac{1}{r_n}\sum_{i=1}^{r_n}\left[\left(W_{n,i}-\tilde{m}_{A,n}\right)^2-\sigma_W^2\right] &= 0\\
\sum_{i=1}^{r_n}\sqrt{Z_{n-1,i}}\left(\frac{Z_{n,i}}{\sqrt{Z_{n-1,i}}}-m_o\sqrt{Z_{n-1,i}}\right) &= 0\\
\frac{1}{r_n}\sum_{i=1}^{r_n}\left[Z_{n-1,i}\left(\frac{Z_{n,i}}{Z_{n-1,i}}-m_o\right)^2-\sigma_o^2\right] &= 0.
\end{aligned}$$

Solving these estimating equations yield,

$$\hat{m}_{o,n}\equiv\frac{\sum_{i=1}^{r_n}Z_{n,i}}{\sum_{i=1}^{r_n}Z_{n-1,i}},$$

$$\hat{\sigma}_{o,n}^2\equiv\frac{1}{r_n}\sum_{i=1}^{r_n}Z_{n-1,i}\cdot\left(\frac{Z_{n,i}}{Z_{n-1,i}}-\hat{m}_{o,n}\right)^2,$$

$$\hat{m}_{A,n}\equiv\frac{1}{r_n}\sum_{i=1}^{r_n}\hat{W}_{n,i},$$

and

$$\hat{\sigma}_{W,n}^2=\frac{1}{r_n}\sum_{i=1}^{r_n}\left(\hat{W}_{n,i}-\hat{m}_{A,n}\right)^2,$$

where

$$\hat{W}_{n,i}\equiv\frac{Z_{n,i}}{\hat{m}_{o,n}{}^{n}}.$$

Finally, based on (4.1), define the estimator of the ancestor variance as

$$\hat{\sigma}_{A,n}^2=\hat{\sigma}_{W,n}^2-\frac{\hat{m}_{A,n}\,\hat{\sigma}_{o,n}^2}{\hat{m}_{o,n}\left(\hat{m}_{o,n}-1\right)}.$$

4.2.3 Consistency and asymptotic normality results

This section gives the consistency and asymptotic normality of the estimators defined above. The results require that we control the rate at which r_n goes to ∞, basically requiring that r_n does not increase exponentially. We make this explicit with the following assumption.

Assumption 3. *$r_n \to \infty$ as $n \to \infty$, such that $\frac{r_n}{r_{n-1}} \to 1$.*

First we state our consistency result.

Theorem 8. *Let Assumption 3 hold and assume $EZ_0^2 < \infty$ and $E\xi^2 < \infty$. Then as $n \to \infty$*

$$(\hat{m}_{o,n}, \hat{m}_{A,n}) \xrightarrow{a.s.} (m_o, m_A), \qquad (\hat{\sigma}^2_{o,n}, \hat{\sigma}^2_{A,n}) \xrightarrow{P} (\sigma^2_o, \sigma^2_A).$$

The next result gives the joint asymptotic normality.

Theorem 9. *Let Assumption 3 hold and $EZ_0^4 < \infty$ and $E\xi^4 < \infty$. Define*

$$\begin{aligned} T_{n,1} &= \sqrt{r_n}\left(\hat{m}_{A,n} - m_A\right), \\ T_{n,2} &= \sqrt{r_n}\left(\hat{\sigma}^2_{W,n} - \sigma^2_W\right), \\ T_{n,3} &= \sqrt{\sum_{i=1}^{r_n} Z_{n-1,i}}\left(\hat{m}_{o,n} - m_o\right), \\ T_{n,4} &= \sqrt{r_n}\left(\hat{\sigma}^2_{o,n} - \sigma^2_o\right), \end{aligned}$$

and $\mathbf{T}_n^t = \left(T_{n,1}, T_{n,2}, T_{n,3}, T_{n,4}\right)$. As $n \to \infty$, $\mathbf{T}_n \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma})$, where

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma^2_W & \mu_{W,3} & 0 & 0 \\ \mu_{W,3} & \mu_{W,4} - \sigma^4_W & 0 & 0 \\ 0 & 0 & \sigma^2_o & 0 \\ 0 & 0 & 0 & 2\sigma^4_o \end{pmatrix}$$

Remark 4. *Under a two moment hypothesis, $EZ_0^2 < \infty$ and $E\xi^2 < \infty$, we obtain the joint asymptotic normality of the centered and scaled means $\hat{m}_{o,n}$ and $\hat{m}_{A,n}$.*

The asymptotic normality of $\hat{\sigma}_{A,n}^2$ follows from the above theorem.

Corollary 3. *Assume the conditions of Theorem 9. Then,*

$$\sqrt{r_n}\left(\hat{\sigma}_{A,n}^2 - \sigma_A^2\right) \xrightarrow{d} N(0, v),$$

where

$$v = \left(\frac{\sigma_o^2}{m_o\left(m_o - 1\right)}\right)^2 \sigma_W^2 + \mu_{W,4} - \sigma_W^4 - 2\left(\frac{\sigma_o^2}{m_o\left(m_o - 1\right)}\right)\mu_{W,3}\,.$$

4.3 Initial estimates

In this section we provide results which are needed in the proofs below. The following result is standard (Athreya and Ney, 1972); we quote it here for convenience.

Proposition 7. *Let $EZ_0^2 < \infty$ and $E\xi^2 < \infty$. Then*

$$Var\left(W_{n,1} - W_{(1)}\right) = E\left(W_{n,1} - W_{(1)}\right)^2 = \frac{C}{m_o^n}.$$

The following harmonic moment result is used to prove the other two results in this section.

Lemma 22. *Let $EZ_0^2 < \infty$ and $E\xi^2 < \infty$. Then, for some $\gamma \in (0,1)$ and some $C \in (0,\infty)$,*

$$E\left(\frac{1}{\sum_{i=1}^{r_n} Z_{n,i}}\right) \le E\left(\frac{1}{Z_{n,1}}\right) \le C\gamma^n.$$

Proof. $\sum_{i=1}^{r_n} Z_{n,i} > Z_{n,1}$ a.s. so that the first inequality is trivial. Also $EZ_{n,1} \geq E\left(Z_{n,1} \middle| Z_{0,1} = 1\right)$. But Theorem 1 in Ney and Vidyashankar (2003) gives, for some $\gamma \in (0,1)$ and some $C \in (0,\infty)$, $E\left(Z_{n,1} \middle| Z_{0,1} = 1\right) \le C\gamma^n$. □

We consider the deviation of the mean estimators $\frac{Z_{n,i}}{Z_{n-1,i}}$ from m_o, namely define

$$M_n^\star \equiv \max_{1\leq i\leq r_n} \left| \frac{Z_{n,i}}{Z_{n-1,i}} - m_o \right| .$$

Lemma 23. *Let Assumption 3 hold. Let $EZ_0^2 < \infty$ and $E\xi^2 < \infty$. Then, with probability one,*

$$\lim_{n\to\infty} M_n^\star = 0.$$

Proof. Fix $\epsilon > 0$. Let $\alpha(n) = P\left(\left|\frac{Z_{n+1,1}}{Z_{n,1}} - m_o\right| > \epsilon\right)$. Then, using the fact $a^n - b^n = (a-b)\sum_{k=0}^{n-1} a^k b^{(n-1)-k}$,

$$\begin{aligned}
\sum_{n\geq 1} P\left(M_{n+1}^\star > \epsilon\right) &= \sum_{n\geq 1} 1 - \left[1-\alpha(n)\right]^{r_n} \\
&= \sum_{n\geq 1} \alpha(n) \sum_{k=0}^{r_n-1} \left(1-\alpha(n)\right)^k \\
&\leq \sum_{n\geq 1} C r_n \gamma^n,
\end{aligned}$$

for some $\gamma \in (0,1)$, using Lemma 22. But, using the ratio test, $\sum_{n\geq 1} C r_n \gamma^n < \infty$, hence the desired result follows from the Borel-Cantelli lemma. □

It is also important to consider the ratio of the offspring mean to the estimator of the offspring mean, raised to the n^{th} power. Specifically, we consider the behavior of θ_n^n, where $\theta_n \equiv \left(\frac{m_o}{\hat{m}_{o,n}}\right)$.

Lemma 24. *Let Assumption 3 hold. Let $EZ_0^2 < \infty$ and $E\xi^2 < \infty$. Then as $n \to \infty$,*

$$\sqrt{r_n}\,|\theta_n^n - 1| \xrightarrow{a.s.} 0$$

Proof. Fix $\epsilon > 0$. Notice that

$$\begin{aligned}
P\left(\sqrt{r_n}\,|\theta_n^n - 1| > \epsilon\right) &= P\left(\theta_n^n > 1 + \frac{\epsilon}{\sqrt{r_n}}\right) + P\left(\theta_n^n < 1 - \frac{\epsilon}{\sqrt{r_n}}\right) \\
&\equiv a_{n,1} + a_{n,2}\,.
\end{aligned}$$

From the Borel-Cantelli lemma it is sufficient to prove $\sum_{n\geq 1} a_{n,i} < \infty$. To this end, define $b_n = (1+\frac{\epsilon}{\sqrt{r_n}})^{-1/n}$ and use Chebychev's inequality to obtain

$$\begin{aligned} P\left(\theta_n^n > 1+\frac{\epsilon}{\sqrt{r_n}}\right) &= P\left(\theta_n^{-1} < b_n\right) \\ &\leq P\left(\left|\hat{m}_{o,n} - m_o\right| > m_o(1-b_n)\right) \\ &\leq \frac{var(\hat{m}_{o,n})}{(m_o(1-b_n))^2} \\ &\leq \frac{C\gamma^n}{(m_o(1-b_n))^2}, \end{aligned}$$

for some $\gamma \in (0,1)$, using the fact that $var\left(\hat{m}_{o,n}\right) = CE\left(\sum_{i=1}^{r_n} Z_{n,i}\right)^{-1}$ and Lemma 22. But, using the ratio test, $\sum_{n\geq 1}\frac{\gamma^n}{(1-b_n)^2} < \infty$. $\sum_{n\geq 1} a_{n,2} < \infty$ follows similarly. □

We use the following (minor) generalization of the central limit theorem for i.i.d. random vectors in our asymptotic normality proofs.

Lemma 25. *Let $\{\mathbf{X}_{n,i}\}$ be a collection random vectors with $\mathbf{X}_{n,i} \in \mathbb{R}^p$ and $E\mathbf{X}_{n,i} = \mathbf{0}$. Additionally, for each fixed n, $\left\{\mathbf{X}_{n,i} : i \geq 1\right\}$ are i.i.d. with finite covariance $cov\left(\mathbf{X}_{n,i}\right) = \mathbf{\Sigma}_n$. If $\mathbf{\Sigma}_n \to \mathbf{\Sigma}$ and $k_n \to \infty$ as $n \to \infty$, then*

$$\frac{1}{\sqrt{k_n}}\sum_{i=1}^{k_n}\mathbf{X}_{n,i} \xrightarrow{d} N_p(\mathbf{0},\mathbf{\Sigma}).$$

Proof. Fix $\lambda \in \mathbb{R}^p$. By the Cramér-Wold device, it is sufficient to prove

$$\lambda^t\left(\frac{1}{\sqrt{k_n}}\sum_{i=1}^{k_n}\mathbf{X}_{n,i}\right) \xrightarrow{d} N(0,\lambda^t\mathbf{\Sigma}\lambda).$$

Let $X_{n,i,j}$ be the j^{th} component of $\mathbf{X}_{n,i}$, then

$$\lambda^t\left(\frac{1}{\sqrt{k_n}}\sum_{i=1}^{k_n}\mathbf{X}_{n,i}\right) = \frac{1}{\sqrt{k_n}}\sum_{i=1}^{k_n} T_{n,i},$$

where $T_{n,i} \equiv \sum_{j=1}^{p} \lambda_j X_{n,i,j}$. Hence, for each fixed n, $T_{n,i}$ are i.i.d. random variables with $E(T_{n,1}) = 0$ and $var(T_{n,1}) = \lambda^t \mathbf{\Sigma_n} \lambda$.

Let ϕ_{T_n} be the characteristic function of $\frac{1}{\sqrt{k_n}} \sum_{i=1}^{k_n} T_{n,i}$ and ϕ_n be the characteristic function of $T_{n,1}$. Expanding ϕ_n in terms of its first two moments yields yields (Chow and Teicher (1997) p. 295)

$$\phi_{T_n}(t) = \left(\phi_n\left(t/\sqrt{k_n}\right)\right)^{k_n} = \left(1 + \frac{-t^2 E T_{n,1}^2}{2k_n} + o\left(k_n^{-1}\right)\right)^{k_n} \rightarrow \exp\left(\frac{-t^2 \lambda^t \mathbf{\Sigma} \lambda}{2}\right),$$

because $ET_{n,1} = 0$ and $ET_{n,1}^2 = \lambda^t \mathbf{\Sigma_n} \lambda \rightarrow \lambda^t \mathbf{\Sigma} \lambda$. □

4.4 Moment estimation for the martingale limit

We present results concerning the estimation of the first two moments of W. These results are of interest in themselves and are also needed for proving the consistency and asymptotic normality results stated in Section 4.2.3. First consider the quantities based on the i.i.d. martingale limits $W_{(i)}$. Define the sample average and variance as

$$m_{A,n} \equiv \frac{1}{r_n} \sum_{i=1}^{r_n} W_{(i)}, \qquad \sigma_{W,n}^2 \equiv \frac{1}{r_n} \sum_{i=1}^{r_n} \left(W_{(i)} - m_{A,n}\right)^2.$$

The law of large numbers and central limit theorem for i.i.d. data yield (under the correct moment assumptions for Z_0 and ξ) consistency and asymptotic normality of $m_{A,n}$ and $\sigma_{W,n}^2$. Next consider the analogous quantities based on $W_{n,i}$,

$$\tilde{m}_{A,n} \equiv \frac{1}{r_n} \sum_{i=1}^{r_n} W_{n,i}, \qquad \tilde{\sigma}_{W,n}^2 \equiv \frac{1}{r_n} \sum_{i=1}^{r_n} \left(W_{n,i} - \tilde{m}_{A,n}\right)^2.$$

The next lemmas gives consistency results for these moment estimators.

Lemma 26. *Let $EZ_0^2 < \infty$ and $E\xi^2 < \infty$. As $n \rightarrow \infty$,*

$$\left(\tilde{m}_{A,n}, \frac{1}{r_n} \sum_{i=1}^{r_n} W_{n,i}^2, \tilde{\sigma}_{W,n}^2\right) \xrightarrow{a.s.} \left(m_A, EW^2, \sigma_W^2\right).$$

Proof. We begin by proving $\tilde{m}_{A,n} \xrightarrow{a.s.} m_A$. First note that,

$$\tilde{m}_{A,n} = \frac{1}{r_n} \sum_{i=1}^{r_n} \left(W_{n,i} - W_{(i)} \right) + m_{A,n} \,.$$

As discussed above, the law of large numbers yields $m_{A,n} \xrightarrow{a.s.} m_A$. Thus, to complete the result it is sufficient to show that as $n \to \infty$,

$$\frac{1}{r_n} \sum_{i=1}^{r_n} \left(W_{n,i} - W_{(i)} \right) \xrightarrow{a.s.} 0.$$

We have

$$\begin{aligned} P\left(\frac{1}{r_n} \left| \sum_{i=1}^{r_n} W_{n,i} - W_{(i)} \right| > \epsilon \right) &\leq \frac{var\left(\sum_{i=1}^{r_n} W_{n,i} - W_{(i)} \right)}{r_n^2 \epsilon^2} \\ &= \frac{var\left(W_{n,i} - W_{(i)} \right)}{r_n \epsilon^2} = \frac{C}{r_n \, m_o^n \, \epsilon^2}. \end{aligned}$$

The result follows from Borel-Cantelli.

We prove the result for $\frac{1}{r_n} \sum_{i=1}^{r_n} W_{n,i}^2$ using the same basic argument. By the law of large numbers, $\frac{1}{r_n} \sum_{i=1}^{r_n} W_{(i)}^2 \xrightarrow{a.s.} EW^2$, so it is sufficient to prove

$$\frac{1}{r_n} \sum_{i=1}^{r_n} \left(W_{n,i}^2 - W_{(i)}^2 \right) \xrightarrow{a.s.} 0.$$

To this end, fix $\epsilon > 0$ and note that

$$\begin{aligned} P\left(\left| \frac{1}{r_n} \sum_{i=1}^{r_n} \left(W_{n,i}^2 - W_{(i)}^2 \right) \right| > \epsilon \right) &\leq P\left(\frac{1}{r_n} \sum_{i=1}^{r_n} \left| W_{n,i}^2 - W_{(i)}^2 \right| > \epsilon \right) \\ &\leq \frac{1}{\epsilon} E \left| W_{n,1}^2 - W_1^2 \right| \\ &\leq C_\epsilon \left(\frac{1}{m_o^n} + \frac{1}{m_o^{n/2}} \right). \end{aligned}$$

Thus, the result follows from Borel-Cantelli.

Finally, note that $\tilde{\sigma}_{W,n}^2 \xrightarrow{a.s} \sigma_W^2$ follows immediately from the results for $\tilde{m}_{A,n}$ and $\frac{1}{r_n} \sum_{i=1}^{r_n} W_{n,i}^2$. □

Now we are ready to prove the consistency of the estimators.

Lemma 27. *Let Assumption 3 hold. Let $EZ_0^2 < \infty$ and $E\xi^2 < \infty$. As $n \to \infty$,*

$$\hat{m}_{A,n} \xrightarrow{a.s.} m_A \, .$$

Proof. Notice $\hat{m}_{A,n} = \frac{1}{r_n} \sum_{i=1}^{r_n} W_{n,i} \, \theta_n^n = \theta_n^n \cdot \sum_{i=1}^{r_n} W_{n,i}$. Thus, consistency follows from Lemmas 24 and 26. □

Lemma 28. *Let Assumption 3 hold and $EZ_0^2 < \infty$ and $E\xi^2 < \infty$. As $n \to \infty$,*

$$\hat{\sigma}^2_{W,n} \xrightarrow{a.s.} \sigma^2_W \, .$$

Proof. Basic algebra yields,

$$\begin{aligned} \hat{\sigma}^2_{W,n} &= \frac{1}{r_n}\sum_{i=1}^{r_n} \left(W_{n,i} - \tilde{m}_{A,n}\right)^2 \quad + \quad \frac{1}{r_n}\sum_{i=1}^{r_n} \left(\hat{W}_{n,i} - W_{n,i}\right)^2 \\ &+ \frac{2}{r_n}\sum_{i=1}^{r_n} \left(\hat{W}_{n,i} - W_{n,i}\right)\left(W_{n,i} - \tilde{m}_{A,n}\right) - \left(\tilde{m}_{A,n} - \hat{m}_{A,n}\right)^2 \\ &\equiv \tilde{\sigma}^2_{W,n} + J_{n,1} + J_{n,2} - J_{n,3} \, . \end{aligned}$$

From Lemma 26, $\tilde{\sigma}^2_{W,n} \xrightarrow{a.s.} \sigma^2_W$; combining Lemmas 26 and 27 $J_{n,3} \xrightarrow{a.s.} 0$. Thus, to complete the proof, we show $J_{n,i} \xrightarrow{a.s.} 0$, $i = 1, 2$. We begin with $J_{n,1}$,

$$J_{n,1} \le \left(\theta_n^n - 1\right)^2 \frac{1}{r_n}\sum_{i=1}^{r_n} W_{n,i}^2 \xrightarrow{a.s.} 0,$$

using Lemmas 24 and 26.

Finally, consider $J_{n,2}$. Applying Cauchy-Schwartz yields,

$$\frac{1}{2}\left|J_{n,2}\right| \le \left|\theta_n^n - 1\right| \left(\frac{1}{r_n}\sum_{i=1}^{r_n} W_{n,i}^2\right)^{1/2} \left(\frac{1}{r_n}\sum_{i=1}^{r_n} \left(W_{n,i} - \tilde{m}_{A,n}\right)^2\right)^{1/2} \xrightarrow{a.s.} 0,$$

using Lemmas 24 and 26. □

4.5 Proof of Theorem 8

We proceed to prove consistency for the offspring moment estimators.

In the following lemma we prove a consistency and asymptotic normality result for $\hat{m}_{o,n}$. The asymptotic normality result is used to prove the consistency of $\hat{\sigma}^2_{o,n}$.

Lemma 29. *Let Assumption 3 hold. Additionally, assume $EZ_0^2 < \infty$ and $E\xi^2 < \infty$. Then as $n \to \infty$,*

$$\hat{m}_{o,n} \xrightarrow{a.s.} m_o$$

and

$$\sqrt{\sum_{i=1}^{r_n} Z_{n-1,i}}\,\left(\hat{m}_{o,n} - m_o\right) \xrightarrow{d} G,$$

where $G \sim N(0, \sigma_o^2)$.

Proof. We begin by proving consistency. We have that

$$\begin{aligned} 0 \;&\leq\; \left|\hat{m}_{o,n+1} - m_o\right| = \left|\sum_{i=1}^{r_n} \frac{Z_{n,i}}{\sum_{j=1}^{r_n} Z_{n,j}} \left(\frac{Z_{n+1,i}}{Z_{n,i}} - m\right)\right| \\ &\leq\; M^{\star}_{n+1}\,. \end{aligned}$$

The result now follows from Lemma 23.

Next consider the asymptotic normality. Define

$$Y_{n,i} \equiv \sqrt{W_{n,i}}\sqrt{Z_{n,i}}\left(\frac{Z_{n+1,i}}{Z_{n,i}} - m_o\right).$$

Notice that for each fixed n, $Y_{n,1}, Y_{n,2}, \ldots$ are i.i.d. with $EY_{n,1} = 0$ and $EY^2_{n,1} = \sigma_o^2\,EW$. Thus,

$$\sqrt{\sum_{i=1}^{r(n+1)} Z_{n,i}}\,\left(\hat{m}_{o,n+1} - m_o\right) = \sqrt{\frac{1}{\tilde{m}_{A,n}}}\left[\frac{1}{\sqrt{r(n+1)}}\sum_{i=1}^{r(n+1)} Y_{n,i}\right] \xrightarrow{d} N(0, \sigma_o^2),$$

using Lemmas 25 and 26 and Slutsky's Theorem. □

Lemma 30. *Let Assumption 3 hold. Additionally, assume $EZ_0^2 < \infty$ and $E\xi^2 < \infty$. Then as $n \to \infty$,*

$$\hat{\sigma}_{o,n}^2 \xrightarrow{P} \sigma_o^2 \, .$$

Proof. Begin by considering the consistency result. Algebra yields,

$$\begin{aligned}
\hat{\sigma}_{o,n+1}^2 - \sigma_o^2 &= \frac{1}{r_n}\sum_{i=1}^{r_n}\left[Z_{n,i}\left(\frac{Z_{n+1,i}}{Z_{n,i}} - m_o\right)^2 - \sigma_o^2\right] + (\hat{m}_{o,n+1} - m_o)^2\,\frac{1}{r_n}\sum_{i=1}^{r_n} Z_{n,i} \\
&- 2(\hat{m}_{o,n+1} - m_o)\,\frac{1}{r_n}\sum_{i=1}^{r_n} Z_{n,i}\left(\frac{Z_{n+1,i}}{Z_{n,i}} - m_o\right) \\
&\equiv J_{n,1} + J_{n,2} + J_{n,3} \, .
\end{aligned}$$

Lemma 29 (combined with similar arguments used in the proof of the lemma) imply $J_{n,2} \xrightarrow{P} 0$ and $J_{n,3} \xrightarrow{P} 0$.

Finally we prove $J_{n,1} \xrightarrow{P} 0$ by showing its characteristic function converges to unity. Notice that $J_{n,1} = \frac{1}{r_n}\sum_{i=1}^{r_n} Y_{n,i}$, where

$$Y_{n,i} \equiv \left[Z_{n,i}\left(\frac{Z_{n+1,i}}{Z_{n,i}} - m_o\right)^2 - \sigma_o^2\right] .$$

Notice that for each fixed n, $Y_{n,1}, Y_{n,2}, ...$ are i.i.d. with $EY_{n,1} = 0$. Let ϕ_n be the characteristic function of $Y_{n,1}$, and ϕ_{J_n} be the characteristic function of $J_{n,1}$. Expanding ϕ_n in terms of $EY_{n,1}$ (Chow and Teicher (1997) p. 295) yields

$$\phi_{J_n}(t) = \left(\phi_n\left(\frac{t}{r_n}\right)\right)^{r_n} = \left(1 + o\left(r_n^{-1}\right)\right)^{r_n} \to 1.$$

□

Proof of Theorem 8. The result follows from Lemmas 27, 28, 29, 30. □

4.6 Proof of Theorem 9

We now consider the joint asymptotic normality results. To ease notation we define the properly centered and scaled quantities that we will consider, namely,

$$\begin{aligned} X_{n,1} &\equiv \sqrt{r_n}\left(\hat{m}_{A,n} - m_A\right) \\ X_{n,2} &\equiv \sqrt{r(n)}\left(\hat{\sigma}^2_{W,n} - \sigma^2_W\right) \\ X_{n,3} &\equiv \sqrt{\sum_{j=1}^{r_n} Z_{n-1,j}} \cdot \left(\hat{m}_{o,n} - m_o\right) \\ X_{n,4} &\equiv \sqrt{r(n)}\left(\hat{\sigma}^2_{o,n} - \sigma^2_o\right). \end{aligned}$$

The necessary calculations are greatly simplified if the ancestor estimators are expressed in terms of data from generations $(n-1, n)$, while the offspring estimators are expressed in terms of data from generations $(n, n+1)$. The following lemma facilitates this idea.

Lemma 31. *Assume the conditions of Theorem 9. Then as n $\to \infty$,*

$$\begin{aligned} X_{n+1,1} - X_{n,1} &\xrightarrow{P} 0 \\ X_{n+1,2} - X_{n,2} &\xrightarrow{P} 0. \end{aligned}$$

Proof. First consider $X_{n,1}$. We begin by proving,

$$X_{n,1} = \frac{1}{\sqrt{r_n}} \sum_{i=1}^{r_n} \left(W_{(i)} - m_A\right) + o_p(1). \tag{4.2}$$

To this end note that

$$\begin{aligned} \sqrt{r_n}\left(\hat{m}_{A,n} - m_A\right) &= \frac{1}{\sqrt{r_n}} \sum_{i=1}^{r_n} \left(W_{(i)} - m_A\right) + \frac{1}{\sqrt{r_n}} \sum_{i=1}^{r_n} \left(W_{n,i} - W_{(i)}\right) \theta_n^n \\ &+ \frac{1}{\sqrt{r_n}} \sum_{i=1}^{r_n} W_{(i)} \left(\theta_n^n - 1\right) \\ &= \frac{1}{\sqrt{r_n}} \sum_{i=1}^{r_n} \left(W_{(i)} - m_A\right) + T_{n,2} + T_{n,3}\,. \end{aligned}$$

We proceed to show $T_{n,i} \xrightarrow{P} 0$, $i = 2, 3$. Consider $T_{n,2}$,

$$\left|T_{n,2}\right| \leq \theta_n^n \frac{1}{\sqrt{r_n}} \sum_{i=1}^{r_n} \left|W_{n,i} - W_{(i)}\right|.$$

From Lemma 24, $\theta_n^n \xrightarrow{a.s.} 1$ and

$$\begin{aligned} P\left(\sum_{i=1}^{r_n} \left|W_{n,i} - W_{(i)}\right| > \epsilon\sqrt{r_n}\right) &\leq P\left(\left|W_{n,1} - W_1\right| > \frac{\epsilon}{\sqrt{r_n}}\right) \\ &\leq \frac{r_n}{\epsilon^2} Var\left(W_{n,1} - W_1\right). \end{aligned}$$

Applying Proposition 7 gives $\frac{1}{\sqrt{r_n}} \sum_{i=1}^{r_n} \left|W_{n,i} - W_{(i)}\right| \xrightarrow{P} 0$. Hence, $T_{n,2} \xrightarrow{P} 0$.

Now consider $T_{n,3}$. Applying the law of large numbers and Lemma 24 yields

$$\left|T_{n,3}\right| \leq \sqrt{r_n}\left|\theta_n^n - 1\right| \cdot \frac{1}{r_n} \sum_{i=1}^{r_n} W_{(i)} \xrightarrow{a.s.} 0.$$

Using (4.2), to prove $X_{n+1,1} - X_{n,1} \xrightarrow{P} 0$, it is sufficient to show that

$$D_n \equiv \frac{1}{\sqrt{r(n+1)}} \sum_{i=1}^{r(n+1)} \left(W_{(i)} - m_A\right) - \frac{1}{\sqrt{r_n}} \sum_{i=1}^{r_n} \left(W_{(i)} - m_A\right) \xrightarrow{P} 0.$$

To this end note that

$$\begin{aligned} D_n &= \left(\frac{1}{\sqrt{r(n+1)}} - \frac{1}{\sqrt{r(n)}}\right) \sum_{i=1}^{r_n} \left(W_{(i)} - m_A\right) \\ &+ \frac{1}{\sqrt{r(n+1)}} \sum_{i=r(n)+1}^{r(n+1)} \left(W_{(i)} - m_A\right). \end{aligned}$$

It is straightforward to show that these two sums converge in probability to zero.

Using the same argument given above, to prove $X_{n+1,2} - X_{n,2} \xrightarrow{P} 0$, it is sufficient to prove

$$X_{n,2} = \frac{1}{\sqrt{r_n}} \sum_{i=1}^{r_n} \left[\left(W_{(i)} - m_A\right)^2 - \sigma_W^2\right] + o_p(1).$$

To this end note that

$$\sqrt{r_n}\left(\hat{\sigma}^2_{W,n} - \sigma^2_W\right) = \sqrt{r_n}\left(\tilde{\sigma}^2_{W,n} - \sigma^2_W\right) + J_{n,1} + J_{n,2} + J_{n,3},$$

where

$$\begin{aligned} J_{n,1} &\equiv \sqrt{r_n}\left(\tilde{m}_{A,n} - \hat{m}_{A,n}\right)^2, \\ J_{n,2} &\equiv \frac{1}{\sqrt{r_n}} \sum_{i=1}^{r_n} \left(\hat{W}_{n,i} - W_{n,i}\right)^2, \\ J_{n,3} &\equiv \frac{2}{\sqrt{r_n}} \sum_{i=1}^{r_n} \left(\hat{W}_{n,i} - W_{n,i}\right)\left(W_{n,i} - \tilde{m}_{A,n}\right). \end{aligned}$$

We proceed to show that $J_{n,i} \xrightarrow{P} 0$, $i = 1, 2, 3$. $J_{n,1} \xrightarrow{P} 0$ follows easily from Lemmas 24 and 26.

Consider $J_{n,2}$.

$$J_{n,2} \leq \sqrt{r_n}\left(\theta^n_n - 1\right)^2 \frac{1}{r_n} \sum_{i=1}^{r_n} W^2_{n,i} \xrightarrow{P} 0,$$

using Lemmas 24 and 26.

Finally, $J_{n,3} \xrightarrow{P} 0$ using the same arguments as above.

Finally, note that

$$\begin{aligned} \sqrt{r_n}\left(\tilde{\sigma}^2_{W,n} - \sigma^2_W\right) &= \sqrt{r_n}\left(\sigma^2_{W,n} - \sigma^2_W\right) \\ &+ \sqrt{r_n}\left(m_{A,n} - \tilde{m}_{A,n}\right)^2 + \frac{1}{\sqrt{r_n}} \sum_{i=1}^{r_n} \left(W_{n,i} - W_{(i)}\right)^2 \\ &+ \frac{2}{\sqrt{r_n}} \sum_{i=1}^{r_n} \left(W_{(i)} - \tilde{m}_{A,n}\right)\left(W_{n,i} - W_{(i)}\right) \\ &\equiv \sqrt{r_n}\left(\sigma^2_{W,n} - \sigma^2_W\right) + J_{n,4} + J_{n,5} + J_{n,6}. \end{aligned}$$

$J_{n,i} \xrightarrow{P} 0$, $i = 4, 5, 6$, using similar arguments given above. □

The next lemma expresses the four main quantities in terms which allow the application of Lemma 25.

Lemma 32. *Assume the conditions of Theorem 9. Then as $n \to \infty$,*

$$\begin{aligned}
X_{n+1,1} &= \frac{1}{\sqrt{r_n}} \sum_{i=1}^{r_n} \big(W_{n,i} - m_A\big) + o_p(1) \\
X_{n+1,2} &= \frac{1}{\sqrt{r_n}} \sum_{i=1}^{r_n} \Big[\big(W_{n,i} - m_A\big)^2 - var(W_{n,i})\Big] + o_p(1) \\
X_{n+1,3} &= \frac{1}{\sqrt{r_n}} \sum_{i=1}^{r_n} \left[\sqrt{\frac{W_{n,i}}{m_A}} \sqrt{Z_{n,i}} \left(\frac{Z_{n+1,i}}{Z_{n,i}} - m_o\right)\right] + o_p(1) \\
X_{n+1,4} &= \frac{1}{\sqrt{r_n}} \sum_{i=1}^{r_n} \left[Z_{n,i} \left(\frac{Z_{n+1,i}}{Z_{n,i}} - m_o\right)^2 - \sigma_o^2\right] + o_p(1).
\end{aligned}$$

Proof. Using a similar argument given in the proof of Lemma 31, gives

$$X_{n,1} = \frac{1}{\sqrt{r_n}} \sum_{i=1}^{r_n} \big(W_{n,i} - m_A\big) + o_p(1).$$

The desired result now follows from Lemma 31. A similar argument yields the result for $X_{n+1,2}$.

Next consider $X_{n+1,3}$. Algebra yields,

$$X_{n+1,3} = \sqrt{\frac{1}{\tilde{m}_{A,n}}} \frac{1}{\sqrt{r(n+1)}} \sum_{i=1}^{r(n+1)} \left[\sqrt{W_{n,i}} \sqrt{Z_{n,i}} \left(\frac{Z_{n+1,i}}{Z_{n,i}} - m_o\right)\right].$$

But $\tilde{m}_{A,n} \xrightarrow{a.s.} m_A$ (Lemma 27), and using Lemma 25,

$$\frac{1}{\sqrt{r(n+1)}} \sum_{i=1}^{r(n+1)} \left[\sqrt{W_{n,i}} \sqrt{Z_{n,i}} \left(\frac{Z_{n+1,i}}{Z_{n,i}} - m_o\right)\right] \xrightarrow{d} N(0, m_A\, \sigma_o^2)$$

therefore as $n \to \infty$,

$$X_{n+1,3} = \frac{1}{\sqrt{r(n+1)}} \sum_{i=1}^{r(n+1)} \left[\sqrt{\frac{W_{n,i}}{m_A}} \sqrt{Z_{n,i}} \left(\frac{Z_{n+1,i}}{Z_{n,i}} - m_o\right)\right] + o_p(1).$$

Define,

$$J_{n,i} \equiv \sqrt{\frac{W_{n,i}}{m_A}} \sqrt{Z_{n,i}} \left(\frac{Z_{n+1,i}}{Z_{n,i}} - m_o\right).$$

The desired result follows because

$$\frac{1}{\sqrt{r(n+1)}}\sum_{i=1}^{r(n+1)} J_{n,i} - \frac{1}{\sqrt{r_n}}\sum_{i=1}^{r_n} J_{n,i} \xrightarrow{P} 0,$$

which is proven using similar arguments given in the proof of Lemma 31.

Finally we consider $X_{n+1,4}$. Algebra yields,

$$\begin{aligned} X_{n+1,4} &= \frac{1}{\sqrt{r(n+1)}}\sum_{i=1}^{r(n+1)}\left[Z_{n,i}\left(\frac{Z_{n+1,i}}{Z_{n,i}} - m_o\right)^2 - \sigma_o^2\right] \\ &+ (\hat{m}_{o,n} - m_o)^2\frac{1}{\sqrt{r(n+1)}}\sum_{i=1}^{r(n+1)} Z_{n,i} \\ &- 2(\hat{m}_{o,n} - m_o)\frac{1}{\sqrt{r(n+1)}}\sum_{i=1}^{r(n+1)} Z_{n,i}\left(\frac{Z_{n+1,i}}{Z_{n,i}} - m_o\right)^2 \\ &\equiv T_{n,1} + T_{n,2} + T_{n,3}\,. \end{aligned}$$

Using similar arguments given in the proof of Lemma 30, $T_{n,2} \xrightarrow{P} 0$ and $T_{n,3} \xrightarrow{P} 0$. Therefore,

$$X_{n+1,4} = \frac{1}{\sqrt{r(n+1)}}\sum_{i=1}^{r(n+1)}\left[Z_{n,i}\left(\frac{Z_{n+1,i}}{Z_{n,i}} - m_o\right)^2 - \sigma_o^2\right] + o_p(1).$$

Again, the desired result follows using similar arguments given in the proof of Lemma 31.

□

The next lemma supplies the necessary moment calculations needed for the application of Lemma 25.

Lemma 33. *Assume the conditions of Theorem 9. Define*

$$\mathbf{V}_{n,i} = \left(V_{n,i,1}, V_{n,i,2}, V_{n,i,3}, V_{n,i,4}\right)^t,$$

where

$$\begin{aligned} V_{n,i,1} &\equiv W_{n,i} - m_A \\ V_{n,i,2} &\equiv \left(W_{n,i} - m_A\right)^2 - var\left(W_{n,i}\right) \\ V_{n,i,3} &\equiv \sqrt{\frac{W_{n,i}}{m_A}}\sqrt{Z_{n,i}}\left(\frac{Z_{n+1,i}}{Z_{n,i}} - m_o\right) \\ V_{n,i,4} &\equiv \left[Z_{n,i}\left(\frac{Z_{n+1,i}}{Z_{n,i}} - m_o\right)^2 - \sigma_o^2\right]. \end{aligned}$$

For each fixed n, $\left\{\mathbf{V}_{n,i} : i \geq 1\right\}$ *are i.i.d. with* $E\mathbf{V}_{n,1} = \mathbf{0}$ *and*

$$cov\left(\mathbf{V}_{n,1}\right) \to \begin{pmatrix} \sigma_W^2 & \mu_{W,3} & 0 & 0 \\ \mu_{W,3} & \mu_{W,4} - \sigma_W^4 & 0 & 0 \\ 0 & 0 & \sigma_o^2 & 0 \\ 0 & 0 & 0 & 2\sigma_o^4 \end{pmatrix}$$

Proof. Standard calculations give $E\mathbf{V}_{n,1} = \mathbf{0}$ and $var\left(V_{n,1,3}\right) = \sigma_o^2$. The L_p martingale convergence theorem yields $var\left(V_{n,1,1}\right) \to \sigma_W^2$, $var\left(V_{n,1,2}\right) \to \mu_{W,4} - \sigma_W^4$. A conditioning argument gives $var\left(V_{n,1,4}\right) = 2\sigma_o^4 + CE\left(Z_n^{-1}\right)$, where

$$C = var\left(\left(\xi - m_o\right)^2\right) - 2\sigma_o^4$$

(see the proof of Theorem 1 in Dion (1975) for details). Therefore, $var\left(V_{n,1,4}\right) \to 2\sigma_o^4$.

We proceed to compute each pair of covariances. Again using, the L_p martingale convergence theorem yields $cov(V_{n,1,1}, V_{n,1,2}) \to \mu_{W,3}$. Using the branching property, $cov(V_{n,1,1}, V_{n,1,3}) = cov(V_{n,1,1}, V_{n,1,4}) = cov(V_{n,1,2}, V_{n,1,3}) = cov(V_{n,1,2}, V_{n,1,4}) = 0$. Finally, using the branching property,

$$\begin{aligned} cov(V_{n,1,3}, V_{n,1,4}) &= \sqrt{\frac{1}{m_A}} m_o^{-n/2} E\left(Z_{n,1}^2\left(\frac{Z_{n+1,1}}{Z_{n,1}} - m_o\right)^3\right) \\ &= \sqrt{\frac{1}{m_A}} m_o^{-n/2} E\left(\xi - m_o\right)^3 \to 0. \end{aligned}$$

□

Finally, we prove Theorem 9.

Proof of Theorem 9. The result follows from Lemmas 25, 32, and 33 and Slutsky's Theorem. □

4.7 Proof of Corollary 3

Now we prove Corollary 3.

Proof of Corollary 3. Basic algebra yields,

$$\begin{aligned}\sqrt{r_n}\left(\hat{\sigma}^2_{A,n}-\sigma^2_A\right) &= \sqrt{r_n}\left(\hat{\sigma}^2_{W,n}-\sigma^2_W\right)-C\eta_{n,1}\eta_{n,2}\sqrt{r_n}\left(\hat{m}_{A,n}-m_A\right)\\ &- \quad C\,m_A\,\sqrt{r_n}\left(\eta_{n,1}\eta_{n,2}-1\right),\end{aligned}$$

where

$$C\equiv\frac{\sigma^2_o}{m_o\left(m_o-1\right)},\quad \eta_{n,1}\equiv\frac{\hat{\sigma}^2_{o,n}}{\sigma^2_o},\quad \eta_{n,2}\equiv\frac{\hat{m}_{o,n}\left(\hat{m}_{o,n}-1\right)}{m_o\left(m_o-1\right)}.$$

We finish the proof by showing

$$\sqrt{r_n}\left(\eta_{n,1}\eta_{n,2}-1\right)\xrightarrow{P}0 \tag{4.3}$$

and

$$\sqrt{r_n}\left(\hat{\sigma}^2_{W,n}-\sigma^2_W\right)-C\eta_{n,1}\eta_{n,2}\sqrt{r_n}\left(\hat{m}_{A,n}-m_A\right)\xrightarrow{d}N(0,v), \tag{4.4}$$

where $v=\left(\frac{\sigma^2_o}{m_o(m_o-1)}\right)^2\sigma^2_W+\mu_{W,4}-\sigma^4_W-2\left(\frac{\sigma^2_o}{m_o(m_o-1)}\right)\mu_{W,3}$.

First consider (4.3). Define $R_{n,1} \equiv \frac{\hat{m}_{o,n}-1}{m_o-1}$ and $R_{n,2} \equiv \frac{\hat{m}_{o,n}}{m_o}$. Algebra yields,

$$\begin{aligned}\sqrt{r_n}\left(\eta_{n,1}\eta_{n,2}-1\right) &= \left(\eta_{n,1}R_{n,1}-\eta_{n,1}R_{n,2}\right)\sqrt{r_n}\left(\hat{m}_{o,n}-m_o\right)\\ &+ m_o\,\eta_{n,1}\sqrt{r_n}\left(R_{n,1}-R_{n,2}\right)\\ &+ \eta_{n,1}\sqrt{r_n}\left(R_{n,2}-1\right).\end{aligned}$$

Standard arguments show that all three of these terms converge in probability to zero.

Because $\eta_{n,1}\eta_{n,2} \xrightarrow{P} 1$ to prove (4.4), it is sufficient to prove

$$\sqrt{r_n}\left(\hat{\sigma}^2_{W,n}-\sigma^2_W\right) - C\sqrt{r_n}\left(\hat{m}_{A,n}-m_A\right) \xrightarrow{d} N(0,v).$$

But this follows immediately from Theorem 9 and the Cramér-Wold device. □

CHAPTER 5

MULTIVARIATE METHODS FOR DIFFERENTIALLY EXPRESSED GENES

5.1 Introduction

Gene expression microarray data are difficult to analyze because they are characterized by high dimensions and small sample sizes (Leung and Cavalieri, 2003). Because standard notation denotes the number of arrays (available samples) as n and the number of genes (dimension of the data) as p, this problem is often referred to as the *large p, small n* or *high-dimensional, low sample size* problem. While microarrays are perhaps the prototypical large p, small n, problem, it is in fact a problem encountered in several scientific areas (Donoho, 2000; Johnstone, 2001; Kosorok and Ma, 2007).

A variety of procedures have been proposed for identifying differentially expressed genes (DEGs). There are methods based on modified t-statistics, fold change methods, linear models, and Bayesian analysis. Dudoit et al. (2002b) provides a nice survey of these commonly used statistical methods; Dudoit and van der Laan (2008) is also a useful resource for this material.

A major drawback of the methods given above is that they are all essentially univariate. Multivariate analysis seems to be the right framework for analyzing gene expression data because it allows the statistician to account for correlation among the genes. In fact, Szabo et al. (2003), Lu et al. (2005), and Kim et al. (2005) have all utilized this idea and developed multivariate procedures based on Hotelling's T^2 statistic to identify DEGs in two-sample problems. Recently, Tsai and Chen (2009) extended these ideas to the k sample problem ($k \geq 2$) by a proposing a modified multivariate analysis of variance solution to the problem; additionally, their work addresses the important question of identifying associations in gene pathways.

In the present paper we develop a multivariate procedure for identifying DEGs in both the one- and two-sample settings. Our procedure is based on a multivariate test for the mean vector suggested by Kuelbs and Vidyashankar (2009). In simulation studies, their test works well (in terms of both power and size) in the large p, small n setting. Using this test we develop a screening algorithm for identifying DEGs. For concreteness, assume we have a data set consisting of 2000 genes. The basic idea is that we replace 2000 univariate tests with tens of multivariate tests, thus mitigating the problem of multiple comparisons.

The remainder of this paper is organized as follows. The next section discusses the basic testing procedure for the one- and two-sample problems and Section 5.3 describes the screening algorithm. The simulations are described in Sections 5.4, 5.5, and 5.6. Finally, we analyze the ApoAI knockout data (Callow et al., 2000) in Section 5.7.

5.2 Sup-Norm Test Statistic

We first present the one-sample formulation of the problem. This then easily extends to the two-sample problem.

5.2.1 One-Sample Formulation

To explain the multivariate test developed in Kuelbs and Vidyashankar (2009) we first recall some definitions from vector analysis. Let $\mathbf{x}$ be a vector in $\mathbb{R}^p$. For $1 \leq \rho \leq \infty$, the ℓ^ρ norms are defined by

$$\|\mathbf{x}\|_\rho = \begin{cases} \left(\sum_{j\geq 1} |x_j|^\rho\right)^{\frac{1}{\rho}} & \text{if } 1 \leq \rho < \infty, \\ \max_{1\leq j\geq p} |x_j| & \text{if } \rho = \infty \end{cases}$$

(see for example Friedman (1982)). As is common, we refer to the ℓ^∞ norm as the sup-norm.

The statistic is based on the following intuitively appealing idea. For each gene, compute the average expression level across the patients; assuming there are p genes the resulting mean vector will be an element of $\mathbb{R}^p$. When concerned with finding differentially expressed genes it is natural to compute the maximum of suitable "averages" of gene expressions. This argument suggests using the sup-norm. In fact, simulation results presented in Kuelbs and Vidyashankar (2009) suggest the superiority of the sup-norm over other ℓ^ρ norms.

We introduce notation to make the above idea mathematically precise. We assume that there are n arrays and p genes. Let $X_{i,j}$ represent the expression level of gene j from array i. Then, $\mathbf{X}_i = (X_{i,1}, ..., X_{i,p})^t$ represents the expression data for array i. We assume that $\mathbf{X}_1, ..., \mathbf{X}_n$ are independent and identically distributed (iid) random vectors with mean μ. Furthermore, let $\bar{\mathbf{X}}$ denote the p dimensional vector of averaged expression levels; more precisely, $\bar{\mathbf{X}} = \left(\bar{X}_1, ..., \bar{X}_p\right)^t$, where $\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{i,j}$, $j = 1, ..., p$. To test the null hypothesis, $H_0 : \mu = \mu_0$, Kuelbs and Vidyashankar (2009) consider statistics of the form, $T_\rho \equiv \|\sqrt{n}\left(\bar{\mathbf{X}} - \mu_0\right)\|_\rho$. We focus on T_∞, which we refer to as the sup-norm (SN) statistic.

Under suitable regularity conditions, Kuelbs and Vidyashankar (2009) prove the asymptotic normality (in large p, small n framework) of $\sqrt{n}\left(\bar{\mathbf{X}} - \mu_0\right)$. Now, let $\boldsymbol{\Sigma}$ denote the covariance matrix of the data, that is $cov\left(\mathbf{X}_1\right) = \boldsymbol{\Sigma}$. Then, informally, the asymptotic normality result gives $\sqrt{n}\left(\bar{\mathbf{X}} - \mu_0\right) \approx N_p\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$. Using the continuous mapping theorem gives the asymptotic normality of T_ρ,

$$T_\rho \approx \|N_p\left(\mathbf{0}, \boldsymbol{\Sigma}\right)\|_\rho. \tag{5.1}$$

Of course, if the $\mathbf{X}_i$ are i.i.d. realizations from a multivariate normal distribution, then these statements are no longer approximate; the distributions are exactly equal to the specified norm of the corresponding normal distribution. One of the strengths of the SN procedure is that the results continue to hold (in some approximate sense) even if the underlying distribution is non-normal.

In the context of testing for DEGs, we are interested in the null hypothesis, $H_0 : \mu = \mathbf{0}$. Clearly, for (6.4) to be useful in testing for DEGs we need to estimate $\boldsymbol{\Sigma}$ accurately. Several authors have discussed the difficulty in estimating $\boldsymbol{\Sigma}$ in large p, small n settings (Tsai and Chen, 2009). We use the shrinkage based estimator developed by Strimmer and his students (Schafer and Strimmer, 2005; Opgen-Rhein and Strimmer, 2007). Using the idea of Ledoit and Wolf (2004), they define a covariance estimator which is guaranteed to be positive definite, even with $p > n$. Their algorithm is implemented in both R code (corpcor) and Matlab code (covshrink), which are freely available at `http://strimmerlab.org/software.html`.

Using (6.4) and the shrinkage estimator for the covariance matrix, we provide an algorithm for testing the null hypothesis that the mean of $\mathbf{X}_1$ is zero, that is that the genes are not differentially expressed. This is a Monte-Carlo algorithm used to approximate the distribution of $\|N_p\left(\mathbf{0}, \boldsymbol{\Sigma}\right)\|_\rho$. The user would first decide on a value of ρ and level of significance α to use. Again, for testing for DEGs,

$\mu_0 = \mathbf{0}$. When $\rho = \infty$, we refer to this procedure as the sup-norm (SN) test.

1. Compute the observed test statistic, T_ρ.
2. Estimate the covariance matrix S, using shrinkage.
3. Generate B random vectors $\mathbf{Y}_1,...,\mathbf{Y}_n \sim N_p(\mathbf{0}, S)$; compute the norm of these vectors, $T_i^\star \equiv \|\mathbf{Y}_i\|_\rho$; finally compute the $(\alpha/2)$ sample quantile $\hat{q}_{\alpha/2}$ and the $(1-\alpha/2)$ sample quantile $\hat{q}_{1-\alpha/2}$ from $T_1^\star, ..., T_B^\star$.
4. Reject if $T_\rho < \hat{q}_{\alpha/2}$ or if $T_\rho > \hat{q}_{1-\alpha/2}$.

5.2.2 Two-Sample Formulation

The two-sample problem is a straight forward generalization of the one-sample problem given above. In this case we have two independent samples

$$\{\mathbf{X}_{i1} : 1 \leq i \leq n_1\} \quad \text{and} \quad \{\mathbf{X}_{i2} : 1 \leq i \leq n_2\}.$$

For fixed k, $\{\mathbf{X}_{i1} : 1 \leq i \leq n_k\}$ are i.i.d. random vectors with mean μ_k and covariance matrix $\mathbf{\Sigma}_k$. Here X_{ijk} represents the expression level of gene j from array i in sample k and $\bar{\mathbf{X}}_k = \left(\bar{X}_{1k}, ..., \bar{X}_{pk}\right)^t$, where $\bar{X}_{jk} = \frac{1}{n_k}\sum_{i=1}^{n_k} X_{ijk}$, $j = 1, ..., p$. To test the null hypothesis of equal sample means, $H_0 : \mu_1 = \mu_2$, we consider statistics of the form, $T_{\rho,2} \equiv \|\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2\|_\rho$. Again, using results from Kuelbs and Vidyashankar (2009), we have that $T_{\rho,2} \approx \|N_p\left(\mathbf{0}, \frac{1}{n_1}\mathbf{\Sigma}_1 + \frac{1}{n_2}\mathbf{\Sigma}_2\right)\|_\rho$.

To the test the null hypothesis we carry out the same basic steps as described in the subsection above. First we use the shrinkage algorithm described in Opgen-Rhein and Strimmer (2007) to estimate the covariance matrices and then approximate the null distribution of the test statistic using the Monte-Carlo algorithm given above.

5.3 Screening Algorithm

The idea of our screening algorithm is to repeatedly apply the SN test, 'throwing out' genes that result in a tests of accepting the null hypothesis and keeping genes that result in a rejecting the null hypothesis. In the algorithm we use the notation I_i to denote the indicator for test i, indicating whether the test detected a DEG among all of the genes in group i. Specifically, $I_i = 1$ means the SN test rejected the null hypothesis for group i ($I_i = 0$ otherwise). There are certain parameters one needs to set to run the algorithm: the (expected) initial dimension size (d_0), the reduction factor (r), and the final cutoff (p_f). We describe an example which explains these values. Assume the data consists of $p = 2000$ genes; set $d_0 = 100$, $r = 2$, and $p_f = 30$. The value $d_0 = 100$ means that in the first round of tests we will divide the genes into $2000/d_0 = 20$ groups with an expected group size of $d_0 = 100$; the value of $r = 2$ means in each subsequent stage the expected group size will be reduced by a factor of 2; finally, the value $p_f = 30$ means that the algorithm will run until the total number of remaining genes is less than or equal to 30. In the first stage, randomly subset the genes into 20 groups (with an average of 100 genes a group) and perform 20 SN tests on these groups. Keep all of the genes in groups with $I_i = 1$, and throw out the others. To start stage 2 take these remaining genes and divide them into groups with expected size $d_0/r = 50$; now repeat the process.

In practice one would set the parameters of the algorithm based on the characteristics of the observed data, for instance, the number of samples, the number of genes, and the variance of the data.

The structure of our algorithm is outlined below. The algorithm outputs a reduced set of genes that ideally will contain all of the differentially expressed genes. Notice in the update step, there is a check for the case $p_u = p_a$. This case comes about if in the current stage the SN test for each group rejects the null hypothesis, and no genes can be removed. If the true number of DEGs is greater than the inputted cutoff, i.e. $p_d > p_f$, then it is desirable for the algorithm to halt before the number of genes is reduced below p_f. But it is also possible that this case arises because of the particular assignment of genes. For example, suppose that there are ten groups and 10 DEGs. If the assignment is such that one DEG is placed in each of the 10 groups, then all 10 tests may (correctly) reject the null hypothesis. To account for this situation, when $p_u = p_a$, we do a second allocation of the genes and test if any genes can be removed after this second allocation. In principle, a user could re-allocate any number of times before deciding the set of genes cannot be reduced further.

Screening Algorithm

Input. Set $p_a = p$, $d_a = d_0$, $count = 0$. Continue to Step 1.

Step 1. (Random Allocation) Randomly allocate the p_a genes to $K \equiv \lceil p_a/d_a \rceil$ groups. Continue to Step 2.

Step 2. (Test). Perform the SN procedure on each of the K groups. Continue to Step 3.

Step 3. (Update) Remove all genes in groups with $I_i = 0$. Let p_u denote the updated number of genes (after removal).

if $p_u = 0$

Stop. Output $\emptyset$ (declare that none of the genes are differentially expressed).

```
elseif $p_u \leq p_f$.
    Stop. Output the set $G_f$, which consists of the labels for the remaining $p_u$ genes.
elseif $p_u = p_a$.
    if $count = 0$.
        Update $count = 1$. Return to Step 1.
    elseif $count = 1$.
        Stop. Output the set $G_f$, which consists of the labels for the remaining $p_u$ genes.
else
    Update $d_u = d_a/r$. Set $p_a = p_u$, $d_a = d_u$, $count = 0$. Return to Step 1.
```

5.4 Specifications for the Simulation Studies

In this section we detail the specifications used in the simulation experiments presented in Sections 5.5 and 5.6. All of the Monte-Carlo experiments presented below are based on 5000 simulated data sets. In all cases the level of the test is fixed at $\alpha = .05$ and $B = 2000$ samples are used to approximate the null distribution. We generate data from multivariate normal distributions. For the one-sample simulations, data is simulated from $N_p(\mu, \mathbf{\Sigma})$; for the two-sample problem, sample k is simulated from $N_p(\mu_k, \mathbf{\Sigma}_k)$, for $k = 1, 2$. We proceed to describe our choices for the mean vector and covariance matrix.

In the one-sample problem the mean vector is chosen as follows. Let p_d be the number of DEGs in the data set, which we will assume all have a common mean $\mu_d \neq 0$. The mean vector contains p_d non-zero elements and $p - p_d$ zeros, $\mu = (\mu_d, ..., \mu_d, 0, ..., 0)^t$. Note that $p_d = 0$ corresponds to the case of no DEGs; with the corresponding mean vector $\mu = \mathbf{0}$. For the two-sample problem, only sample one contains DEGs, while sample two contains all null genes. Thus, $\mu_1 = (\mu_d, ..., \mu_d, 0, ..., 0)^t$, while $\mu_2 = \mathbf{0}$.

We use experimental data to set the covariance matrices. Specifically, we use the leukemia dataset described by Golub et al. (1999), which studies the gene expression in two types of leukemia, acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). We use the same pre-processing step as described in Dudoit et al. (2002a) Section 3.1; leaving 3571 genes from 72 patients, 38 ALL and 25 AML. On the remaining 3571 genes, we apply the standardization technique described in Section 3.3 of the same paper. We then separately estimated a covariance matrix from the ALL group and the AML group, denoted $\mathbf{\Sigma}_L$ and $\mathbf{\Sigma}_M$, respectively. Specifically, we randomly permuted the 3571 genes and then estimated the covariance matrix using the shrinkage algorithm (for correlations) of Schafer and Strimmer (2005). This method produces two 3571 × 3571 covariance matrices which are fixed throughout the paper. For a simulation study based on $p \leq 3571$ genes, we first fix the covariance matrix of appropriate dimensions by considering the $p \times p$ upper sub-matrix of $\mathbf{\Sigma}_I$ denoted $\mathbf{\Sigma}_{I,p}$, $I = L, M$. More precisely, $\left(\mathbf{\Sigma}_{I,p}\right)_{i,j} = \left(\mathbf{\Sigma}_I\right)_{i,j}$, for $1 \leq i, j \leq p$. Finally, for each simulated data set, we simulate n random vectors from the $p-$dimensional normal distribution $N_p\left(\mu, \mathbf{\Sigma}_{I,p}\right)$, where μ is a specified $p \times 1$ vector which represents the mean expression level of the genes.

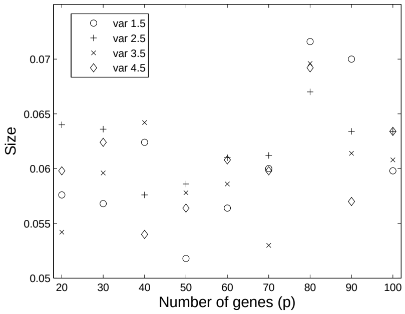


Figure 5.1: One-sample size analysis. Simulated size versus the number of genes (p), for different covariance matrices $v\mathbf{\Sigma}_{L,p}$.

Throughout, $\mathbf{\Sigma}$ is set to be a constant multiple of $\mathbf{\Sigma}_{L,p}$, $\mathbf{\Sigma} = v\mathbf{\Sigma}_{G,p}$, where $v > 1$ gives a simple way for examining increased variance in the data. Similarly, for the two-sample simulations, $\mathbf{\Sigma}_1 = v_1\mathbf{\Sigma}_{L,p}$ and $\mathbf{\Sigma}_2 = v_2\mathbf{\Sigma}_{M,p}$.

5.5 One-Sample Simulation Studies

In this section we consider simulation experiments related to the one-sample problem. First we present results which study the size and power of the SN test. We then present results for the screening algorithm. Recall that the specifications for the simulations are described in Section 5.4.

5.5.1 SN Test

First we evaluate the size of the SN test under different conditions. For size experiments, all of the genes are null, and thus we set $\mu = \mathbf{0}$. We consider data based on $n = 20$ replicates and examine the size as the number of genes increases from $p = 20$ to 100. The result of the simulations are displayed in Figure 5.1.

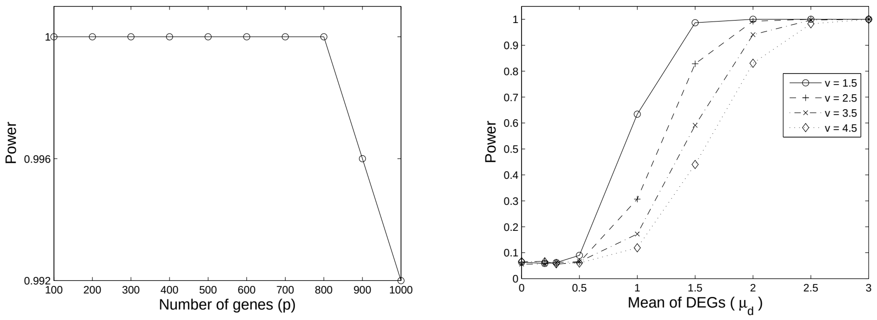


Figure 5.2: One-sample power analysis. The second graph displays results for different covariance matrices $v\mathbf{\Sigma}_{L,p}$.

Notice that in all cases the test achieves close the nominal size of $\alpha = .05$.

Next we consider the power of the SN test; for power experiments, the set of genes includes at least one DEG. We consider experiments that examine the power under increasing variance and increasing number of genes. The result of the simulations are displayed in Figure 5.2. Notice that the SN test is very powerful in detecting a single DEG. With $p = 800$ total genes (and only one DEG) the test correctly rejects the null hypothesis in all 5000 experiments; with $p = 1000$ genes the test rejects the null in 4960 of the experiments.

5.5.2 Screening Algorithm

In this section we present, numerical results obtained by performing the screening algorithm on simulated data sets. Recall that the screening algorithm repeatedly applies the SN test reducing the original set of genes to a set G_f. For a single data set, we record two performance measures of the screening algorithm: the number of retained DEGs (R) and the total number of genes after the final run, $|G_f|$. For each experiment, we report the minimum, maximum, and mean of R and $|G_f|$

over the 5000 simulations. We will clarify these ideas with a concrete example. Assume that there are 2000 total genes, 10 DEGs, and that we set the cutoff at $p_f = 30$. Furthermore, assume that the algorithm continues to run until $|G_f| < p_f$. Ideally, after the algorithm has run, all 10 DEGs remain in the final set G_f. We record the size of G_f and the number of DEGs which remain in G_f. Recall that the screening algorithm can end in three different ways: *exit one* occurs when the algorithm runs until the cutoff is reached, $1 \leq |G_f| \leq p_f$; *exit two* occurs when the algorithm cannot reduce the number of genes below p_f, $|G_f| > p_f$; and *exit three* occurs when the algorithm declares that all of the genes are null, $G_f = \emptyset$. We only report R and $|G_f|$ for those data sets which result in exit one or exit two. In all of the simulations there are 10 DEGs; additionally, the parameters of the screening algorithm are fixed at $(r, d_0, p_f) = (2, 100, 30)$.

First we consider the impact of changing the mean for the DEGs. In this experiment there are $p = 2000$ genes, 1990 of the genes have mean zero while the remaining 10 DEGs have mean μ_d; we consider $\mu_d = .5, 1, 1.5$, and 2. With $\mu_d = .5$, 71.82% of the simulated data sets resulted in exit one, the remaining 28.18% resulted in exit three; for the other values of μ_d all 5000 simulated data sets resulted in exit one. The results of the simulations are displayed in Figure 5.3. If $\mu_d = .5$, the algorithm does not perform well; on average it only retains one of the DEGs. However, with $\mu_d = 1.5$, the algorithm, on average, is retaining all 10 of the DEGs. With $\mu_d = 2$, in all 5000 simulations, the algorithm retains all 10 of the DEGs.

Next we consider the impact of the total number of genes present. In this case, $\mu_d = 2$ is fixed and we considered $p = 1000, 1500, 2000, 2500$, and 3000 genes. In this experiment, the algorithm ended in exit one and retained all 10 DEGs for

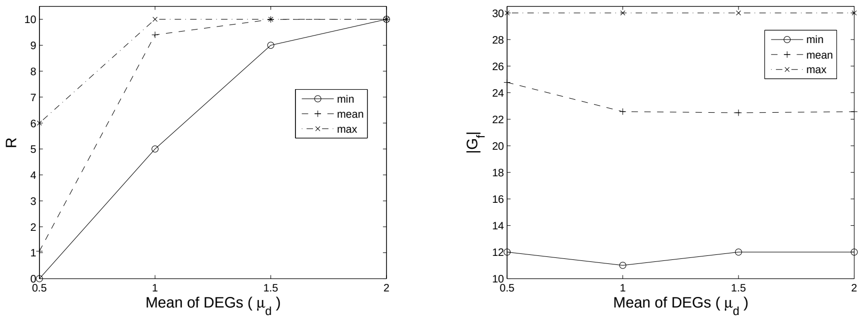


Figure 5.3: One-sample screening analysis. Plots of the minimum, mean, and maximum values over the 5000 simulations for the number of retained DEGs, R, and the final number of genes, $|G_f|$.

every simulated data set. Evidently, with $\mu_d = 2$, the algorithm can handle very high dimensions.

5.6 Two-Sample Simulation Studies

In this section we consider simulation experiments related to the two-sample problem. First we present results which study the size and power of the SN test. We then present results for the screening algorithm.

5.6.1 SN Test

All of the experiments presented in this section use $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_{L,p}$ and $\mathbf{\Sigma}_2 = \mathbf{\Sigma}_{M,p}$. First we evaluate the size of the SN test under different conditions. In the first experiment, we consider data based on $n_1 = 10$ and $n_2 = 15$ replicates and examine the size as the number of genes increases from $p = 100$ to 500. In the second experiment, we have $n_1 = 22$ and $n_2 = 25$ replicates and examine the size as the

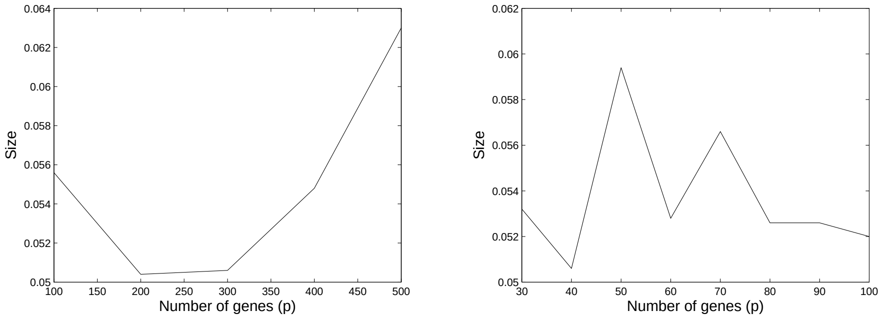


Figure 5.4: Two-sample size analysis. In the first graph, $n_1 = 10$, $n_2 = 15$; in the second graph, $n_1 = 22$, $n_2 = 25$.

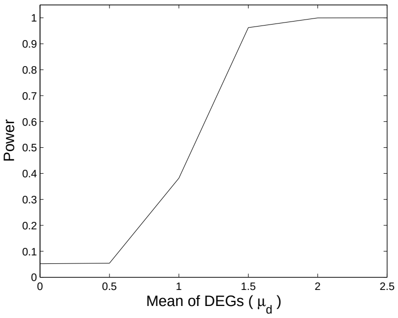


Figure 5.5: Two-sample power analysis. Simulated power versus the DEG mean μ_d.

number of genes increases from $p = 30$ to 100. The result of the simulations are displayed in Figure 5.4. Notice that in all cases the test achieves close the nominal size of $\alpha = .05$.

Next we consider the power of the SN test; in these experiments the sample sizes are fixed at $n_1 = 10$ and $n_2 = 15$. We consider experiments that examine the impact of the total number genes, the mean of the DEGs, and the total number of DEGs. The result of the simulations are displayed in Figure 5.5. Notice that the SN test is very powerful in detecting a single DEG.

5.6.2 Screening Algorithm

We consider the performance of the screening algorithm in the two-sample setting. All of the experiments presented in this section use $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_{L,p}$ and $\mathbf{\Sigma}_2 = \mathbf{\Sigma}_{M,p}$, and $n_1 = 10$, $n_2 = 15$.

First we consider the impact of changing the mean for the DEGs. In this experiment there are $p = 2000$ genes, 1990 of the genes have mean zero while the remaining 10 DEGs have mean μ_d; we consider $\mu_d = .5, 1, 1.5$, and 2. With $\mu_d = .5$, 65.42% of the simulated data sets resulted in exit one, the remaining 34.58% resulted in exit three; for the other values of μ_d all 5000 simulated data sets resulted in exit one. The results of the simulations are displayed in Figure 5.6. The results of this experiment are almost identical to the one-sample analog. Specifically, with $\mu_d = .5$ the algorithm does not perform well; however, with $\mu_d = 2$, in all 5000 simulations, the algorithm retains all 10 of the DEGs.

Next we consider the impact of the total number of genes present. In this case, $\mu_d = 2$ is fixed and we considered $p = 1000, 1500, 2000, 2500$, and 3000 genes. Just as in the one-sample analog, in this experiment, the algorithm ended in exit one and retained all 10 DEGs for every simulated data set.

5.7 Data Analysis

In this section we analyze data from a study of the apolipoprotein AI (ApoAI) gene described in Callow et al. (2000). This data has been previously analyzed by Smyth (2004); a tutorial for analyzing the data set is available online as part of the LIMMA user's manual (Smyth et al., 2003). We normalize the data using the

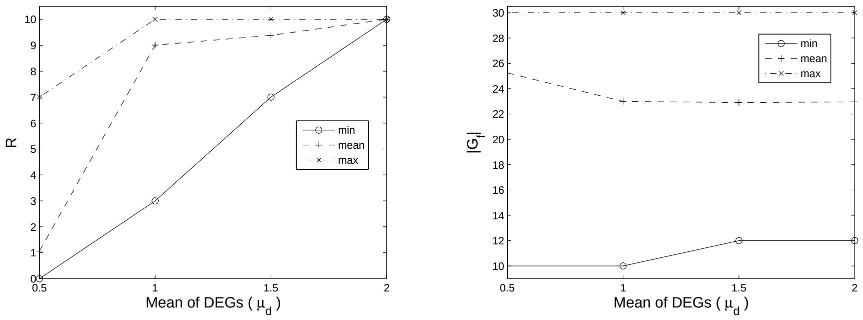


Figure 5.6: Two-sample screening analysis. Plots of the minimum, mean, and maximum values over the 5000 simulations for the number of retained DEGs, R, and the final number of genes, $|G_f|$.

LIMMA package as described in Smyth et al. (2003).

The ApoAI gene plays a central role in high density lipoprotein (HDL) metabolism; see Williamson et al. (1992) and Plump et al. (1996) for more detailed discussions of the ApoAI knockout model. The Callow et al. (2000) experiment was designed to study the effect of ApoAI deficiency on other genes in the liver. To this end, data was collected on 8 ApoAI knockout mice and 8 control mice. For each of these 16 mice, mRNA measurements were collected from liver tissue. The RNA from each mouse was hybridized to a separate array. The data set consists of 16 arrays with measurements on 5548 expressed sequence tags (ESTs).

Callow et al. (2000) identified 8 ESTs (representing four different genes) which are differentially expressed in the knockout group versus the control group. Smyth (2004) lists the top 15 differentially expressed ESTs based on his LIMMA approach. Of these 15, the top 8 coincide with the ones identified in Callow et al. (2000). In fact, Smyth writes “the top eight genes stand out clearly from the other genes and all methods clearly separate these genes from the others” (note that Smyth uses “gene” instead of “EST”).

We ran our screening algorithm twice of this data set. The overlap of the two runs consisted of 11 ESTs, which included the 8 ESTs identified in Callow et al. (2000).

CHAPTER 6

ROBUST ASYMPTOTIC INFERENCE FOR HIGH DIMENSIONAL DATA

6.1 Introduction

Modern scientific technology is providing a class of statistical problems that involve high dimensional data. These data are typically characterized by small sample size (n) and a large number of parameters (p); hence, the terminology *large p, small n problems.* While data from gene expression microarray experiments are a canonical example for such data, other examples arise in diverse fields such as proteomics, chemometrics, functional magnetic resonance imaging, and astronomy (Donoho, 2000; Johnstone, 2001; Varmuza and Filzmoser, 2008).

Several authors have studied the problem of robust methods for high dimensional data (Aggarwal and Yu, 2001; Hubert and Engelen, 2004; Hubert et al., 2005; Filzmoser et al., 2008). Kadota et al. (2003), Oh and Gao (2009), and Shieh and Hung (2009) study the problem of outlier detection in the specific context of microarray data. This work focuses on detecting and removing outliers, and then proceeding with standard analysis on the remaining data points. However, it is desirable to develop procedures which are robust to distributional assumptions, in addition to being robust to outliers. In fact, what is needed is a framework for developing inferential tools in the large p, small n setting which are "robust" and "efficient." The meaning of robustness and efficiency needs to be better understood in this context. This chapter is a first step in that direction, developing a robust, adaptive procedure for multivariate problems.

Bickel (1982) reviews the work of adaptive inference for independent and identically distributed (i.i.d.) data. Additionally, Beran (1978) considered robust and adaptive inference in the i.i.d. setting and established asymptotic efficiency and a form of asymptotic robustness of his procedure. In the context of regression problems, Stone (1982) studied adaptive estimation and established rates of convergence. More recently, Bickel et al. (1993) studied adaptive inference for semi-parametric models.

The rest of the paper is organized as follows. Section 2 introduces basic notation and assumptions, while Section 3 is devoted to a brief literature review concerning adaptive inference for one-dimensional data. Section 4 describes the methodology for our model while Section 5 is devoted to simulation results. Finally, Section 6 describes a plan for future work.

6.2 Notation and Assumptions

We shall denote the data as $\big\{X_{i,j} : 1 \leq i \leq n, 1 \leq j \leq p(n)\big\}$. In the context of microarrays, the data is interpreted as a collection of gene expression data for $p(n)$ genes from n replicates, where $X_{i,j}$ represents the expression level of the i^{th} replicate, for gene j. More generally, the number of replicates could be represented as some function of n, say $r(n)$, but this comes at the price of more cumbersome notation. We assume that for fixed n, $\mathbf{X}_i \equiv \big(X_{i,1}, ..., X_{i,p(n)}\big)$, $1 \leq i \leq n$, are i.i.d. random vectors. We additionally assume that $p(n)$ is non-random and that it is the same for all replicates. This assumption is restrictive, for example it does not cover the case of missing data, but it can be removed using the techniques of Kuelbs and Vidyashankar (2009). We specify the following model assumptions.

Assumption 4. *The marginal distributions of $X_{i,j}$ are symmetric about $\theta_{0,j}$ with density f_j, which belongs to some location family. Additionally, the covariance matrix of $\mathbf{X}_i$ is $\mathbf{\Sigma}_n$, which is a $p(n) \times p(n)$, non-singular matrix.*

6.3 Adaptive, Robust Estimation for Univariate Data

We recall the robust, adaptive location estimator for univariate data studied in Beran (1978). Throughout this section, any reference to Beran without further date information refers to Beran (1978).

Let $\xi_1, \xi_2, ..., \xi_n$ be i.i.d. real valued random variables with density g. As a model for the data, assume that g belongs to the location family $\left\{f(x-\theta) : \theta \in \mathbb{R}\right\}$, where f is symmetric about zero, absolutely continuous, and has finite Fisher information $I(f) \equiv \int \left(f'(x)\right)^2 / f(x) dx < \infty$. Beran proves that his location estimator $\hat{\theta}_n$ is asymptotically efficient under the model. More precisely, he proves, as $n \to \infty$,

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) = n^{-1/2} I^{-1}(f) \sum_{i=1}^{n} -\frac{f'\left(X_i - \theta\right)}{f\left(X_i - \theta\right)} \quad + \quad o_p(1),$$

under every symmetric density $f(x-\theta)$ belonging to the model. Beran additionally proves 'robustness' to the symmetry assumption. Informally, the distribution of $\hat{\theta}_n$ does not change much if the distribution of each X_i is perturbed from a symmetric shape to an arbitrary nearby shape. He formulates robustness using an extension of the regularity concept discussed in Hájek (1970). We do not pursue robustness to the symmetry assumption in this paper.

Let g_n be a non-parametric kernel density estimator (KDE) of the form,

$$g_n(x) \equiv \frac{1}{nc_n} \sum_{i=1}^{n} K\left(\frac{x - \xi_i}{c_n}\right).$$

Utilizing the symmetry of the data, Beran proposes an estimator which minimizes the Hellinger distance between $g_n(x)$ and $g_n(-x+2\theta)$. To be more precise, first recall that the *Hellinger distance*, denoted H, between two densities f and g is given by

$$H(f,g) \equiv ||f^{\frac{1}{2}} - g^{\frac{1}{2}}||_2 = 2 - 2\gamma(f,g),$$

where

$$\gamma(f,g) \equiv \int (f(x))^{1/2}(g(x))^{1/2}dx.$$

Therefore, the location estimator $\hat{\theta}_n$ which minimizes the the Hellinger distance between $g_n(x)$ and $g_n(-x+2\theta)$, can be obtained as

$$\hat{\theta}_n \equiv \underset{\theta\in\mathbb{R}}{\text{argmax}} \int (g_n(x))^{1/2}(g_n(-x+2\theta))^{1/2}dx. \tag{6.1}$$

Beran assumes the following the conditions.

B 1. *$K(x)$ is a non-vanishing density, symmetric about zero, absolutely continuous, and the ratio $K'(x)/K(x)$ is bounded over the real line.*

B 2. *The density g is symmetric about θ, absolutely continuous, and has finite Fisher information $I(g) \equiv \int \left(g'(x)\right)^2/g(x)dx < \infty$.*

Remark 5. *Using arguments similar to those given in Cheng and Vidyashankar (2006), the assumptions on the density K can be weakened to allow K to be symmetric about zero and absolutely continuous.*

Beran uses a standard technique (see also Beran (1977)) to address the existence and consistency of the estimator $\hat{\theta}_n$. Namely, he studies the functional related to (6.1). To be precise let $\mathcal{G}$ be the class of densities metrized by the L_1 distance. Define the (possibly multi-valued) functional T by the requirement that for every

$k \in \mathcal{G}$,

$$\int \left(k(-x+2T(k))\right)^{1/2} \left(k(x)\right)^{1/2} dx = \max_{\theta \in \mathbb{R}} \int \left(k(-x+2\theta)\right)^{1/2} \left(k(x)\right)^{1/2} dx. \quad (6.2)$$

We summarize the results of Beran's Lemma 1 and Lemma 2 in the following proposition. The result shows that (6.1) is well-defined, and, in particular, if k is symmetric then $T(k)$ is uniquely defined. Additionally, the result shows the continuity of T.

Proposition 8. *Let T be defined by* (6.2)*.*

1. *For $k \in \mathcal{G}$, the set of values $T(k)$ which satisfies* (6.2) *is non-empty and compact. If k is symmetric, then $T(k)$ is uniquely defined as the center of symmetry.*
2. *Let $\{k_n \in \mathcal{G}\}$ be a sequence converging to $k \in \mathcal{G}$ in L_1. If $T(k)$ is uniquely defined, then every value of $T(k_n)$ converges to $T(k)$.*

Now, Proposition 8 together with L_1 consistency of the KDE yields the consistency of $\hat{\theta}_n$ (see Beran's Theorem 1). Under the additional finite Fisher information assumption, Beran also proves the asymptotic efficiency of $\hat{\theta}_n$ (see Theorem 3). We summarize these results in the next proposition.

Proposition 9. *Let $c_n \to 0$ and $nc_n \to \infty$.*

1. *(Consistency). Assume **B1** holds and that $T(g)$ is uniquely defined. Then every sequence $\hat{\theta}_n$ converges to $T(g)$ in probability, as $n \to \infty$.*
2. *(Asymptotic efficiency). Assume **B1** and **B2** hold. Then, as $n \to \infty$,*

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} N\left(0, I^{-1}(g)\right).$$

We make a final comment concerning the functional T. As alluded to above, a standard proof technique for minimum distance estimators based on density estimators, is to prove the consistency of the corresponding functional T. And then use this consistency result to show that the sequence of minimizers ($\hat{\theta}_n$ in this case) 'inherits' the properties of the density estimator. For example, Devroye (1987) (see Theorem 4.2) proves that the robustness of Beran's parametric MHDE (Beran, 1977) follows from the robustness of the non-parametric density estimator. In fact, we use this heuristic to argue that the plausibility that the joint consistency of the minimizers follows from the joint consistency of the corresponding density estimators.

6.4 Robust and Adaptive Inference for Large p, Small n Problems

In this section we describe the statistical methodology to test the null hypothesis

$$H_0 : \theta_1 = \theta_{0,1}, \theta_2 = \theta_{0,2}, ..., \theta_{p(n)} = \theta_{0,p(n)}.$$

Following Kuelbs and Vidyashankar (2009), we propose to estimate the parameters 'component-wise' and then estimate the covariance matrix to perform the test. We begin by describing the estimation methodology.

Assumption 5. *For each $j \geq 1$, K_j is a density satisfying the following properties.*

1. $\int xK_j(x)dx = 0$.
2. $k_j^2 \equiv \int x^2 K_j(x)dx < \infty$. *Moreover,* $\sup_{j\geq 1} k_j^2 < \infty$.

Let $g_{n,j}$ be kernel density estimator (KDE) for the j^{th} component, namely

$$g_{n,j}(x) \equiv \frac{1}{nc_{n,j}} \sum_{i=1}^{n} K_j \left(\frac{x - X_{i,j}}{c_{n,j}} \right).$$

For each fixed j, it is known (Devroye, 1983) that if $c_{n,j} \to 0$ and $nc_{n,j} \to \infty$, then the KDE is pointwise and L_1 strongly consistent; namely, as $n \to \infty$,

$$g_{n,j}(x) \xrightarrow{a.s.} f_j(x - \theta_{0,j}) \quad \text{[a.e. x]}, \qquad \int \left| g_{n,j}(x) - f_j(x - \theta_{0,j}) \right| dx \xrightarrow{a.s.} 0.$$

To obtain joint inferential results (across all p_n components) requires a uniform consistency result for the density estimators. This result is related to uniform large deviations and has recently been investigated by Louani (2005). We now state his result.

Proposition 10. *[Louani 2005] Let g_n be a kernel density estimator and $\mathcal{F}$ a class of densities. Assume that*

$$\lim_{n\to\infty} \sup_{f\in\mathcal{F}} \|E(g_n) - f\|_1 = 0. \tag{6.3}$$

Then, if $c_n \to 0$ and $nc_n \to \infty$,

$$\sup_{f\in\mathcal{F}} P\left(\|g_n - f\|_1 > \epsilon\right) \leq e^{-C_\epsilon n}$$

where C_ϵ is independent of n and K.

As an immediate corollary of the Proposition, we obtain joint consistency for the density estimates.

Corollary 4. *Assume, for each j, $f_j \in \mathcal{F}$, which satisfies condition* (6.3). *If $c_{n,j} \to 0$, $nc_{n,j} \to \infty$, and $\frac{\log p(n)}{n} \to 0$, then*

$$\sup_{1\leq j\leq p(n)} \|g_{n,j}(x) - f_j(x - \theta_{0,j})\|_1 \xrightarrow{a.s.} 0.$$

Louani (2005) discusses classes of densities $\mathcal{F}$ which satisfy the property (6.3).

Following Beran's univariate work (Beran, 1978), we define componetwise adaptive estimators as

$$\hat{\theta}_{n,j} = \underset{\theta \in \mathbb{R}}{\text{argmax}} \int g_{n,j}^{1/2}(x) g_{n,j}^{1/2}(-x + 2\theta) dx.$$

Using Beran's arguments, it follows that $\hat{\theta}_{n,j}$ is an efficient estimator of $\theta_{0,j}$ under the assumed model. Motivated by Corollary 4, we conjecture that consistency holds jointly,

$$\max_{1 \le j \le p(n)} \left| \hat{\theta}_{n,j} - \theta_j \right| \xrightarrow{a.s.} 0.$$

We are currently investigating this result as well as the joint asymptotic normality of the vector

$$\left(\sqrt{n} \left(\hat{\theta}_{n,j} - \theta_j \right) : 1 \le j \le p(n) \right).$$

For testing the null hypothesis H_0, Kuelbs and Vidyashankar (2009) propose an estimator based on the sample mean vector $\mathbf{X}_n$. To be precise, fix the dimension p. Assuming regularity conditions, they prove, under H_0,

$$\| \sqrt{n} \left(\bar{\mathbf{X}}_n - \theta_0 \right) \|_\infty \approx \| N_p \left(\mathbf{0}, \mathbf{\Sigma} \right) \|_\infty .$$

They approximate the null distribution with a Monte-Carlo procedure.

We propose an analogous procedure, replacing the center and scaled sample mean with the vector of center and scaled adaptive location estimates

$$\left(\sqrt{n} \left(\hat{\theta}_{n,j} - \theta_{0,j} \right) : 1 \le j \le p \right).$$

This procedure requires estimating $\mathbf{\Sigma}$ in the large p, small n setting which is known to be a difficult problem (Tsai and Chen, 2009). We use the shrinkage based algorithm developed by Strimmer and his students (Schafer and Strimmer, 2005;

Opgen-Rhein and Strimmer, 2007). Using the idea of Ledoit and Wolf (2004), they define a covariance estimator which is guaranteed to be positive definite, even with $p > n$.

6.5 Numerical Implementation

In this section we describe the numerical implementation of our procedure. Because we adopt a component-wise approach it is sufficient to describe the procedure in the univariate case. Note that the last paragraph of this section gives a precise summary of the numerical implementation used for our simulation results.

We recall the basic the univariate problem. Let $\xi_1, ..., \xi_n$ be i.i.d. random variables with symmetric density g and location parameter θ_0. Define the KDE $g_n = \frac{1}{nc_n} \sum_{i=1}^{n} K\left(\frac{x-\xi_i}{c_n}\right)$. The objective function for the estimation problem is given by

$$\max_{\theta \in \mathbb{R}} \int g_n^{1/2}(x) g_n^{1/2}(-x + 2\theta) dx. \tag{6.4}$$

First we discuss the choice of kernel K and window width c_n for the KDE. Silverman (1986) (see Section 3.3) discusses the 'optimal' choice of kernel K and window width c_n in terms of minimizing the (approximate) integrated mean square error between a kernel density estimator and the true density g. His discussion is based on Lemma 4a of Parzen (1962) which provides an approximate expression for the integrated mean square error. This leads to the so-called Silverman rule-of-thumb for window widths

$$c_{n,S}(K) \equiv s_n C(K) n^{-1/5},$$

where $C(K)$ is a constant which depends on the kernel and s_n is a scale estimate.

Because the Epanechnikov kernel possesses certain optimality properties in this context (Silverman, 1986) we set $K(t) = \frac{3}{4\sqrt{5}} \left(1 - t^2/5\right) I \left(|t| < \sqrt{5}\right)$; for this kernel the constant is $C(K) = 2.34$. For s_n we use a standard robust estimator of scale, namely the normalized median absolute deviation (MAD) $s_n = 1.4826\hat{m}_n$, where $\hat{m}_n$ is the sample MAD (Maronna et al., 2006). Putting these terms together, we use the window width $c_n = 2.34(1.4826\hat{m}_n)n^{-1/5} = 3.4693\hat{m}_n n^{-1/5}$.

We now discuss the numerical optimization of (6.4). Adopting the technique of Cheng and Vidyashankar (2006), we use a Monte-Carlo algorithm to approximate the integral in (6.4). Namely,

$$\begin{aligned}\int g_n^{1/2}(x) g_n^{1/2}(-x+2\theta)dx &= \int \sqrt{\frac{g_n(-x+2\theta)}{g_n(x)}} g_n(x)dx \\ &\approx \frac{1}{M} \sum_{j=1}^{M} \sqrt{\frac{g_n(-y_j+2\theta)}{g_n(-y_j)}},\end{aligned}$$

where $y_1, ..., y_n$ are i.i.d. samples from g_n. Using this approximation we replace (6.4) with approximate objective function

$$\max_{\theta \in \mathbb{R}} \sum_{j=1}^{M} \sqrt{\frac{g_n(-y_j+2\theta)}{g_n(-y_j)}}. \tag{6.5}$$

Both Silverman (1986) (see Section 6.4) and Cheng and Vidyashankar (2006) provide algorithms for generating pseudo-random variates from a KDE.

For the simulation results presented in this paper we use following numerical implementation. We use the Epanechnikov kernel with window width $c_n = 3.4693\hat{m}_n n^{-1/5}$, where $\hat{m}_n$ is the sample MAD. The estimator is obtained by using the Monte-Carlo approximation to the objective function (6.5), with $M = 2000$ Monte-Carlo samples. This optimization is carried out in Matlab using the function fminunc with the initial value set as the sample median.

6.6 Simulation Section

In this section we test our methodology using simulated data. All of the experiments are based on 500 simulated data sets. The level of the test is fixed at $\alpha = .05$ and $B = 2000$ samples are used to approximate the null distribution.

We are interested in using the methodology described in Section 4 to test the null hypothesis $H_0 : \mu = \mathbf{0}$. The simulated data sets are generated from a 100 dimensional multivariate normal distribution. To examine the effect of data contamination, we use the following model: $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_9 \sim N_{100}\left(\mathbf{0}, \mathbf{\Sigma}\right)$ and $\mathbf{X}_{10} \sim N_{100}\left(\mu, \mathbf{\Sigma}\right)$, where $\mu = \left(m, m, ..., m\right)^t$. The covariance matrix $\mathbf{\Sigma}$ is estimated from 100 randomly selected genes from the ALL group in the leukemia dataset described by Golub et al. (1999).

We test the null hypothesis using two different procedures: the procedure described in Section 4, which is based on the AMHDE, and the related procedure described in Kuelbs and Vidyashankar (2009), which is based on the sample mean vector $\mathbf{X}_n$. Additionally, we consider two different experiments. In the first experiment (Experiment A) $\mathbf{\Sigma}$ is known. In the second experiment (Experiment B), $\mathbf{\Sigma}$ is estimated using the shrinkage estimator described in Section 4. Experiment A, which is admittedly unrealistic, allows us to study the effect of contamination on the location estimator, without confounding it with the effect on the covariance estimator.

The results of the simulations are displayed in Table 6.1. Both procedures achieve close to nominal size in the absence of contamination. In Experiment A, the AMHDP displays robustness to contamination while the sample mean procedure does not. In Experiment B, both procedures break down in the presence of

Table 6.1: Simulation results for testing $H_0 : \mu = \mathbf{0}$ in the presence of data contamination

	$m = 0$	$m = 1$	$m = 5$	$m = 10$
MLE (Exp A)	0.0520	0.0560	0.3180	0.9920
MHDE (Exp A)	0.048	0.0520	0.07	0.056
MLE (Exp B)	0.046	0.084	0.508	0.97
MHDE (Exp B)	0.056	0.102	0.954	1

contamination. This result is not surprising because we are not using a robust estimator for the covariance matrix.

6.7 Future Research

This section outlines our plan for future work. The results in the simulation section suggest the need to develop a robust, high dimensional covariance estimator. We are currently studying a modified version of the Strimmer shrinkage based covariance estimator. The basic idea is to shrink towards a robust target. In the asymptotic setting, we want to formalize the ideas outlined in Section 4 to prove the joint consistency of our estimator; the next step is to prove asymptotic normality. Finally, we want to study 'efficiency' and 'robustness' in the large p, small n setting.

BIBLIOGRAPHY

Aggarwal, C. C. and Yu, P. (2001). Outlier detection for high dimensional data. In *ACM SIGMOD Conference*.

Altman, R. M. (2007). Mixed hidden markov models: An extension of the hidden markov model to the longitudinal data setting. *Journal of the American Statistical Association*, **102** 201–210.

Alvarez, M. J., Vila-Ortiz, G. J., Salibe, M. C., Podhajcer, O. L. and Pitossi, F. J. (2007). Model based analysis of real-time PCR data from DNA binding dye protocols. *BMC Bioinformatics*, **8**.

Argaç, D. (2004). Testing for homogeneity in a general one-way classification with fixed effects: power simulations and comparative study. *Comput. Statist. Data Anal.*, **44** 603–612.

Arnold, S. F. (1980). Asymptotic validity of F tests for the ordinary linear model and the multiple correlation model. *J. Amer. Statist. Assoc.*, **75** 890–894.

Athreya, K. (1971). A note on a functional equation arising in Galton-Watson branching processes. **8** 589–598.

Athreya, K. and Ney, P. (1972). *Branching Processes.* Springer-Verlag, Berlin.

Bedall, F. K. (1978). Test statistics for simple Markov chains. A Monte Carlo study. *Biometrical Journal*, **20** 41–49.

Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *Ann. Statist.*, **5** 445–463.

Beran, R. (1978). An efficient and robust adaptive estimator of location. *Ann. Statist.*, **6** 292–313.

Bickel, P. J. (1982). On adaptive estimation. *Ann. Statist.*, **10** 647–671.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models.* Johns Hopkins Series in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD.

Brillinger, D. R. (1980). Analysis of variance and problems under time series models. In *Handbook of Statistics (Vol. 1)* (P. R. Krishnaiah, ed.). New York: North-Holland, 237–278.

Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P. and Rubin, E. M. (2000). Microarray expression profiling identifies genes with altered expression in HDL deficient mice. *Genome Research*, **10** 2022 – 2029.

Cheng, A.-l. and Vidyashankar, A. N. (2006). Minimum Hellinger distance estimation for randomized play the winner design. *J. Statist. Plann. Inference*, **136** 1875–1910.

Chow, Y. S. and Teicher, H. (1997). *Probability theory.* 3rd ed. Springer Texts in Statistics, Springer-Verlag, New York. Independence, interchangeability, martingales.

De Iorio, M., Müller, P., Rosner, G. L. and MacEachern, S. N. (2004). An ANOVA model for dependent random measures. *J. Amer. Statist. Assoc.*, **99** 205–215.

Devroye, L. (1983). The equivalence of weak, strong, and complete convergence in l_1 for kernel density estimates. *The Annals of Statistics*, **11** 896–904.

Devroye, L. (1987). *A course in density estimation*, vol. 14 of *Progress in Probability and Statistics*. Birkhäuser Boston Inc., Boston, MA.

Dion, J. (1975). Estimation of the variance of a branching process. **3** 1183–1187.

Dion, J.-P. and Yanev, N. M. (1994). Statistical inference for branching processes with an increasing random number of ancestors. *J. Statist. Plann. Inference*, **39** 329–351.

Dion, J.-P. and Yanev, N. M. (1995). Central limit theorem for martingales in BGWR branching processes with some statistical applications. *Math. Methods Statist.*, **4** 344–358.

Dion, J. P. and Yanev, N. M. (1997). Limit theorems and estimation theory for branching processes with an increasing random number of ancestors. *J. Appl. Probab.*, **34** 309–327.

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. Lecture on August 8, 2000 to the American Mathematical Society "Math Challenges of the 21st Century".

Duby, C. and Rouault, A. (1982). Estimation non paramétrique de l'espérance et de la variance de la loi de reproduction d'un processus de ramification. *Ann. Inst. H. Poincaré Sect. B (N.S.)*, **18** 149–163.

Dudoit, S., Fridlyand, J. and Speed, T. P. (2002a). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Statist. Assoc.*, **97** 77–87.

Dudoit, S. and van der Laan, M. J. (2008). *Multiple testing procedures with applications to genomics*. Springer Series in Statistics, Springer, New York.

Dudoit, S., Yang, Y. H., Callow, M. J. and Speed, T. P. (2002b). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statist. Sinica*, **12** 111–139. Special issue on bioinformatics.

Ferré, F. (1998). *Gene Quantification.* Birkhauser, Boston.

Filzmoser, P., Maronna, R. and Werner, M. (2008). Outlier identification in high dimensions. *Computational Statistics and Data Analysis*, **52** 1694–1711.

Friedman, A. (1982). *Foundations of modern analysis.* Dover Publications Inc., New York. Reprint of the 1970 original.

Goll, R., Olsen, T., Cui, G. and Florholmen, J. (2006). Evaluation of absolute quantitation by nonlinear regression in probe-based real-time PCR. *BMC Bioinformatics*, **7** 107–118.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286** 531–537.

Guttorp, P. (1991). *Statistical Inference for Branching Processes.* Wiley, New York.

Hájek, J. (1970). A characterization of limiting distributions of regular estimates. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, **14** 323–330.

Harris, T. E. (1948). Branching processes. *Ann. Math. Statistics*, **19** 474–494.

Harris, T. E. (2002). *The Theory of Branching Processes.* 2nd ed. Dover Publications, New York.

Hayashi, K. (1990). Mutations induced during polymerase chain reaction. *Technique*, **2** 216–217.

Heyde, C. (1974). On estimating the variance of the offspring distribution in a simple branching process. **3** 421–433.

Hoppe, F. M. (1980). On a Schröder equation arising in branching processes. *Aequationes Math.*, **20** 33–37.

Hubert, M. and Engelen, S. (2004). Robust PCA and classification in biosciences. *Bioinformatics*, **20** 1728–1736.

Hubert, M., Rousseeuw, P. J. and Vanden Branden, K. (2005). ROBPCA: a new approach to robust principal component analysis. *Technometrics*, **47** 64–79.

Hyrien, O. and Zand, M. (2008). A mixture model with dependent observations for the analysis of CSFE-labeling experiments. *Journal of the American Statistical Association*, **103** 222–239.

Ito, P. K. (1980). Robustness of ANOVA and MANOVA test procedures. In *Handbook of Statistics (Vol. 1)* (P. R. Krishnaiah, ed.). New York: North-Holland, 199–236.

Jacob, C. and Peccoud, J. (1998). Estimation of the parameters of a branching process from migrating binomial observations. **30** 948–967.

Jagers, P. and Klebaner, F. (2003a). Random variation and concentration effects in PCR. *Journal of Theoretical Biology*, **224** 299–304.

Jagers, P. and Klebaner, F. (2003b). Random variation and concentration effects in PCR. *J. Theoret. Biol.*, **224** 299–304.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, **29** 295–327.

Kadota, K., Tominaga, D., Akiyama, Y. and Takahashi, K. (2003). Detecting outlying samples in microarray data: A critical assessment of the effect of outliers on sample classification. *Chem-Bio Informatics Journal*, **3** 30–45.

Kesten, H. and Stigum, B. P. (1966). A limit theorem for multidimensional Galton-Watson processes. *Ann. Math. Statist.*, **37** 1211–1223.

Kim, B. S., Kim, I., Lee, S., Kim, S., Rha, S. Y. and Chung, H. C. (2005). Statistical methods of translating microarray data into clinically relevant diagnostic information in colorectal cancer. *Bioinformatics*, **21** 517–528.

Kosorok, M. R. and Ma, S. (2007). Marginal asymptotics for the "large p, small n" paradigm: with applications to microarray data. *Ann. Statist.*, **35** 1456–1486.

Krawczak, M., Reiss, J., Schmidtke, J. and Rosler, U. (1989). Polymerase chain reaction: replication errors and reliability of gene diagnosis. **17** 2197–2201.

Krishnamoorthy, K., Lu, F. and Mathew, T. (2007). A parametric bootstrap approach for ANOVA with unequal variances: fixed and random models. *Comput. Statist. Data Anal.*, **51** 5731–5742.

Krutchkoff, R. G. (1988). One-way fixed effects analysis of variance when the error variances may be unequal. *J. Stat. Comput. Simul.*, **30** 259–271.

Krutchkoff, R. G. (1989). Two-way fixed effects analysis of variance when the error variances may be unequal. *J. Stat. Comput. Simul.*, **32** 177–183.

Kubista, M., Andrade, J., Bengtsson, M., Forootan, A., Jonak, J., Lind, K., Sindelka, R., Sjoback, R., Sjogreen, B., Strombom, L., Stahlberg, A. and Zoric, N. (2006). The real-time polymerase chain reaction. *Molecular Aspects of Medicine*, **27** 95 – 125.

Kuelbs, J. and Vidyashankar, A. N. (2009). Asymptotic inference for high dimensional data. *Ann. Statist.* In Press. Preprint available: http://www.stat.cornell.edu/˜vidyashankar/.

Kulinskaya, E., Staudte, R. G. and Gao, H. (2003). Power approximations in testing for unequal means in a one-way ANOVA weighted for unequal variances. *Comm. Statist. Theory Methods*, **32** 2353–2371.

Lalam, N. (2007). Statistical inference for quantitative polymerase chain reaction using a hidden Markov model: a Bayesian approach. *Stat. Appl. Genet. Mol. Biol.*, **6** Art. 10, 35 pp. (electronic).

Lalam, N. and Jacob, C. (2007). Bayesian estimation for quantification by real-time polymerase chain reaction under a branching process model of the DNA molecules amplification process. *Math. Popul. Stud.*, **14** 111–129.

Lalam, N., Jacob, C. and Jagers, P. (2004a). Modelling the PCR amplification process by a size-dependent branching process and estimation of the efficiency. *Adv. Appl. Probab*, **36** 602–615.

Lalam, N., Jacob, C. and Jagers, P. (2004b). Modelling the PCR amplification process by a size-dependent branching process and estimation of the efficiency. **36** 602–615.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.*, **88** 365–411.

Lee, J., Khuri, A. I., Kim, K. W. and Lee, S. (2007). On the size of the F-test for the one-way random model with heterogeneous error variances. *J. Stat. Comput. Simul.*, **77** 443–455.

Lee, S. and Ahn, C. H. (2003). Modified ANOVA for unequal variances. *Comm. Statist. Simulation Comput.*, **32** 987–1004.

Leung, Y. F. and Cavalieri, D. (2003). Fundamentals of cdna microarray data analysis. *Trends in Genetics*, **19** 649 – 659.

Livak, K. (2001). ABI prism 7700 sequence detection system. user bulletin two. *PE Applied Biosystems*.

Louani, D. (2005). Uniform L_1-distance large deviations in nonparametric density estimation. *Test*, **14** 75–98.

Lu, Y., Liu, P.-Y., Xiao, P. and Deng, H.-W. (2005). Hotelling's T2 multivariate profiling for detecting differential expression in microarrays. *Bioinformatics*, **21** 3105–3113.

Marden, J. I. (1995). *Analyzing and modeling rank data*, vol. 64 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006). *Robust statistics*. Wiley Series in Probability and Statistics, John Wiley & Sons Ltd., Chichester. Theory and methods.

Maruyama, I. (1990). Estimation of errors in polymerase chain reaction. *Technique*, **2** 216–217.

Mathai, A. M. and Provost, S. B. (1992). *Quadratic forms in random vari-*

ables, vol. 126 of *Statistics: Textbooks and Monographs.* Marcel Dekker Inc., New York. Theory and applications.

Ming, X. and Kwok, P. (2003). DNA analysis by fluorescence quenching detection. *Genome Research*, **13** 932–939.

Monis, P., Giglio, S. and Saint, C. (2005). Comparison of SYTO9 and SYBR Green I for real-time polymerase chain reaction and investigation of the effect of dye concentration on amplification and DNA melting curve analysis. *Analytical Biochemistry*, **340** 24–34.

Moser, B. K. (1996). *Linear models.* Probability and Mathematical Statistics, Academic Press Inc., San Diego, CA. A mean model approach.

Mullis, K. B., Ferre, F. and Gibbs, R. A. (1994). *The Polymerase chain reaction.* Birkhauser, Boston.

Mykland, P. A. and Zhang, L. (2006). ANOVA for diffusions and Itô processes. *Ann. Statist.*, **34** 1931–1963.

Nagaev, A. V. (1967). Estimation of the mean number of direct descendants of a particle in a branching random process. *Teor. Verojatnost. i Primenen*, **12** 363–369.

Nedelman, J., Heagerty, P. and Lawrence, C. (1992). Quantitative PCR - procedures and precisions. *Bulletin of Mathematical Biology*, **54** 477 – 502.

Ney, P. and Vidyashankar, A. (2003). Harmonic moments and large deviation rates for supercritical branching processes. **13** 475–489.

Nolan, T., Hands, R. and Bustin, S. (2006). Quantification of mRNA using real-time RT-PCR. *Nature Protocols*, **1** 1559 – 1582.

Oh, J. and Gao, J. (2009). A kernel-based approach for detecting outliers of high-dimensional biological data. *BMC Bioinformatics*, **10** S7.

Opgen-Rhein, R. and Strimmer, K. (2007). Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Statistical Applications in Genetics and Molecular Biology*, **6** Article 9.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.*, **33** 1065–1076.

Piau, D. (2002). Mutation-replication statistics for polymerase chain reactions. *J. Comput. Biol*, **9** 831–847.

Piau, D. (2004). Immortal branching Markov processes: averaging properties and PCR applications. *Ann. Probab.*, **32** 337–364.

Piau, D. (2005). Confidence intervals for nonhomogeneous branching processes and polymerase chain reactions. *Ann. Probab.*, **33** 674–702.

Piau, D. (2008). Asymptotics of posteriors for binary branching processes. **45** 727–742.

Plump, A., Erickson, S., Weng, W., Partin, J., Breslow, J. and Williams, D. (1996). Apolipoprotein A-I is required for cholesteryl ester accumulation in steroidogenic cells and for normal adrenal steroid production. *J. Clin. Invest.*, **97** 2660 – 2671.

Reiss, J., Krawczak, M., Schloesser, M., Wagner, M. and Cooper, D. (1990). The effects of replication errors on the mismatch analysis of PCR-amplified DNA. **18** 973–978.

Rubin, H. and Vere-Jones, D. (1968). Domains of attraction for the subcritical Galton-Watson branching process. *J. Appl. Probability*, **5** 216–219.

Rutledge, R. (2004). Sigmoidal curve-fitting redefines quantitative real-time PCR with the prospective of developing automated high-througput applications. *Nucleic Acids Research*, **32** 178–186.

Schafer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, **4** Article 32.

Schnell, S. and Mendoza, C. (1997). Theoretical description of the polymerase chain reaction. *Journal of Theoretical Biology*, **188** 313–318.

Shieh, A. and Hung, Y. (2009). Detecting outlier samples in microarray data. *Statistical Applications in Genetics and Molecular Biology*, **8** 13.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis.* Monographs on Statistics and Applied Probability, Chapman & Hall, London.

Smith, W. L. and Wilkinson, W. E. (1969). On branching processes in random environments. *Ann. Math. Statist.*, **40** 814–827.

Smyth, G. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**. Article 3.

Smyth, G. K., Thorne, N. and Wettenhall, J. (2003). Limma: Linear Models for Microarray, User's Guide. Software manual available from http://bioinf.wehi.edu.au/limma.

Stoimenova, V. (2005). Robust parametric estimation of branching processes with a random number of ancestors. *Serdica Math. J.*, **31** 243–262.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, **10** 1040–1053.

Strawderman, R. L. (2004). Computing tail probabilities by numerical Fourier inversion: the absolutely continuous case. *Statist. Sinica*, **14** 175–201.

Styan, G. P. (1970). Notes on the distribution of quadratic forms in singular normal variables. *Biometrika*, **57** 567–572.

Sun, F. (1995). The polymerase chain reaction and branching processes. *J. Comput. Biol*, **2** 63–86.

Szabo, A., Boucher, K., Jones, D., Tsodikov, A. D., Klebanov, L. B. and Yakovlev, A. Y. (2003). Multivariate exploratory tools for microarray data analysis. *Biostat*, **4** 555–567.

Tsai, C.-A. and Chen, J. J. (2009). Multivariate analysis of variance test for gene set analysis. *Bioinformatics*, **25** 897–903.

Varmuza, K. and Filzmoser, P. (2008). *Introduction to multivariate statistical analysis in chemometrics.* Taylor & Francis,, Boca Raton.

Williamson, R., Lee, D., Hagaman, J. and Maeda, N. (1992). Marked reduction of high density lipoprotein cholesterol in mice genetically modified to lack apolipoprotein A-I. *Proceedings of the National Academy of Sciences of the United States of America*, **89** 7134–7138.

Yanev, N. M. (1975). The statistics of branching processes. *Teor. Verojatnost. i Primenen.*, **20** 623–633.

Yanev, N. M. (1985). Limit theorems for estimators in Galton-Watson branching processes. *C. R. Acad. Bulgare Sci.*, **38** 683–686.