ESTIMATION BASED ON POOLED DATA IN HUMAN BIOMONITORING AND STATISTICAL GENETICS

LI XIANG

(B.Sc., UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2014

DECLARATION

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Li Xiang

1st May 2014

Thesis Supervisors

Anthony Kuk Yung Cheung Professor; Department of Statistics and Applied Probability, National University of Singapore, Singapore, 117546, Singapore (Main)

Xu Jinfeng Assistant Professor; Division of Biostatistics, Department of Population Health, New York University School of Medicine, New York, NY 10016, USA (Co-supervisor)

Papers and Manuscript

Kuk, A. Y., Li, X., and Xu, J. (2013a). A fast collapsed data method for estimating haplotype frequencies from pooled genotype data with applications to the study of rare variants. *Statistics in medicine*, 32(8):1343–1360.

Kuk, A. Y., Li, X., and Xu, J. (2013b). An em algorithm based on an internal list for estimating haplotype distributions of rare variants from pooled genotype data. *BMC genetics*, 14(1):1–17.

Li, X., Kuk, A. Y., and Xu, J. (2014). Empirical bayes gaussian likelihood estimation of exposure distributions from pooled samples in human biomonitoring. *In second revision: Statistics in medicine*.

Acknowledgements

There are many people who have supported and guided me through the journey. I would like to express my sincere gratitude and appreciation to my supervisor, Professor Anthony Kuk for his unwavering support, continual guidance and many opportunities that broadened my experience in Statistics. I would also like to thank my co-supervisor, Dr. Xu Jinfeng who is very helpful and encouraging. I am thankful to Associate Professors Li Jialiang and David Nott in my pre-qualifying exam committee for providing critical insights and suggestions.

I want to take this opportunity to thank Associate Professor Zhang Jin-Ting for his support in my PhD application. I am thankful to Professor Loh Wei Liem for his kind advice and encouragement. I would like to express special thanks to other faculty members and support staffs. I am grateful to NUS for awarding me the Graduate Research Scholarship to pursue research in my area of interest with financial independence.

I would also like to express my sincere thanks to my classmates and friends, Tian Dechao, Huang Lei and Huang Zhipeng for their friendship and encouragement in the journey. Finally, I am grateful to my family for their moral support, especially my wife Wan Ling for her unconditional love, support and encouragement without which this thesis would not have been possible.

Contents

Summary

Pooling is a cost-effective way to collect data. However, estimation is complicated by the often intractable distributions of the observed pool averages. In this thesis, we consider two applications involving pooled data. The first is to use aggregate data collected from pools of individuals to estimate the levels of individual exposure for various environmental biochemicals. We propose a quasi empirical Bayes estimation approach based on a Gaussian working likelihood which enables pooling of information across different demographic groups. The new estimator out-performs an existing estimator in simulation studies. We consider haplotype frequency estimation from pooled genotype data in our second application. A quick collapsed data estimator is proposed which does not lose much efficiency for rare genetic variants. For more efficient estimates, we propose a way to construct a data-based list of possible haplotypes to be used in conjunction with the expectation maximization (EM) algorithm to make it more feasible computationally. For non-rare alleles, haplotype distributions cannot be estimated well from pooled data, and a sensible strategy is to collect individual as well as pooled genotype data. A calibration type estimator based on the combined data is proposed which is more efficient than the estimator based on individual data alone.

List of Tables

List of Figures

List of Abbreviations

AIC Akaike information criterion.

BIC Bayesian information criterion.

EM Expectation maximization.

GLE Gaussian likelihood estimator.

MCEM Monte Carlo expectation maximization.

MCMC Markov chain Monte Carlo.

MLE Maximum likelihood estimate.

Chapter 1

Introduction

Pooling of samples is a cost effective and often efficient way to collect data. The pooling design allows a large number of individuals from the population to be sampled at reduced analytical costs. Estimation is, however, complicated by the fact that the individual values within each pool are not observed but are only known up to their average. In this thesis, we consider two applications involving pooled data, i.e. human biomonitoring and statistical genetics.

This chapter is organized as follows. Section 1.1 introduces the background of human biomonitoring (section 1.1.1), reviews the existing methods (section 1.1.3) and highlights the focus of this topic (section 1.1.4); Section 1.2 briefly describes the haplotype frequency estimation (section 1.2.1), reviews some existing methods (section 1.2.3) and highlights the focus of this topic (section 1.2.4).

1.1 Human Biomonitoring

1.1.1 Background

Human biomonitoring offers a way to better understand population exposure to environmental chemicals by directly measuring the chemical compounds or their metabolites in human specimens, such as blood and urine (Sexton et al., 2004; Angerer et al., 2007). The early examples of biomonitoring could be traced back to the determination of lead in Kehoe et al. (1933) or benzene metabolites in Yant et al. (1936), which were mainly used to control the exposure to contaminants at the workplace. A more recent example arose when blood and urine samples were taken from rescuers and examined for exposure to potentially toxic smoke from the rubble after the World Trade Center collapse on 11 September 2001 (Erik, 2004). Nowadays, more regular survey studies are conducted in various countries or regions to determine a broad range of internal chemical concentrations in general populations, like the National Health and Nutrition Examination Surveys (NHANES) in the U.S. and the German Environmental Survey (GerES) in Germany. The data from biomonitoring are used to characterize the concentration distributions of compounds among the general population and to identify vulnerable groups with high exposure (Thornton et al., 2002). Uncertainties in characterizing concentrations arise when exposure measurements approach the limit of detection (LOD) or with insufficient volume of material (Caudill, 2010; Caudill et al., 2007b). Despite continuous improvement in analytical techniques, Caudill (2010) pointed out that "the percentage of results below the LOD is not declining and may actually be increasing concurrently with decreasing exposure levels". Another

problem in evaluating environmental exposures is the expense of measuring some compounds as the cost generally increases with the accuracy of the chemical assessment (Sexton et al., 2004). In the U.S., cost varies widely from a few U.S. dollars for lead metals to thousands of U.S. dollars for dioxins and polychlorinated biphenyls (PCBs). When evaluating communities or populations, the cost of biomonitoring can increase exponentially.

Pooling of samples can provide one possible solution to both problems by yielding larger sample volumes and reducing the number of analytic measurements to save cost (Bates et al., 2004, 2005; Caudill, 2011, 2012). A weighted pooled sample design was first implemented in NHANES 2005-06 (Caudill, 2012). The number of chemical measurements required was reduced from 2201 to 228 and hence the study saved approximately $2.78 million at a cost of $1400 per testing. Estimation is, however, complicated by the fact that the individual values within each pool are not observed but are only known up to their average or weighted average. The distribution of such averages is intractable when the individual measurements are log-normally distributed, which is a common and realistic assumption (Caudill, 2010). Furthermore, pooled samples may lose valuable information on dispersion (Bignert et al., 1993) and lead to biased estimates of central tendency (Caudill, 2011). Caudill et al. (2007a) proposed a method to correct the bias of estimates obtained using pooled data from a log-normal distribution. Caudill (2010) extended their method to characterize the population distribution by using percentiles. More recently, Caudill addressed estimation using information from an auxiliary source (Caudill, 2011) and extended the method to a weighted pooled sample design in a special issue of *Statistics in Medicine* (Caudill, 2012). But Caudill's esti-

mator is quite ad hoc, and its latest version (Caudill, 2012) relies on the fitting of two straight lines with unexplained weights to perform some kind of smoothing across demographic groups.

1.1.2 Notation

Suppose individual samples were grouped into n_i pools of equal size K in the i^{th} demographic group, $i = 1, \cdots, d$. Denote by X_{ijk} the pollutant concentration of individual k in the j^{th} pool of the i^{th} demographic group with $Y_{ijk} = \log X_{ijk} \sim N(\mu_i, \sigma_i^2)$ independently, where $i = 1, \cdots, d,\ j = 1, \cdots, n_i,\ k = 1, \cdots, K$. Assume the unweighed average $A_{ij} = \sum_{k=1}^{K} X_{ijk}/K$ is recorded for the j^{th} pool in the i^{th} group. All the methods using unweighed average can be easily extended to unequal weights ω_{ijk}, $A_{ij,\omega} = \sum_{k=1}^{K} \omega_{ijk} X_{ijk}$. The mean α_i and variance β_i^2 of X_{ijk} is given by

$$\alpha_i = \mathrm{E}[X_{ijk}] = \exp\left(\mu_i + \sigma_i^2/2\right), \tag{1.1}$$

$$\beta_i^2 = \mathrm{var}[X_{ijk}] = \exp\left(2\mu_i + \sigma_i^2\right)\left[\exp\left(\sigma_i^2\right) - 1\right] = \alpha_i^2\left[\exp\left(\sigma_i^2\right) - 1\right]. \tag{1.2}$$

For the case of unweighed average, we can obtain the mean and variance of A_{ij}

$$\mathrm{E}[A_{ij}] = \mathrm{E}[X_{ijk}] = \alpha_i, \tag{1.3}$$

$$\mathrm{var}[A_{ij}] = \mathrm{var}[X_{ijk}]/K = \beta_i^2/K. \tag{1.4}$$

1.1.3 Existing methods

In this section, we briefly review the existing methods.

- Caudill et al. (2007a) noticed that the measured value of a pooled

sample A_{ij} was an estimate of $\exp(\mu_i + \sigma_i^2/2)$, based on Equations (1.1) and (1.3), but there was a positive bias when estimating μ_i using $\log A_{ij}$ alone. They proposed a way to correct this bias, which was equal to one-half the variance of the logarithm of the individual samples constituting the pool. The squared coefficient of variation (CV_i^2) of A_{ij} is given by

$$\mathrm{CV}_i^2 = \frac{\mathrm{var}[A_{ij}]}{\mathrm{E}[A_{ij}]^2} = \left[\exp\left(\sigma_i^2\right) - 1\right]/K. \tag{1.5}$$

which could be used to calculate σ_i^2 after estimating CV_i^2. The CV_i^2 can be estimated as the ratio between sample variance and squared sample mean of A_{ij} for each demographic group. Due to the small number of pools in some demographic groups, they estimated $\mathrm{var}[A_{ij}]$ by using the range based on $\mathrm{var}[A_{ij}] = w_K(A_{i,\max} - A_{i,\min})$, where w_K was the factor used to convert an observed range for K samples to a variance estimate on the basis of the distribution of the range of normally distributed samples (Gosset, 1927), and $A_{i,\max}$ and $A_{i,\min}$ were the maximum and minimum values in the i^{th} demographic group respectively. Furthermore, they fit a weighted least squares regression of CV_i on the logarithm of the median in the corresponding demographic group with weights n_i^2. The fitted value $\widehat{\mathrm{CV}}_i$ was used to estimate σ_i^2 according to Equation (1.5). Then the estimate of μ_i was given by the average of the bias-corrected values

$$\hat{\mu}_i = \frac{\sum_{j=1}^{n_i} \log A_{ij}}{n_i} - \frac{\hat{\sigma}_i^2}{2} = \frac{\sum_{j=1}^{n_i} \log A_{ij}}{n_i} - \frac{\log\left(K\widehat{\mathrm{CV}}_i^2 + 1\right)}{2}.$$

However, there is a lack of explanation for the use of weighted least squares and its choice of weights.

- Caudill (2010) extended their method (Caudill et al., 2007a) to char-

acterize the population distribution by using percentiles and also provided formulas of calculating confidence limits around the percentile estimate. The p^{th} percentile for log-normal populations was given by

$$P_{i,p} = \exp\left(\mu_i + f_p \sigma_i^*\right) \tag{1.6}$$

where f_p was the p^{th} percentile of the standard normal distribution. Similar method was used to estimate μ as described in Caudill et al. (2007a), excepting that in this paper he suggested using sample coefficient of variation as a natural estimator instead (Caudill, 2010). He suggested several ways to estimate σ_i^* in the Equation 1.6. One of them was to simply compute the sample standard deviation of the bias-corrected values $\log A_{ij} - \hat{\sigma}_i^2/2$. Two-sided $100(1-\alpha)\%$ confidence limits (LL_P, UL_P) around a percentile estimate was computed by using a noncentral t distribution that can be obtained from Table 1 of Odeh and Owen (1980).

- Caudill (2011) investigated ways to further reduce the bias in the estimation by augmenting variance information from other studies. Similar technique was applied as in Caudill et al. (2007a), by using a weighted least squares regression of CV_i on the logarithm of the median in the corresponding demographic group with weights n_i^2. Augmentation can be made by taking into account the data from other studies or other groups. They found the increase in number of pools may help reduce the bias using the same number of individuals, while the increase in the number of samples in each pool may not.

- More recently, Caudill (2012) extended his own methods to a weighted pooled sample design in a special issue of *Statistics in Medicine*. For simplicity of the presentation, only the case of unweighed average is reviewed

here. In this paper, he slightly changed the assumption of the distribution of individual measurement to $Y_{ijk} = \log X_{ijk} \sim N\left(\mu_{ij}, \sigma_{ij}^2\right)$, with various means and variances for each pool. The bias-corrected values changed to $\log A_{ij} - \hat{\sigma}_{ij}^2/2$, and hence the estimate of μ_i was given by the average of the bias-corrected values

$$\hat{\mu}_i = \frac{\sum_{j=1}^{n_i}\left(\log A_{ij} - \hat{\sigma}_{ij}^2/2\right)}{n_i}$$

According to Equation 1.5, $\hat{\sigma}_{ij}^2 = \log\left(K\widehat{\mathrm{CV}}_{ij}^2 + 1\right)$. $\widehat{\mathrm{CV}}_{ij}^2$ was estimated as the ratio between $\hat{\sigma}_{A_{ij}}$ and A_{ij}, where $\hat{\sigma}_{A_{ij}}$ was the estimated standard deviation of A_{ij}. In order to obtain $\hat{\sigma}_{A_{ij}}$, he fit a weighted least squares regression of logarithm of $\hat{\sigma}_{A_i}$ on the logarithm of the median of A_{ij} in the corresponding demographic group with weights n_i^2, and estimated $\hat{\sigma}_{A_{ij}}$ from the weighted least squares model by the corresponding pool measured value A_{ij}.

Equation (1.6) was used to estimate the percentile. He estimated σ_i^{*2} as the total (i.e. within-pool and among-pool) variance associated with logarithm of the unmeasured individual samples. The within-pool component of the variance was calculated as $\sigma_{i,\text{within}}^2 = \sum_{j=1}^{n_i} \hat{\sigma}_{ij}^2/n_i$ and the between-pool component as the sample variance of the bias-corrected values $\log A_{ij} - \hat{\sigma}_{ij}^2/2$ in the demographic group. Furthermore, he fit another weighted least squares regression of $\log\left(\hat{\sigma}_i^*\right)$ on $\hat{\mu}_i$ with weights n_i^2 and used the estimated $\hat{\sigma}_i^{**}$ from the regression model as input to the percentile estimate $\hat{P}_{i,p} = \exp\left(\hat{\mu}_i + f_p\hat{\sigma}_i^{**}\right)$.

1.1.4 The focus of this topic

Caudill proposed a few ways to characterize the concentration distributions of compounds based on pooled samples (Caudill, 2010, 2011, 2012). However, Caudill's estimator is quite ad hoc, and its latest version (Caudill, 2012) relies on the fitting of two straight lines with unexplained weights to perform some kind of smoothing across demographic groups.

In chapter 2, we propose to replace the intractable distribution of the pool averages by a Gaussian likelihood. An empirical Bayes Gaussian likelihood approach, as well as its Bayesian analogue, are developed to pool information from various demographic groups by a mixed effect formulation. Also discussed are methods to estimate the underlying mean-variance relationship, and to select a good model for the means.

1.2 Haplotype Frequency Estimation

1.2.1 Background

In statistical genetics, the haplotype distribution is the joint distribution of the allele types at, say, L loci. We will focus on bi-allelic loci in this study so that each haplotype vector is a vector of binary values, and the haplotype distribution is a multivariate binary distribution. The importance of haplotypes is well documented (Morris and Kaplan, 2002; Clark, 2004; Schaid, 2004) and reinforced more recently by the works of Muers (2010) and Tewhey et al. (2011). By incorporating linkage disequilibrium information from multiple loci, haplotype-based inference can lead to more powerful tests of genetic association than single-locus analyses. Haplotype distributions are usually estimated from individual genotype data which is

the sum of the maternal and paternal haplotype vectors of an individual. As reviewed by Niu (2004) and Marchini et al. (2006), statistical approaches to haplotype inference based on individual genotype data are effective and cost-efficient. These include the expectation-maximization (EM) type algorithms for finding maximum likelihood estimates (MLE) (Excoffier and Slatkin, 1995), and the Bayesian PHASE algorithm (Stephens and Scheet, 2005). Since DNA pooling is a popular and cost-effective way of collecting data in genetic association studies (Sham et al., 2002; Norton et al., 2004; Meaburn et al., 2006; Homer et al., 2008; Macgregor et al., 2008), the EM algorithm and its variants have been extended by various authors (Ito et al., 2003; Kirkpatrick et al., 2007; Zhang et al., 2008; Kuk et al., 2009) to handle pooled genotype data (i.e., the sum of all $K = 2k$ haplotype vectors of all k individuals in a pool), whereas Pirinen et al. (2008), Gasbarra et al. (2011) and Pirinen (2009) have extended Bayesian algorithms using Markov Chain Monte Carlo (MCMC) or reversible jump MCMC schemes. Also from a Bayesian perspective, Iliadis et al. (2012) conduct deterministic tree-based sampling instead of MCMC sampling, but their algorithm is feasible for small pool sizes only, even though the block size can be arbitrary. Despite the falling costs of genotyping, the popularity of the pooling strategy has not waned, with Kim et al. (2010) and Liang et al. (2012) advocating the use of pooling for next-generation sequencing data. The importance of pooling increases with the recent surge of interest in rare variant analysis based on re-sequencing data (Mardis, 2008) to explain missing heritability (Eichler et al., 2010) and diseases that cannot be explained by common variants. Roach et al. (2011) predict that "haplotypes that include rare alleles ... will play an increasingly important role in

understanding biology, health, and disease". Perhaps more so than in the analysis of common variants, pooling has an important role to play in the analysis of rare variants. This is because the standard methods for testing genetic association are underpowered for rare variants due to insufficient sample size as only a small percentage of study subjects would carry a rare mutation, and pooling is a way to increase the chance of observing a rare mutation. By using a pooling design, we could include more individuals in a study at the same genotyping cost. The study by Kuk et al. (2010) shows that pooling does not lead to much loss of estimation efficiency relative to no pooling when the alleles are rare.

1.2.2 Notation

Focusing on bi-allelic loci, the two possible alleles at each locus can be represented by "1" (the minor or variant allele) and "0" (the major allele). As a result, the alleles at selected loci of a chromosome can be represented by a binary haplotype vector. Since human chromosomes come in pairs, there are 2 haplotype vectors for each individual, one maternal, and one paternal. Suppose we have n pools of k individuals each so that there are $K = 2k$ haplotypes within each pool. Denote by $Y_{ij} = (Y_{1ij}, \cdots, Y_{Lij})'$ the j^{th} haplotype in the i^{th} pool, where $i = 1, \cdots, n$, $j = 1, \cdots, K$, and L is the number of loci to be genotyped. Assuming Hardy-Weinberg equilibrium, the nK haplotype vectors are independent and identically distributed with probability function

$$f(y_1, \cdots, y_L) = P\left(Y_{1ij} = y_1, \cdots, Y_{Lij} = y_L\right)$$

for every L-tuple $y = (y_1, \cdots, y_L)'$ belonging to the Cartesian product $\Omega = \{0,1\}^L$. With pooling, the observed data are the pool totals

$$T_i = \sum_{j=1}^{K} Y_{ij} = \left(\sum_{j=1}^{K} Y_{1ij}, \cdots, \sum_{j=1}^{K} Y_{Lij} \right)' = (T_{1i}, \cdots, T_{Li})', \ i = 1, \cdots, n.$$

The probability function $p(t_1, \cdots, t_L)$ of each pool total is given by the K-fold convolution of the haplotype probability function $f(y_1, \cdots, y_L)$ and so the likelihood based on the observed pooled data is highly intractable and not easy to maximize directly.

In Zhang et al. (2008) and Kuk et al. (2009), Gaussian approximation was applied to the observed pooled genotype data T_i. Denote by the L-tuple $y^{(i)} = \left(y_1^{(i)}, \cdots, y_L^{(i)}\right)'$ the corresponding haplotype i with haplotype frequency $f^{(i)}$. Let $\mathbf{f} = \left(f^{(1)}, \cdots, f^{(r)}\right)'$ be the vector containing frequencies of all possible haplotypes, where $r = 2^L$ is the total number of haplotypes for L loci, and $\omega = (\omega_1, \cdots, \omega_L)'$ be the vector of allele frequencies for allele 1's and Σ_0 be the variance-covariance matrix for the L loci. Multivariate normal distribution was used to approximate the distribution of the pooled genotype data guaranteed by the Central Limit Theorem

$$T_i = (T_{1i}, \cdots, T_{Li})' \,|\mathbf{f} \sim N(\mu_{\mathbf{f}}, \Sigma_{\mathbf{f}}), \quad \text{as } K \to \infty \tag{1.7}$$

where $\mu_{\mathbf{f}} = K\omega, \ \Sigma_{\mathbf{f}} = K\Sigma_0$.

1.2.3 Existing methods

In this section, we briefly review the existing EM-type algorithms to estimate haplotype frequencies, which is the focus of this thesis.

- **Standard EM algorithm**. Excoffier and Slatkin (1995) were the first

to apply standard EM algorithm for individual genotype data, and then Ito et al. (2003) extended EM algorithm to deal with pooled genotype data in the computer program LDPooled. If the individual haplotypes $Y_{ij},\ i = 1, \cdots, n,\ j = 1, \cdots, K$, were actually observed, the complete data MLE of $f(y),\ y \in \Omega$, was given by the sample proportion of haplotype

$$\hat{f}_C(y) = \frac{m(y)}{nK}, \tag{1.8}$$

where $m(y) = \sum_{i=1}^{n} \sum_{j=1}^{K} I\left(Y_{ij} = y\right)$ was the number of times y appears in Y_{ij}. The E-step of the EM algorithm involved taking conditional expectation of $m(y)$ given the observed data and current estimates $\hat{f}^{(t)}(y),\ y \in \Omega$, to get

$$\begin{aligned}
\hat{m}^{(t)}(y) =& \mathrm{E}\left[m(y)|T_1 = t_1, \cdots, T_n = t_n\right] \\
=& \sum_{i=1}^{n} \sum_{j=1}^{K} P\left(Y_{ij} = y|T_i = t_i\right) \\
=& \sum_{i=1}^{n} KP\left(Y_{i1} = y|T_i = t_i\right),
\end{aligned}$$

where

$$\begin{aligned}
P\left(Y_{i1} = y|T_i = t_i\right) =& \frac{P\left(Y_{i1} = y, T_i = t_i\right)}{P\left(T_i = t_i\right)} \\
=& \frac{\sum\limits_{\substack{y_2 \in \Omega, \cdots, y_K \in \Omega \\ y + y_2 + \cdots + y_K = t_i}} \left[\hat{f}^{(t)}(y) \prod_{j=2}^{K} \hat{f}^{(t)}(y_j)\right]}{\sum\limits_{\substack{y_1 \in \Omega, \cdots, y_K \in \Omega \\ y_1 + \cdots + y_K = t_i}} \left[\prod_{j=1}^{K} \hat{f}^{(t)}(y_j)\right]}.
\end{aligned} \tag{1.9}$$

Since the complete data multinomial likelihood belongs to the exponential family, the M-step can be carried out analytically to yield the updating

formula

$$\hat{f}^{(t+1)}(y) = \frac{\hat{m}^{(t)}(y)}{nK}$$

which was just Equation (1.8) with $m(y)$ replaced by the imputed value $\hat{m}^{(t)}(y)$.

Excoffier and Slatkin (1995) derived the approximate estimates of the variance-covariance matrix for large samples by inverting the estimated information matrix. However, they found this approach may not lead to the desired results because the information matrix may be impossible to invert for one of the following reasons: the number of possible haplotypes may be extremely large; some haplotypes may have MLE equal or close to zero; a particular estimated information matrix may be singular or nearly singular even when all haplotypes have nonzero frequencies. In the case of individual genotype data, their method was limited in practice by the number of possible genotypes, which grows exponentially with the haplotype length. They considered only when all individuals were heterozygous for fewer than 16 loci and when the total number of possible haplotypes in the sample did not exceed 16,384.

Ito et al. (2003) applied nonparametric bootstrap method to estimate empirically the standard errors of the frequencies of haplotype. The real data analysis showed that the frequencies of haplotypes could be inferred rather accurately from the pooled DNA data when the frequencies were bigger than 0.1, while the estimated haplotype frequencies with lower frequencies were not reliable as shown by the large standard errors calculated by the bootstrap method. The performance of their program depended on the number of combinations, which increased by a power function of the number of alleles at a locus and also by a factorial of the number of subjects

in a pool. They commented that their program could work for genotype data with 6 loci and pool size 6, 13 loci and pool size 2, or 25 loci and pool size 1 (i.e. individual).

- **HaploPool**. Kirkpatrick et al. (2007) proposed a method of estimating haplotype frequencies from blocks of consecutive single-nucleotide polymorphisms (SNPs). They suggested searching for a set of potential haplotypes of size D, $\mathcal{H}_c = \left\{\widetilde{Y}_d = \left(\widetilde{Y}_{1d}, \cdots, \widetilde{Y}_{Ld}\right)', \ d = 1, \cdots, D\right\}$, with corresponding frequencies $\tilde{f}_d, \ d = 1, \cdots, D$, by using the perfect phylogeny model (Kingman, 1982). Since the tree $\widetilde{T}$ generated from the perfect phylogeny model may not include all the valid haplotypes which were compatible with the observed pooled data, they proposed adding a penalty factor to the likelihood function, called the mutation number of the configuration, which measured how the observed data deviated from the set of haplotypes $\left\{\widetilde{Y}_d, \ d = 1, \cdots, D\right\}$. This mutation number was defined as the difference between the observed data and the generated haplotypes with configuration c_i for pool i, $mut(c_i, i) = \sum_{l=1}^{L} \left| t_{li} - \sum_{d=1}^{D} \tilde{y}_{ld} c_{ld} \right|$, where $c_i = (c_{i1}, \cdots, c_{iD})$. With this definition, the likelihood function can be written as

$$L(t_1, \cdots, t_n | \widetilde{T}) = \prod_{i=1}^{n} \left(\max_{c_i} \epsilon^{mut(c_i, i)} \prod_{d=1}^{D} \frac{\tilde{f}_d^{c_{id}}}{c_{id}!} \right).$$

where ϵ was the given probability for a mutation. A bottom-up dynamic programming algorithm was used on the tree to find the most likely configuration.

However, when the mutation number was large, the observed data cannot be explained by the generated haplotype from the perfect phylogeny model. They suggested using a greedy approach (Halperin and Karp, 2004)

to obtain another potential set of haplotypes $\mathcal{H}_g$. For each pool, this algorithm can provide a valid configuration for all pooled data. Thus, the observed pool can be eventually explained by haplotypes in the configuration. A plausible set of haplotypes needed to be assessed was a combination of sets from the perfect phylogeny model together with the greedy algorithm, $\mathcal{H} = \mathcal{H}_c \cup \mathcal{H}_g$. Then the standard EM algorithm can be applied to this set of haplotypes $\mathcal{H}$.

They expected the number of valid haplotypes D was very small (no more than 20) with small pool sizes (typically be 1, 2 or 3), and therefore their algorithms can run efficiently. So they needed to partition the region into small blocks. Each subset of SNPs was analyzed separately and can be treated as a linear combination of the entire region, in the form of $C_i x = b_i$, where the $\{0, 1\}$ matrix C_i denotes the combination of subset i, the vector b_i denotes the haplotype frequencies of the subset i and the vector x denotes the frequencies of entire haplotypes. The aim is to find

$$x^* = \arg\min_{x \geq 0} \left(C_i x - b_i\right)^2,$$

• **PoooL**. Zhang et al. (2008) proposed a constrained EM algorithm to estimate haplotype frequencies from large pooled genotype data. Calculation of the expected number of haplotypes that are compatible with the pooled genotypes in the Equation (1.9), was the most time-consuming part of the EM algorithm. A multivariate normal distribution was used to approximate the distribution of the pooled genotype T_i. Under the normality assumption (1.7), they showed that Equation (1.9) depended on $\mathbf{f}$ only through ω and Σ_0, which can be estimated in the t^{th} step, as $\hat{\omega}^{(t)} = \sum_{j=1}^{r} \hat{f}^{(j),(t)} y^{(j)}$ and $\hat{\Sigma}_0^{(t)} = \sum_{j=1}^{r} \hat{f}^{(j),(t)} y^{(j)} y^{(j)\prime} - \hat{\omega}^{(t)} \hat{\omega}^{(t)\prime}$, respec-

tively. Then they suggested applying a constrained maximization method to estimate the haplotype frequencies $\mathbf{f}$.

They suggested a computational efficiency for large pools via the use of asymptotic normality of the pooled allele frequencies. Their approach cannot work properly when the number of loci was large. Hence, they suggested to incorporate sliding window method (Yang et al., 2006) and partition-ligation method (Niu et al., 2002) when the number of loci was large.

- **Approximate EM algorithm**. Instead of applying a constrained maximization method (Zhang et al., 2008), Kuk et al. (2009) proposed to revert to the usual EM algorithm to obtain MLE via the use of asymptotic normality of the pooled genotype data. The denominator in the Equation (1.9) can be approximated by the normal density functions

$$P\left(T_i = t_i\right) \approx \Phi\left(T_i;\ K\hat{\omega}^{(t)}, K\hat{\Sigma}_0^{(t)}\right).$$

where Φ is normal density functions. When $y = y^{(i)}$, the numerator can be written as $P\left(Y_{i1} = y^{(i)}, T_i = t_i\right) = P\left(T_i = t_i | Y_{i1} = y^{(i)}\right) \hat{f}^{(i),(t)}$, where

$$P\left(T_i = t_i | Y_{i1} = y^{(i)}\right) \approx \Phi\left[T_i - y;\ (K-1)\hat{\omega}^{(t)}, (K-1)\hat{\Sigma}_0^{(t)}\right].$$

So the ratio in the Equation (1.9) can be approximated by

$$\begin{aligned} P\left(Y_{i1} = y | T_i = t_i\right) =& \frac{P\left(Y_{i1} = y, T_i = t_i\right)}{P\left(T_i = t_i\right)} \\ \approx& \hat{f}^{(i),(t)} \frac{\Phi\left[T_i - y;\ (K-1)\hat{\omega}^{(t)}, (K-1)\hat{\Sigma}_0^{(t)}\right]}{\Phi\left(T_i;\ K\hat{\omega}^{(t)}, K\hat{\Sigma}_0^{(t)}\right)} \end{aligned}$$

Compared with PoooL algorithm (Zhang et al., 2008), the proposed

method was much simpler to implement since there was no need to invoke sophisticated iterative scaling methods. Simulation study showed that the proposed approach lead to estimates with substantially smaller SDs than PoooL while retaining the advantage of computational efficiency over the EM algorithm. Similar to most of other haplotype estimates, the major limitation of this approach was that it cannot work properly when the number of loci was large. Like in Zhang et al. (2008), sliding window method (Yang et al., 2006) or partition-ligation method (Niu et al., 2002) were suggested to be incorporated when the number of loci was very large.

1.2.4 The focus of this topic

Our focus is on computationally fast non-Bayesian methods of estimating haplotype frequencies from individual or pooled genotype data with applications to case-control studies involving rare variants (RVs). There are two main impediments to the use of EM algorithm in estimating haplotype distribution from pooled genotype data. First, the number of putative haplotypes grows exponentially with the number of loci. Secondly, things get worse when pool size increases as the number of individual haplotype configurations compatible with the observed pool totals becomes astronomical quickly. As a result, the EM algorithm can only be applied to data with small to moderate number of markers and pool size.

In chapter 3, we propose a collapsed data MLE that does not suffer from the two aforementioned drawbacks of the EM algorithm. This desirable algorithm is made possible by collapsing the pool total at each marker to just "0" or "at least 1", as carried out in the literature of group testing (Dorfman, 1943; Gastwirth and Hammick, 1989). We show our proposed

method can be calculated very fast regardless of pool size and haplotype length. We provide theoretical and empirical evidence to suggest that the proposed estimation method will not suffer much loss in efficiency if the variants are rare.

However, if the pool size is moderate or large, which is recommended from the cost saving point of view, an estimator based on the original pooled data without collapsing can be substantially more efficient than the collapsed data MLE. This is why we want to modify the EM algorithm for finding the pooled data MLE to make it computationally feasible. Gasbarra et al. (2011) commented that without prior knowledge or restriction on the possible haplotypes, existing algorithms cannot handle the case of 21 loci with pool size 6. We have recorded running times of 1862 and 2900 seconds on an intel (R) Core (TM) desktop when the traditional EM algorithm is applied to pooled genotype data with 12 loci for 74/37 pools of size 2/4 each. Gasbarra et al. (2011) advocate the use of database information to create a list of frequently occurring haplotypes. By combining this idea of using database information to create a list with a normal approximation (Zhang et al., 2008) for the density of the pooled allele frequencies, Pirinen (2009) proposed an AEML (Approximate EM with List) algorithm which runs much faster than the unrestricted EM algorithm.

In chapter 4, we propose using collapsed data list to create an internal list from the data at hand, and then restrict the haplotypes to come from this list only in implementing the EM algorithm. We do not assume the existence of an external list for two reasons. First, database information for rare alleles is currently still lacking. Secondly, an EM type algorithm restricted to a list is sensitive to the correct choice and completeness of the

external list used. Our collapsed data list is shown to have the desirable effect of amplifying the haplotype frequencies. To improve coverage, we propose ways to add and remove haplotypes from the list, and a benchmarking method to determine the frequency threshold for removing haplotypes.

Chapter 2

Human Biomonitoring

This chapter is organized as follows. Section 2.1 highlights the main findings of our method; Section 2.2 describes a group-specific Gaussian likelihood estimator (GLE) and section 2.3 demonstrates the usefulness of pooling by a real data example; Section 2.4 considers an empirical Bayesian Gaussian likelihood approach (EB-GLE) to pool information across demographic groups, by using a mixed effect formulation, followed by an adaptive version of EB-GLE to accommodate a more general mean-variance relationship in section 2.5; Section 2.6 describes a way to select mean model via Gaussian likelihood Akaike Information Criterion (AIC) (Akaike, 1974) and Bayesian Information Criterion (BIC) (Gideon, 1978), and provides further analyses on NHANES 2003-04 data based on various estimators with smoothing across demographic groups; Section 2.7 describes Bayesian analogues of the empirical Bayes Gaussian likelihood estimators; Section 2.8 considers a simulation study and section 2.9 concludes this chapter with some discussion.

The materials presented in this chapter have been submitted to *Statistics in Medicine* for the first revision.

2.1 Summary

Motivated by Caudill's papers, we propose a more efficient method to estimate the log-normal distribution of the concentration using pooled samples. The single measurement from each pool is an average (with equal or unequal weights) of log-normal results, which is approximately normally distributed if the pool size K is large enough by the Central Limit Theorem. So it is tempting to approximate the true distribution by a Gaussian likelihood and use it to obtain estimates. Even though the pool size K is required to be large to justify the Gaussian likelihood approximation, K does not need to be large to produce consistent estimates as the number of pools increases. This is because Gaussian estimation is based on unbiased estimating equations. We further suggest using a mixed effect formulation by treating the means of the log-normal distributions across demographic groups as fixed effects and the squared coefficients of variation as random effects. By assuming a common distribution for the random effects, we are able to use an empirical Bayes approach in conjunction with the Gaussian working likelihood to pool information across demographic groups. Underlying this suggestion of treating the squared coefficients of variation as random effects is the belief that the variance of the exposure distribution is roughly proportional to the square of the mean exposure. More generally, we can postulate that the variance is proportional to the mean raised to a power. We describe a weighted least squares method to estimate the power coefficient from pooled data, which leads to an adaptive version of the empirical Bayes Gaussian likelihood estimator. Gaussian likelihood versions of the AIC and BIC are used as an exploratory tool to select a model of the mean exposure as a function of the demographic variables. One could

also use the selected mean model in place of the saturated model in the proposed quasi empirical Bayes approach. Bayesian analogues of the empirical Bayes Gaussian likelihood estimators can also be obtained using the software JAGS (Plummer, 2003). We use the 2003-04 NHANES exposure data for the biochemical 2,2',4,4',5,5'-hexachlorobiphenyl (PCB153) as the main data set to illustrate all these techniques. The advantage of using the 2003-04 data, which were collected at the individual level, is that we can form our own pools to compare the estimates based on individual and pooled data.

To assess the performance of the various estimators proposed, we apply them to data simulated from 24 log-normal distributions, one for each demographic group, with parameters in each group set to values compatible with the 2003-04 data. The simulation results show that the proposed empirical Bayes Gaussian likelihood estimators outperform Caudill's (Caudill, 2012) estimators for most demographic groups with much smaller bias and better coverage in interval estimation, particularly after bias correction. It also has smaller mean squared error than the group-specific Gaussian likelihood estimator, which highlights the benefit of borrowing strength from other groups. Our study also shows that the reduction in variance which arises from the use of a more parsimonious model of the means is offset by an increase in bias, leading to poor confidence interval coverage in a few groups, and the empirical Bayes estimator based on a saturated mean model actually has better performance. The empirical Bayes Gaussian likelihood estimator and its Bayes analogue perform similarly, but the former is less computing intensive because it does not require Markov chain Monte Carlo (MCMC) sampling, which opens the possibility of further improve-

ment by using the bootstrap to estimate the bias and mean squared error of estimators.

2.2 Gaussian Estimation

In this section, we consider parameter estimation one demographic group at a time, and derive sandwich type variance formulae for the Gaussian likelihood estimators. Smoothing of estimates across demographic groups will be dealt with in section 2.4. Suppose individual samples were grouped into n_i pools of equal size K in the i^{th} demographic group, $i = 1, \cdots, d$. Denote by X_{ijk} the pollutant concentration of individual k in the j^{th} pool of the i^{th} demographic group with $\log X_{ijk} \sim N(\mu_i, \sigma_i^2)$ independently, where $i = 1, \cdots, d,\ j = 1, \cdots, n_i,\ k = 1, \cdots, K$. Assume for the time being that the unweighed average $A_{ij} = \sum_{k=1}^{K} X_{ijk}/K$ is recorded for the j^{th} pool in the i^{th} group. This can be extended to unequal weights (see section 2.9). The probability density of each A_{ij} is given by the K-fold convolution of log-normal densities which is highly intractable and not easy to maximize directly. In principle, one could use the expectation maximization (EM) algorithm (Dempster et al., 1977) to obtain the maximum likelihood estimators, but the conditional expectations of the sufficient statistics under the log-normal assumption

$$\mathrm{E}\left[\sum_{k=1}^{K} \log X_{ijk} \middle| \sum_{k=1}^{K} X_{ijk}\right], \quad \mathrm{E}\left[\sum_{k=1}^{K} (\log X_{ijk})^2 \middle| \sum_{k=1}^{K} X_{ijk}\right]$$

required in the E-step of the EM algorithm have no closed-form expressions. In our real data example in section 2.3, we consider a Monte Carlo implementation of the EM algorithm by approximating the above condi-

tional expectations using simulations at each step. The resulting algorithm is computing intensive and is not amenable to hierarchical modeling to facilitate pooling of information across demographic groups. For these reasons, we will consider the EM algorithm no further in this study and advocate Gaussian estimation instead as detailed below.

According to the Central Limit Theorem, the pool average A_{ij} is approximately normally distributed if the pool size K is large with mean α_i and variance β_i^2/K, $i = 1, \cdots, d$, where

$$\alpha_i = \mathrm{E}\left[X_{ijk}\right] = \exp\left(\mu_i + \sigma_i^2/2\right), \quad \beta_i^2 = \mathrm{var}\left[X_{ijk}\right] = \alpha_i^2\left[\exp\left(\sigma_i^2\right) - 1\right]. \tag{2.1}$$

So it is tempting to approximate the pooled data likelihood by a Gaussian likelihood. This is a special case of Gaussian estimation (Whittle, 1962; Crowder, 1985) applied to pooled data. Whereas Gaussian estimation is usually used for non-Gaussian data, the pool average A_{ij} is asymptotically normally distributed if K is large and so the maximum Gaussian likelihood estimator (GLE) can be expected to be asymptotically equivalent to the pooled data maximum likelihood estimate (MLE) for large pool size K. However, K does not need to be large for the method to produce consistent estimates as the number of pools increases. This is a property of Gaussian estimation. As long as the mean and variance-covariance structures of the model are not misspecified, the score functions of the Gaussian working likelihood yield unbiased estimating equations (Crowder, 2001), and the usual sandwich type standard error estimates can be used to assess the precision of the estimates. The Gaussian likelihood in the present case can be maximized easily to yield $\hat{\alpha}_{i,G} = a_i$ and $\hat{\beta}_{i,G}^2 = Kb_i^2$, where $a_i =$

$\sum_{j=1}^{n_i} A_{ij}/n_i = \bar{A}_{i\cdot}$ and $b_i^2 = \sum_{j=1}^{n_i} \left(A_{ij} - \bar{A}_{i\cdot}\right)^2 / n_i$ are the sample mean and sample variance of the pool averages A_{ij} in the i^{th} demographic group. The GLE of μ_i and σ_i^2 can be obtained by substituting $\hat{\alpha}_{i,G}$ and $\hat{\beta}_{i,G}^2$ into (2.1), which can be inverted to give

$$\begin{aligned} \hat{\sigma}_{i,G}^2 &= \log\left(\hat{\beta}_{i,G}^2/\hat{\alpha}_{i,G}^2 + 1\right), \\ \hat{\mu}_{i,G} &= \log\hat{\alpha}_{i,G} - \hat{\sigma}_{i,G}^2/2 = \log a_i - \hat{\sigma}_{i,G}^2/2. \end{aligned} \tag{2.2}$$

Note that $\hat{\beta}_{i,G}^2/\hat{\alpha}_{i,G}^2 = K(b_i/a_i)^2$, where $(b_i/a_i)^2$ is the square of the sample coefficient of variation. For the purpose of comparison, Caudill's estimator of μ_i in the equal weight case without smoothing across demographic groups is

$$\hat{\mu}_{i,C} = \sum_{j=1}^{n_i} \log A_{ij}/n_i - \hat{\sigma}_{i,G}^2/2 = \log g_i - \hat{\sigma}_{i,G}^2/2$$

where $g_i = \left(\prod_{j=1}^{n_i} A_{ij}\right)^{1/n_i}$ is the geometric mean in the i^{th} demographic group. Comparing $\hat{\mu}_{i,C}$ with (2.2), we can see that $\hat{\mu}_{i,C}$ differs from $\hat{\mu}_{i,G}$ in the use of the geometric mean g_i rather than the arithmetic mean a_i. Since $g_i \leq a_i$, it follows that $\hat{\mu}_{i,C} \leq \hat{\mu}_{i,G}$, which explains the negative bias of $\hat{\mu}_{i,C}$.

We derive the asymptotic variance formulae for the GLE's next. Based on the Gaussian approximation for the distribution of pooled data, we use $N\left(\alpha_i, \beta_i^2/K\right)$ as the working distribution for the pool average A_{ij}. Taking the first order partial derivative of the log Gaussian likelihood based on one pool average $A_{i,1}$ with respect to α_i and β_i^2 yields the score vector $S_{i,1} = \left(S_{\alpha_i}, S_{\beta_i^2}\right)^T$, where $S_{\alpha_i} = K\left(A_{i,1} - \alpha_i\right)/\beta_i^2$ and $S_{\beta_i^2} = K\left(A_{i,1} - \alpha_i\right)^2/(2\beta_i^4) - 1/(2\beta_i^2)$. The variance-covariance matrix of the score

vector $S_{i,1}$ is

$$V_{i,1} = \begin{pmatrix} \frac{K}{\beta_i^2} & \frac{\rho_{i,3}}{2\beta_i^3} \\ \frac{\rho_{i,3}}{2\beta_i^3} & \frac{2K+\rho_{i,4}}{4K\beta_i^4} \end{pmatrix},$$

where $\rho_{i,3} = [\exp(\sigma_i^2) + 2]\,[\exp(\sigma_i^2) - 1]^{1/2}$ and $\rho_{i,4} = \exp(4\sigma_i^2) + 2\exp(3\sigma_i^2) + 3\exp(2\sigma_i^2) - 6$ are the skewness and excess kurtosis of the individual data X_{ijk} in the i^{th} demographic group. Differentiating $S_{i,1}$ and then taking the expectation, we obtain the Hessian matrix

$$H_{i,1} = -\begin{pmatrix} \frac{K}{\beta_i^2} & 0 \\ 0 & \frac{1}{2\beta_i^4} \end{pmatrix}.$$

With n_i pools of data, $H_i = n_i H_{i,1}$ and $V_i = n_i V_{i,1}$, and the sandwich type asymptotic covariance matrix of the maximum Gaussian likelihood estimates of α_i and β_i^2 is given by

$$\mathrm{Cov}(\hat{\alpha}_{i,G}, \hat{\beta}_{i,G}^2) = H_i^{-1} V_i H_i^{-1} = \frac{1}{n_i K} \begin{pmatrix} \beta_i^2 & \beta_i^3 \rho_{i,3} \\ \beta_i^3 \rho_{i,3} & \beta_i^4 (2K + \rho_{i,4}) \end{pmatrix}.$$

Using the delta method, the asymptotic covariance matrix of the GLE of μ_i and σ_i^2 based on pooled samples is given by

$$\begin{aligned} \mathrm{Cov}(\hat{\mu}_{i,G}, \hat{\sigma}_{i,G}^2) =& \begin{pmatrix} \frac{\partial \mu_i}{\partial \alpha_i} & \frac{\partial \mu_i}{\partial \beta_i^2} \\ \frac{\partial \sigma_i^2}{\partial \alpha_i} & \frac{\partial \sigma_i^2}{\partial \beta_i^2} \end{pmatrix} \mathrm{Cov}(\hat{\alpha}_{i,G}, \hat{\beta}_{i,G}^2) \begin{pmatrix} \frac{\partial \mu_i}{\partial \alpha_i} & \frac{\partial \mu_i}{\partial \beta_i^2} \\ \frac{\partial \sigma_i^2}{\partial \alpha_i} & \frac{\partial \sigma_i^2}{\partial \beta_i^2} \end{pmatrix}^T \\ =& \frac{\beta_i^2}{4\alpha_i^2(\alpha_i^2 + \beta_i^2)^2 n_i K} \begin{pmatrix} s_i & t_i \\ t_i & u_i \end{pmatrix}, \end{aligned}$$

where $s_i = \alpha_i^2\beta_i^2(2K + \rho_{i,4}) + (\alpha_i^2 + 2\beta_i^2)(4\alpha_i^2 + 8\beta_i^2 - 4\alpha_i\beta_i\rho_{i,3})$, $u_i = 4\alpha_i^2\beta_i^2(2K+\rho_{i,4})+16\beta_i^4-16\alpha_i\beta_i^3\rho_{i,3}$, and $t_i = -2\alpha_i^2\beta_i^2(2K+\rho_{i,4})+12\alpha_i\beta_i^3\rho_{i,3}+4\alpha_i^3\beta_i\rho_{i,3} - 8\alpha_i^2\beta_i^2 - 16\beta_i^4$.

The 95^{th} percentile of the concentration is $P_{i,95} = \exp(\mu_i + f\sigma_i)$, where $f = 1.645$ is the 95^{th} percentile of the standard normal distribution. The vector of the first order partial derivatives of $P_{i,95}$ with respect to μ_i and σ_i^2 is $d_{P_{i,95}} = (\exp(\mu_i + f\sigma_i), \exp(\mu_i + f\sigma_i)f/(2\sigma_i))^T$. Using the delta method again, we obtain

$$\text{var}\left(\hat{P}_{i,95,G}\right) = d_{P_{i,95}}^T \text{Cov}(\hat{\mu}_{i,G}, \hat{\sigma}_{i,G}^2) d_{P_{i,95}}.$$

2.3 First Analysis of the 2003-04 NHANES Data

In 2001, the U.S. Centers for Disease Control and Prevention (CDC) began to provide an ongoing assessment of the U.S. populations exposure to environmental chemicals by conducting NHANES surveys, and published its findings in the biennial National Report on Human Exposure to Environmental Chemicals. To save cost, a weighted pooled-sample design was first used in NHANES 2005-06, as opposed to the 2003-04 design of taking measurements from individuals. The details of the sampling scheme for NHANES 2005-06 and the estimation method proposed by Caudill has been reported before (Caudill, 2012).

The main example in this study is to use the 2003-04 NHANES data to estimate the exposure levels of PCB153 in 24 demographic groups in the U.S. defined according to race or ethnicity (Non-Hispanic White, Non-Hispanic Black, Mexican American), gender (male, female), and age (12-19 years, 20-39 years, 40-59 years, 60+ years). The 2003-04 data are preferred

over the 2005-06 data because individual data were collected in 2003-04, which allow us to test the log-normal assumption, and provide us with the extra option of pooling the data ourselves to enable comparisons between the individual and pooled data estimates. The data can be obtained online from (http://wwwn.cdc.gov/nchs/nhanes/search/nhanes03_04.aspx).

We will begin by testing the log-normal assumption within the Box-Cox family of distributions. In other words, we assume that the individual measurements taken in each demographic group are normally distributed with group-specific mean and variance after the transformation

$$X^{(\lambda)} = \begin{cases} \frac{X^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \log X & \text{if } \lambda = 0 \end{cases},$$

where the transformation parameter λ is common to all the groups. A point estimate of λ based on the 2003-04 individual PCB153 data is -0.033, and a 95% confidence interval of λ obtained by inverting the likelihood ratio test (i.e., from the profile log-likelihood of λ) is (-0.081, 0.013). Thus the log-normal assumption seems to be supported by the data. We also perform Kolmogorov-Smirnov test and obtain similar results that we cannot reject the log-normal assumption for all the 24 demographic groups (P values, range: 0.229-0.980; median: 0.734), providing further justification for the use of log-normal distribution.

Table 2.1(a) presents the demographic group-specific MLEs of the 95^{th} percentile $P_{i,95}$ $(i = 1, \cdots, 24)$ of individual exposure to PCB153 based on the individual data collected in 2003-04 and the log-normal assumption. We focus on estimation of the 95^{th} percentile in this study since health officials are primarily interested in high exposure, which is also the most

challenging. In comparison, estimation of the mean or median exposure is relatively easier, with little observed difference between different estimators. Also shown in Table 2.1(a) are the group-specific empirical 95^{th} percentiles which do not depend on distributional assumptions and hence are non-parametric. The parametric and non-parametric estimates are reasonably close for most groups. In addition to point estimates, we also report 95% confidence intervals of the 95^{th} percentiles. These confidence intervals are of the form

$$\left(\exp\left[\log \hat{P}_{i,95} - 1.96\,\mathrm{SE}\left(\log \hat{P}_{i,95}\right)\right], \exp\left[\log \hat{P}_{i,95} + 1.96\,\mathrm{SE}\left(\log \hat{P}_{i,95}\right)\right]\right)$$

since the distribution of $\log \hat{P}_{i,95}$ is likely to be less skewed than that of $\hat{P}_{i,95}$. For the empirical 95^{th} percentiles, their standard errors involve the density function which is estimated using kernel method with Gaussian kernel and bandwidth determined by cross-validation.

To demonstrate the usefulness of pooling, we randomly group the individual observations within each demographic group into pools of size 8 each, and we apply the Monte Carlo EM algorithm described briefly in the last section, as well as the Gaussian likelihood method to estimate the parameters separately for each group. The results are summarized in Table 2.1(b). It can be seen that the Gaussian likelihood estimates, which are very simple to compute, are very close to the Monte Carlo EM estimates. This suggests that the Gaussian approximation is adequate in this example with pool size equal to 8. As the Monte Carlo EM algorithm is computing intensive, and not easy to generalize to enable pooling of information across groups, we will consider it no further in this study. With or without pooling, the results suggest that non-Hispanic blacks aged 60 or above,

Table 2.1: Estimates of group-specific 95^{th} percentiles using individual data based on nonparametric method and log-normal assumption, and using pooled data based on Monte Carlo EM (MCEM) and Gaussian likelihood estimator (GLE), with 95% confidence intervals in parentheses.

(a) Individual data

Gender	Race*	Age	$N^{\dagger}$	Nonparametric		Log-normal	
Male	NHW	12-19	77	30.8	(24.4-38.9)	23.4	(17.9-30.5)
		20-39	102	49.4	(41.8-58.5)	40.9	(33.1-50.7)
		40-59	106	113.5	(86.6-148.8)	105.5	(84.4-131.8)
		60+	153	150.3	(136.4-165.6)	168.1	(143.5-196.9)
	NHB	12-19	124	23.3	(16.9-32.2)	23.1	(19.2-27.7)
		20-39	39	65.4	(53.6-79.8)	44.4	(30.6-64.5)
		40-59	44	165.5	(102.6-266.8)	139.7	(99.4-196.1)
		60+	30	441.5	(142.1-1371.8)	532.6	(310.8-912.6)
	MA	12-19	98	17.2	(12.8-23.2)	12.9	(10.7-15.5)
		20-39	44	20.8	(12.8-34.0)	19.7	(15.2-25.5)
		40-59	30	86.3	(35.4-210.1)	61.0	(42.3-88.0)
		60+	37	95.5	(70.5-129.3)	81.3	(62.0-106.6)
Female	NHW	12-19	76	19.5	(14.2-26.7)	17.1	(13.5-21.6)
		20-39	128	42.1	(27.4-64.7)	35.5	(29.3-43.1)
		40-59	101	76.4	(64.1-91.1)	76.4	(64.3-90.8)
		60+	142	145.7	(125.2-169.5)	146.7	(126.8-169.8)
	NHB	12-19	106	20.9	(18.5-23.5)	16.5	(13.6-19.9)
		20-39	46	38.0	(28.7-50.5)	31.6	(23.4-42.6)
		40-59	44	139.0	(53.6-360.5)	119.4	(89.6-159.2)
		60+	31	332.4	(261.6-422.4)	323.2	(230.8-452.7)
	MA	12-19	85	9.4	(7.7-11.6)	8.1	(7.0-9.4)
		20-39	54	15.6	(11.8-20.5)	15.0	(11.9-18.7)
		40-59	32	94.1	(24.1-366.8)	64.1	(40.2-102.2)
		60+	45	110.0	(83.8-144.4)	87.6	(67.1-114.4)

*NHW: non-Hispanic white, NHB: non-Hispanic black, MA: Mexican American.

$^{\dagger}N$: number of individual data.

(b) Pooled data

Gender	Race*	Age	$n^{\dagger}$	MCEM		GLE	
Male	NHW	12-19	9	24.5	(16.1-37.2)	26.1	(15.5-44.0)
		20-39	12	48.4	(33.8-69.3)	48.3	(32.1-72.8)
		40-59	13	113.8	(79.7-162.6)	105.5	(72.5-153.6)
		60+	19	126.6	(104.5-153.5)	123.5	(102.6-148.7)
	NHB	12-19	15	25.8	(19.0-35.0)	25.4	(18.4-35.2)
		20-39	4	48.9	(25.6-93.5)	47.0	(23.7-93.0)
		40-59	5	117.4	(72.4-190.3)	116.7	(71.0-191.9)
		60+	3	341.9	(188.4-620.7)	329.6	(182.4-595.8)
	MA	12-19	12	11.2	(8.8-14.3)	10.8	(8.6-13.7)
		20-39	5	19.7	(12.6-30.8)	18.4	(12.1-28.0)
		40-59	3	35.0	(26.8-45.7)	34.6	(25.9-46.2)
		60+	4	95.6	(52.8-173.1)	102.8	(52.7-200.3)
Female	NHW	12-19	9	22.0	(13.9-34.8)	22.5	(12.6-40.4)
		20-39	16	38.0	(28.1-51.3)	36.7	(26.6-50.5)
		40-59	12	61.7	(49.8-76.6)	61.4	(49.4-76.3)
		60+	17	131.6	(107.1-161.6)	131.1	(106.0-162.1)
	NHB	12-19	13	17.1	(12.7-22.9)	17.6	(12.7-24.4)
		20-39	5	32.9	(19.5-55.4)	33.2	(19.2-57.3)
		40-59	5	129.4	(78.9-212.4)	119.7	(74.3-193.0)
		60+	3	335.2	(176.7-635.9)	335.4	(170.8-658.6)
	MA	12-19	10	7.9	(6.1-10.2)	8.0	(6.1-10.5)
		20-39	6	12.3	(8.9-16.9)	12.1	(8.8-16.6)
		40-59	4	70.7	(34.6-144.3)	71.4	(29.0-175.8)
		60+	5	92.0	(55.8-151.5)	91.1	(55.5-149.4)

*NHW: non-Hispanic white, NHB: non-Hispanic black, MA: Mexican American.

$^{\dagger}n$: number of pools.

regardless of gender, have the highest exposure to PCB153.

2.4 Empirical Bayes GLE

According to the sampling scheme of NHANES 2005-06 (Caudill, 2012), individual samples were collected in 24 demographic groups based on race / ethnicity, gender and age group. These samples were pooled and measured in each group. The number of pools in demographic groups varied depending on the total number of individual aliquots available, with a range from 3 to 17. The estimates may be not very accurate for groups with very few pools. A better estimator can be obtained by borrowing strength from other demographic groups. Hence, Caudill (2012) used weighted linear squares to determine the relationship between different groups and pooled the information in an ad hoc way. We propose using an empirical Bayes approach to pool information across demographic groups by treating certain group-specific parameters as random effects that follow a common distribution.

Since the exact distribution of A_{ij} is intractable, we replace it by a working Gaussian likelihood $\phi(A_{ij}; \alpha_i, \beta_i^2) = \left(\frac{K}{2\pi\beta_i^2}\right)^{1/2} \exp\left[-\frac{K(A_{ij}-\alpha_i)^2}{2\beta_i^2}\right]$. Note that $\beta_i^2 = \gamma_i \alpha_i^2$, where $\gamma_i = \exp(\sigma_i^2) - 1$, $i = 1, \cdots, d$. Instead of estimating separately for each demographic group, we propose using an empirical Bayes approach to estimate α_i and γ_i simultaneously for all demographic groups. We will treat the easier to estimate first moments $\alpha_i = \mathrm{E}[X_{ijk}]$ as fixed effects, and $\gamma_i = (\beta_i/\alpha_i)^2 = \mathrm{CV}_i^2$ as random effects to result in a mixed model. The reason for dividing β_i^2 by α_i^2 is to remove the dependence of $\beta_i^2 = \mathrm{var}[X_{ijk}]$ on $\alpha_i = \mathrm{E}[X_{ijk}]$. Out of convenience, we assume the random effects γ_i follow a conjugate inverse gamma distribution $\pi(\gamma; \kappa, \lambda) = \frac{\lambda^\kappa}{\Gamma(\kappa)} \gamma^{-(\kappa+1)} \exp\left(-\frac{\lambda}{\gamma}\right)$, where $\kappa > 0$ is the shape parameter

and $\lambda > 0$ is the rate parameter. Note that we are using this mixed model merely as a vehicle for combining information across groups. The data that we use in section 2.8 to test the method are not simulated from this mixed model. The joint distribution under this working model is given by

$$\begin{aligned}
&\prod_{i=1}^{d}\left[\prod_{j=1}^{n_i}\phi(A_{ij};\alpha_i,\gamma_i)\right]\pi(\gamma_i;\kappa,\lambda)\\
=&\prod_{i=1}^{d}\left\{\prod_{j=1}^{n_i}\left(\frac{K}{2\pi\gamma_i\alpha_i^2}\right)^{1/2}\exp\left[-\frac{K(A_{ij}-\alpha_i)^2}{2\gamma_i\alpha_i^2}\right]\right\}\frac{\lambda^\kappa}{\Gamma(\kappa)}\gamma_i^{-(\kappa+1)}\exp\left(-\frac{\lambda}{\gamma_i}\right)\\
=&\prod_{i=1}^{d}\left(\frac{K}{2\pi\alpha_i^2}\right)^{n_i/2}\gamma_i^{-(\kappa+n_i/2+1)}\exp\left\{-\left[\frac{K\sum_{j=1}^{n_i}(A_{ij}-\alpha_i)^2}{2\alpha_i^2}+\lambda\right]/\gamma_i\right\}\frac{\lambda^\kappa}{\Gamma(\kappa)}.
\end{aligned}$$

Integrating out the random effects γ_i in the above function yields the marginal likelihood

$$\prod_{i=1}^{d}\left\{\left(\frac{K}{2\pi\alpha_i^2}\right)^{n_i/2}\frac{\Gamma(\kappa+n_i/2)}{\left[\frac{K\sum_{j=1}^{n_i}(A_{ij}-\alpha_i)^2}{2\alpha_i^2}+\lambda\right]^{\kappa+n_i/2}}\frac{\lambda^\kappa}{\Gamma(\kappa)}\right\}.$$

Taking the logarithm of the above yields the marginal log-likelihood function

$$\begin{aligned}
l(\alpha_1,\cdots,\alpha_d,\kappa,\lambda)=\sum_{i=1}^{d}\Bigg\{&2\kappa\log\alpha_i-\left(\kappa+\frac{n_i}{2}\right)\log\left[K\sum_{j=1}^{n_i}(A_{ij}-\alpha_i)^2+2\lambda\alpha_i^2\right]\\
&+\log\Gamma\left(\kappa+\frac{n_i}{2}\right)\Bigg\}+d\kappa\log(2\lambda)-d\log\Gamma(\kappa).
\end{aligned}\tag{2.3}$$

up to an additive constant.

Taking the first order partial derivatives with respect to $\alpha_i,\ i=1,\cdots,d$,

yields

$$\frac{\partial l}{\partial \alpha_i} = \frac{2\kappa}{\alpha_i} - \left(\kappa + \frac{n_i}{2}\right) \frac{-2K \sum_{j=1}^{n_i} (A_{ij} - \alpha_i) + 4\lambda\alpha_i}{K \sum_{j=1}^{n_i} (A_{ij} - \alpha_i)^2 + 2\lambda\alpha_i^2}.$$

Setting $\frac{\partial l}{\partial \alpha_i} = 0$ yields

$$\left(2n_i\lambda + n_i^2 K\right)\alpha_i^2 + \left(2\kappa K \sum_{j=1}^{n_i} A_{ij} - n_i K \sum_{j=1}^{n_i} A_{ij}\right)\alpha_i - 2\kappa K \sum_{j=1}^{n_i} A_{ij}^2 = 0. \tag{2.4}$$

This quadratic equation has two roots and we keep the positive one $\hat{\alpha}_i = \frac{-b+\sqrt{b^2-4ac}}{2a}$ since $\alpha_i > 0$, where $a = 2n_i\lambda + n_i^2 K$, $b = 2\kappa K \sum_{j=1}^{n_i} A_{ij} - n_i K \sum_{j=1}^{n_i} A_{ij}$, $c = -2\kappa K \sum_{j=1}^{n_i} A_{ij}^2$. This positive root is the maxima of the log-likelihood function (2.3) given λ and κ since $\frac{\partial^2 l}{\partial \alpha_i^2}(\hat{\alpha}_i) < 0$. To maximize the log-likelihood with respect to $\theta = (\lambda, \kappa)^T$, we derive the first order derivatives with respect to θ and can obtain the score vector $S_\theta = \left(\frac{\partial l}{\partial \lambda}, \frac{\partial l}{\partial \kappa}\right)^T|_\theta$,

$$\begin{aligned}
\frac{\partial l}{\partial \lambda} &= \sum_{i=1}^{d} \left[\frac{-(2\kappa + n_i)\alpha_i^2}{K \sum_{j=1}^{n_i} (A_{ij} - \alpha_i)^2 + 2\lambda\alpha_i^2}\right] + \frac{d\kappa}{\lambda}, \\
\frac{\partial l}{\partial \kappa} &= \sum_{i=1}^{d} \left\{2\log\alpha_i - \log\left[K \sum_{j=1}^{n_i} (A_{ij} - \alpha_i)^2 + 2\lambda\alpha_i^2\right] + \psi\left(\kappa + \frac{n_i}{2}\right)\right\} + \\
&\qquad d\log(2\lambda) - d\psi(\kappa),
\end{aligned} \tag{2.5}$$

where $\psi(x) = \mathrm{d}\left[\log\Gamma(x)\right]/\mathrm{d}x$ is the first order derivative of the logarithm of the gamma function. To use Newton-Raphson algorithm, we further take the second order derivatives of the log-likelihood function with respect to θ and obtain the observed information matrix $J_\theta = -\begin{pmatrix} \frac{\partial^2 l}{\partial \lambda^2} & \frac{\partial^2 l}{\partial \lambda \partial \kappa} \\ \frac{\partial^2 l}{\partial \lambda \partial \kappa} & \frac{\partial^2 l}{\partial \kappa^2} \end{pmatrix}\Bigg|_\theta$,

with

$$\begin{aligned}
\frac{\partial^2 l}{\partial \lambda^2} &= \sum_{i=1}^{d} \left\{ \frac{(4\kappa + 2n_i)\,\alpha_i^4}{\left[K \sum_{j=1}^{n_i} (A_{ij} - \alpha_i)^2 + 2\lambda\alpha_i^2\right]^2} \right\} - \frac{d\kappa}{\lambda^2}, \\
\frac{\partial^2 l}{\partial \lambda \partial \kappa} &= \sum_{i=1}^{d} \left[-\frac{2\alpha_i^2}{K \sum_{j=1}^{n_i} (A_{ij} - \alpha_i)^2 + 2\lambda\alpha_i^2} \right] + \frac{d}{\lambda}, \\
\frac{\partial^2 l}{\partial \kappa^2} &= \sum_{i=1}^{d} \left[\psi'\left(\kappa + \frac{n_i}{2}\right) \right] - d\psi'(\kappa),
\end{aligned} \tag{2.6}$$

where $\psi'(x) = \mathrm{d}^2 \left[\log \Gamma(x)\right] / \mathrm{d}x^2$ is the second order derivative of the logarithm of the gamma function. The algorithm for solving $S_\theta = 0$ given $\alpha_1, \cdots, \alpha_d$ is as follows:

1. Initialize $\hat{\theta}_r$, $r = 0$;
2. Update $\hat{\theta}_{r+1} = \hat{\theta}_r + J_{\hat{\theta}_r}^{-1} S_{\hat{\theta}_r}$ according to (2.5) and (2.6), and set $r = r + 1$;
3. If the increase in marginal log-likelihood function is small, i.e. $l(\alpha_1, \cdots, \alpha_d, \hat{\theta}_{r+1}) - l(\alpha_1, \cdots, \alpha_d, \hat{\theta}_r) < \epsilon$, stop; Otherwise, back to step 2.

where ϵ is a prespecified tolerance value (e.g. 1e-5). We have described ways to update estimates for $\alpha_1, \cdots, \alpha_d$, λ and κ, and we will present our full algorithm next. EB-GLE (empirical Bayes Gaussian likelihood estimation):

1. Initialize $\hat{\theta}^{(q)} = (\hat{\lambda}^{(q)}, \hat{\kappa}^{(q)})$ and $\hat{\alpha}_i^{(q)}$, $i = 1, \cdots, d$; Set $q = 0$;
2. Update $\hat{\alpha}_i^{(q+1)}$, $i = 1, \cdots, d$, using the positive root obtained from Equation (2.4) given $\hat{\lambda}^{(q)}$ and $\hat{\kappa}^{(q)}$;
3. Update $\hat{\theta}^{(q+1)}$ using Newton-Raphson algorithm given $\hat{\theta}^{(q)}$ and $\hat{\alpha}_i^{(q+1)}$, $i = 1, \cdots, d$;

4. Set $q = q+1$; Stop if $\hat{\theta}^{(q+1)}$ and $\hat{\alpha}_i^{(q+1)}$, $i = 1, \cdots, d$, converge and satisfy the criteria that the increase in marginal log-likelihood function is small, i.e. $l(\hat{\alpha}_1^{(q+1)}, \cdots, \hat{\alpha}_d^{(q+1)}, \hat{\theta}^{(q+1)}) - l(\hat{\alpha}_1^{(q)}, \cdots, \hat{\alpha}_d^{(q)}, \hat{\theta}^{(q)}) < \epsilon$; Otherwise, back to step 2.

Let $\hat{\lambda}_{EB}$, $\hat{\kappa}_{EB}$ and $\hat{\alpha}_{i,EB}$, $i = 1, \cdots, d$, be the resulting empirical Bayes Gaussian likelihood estimates (EB-GLEs), γ_i can be estimated by using the conditional expectation given the data A_{ij}, $j = 1, \cdots, n_i$, and with parameter values set to $\hat{\lambda}_{EB}$, $\hat{\kappa}_{EB}$ and $\hat{\alpha}_{i,EB}$, $i = 1, \cdots, d$. The conditional distribution of γ_i is an inverse gamma distribution, $\Gamma^{-1}\left[\kappa + n_i/2, K\sum_{j=1}^{n_i}(A_{ij} - \alpha_i)^2/(2\alpha_i^2) + \lambda\right]$, and hence γ_i can be estimated by

$$\begin{aligned}\hat{\gamma}_{i,EB} =& E\left[\gamma_i \middle| A_{ij}, j = 1, \cdots, n_i; \hat{\alpha}_{i,EB}, \hat{\lambda}_{EB}, \hat{\kappa}_{EB}\right] \\ =& \frac{K\sum_{j=1}^{n_i}(A_{ij} - \hat{\alpha}_{i,EB})^2/(2\hat{\alpha}_{i,EB}^2) + \hat{\lambda}_{EB}}{\hat{\kappa}_{EB} + n_i/2 - 1}.\end{aligned}$$

Next, μ_i and σ_i can be estimated by substituting $\hat{\alpha}_{i,EB}$ and $\hat{\beta}_{i,EB}^2 = \hat{\gamma}_{i,EB}\hat{\alpha}_{i,EB}^2$ into (2.1), which can be inverted to give

$$\hat{\sigma}_{i,EB} = \sqrt{\log\left(\hat{\gamma}_{i,EB} + 1\right)}, \qquad \hat{\mu}_{i,EB} = \log\hat{\alpha}_{i,EB} - \hat{\sigma}_{i,EB}^2/2.$$

The p^{th} percentile for the i^{th} demographic group can be estimated by

$$\hat{P}_{i,p,EB} = \exp\left(\hat{\mu}_{i,EB} + f_p\hat{\sigma}_{i,EB}\right),$$

where f_p is the p^{th} percentile of the standard normal distribution.

2.5 An Adaptive EB Estimator via Estimating the Mean-Variance Relationship

In section 2.4, we assume that $\gamma_i = \frac{\mathrm{var}\left[X_{ijk}\right]}{\left\{\mathrm{E}\left[X_{ijk}\right]\right\}^2}$, $i = 1, \cdots, 24$, can be treated as identically distributed random effects. This can be considered as a relaxation from constant γ_i to random γ_i. Now, constant γ_i corresponds to the mean-variance relationship $\mathrm{var}\left[X_{ijk}\right] = c\left\{\mathrm{E}\left[X_{ijk}\right]\right\}^2$, which in a way is the natural one for log-normal X_{ijk} because this holds when $\sigma_i^2 = \mathrm{var}\left[\log X_{ijk}\right]$ are constant.

The empirical Bayes Gaussian likelihood estimation method described in section 2.4 can be extended easily to accommodate a more general mean-variance relationship

$$\mathrm{var}\left[X_{ijk}\right] = c\left\{\mathrm{E}\left[X_{ijk}\right]\right\}^{\phi}. \tag{2.7}$$

In fact, the only change is to replace the working Gaussian distribution with $N\left(\alpha_i, \gamma_i \alpha_i^{\phi}/K\right)$. The corresponding marginal log-likelihood is given by

$$\prod_{i=1}^{d}\left\{\left(\frac{K}{2\pi\alpha_i^{\phi}}\right)^{n_i/2}\frac{\Gamma(\kappa+n_i/2)}{\left[\frac{K\sum_{j=1}^{n_i}(A_{ij}-\alpha_i)^2}{2\alpha_i^{\phi}}+\lambda\right]^{\kappa+n_i/2}}\frac{\lambda^{\kappa}}{\Gamma(\kappa)}\right\},$$

which can be maximized to obtain the adaptive version of the estimates of λ, κ and α_i, $i = 1, \cdots, d$, using an algorithm similar to the one described in section 2.4.

We now describe a method to estimate ϕ from pooled data. Taking

logarithm on both sides of (2.7), we have

$$\log\{\text{var}[X_{ijk}]\} = a + \phi \log\{\text{E}[X_{ijk}]\}.$$

Now $\text{E}[X_{ijk}]$ can be estimated by the group average $\bar{A}_i = \sum_{j=1}^{n_i} A_{ij}/n_i$, and $\text{var}[X_{ijk}]$ can be estimated by Ku_i^2, where $u_i^2 = \sum_{j=1}^{n_i} \left(A_{ij} - \bar{A}_i\right)^2/(n_i - 1)$ is the unbiased sample variance of the pool averages A_{ij} in demographic group i. This suggests that ϕ can be estimated by the slope of a weighted least squares regression of $\log(u_i^2)$ on $\log\left(\bar{A}_i\right)$ with weights $n_i - 1$. This choice of weights is suggested by the following argument. If the A_{ij} are truly normal, then $(n_i - 1)u_i^2 = \sum_{j=1}^{n_i} \left(A_{ij} - \bar{A}_i\right)^2 \sim \text{var}[A_{ij}]\,\chi^2_{n_i-1}$, and so $(n_i - 1)^2 \text{var}[u_i^2] = \text{var}[A_{ij}]^2\, 2(n_i - 1)$, which implies $\text{var}[\log(u_i^2)] \approx \frac{\text{var}[u_i^2]}{\text{E}^2[u_i^2]} = \frac{2\,\text{var}^2[A_{ij}]}{(n_i-1)\,\text{var}^2[A_{ij}]} \propto (n_i - 1)^{-1}$.

An adaptive version of the empirical Bayes Gaussian likelihood estimator suggests itself when the weighted least squares estimator $\hat{\phi}$ is used in place of ϕ.

2.6 Further Analysis of the 2003-04 NHANES Data

Based on the pooled version of the 2003-04 data, the weighted least squares estimate of ϕ is 1.843, which is quite near 2. A plot of $\log(u_i^2)$ versus $\log(\bar{A}_i)$ with the weighted least squares line superimposed can be found in Figure 2.1. Thus the estimation method derived in section 2.4 based on a fixed ϕ of 2 should suffice. The resulting empirical Bayes estimates of the 95^{th} percentiles of the 24 demographic groups are displayed in Table 2.2 alongside the estimates obtained using Caudill's method and the group-specific GLEs. Ninety five percent confidence intervals of the form

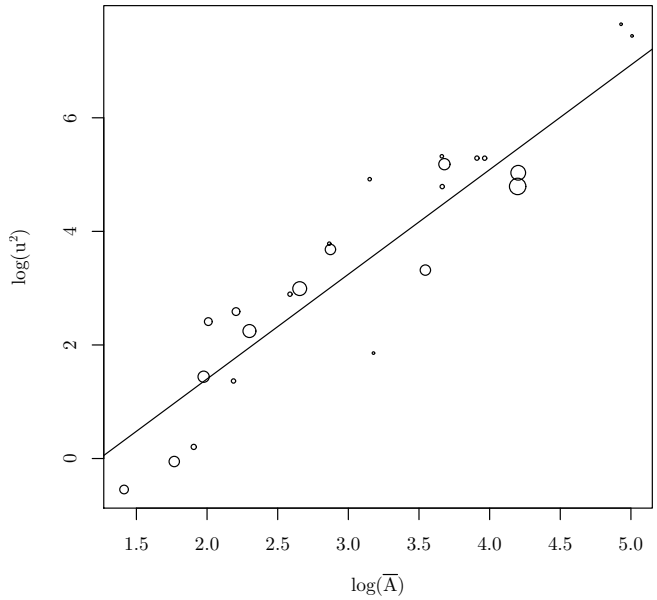

Figure 2.1: Plot of $\log\left(u_i^2\right)$ versus $\log\left(\bar{A}_i\right)$ for the artificially pooled NHANES 2003-04 data. The radius of the circle indicates the relative weight of this data point in the weighted least squares regression and the line represents the weighted least squares fit.

$$\text{CI1} = \left(\exp\left[\log \hat{P}_{i,95} - 1.96\,\text{SE}\left(\log \hat{P}_{i,95}\right)\right], \exp\left[\log \hat{P}_{i,95} + 1.96\,\text{SE}\left(\log \hat{P}_{i,95}\right)\right]\right) \tag{2.8}$$

are given for all estimators. For the group-specific Gaussian likelihood estimates, standard errors are obtained using the formulae given in section 2.2. For the other estimators, parametric bootstrap (Efron, 1979) with 2000 bootstrap replicates is used to obtain the standard errors or root mean squared errors of $\log\left(\hat{P}_{95}\right)$. To be precise, the bootstrap samples are simulated according to log-normal distributions with parameters μ_i and σ_i^2 set to the group-specific GLEs.

We will use the confidence intervals constructed from the group-specific

Table 2.2: Estimates of 95^{th} percentiles using pooled data based on group-specific Gaussian likelihood estimator (GLE), Caudill's estimator (Caudill), empirical Bayes Gaussian likelihood estimator (EB-GLE) and EB-GLE with selected mean model (EB-GLEM), with the 95% confidence intervals (CIs) constructed using three methods.

(a) GLE and Caudill's estimator

					GLE		Caudill			
i	Gender	Race*	Age	$n^{\dagger}$	Estimate	CI1‡	Estimate	CI1‡	CI2‡	CI1-BC‡
1	Male	NHW	12-19	9	26.1	(15.5-44.0)	24.2	(19.1-30.6)	(15.9-36.6)	(22.7-36.4)
2			20-39	12	48.3	(32.1-72.8)	43.8	(36.4-52.9)	(29.6-64.9)	(43.4-63.0)
3			40-59	13	105.5	(72.5-153.6)	90.1	(74.9-108.4)	(56.5-143.6)	(93.2-134.9)
4			60+	19	123.5	(102.6-148.7)	149.3	(130.5-170.9)	(110.3-202.1)	(113.6-148.8)
5		NHB	12-19	15	25.4	(18.4-35.2)	26.7	(22.6-31.6)	(22.1-32.2)	(23.6-33.0)
6			20-39	4	47.0	(23.7-93.0)	43.5	(32.4-58.4)	(28.6-66.2)	(37.8-68.0)
7			40-59	5	116.7	(71.0-191.9)	113.3	(91.1-140.9)	(85.8-149.6)	(99.4-153.9)
8			60+	3	329.6	(182.4-595.8)	307.6	(233.8-404.8)	(216.9-436.3)	(261.0-452.0)
9		MA	12-19	12	10.8	(8.6-13.7)	17.2	(14.6-20.2)	(8.3-35.4)	(10.2-14.1)
10			20-39	5	18.4	(12.1-28.0)	24.6	(20.4-29.7)	(15.9-38.1)	(16.7-24.3)
11			40-59	3	34.6	(25.9-46.2)	60.2	(53.5-67.7)	(23.2-156.1)	(33.0-41.8)
12			60+	4	102.8	(52.7-200.3)	89.6	(67.3-119.3)	(54.7-146.8)	(82.6-146.5)
13	Female	NHW	12-19	9	22.5	(12.6-40.4)	20.0	(15.4-26.1)	(12.2-32.9)	(19.1-32.3)
14			20-39	16	36.7	(26.6-50.5)	36.4	(31.2-42.5)	(28.8-46.0)	(34.2-46.5)
15			40-59	12	61.4	(49.4-76.3)	83.1	(74.2-93.2)	(51.5-134.3)	(58.5-73.5)
16			60+	17	131.1	(106.0-162.1)	149.5	(130.4-171.4)	(121.4-184.1)	(120.4-158.2)
17		NHB	12-19	13	17.6	(12.7-24.4)	20.2	(16.8-24.3)	(16.6-24.7)	(16.2-23.4)
18			20-39	5	33.2	(19.2-57.3)	34.5	(27.2-43.7)	(26.7-44.5)	(28.6-45.9)
19			40-59	5	119.7	(74.3-193.0)	118.9	(95.9-147.5)	(93.1-151.9)	(101.9-156.6)
20			60+	3	335.4	(170.8-658.6)	284.4	(209.3-386.6)	(171.4-471.9)	(257.0-474.8)
21		MA	12-19	10	8.0	(6.1-10.5)	12.6	(10.3-15.5)	(6.2-25.7)	(7.3-10.9)
22			20-39	6	12.1	(8.8-16.6)	19.4	(16.5-22.8)	(9.1-41.3)	(11.3-15.7)
23			40-59	4	71.4	(29.0-175.8)	54.7	(37.4-80.0)	(25.5-117.1)	(52.4-112.0)
24			60+	5	91.1	(55.5-149.4)	90.8	(73.3-112.5)	(70.8-116.5)	(78.2-120.0)
Average length						85.4		36.1	66.6	40.2
$n \geq 9$						30.8		17.6	40.3	17.7
$n \leq 6$						140.0		54.7	92.9	62.6

*NHW: non-Hispanic white, NHB: non-Hispanic black, MA: Mexican American.
$^{\dagger}n$: number of pools.
‡Confidence intervals given by (2.8), (2.9) and (2.10).

GLEs as the benchmark since these estimates are consistent if there is sufficient numbers of pools within each demographic group. To protect against possible bias in the EB-GLE and Caudill's estimates, we also construct confidence intervals of the form

$$\text{CI2} = \left(\exp\left[\log \hat{P}_{i,95} - 1.96\,\text{RMSE}\left(\log \hat{P}_{i,95}\right)\right], \exp\left[\log \hat{P}_{i,95} + 1.96\,\text{RMSE}\left(\log \hat{P}_{i,95}\right)\right]\right) \tag{2.9}$$

where $\text{RMSE}\left(\log \hat{P}_{i,95}\right)$ is the bootstrap estimates of the root mean squared

(b) EB-GLE and EB-GLEM

	EB-GLE				EB-GLEM			
i	Estimate	CI1*	CI2*	CI1-BC*	Estimate	CI1*	CI2*	CI1-BC*
1	25.0	(16.5-38.0)	(15.7-39.9)	(18.3-42.2)	32.2	(24.5-42.4)	(21.1-49.3)	(20.8-35.9)
2	47.0	(33.5-65.9)	(32.3-68.2)	(36.3-71.4)	48.2	(37.1-62.6)	(36.4-63.8)	(39.1-66.0)
3	103.1	(75.3-141.3)	(72.7-146.2)	(81.3-152.6)	98.1	(77.2-124.6)	(71.1-135.3)	(86.2-139.0)
4	137.9	(117.8-161.5)	(116.2-163.7)	(113.8-156.1)	140.1	(118.8-165.3)	(113.2-173.4)	(110.9-154.2)
5	25.1	(19.0-33.2)	(18.7-33.8)	(20.0-35.0)	25.4	(19.9-32.4)	(19.7-32.8)	(20.7-33.8)
6	45.1	(28.2-72.0)	(25.8-78.9)	(33.0-84.2)	50.0	(39.2-63.7)	(39.1-63.8)	(38.6-62.8)
7	119.6	(86.4-165.5)	(84.2-169.8)	(92.5-177.0)	125.6	(99.2-159.0)	(99.2-159.1)	(98.5-157.9)
8	352.7	(253.7-490.2)	(250.2-497.1)	(266.5-515.0)	366.1	(283.6-472.7)	(282.2-475.0)	(276.5-461.0)
9	12.4	(10.4-14.8)	(10.2-15.2)	(10.0-14.2)	12.3	(10.4-14.6)	(10.0-15.2)	(9.7-13.7)
10	20.4	(15.7-26.4)	(15.7-26.4)	(15.6-26.3)	21.9	(18.3-26.3)	(16.5-29.2)	(16.3-23.5)
11	52.5	(46.2-59.8)	(28.3-97.4)	(33.9-43.9)	48.1	(42.1-54.9)	(24.6-94.1)	(30.1-39.2)
12	99.2	(63.2-155.5)	(58.4-168.5)	(73.0-179.4)	108.8	(78.4-150.9)	(77.8-152.0)	(81.1-156.2)
13	21.6	(13.7-34.0)	(12.9-35.9)	(15.4-38.2)	23.9	(17.8-32.2)	(17.8-32.2)	(17.8-32.3)
14	36.2	(27.5-47.5)	(26.9-48.6)	(29.2-50.4)	36.7	(29.2-46.2)	(28.7-46.9)	(30.5-48.3)
15	72.5	(61.4-85.5)	(59.0-89.1)	(57.7-80.3)	70.7	(60.0-83.3)	(53.7-93.1)	(53.6-74.4)
16	143.1	(120.5-169.9)	(120.1-170.4)	(118.5-167.1)	148.6	(123.6-178.6)	(118.0-187.2)	(115.1-166.3)
17	17.6	(13.5-23.0)	(13.2-23.5)	(14.2-24.3)	17.9	(14.1-22.8)	(13.9-23.0)	(14.6-23.6)
18	32.9	(22.6-47.8)	(21.4-50.5)	(25.2-53.3)	38.5	(30.8-48.1)	(28.9-51.3)	(28.1-43.8)
19	124.8	(91.3-170.5)	(89.9-173.3)	(96.2-179.6)	118.2	(91.2-153.1)	(89.7-155.8)	(95.8-160.8)
20	337.6	(226.2-503.9)	(213.2-534.6)	(253.8-565.4)	381.5	(287.4-506.3)	(282.0-516.0)	(272.4-479.9)
21	9.0	(7.3-11.0)	(7.3-11.0)	(7.1-10.7)	9.1	(7.6-11.0)	(7.2-11.6)	(7.0-10.3)
22	14.7	(12.2-17.7)	(11.4-19.0)	(11.2-16.2)	16.7	(14.1-19.7)	(10.1-27.4)	(11.1-15.5)
23	66.6	(35.7-124.2)	(31.3-141.7)	(44.4-154.4)	54.7	(39.8-75.3)	(25.6-116.8)	(56.6-106.9)
24	93.5	(67.3-129.8)	(65.5-133.4)	(72.2-139.2)	106.0	(78.4-143.2)	(77.2-145.5)	(74.6-136.2)
Average length		55.1	64.0	59.9		43.6	53.6	43.2
$n \geq 9$		25.8	28.4	26.7		23.0	29.4	22.6
$n \leq 6$		84.5	99.6	93.0		64.2	77.7	63.7

*Confidence intervals given by (2.8), (2.9) and (2.10).

error (RMSE) of $\log \hat{P}_{i,95}$ as an estimator of $\log P_{i,95}$.

By inspecting Table 2.2, we can see that both Caudill's estimator and EB-GLE lead to confidence intervals shorter than those constructed using group-specific GLE, and this is because of pooling of information across demographic groups. There is a huge difference in length between CI1 and CI2 constructed using Caudill's estimates (the average length is almost doubled from 36.1 to 66.6). As the difference between CI2 and CI1 is due to the use of RMSE instead of SE, this suggests that Caudill's estimates are severely biased and that the resulting CI1s are too short and likely to undercover. In contrast, there is very little difference between CI1 and CI2 based on EB-GLE in 23 out of the 24 groups. The only demographic group with a substantial difference between the two CIs is group 11 (which has only 3 pools of individuals) for which CI1 $= (46.2, 59.8)$ and CI2 $= (28.3, 97.4)$.

Rather than using the RMSE to construct confidence intervals as in (2.9), which is likely to lead to intervals on the wide side, anther strategy is to incorporate bias correction in CI1, leading to

$$\text{CI1-BC} = \left(\exp\left[\log \hat{P}_{i,95} - b - 1.96\,\text{SE}\left(\log \hat{P}_{i,95}\right)\right], \exp\left[\log \hat{P}_{i,95} - b + 1.96\,\text{SE}\left(\log \hat{P}_{i,95}\right)\right]\right) \tag{2.10}$$

where b is the bootstrap estimate of the bias of $\log \hat{P}_{i,95}$ as an estimator of $\log P_{i,95}$. Note that we have ignored the variability of b as an estimator of the bias in (2.10) in the hope that it is relatively smaller than the variance of $\log \hat{P}_{i,95}$. Simulation results to be reported later in section 2.8 suggest that CI1-BC works well for EB-GLE but not for Caudill's estimator. From Table 2.2, we can see that CI1-BC seems to work well for EB-GLE and affects changes only where it matters in correcting the confidence interval from

$(46.2, 59.8)$ to $(33.9, 43.9)$ for group 11 without increasing the width greatly. This is in contrast to the wide CI2 $= (28.3, 97.4)$ obtained when RMSE is used instead of SE. The average length of CI1-BC for EB-GLE is 59.9 which is 30% shorter than the average length of CI1 based on the group-specific GLEs. The average length of CI1-BC for Caudill's estimator is 40.2 which is even shorter, but simulation results to be reported in section 2.8 show that CI1-BC constructed from Caudill's estimator has poor coverage.

Table 2.3: Selection of log-linear model of mean exposure based on pooled 2003-04 NHANES data by Gaussian AIC/BIC*, and parameter estimates under the selected model.

(a) Model selection.

i	Model†	logL‡	AIC*	BIC*
1	gender	-1014.11	2034.22	2044.25
2	race	-979.64	1967.29	1980.66
3	age	-753.87	1513.74	1523.77
4	gender+race	-979.59	1969.19	1985.90
5	gender+age	-749.19	1506.38	1519.75
6	race+age	-697.88	1405.77	1422.48
7	gender×race	-979.26	1972.53	1995.92
8	gender×age	-747.46	1504.92	1521.63
9	race×age	-677.36	1368.71	1392.11
10	gender+race+age	-691.39	1394.78	1414.84
11	age+gender×race	-691.25	1398.50	1425.24
12	race+gender×age	-688.22	1390.45	1413.85
13	gender+race×age	-667.99	1351.99	1378.72
14§	**gender×age+race×age**	-662.91	1343.83	1373.91
15	gender×race+race×age	-667.74	1355.49	1388.91
16	gender×race+gender×age	-688.19	1394.37	1424.45
17	gender×race+gender×age+race×age	-662.73	1347.47	1384.23
18	gender×race×age	-662.22	1350.45	1393.90

*AIC: Akaike information criterion, BIC: Bayesian information criterion.
†Model: '×' indicates interaction with main effects included.
‡logL: Gaussian log-likelihood function.
§Model 14: selected with smallest AIC and BIC.

Thus far, we have treated the means $\alpha_i = \mathrm{E}[X_{ijk}]$, $i = 1, \cdots, 24$, as fixed effects, and $\frac{\mathrm{var}[X_{ijk}]}{\{\mathrm{E}[X_{ijk}]\}^{\phi}}$ with appropriately chosen ϕ (the default is 2) as random effects. A natural next step is to use a more parsimonious model of the means which allows even more pooling of information across demographic groups. Towards this aim, we adopt a generalized linear framework, with log-link for the $\alpha_i = \mathrm{E}[X_{ijk}]$, since they are

(b) Estimates and standard errors (SEs) of the coefficients in the selected model.

	Coefficients in the mean model				
	Estimate	SE		Estimate	SE
Constant	3.2180	0.0484	Age-40	0.0309	0.0018
Gender:Female	-0.1696	0.0468	(Age-40)×Female	0.0058	0.0018
Race:NHB*	0.2308	0.0588	(Age-40)×NHB*	0.0167	0.0027
Race:MA*	-0.5795	0.0846	(Age-40)×MA*	0.0069	0.0034

	Coefficients of Age				
Race*	Male		Race*	Female	
NHW	0.0309	0.0018	NHW	0.0367	0.0019
NHB	0.0476	0.0023	NHB	0.0534	0.0023
MA	0.0378	0.0030	MA	0.0436	0.0033

*NHW: non-Hispanic white, NHB: non-Hispanic black, MA: Mexican American.

(c) Estimates and standard errors (SEs) of the group-specific means

		Male					Female				
			Saturated model		Selected model			Saturated model		Selected model	
Race*	Age	$n^\dagger$	Estimate	SE	Estimate	SE	$n^\dagger$	Estimate	SE	Mean	SE
NHW	12-19	9	9.07	1.15	11.73	0.93	9	7.46	1.05	8.58	0.67
	20-39	12	17.69	1.74	18.06	1.08	16	14.24	1.08	14.34	0.78
	40-59	13	39.64	3.56	33.49	1.44	12	34.65	1.45	29.88	0.90
	60+	19	66.57	2.45	68.10	3.76	17	66.85	2.91	69.48	3.27
NHB	12-19	15	9.97	0.77	9.81	0.67	13	7.21	0.55	7.18	0.47
	20-39	4	17.55	2.87	19.09	0.99	5	13.29	1.70	15.16	0.76
	40-59	5	49.90	5.63	50.62	2.95	5	52.71	5.63	44.10	2.60
	60+	3	149.56	19.47	147.60	14.08	3	138.40	21.58	150.59	14.88
MA	12-19	12	5.86	0.27	5.54	0.26	10	4.11	0.23	4.06	0.22
	20-39	5	8.91	0.79	9.41	0.43	6	6.72	0.41	7.47	0.43
	40-59	3	23.99	1.19	19.29	1.69	4	23.36	5.07	17.49	1.84
	60+	4	38.93	6.20	47.78	7.45	5	39.04	4.38	48.74	8.62

*NHW: non-Hispanic white, NHB: non-Hispanic black, MA: Mexican American.
$^\dagger n$: number of pools.

positive, and with $\text{var}\left[X_{ijk}\right] = c\left\{\text{E}\left[X_{ijk}\right]\right\}^2$, which is supported by the weighted least squares analysis. The quasi AIC/BIC criteria are used as an exploratory tool for model selection based on the Gaussian likelihood computed using the pooled data A_{ij} under various models for $\log \alpha_i = \log\left\{\text{E}\left[X_{ijk}\right]\right\} = \log\left\{\text{E}\left[A_{ij}\right]\right\}$, and with mean-variance relationship given by $\text{var}\left[\bar{A}_{ij}\right] = K^{-1}\,\text{var}\left[X_{ijk}\right] \propto \left\{\text{E}\left[X_{ijk}\right]\right\}^2$. A list of the models considered and their associated Gaussian likelihood AIC/BIC are given in Table 2.3(a). It can be seen that both the AIC and BIC select a model which contains the main effects of race, gender and age, as well as the interaction between age and race, and between age and gender, but no interaction between race

and gender, and no 3-way interaction. The parameter estimates for this selected model are given in Table 2.3(b). It can be seen that the regression coefficients of age are significantly positive under all 6 combinations of race and gender, which suggest that there are age-related accumulation of PCB153. It is also interesting to note that non-Hispanic black females and non-Hispanic black males have the largest two coefficients of age, and this is consistent with the last remark in section 2.3 that non-Hispanic blacks aged 60 or above, regardless of gender, have the highest exposure to PCB153. Table 2.3(c) lists the estimates of the group-specific means obtained under the saturated model and the selected model. The estimates are broadly comparable, but as expected, by using the more parsimonious model selected by AIC/BIC, the estimated standard errors become smaller.

Assuming again an inverse gamma distribution for $\gamma_i = \frac{\mathrm{var}[X_{ijk}]}{\{\mathrm{E}[X_{ijk}]\}^2}$, an empirical Bayes Gaussian likelihood estimator based on the selected mean model can be derived in a way similar to the derivation in section 2.4, where the only difference is in replacing the saturated model for the $\alpha_i = \mathrm{E}[X_{ijk}]$ by the selected model. The empirical Bayes Gaussian likelihood estimates of the 95^{th} percentiles of the 24 demographic groups with the selected mean model incorporated are also shown in Table 2.2 under the name EB-GLEM, together with the associated CI1, CI2 and CI1-BC. As is the case for EB-GLE, CI1 and CI1-BC for EB-GLEM are substantially different only in demographic group 11. The CIs constructed from EB-GLEM are shorter than those based on EB-GLE (43.6 versus 55.1 for CI1; 43.2 versus 59.9 for CI1-BC), due to a more economic modeling of the mean structure.

2.7 Bayesian Estimates

A quasi Bayesian approach could also be used instead of the proposed quasi empirical Bayes approach. We implement this using the software JAGS by specifying independent normal priors with mean 0 and variance 10,000 for the parameters of the log-linear model of the means, and independent gamma priors $\Gamma(0.01, 0.01)$ for the hyper-parameters κ and λ of the inverse gamma mixing distribution of the γ_i. We call this a quasi Bayesian approach because we are treating the A_{ij} as if they are exactly normal. The posterior medians will be used as point estimates, and 95% credible intervals constructed instead of confidence intervals.

An exact Bayesian approach based on the observed pooled data (which is a form of incomplete data) is computing intensive. Akin to exact maximum likelihood based on pooled log-normal data, which requires the use of the EM algorithm to solve a fixed point problem, exact Bayesian inference based on pooled data in this case would require the use of the poor man's data augmentation algorithm by Wei and Tanner (1990) to solve a functional fixed point problem. To digress, consider only one pool and imagine a situation where the posterior distribution $f(\theta|x_1, \cdots, x_K)$ of the generic parameter θ given the complete data $x_1, \cdots, x_K$ is easy to find, but the posterior distribution $f(\theta|\bar{x})$ when only the average $\bar{x}$ of $x_1, \cdots, x_K$ is observed is difficult to obtain. Now

$$\begin{aligned} f(\theta|\bar{x}) &= \int \cdots \int f(\theta|x_1, \cdots, x_K) f(x_1, \cdots, x_K|\bar{x}) dx_1 \cdots dx_K \\ &= \int \cdots \int f(\theta|x_1, \cdots, x_K) \left[\int f(x_1, \cdots, x_K|\bar{x}, \theta^*) f(\theta^*|\bar{x}) d\theta^* \right] dx_1 \cdots dx_K. \end{aligned}$$

Thus $f(\theta|\bar{x})$ is a fixed point to the above functional equation and the poor man's data augmentation algorithm is an iterative simulations based pro-

cedure to solve this functional fixed point problem.

To control the amount of computation at a manageable level, we will not implement the exact Bayes approach. Instead, we compute the quasi Bayes estimates which treat the Gaussian likelihood as if it is the true likelihood. The resulting Bayesian analogue of EB-GLE will be denoted by B-GLE. By making use of MCMC method, which is built into JAGS, we can use other mixing distributions for the random effects γ_i, and not limit ourselves to the conjugate inverse gamma mixing distribution. In our simulation study, we will consider gamma and log-normal mixing distributions in addition to inverse gamma. The flip side of using MCMC method is that it is computing intensive, and so the B-GLE is not amenable to bias correction via the bootstrap.

2.8 Simulation Study

To compare the performance of the various estimators constructed from pooled data, we conduct a simulation study that mimics the NHANES design. Specifically, we simulate $8n_i$ individual samples from demographic group i ($i = 1, \cdots, 24$) according to a log-normal distribution with parameters μ_i and σ_i^2 set to the values estimated by the group specific Gaussian estimation method. These values of μ_i, σ_i^2 and n_i (which varies from 3 to 17) are listed in Table 2.4 and treated as the true parameter values for our simulation study. After simulating the individual values, we group them at random into n_i pools of size 8 each and compute the simple average of each pool, and the pool averages, rather than the underlying individual values, are what we use to construct the various estimates. The methods that we use to construct estimates include (i) group-specific Gaussian

Table 2.4: Mean, percent bias (% bias) and mean squared error (MSE) of the group-specific Gaussian likelihood estimator (GLE), empirical Bayes Gaussian likelihood estimator (EB-GLE) and Caudills estimator of the 95^{th} percentile P_{95} for 24 demographic groups based on 1000 simulations, together with average length (L) and coverage (C) of the 95% confidence intervals (CIs) based on three methods.

(a) Estimation

		TRUE			GLE			EB-GLE			Caudill		
i	n	μ	σ^2	P_{95}	Mean	% bias	MSE	Mean	% bias	MSE	Mean	% bias	MSE
1	9	1.82	0.77	26.08	23.99	-8.02	33.05	23.87	-8.47	35.11	22.04	-15.50	23.69
2	12	2.54	0.66	48.34	45.68	-5.51	71.33	45.41	-6.07	70.28	40.65	-15.90	73.92
3	13	3.38	0.61	105.50	100.30	-4.93	290.11	99.80	-5.40	276.56	85.23	-19.21	473.30
4	19	4.10	0.19	123.54	121.19	-1.90	134.50	127.81	3.46	122.16	141.90	14.86	425.76
5	15	2.03	0.54	25.43	24.21	-4.83	15.64	24.17	-4.95	14.36	24.31	-4.42	5.92
6	4	2.56	0.62	46.96	39.75	-15.37	200.39	41.20	-12.26	143.37	40.58	-13.59	73.97
7	5	3.70	0.41	116.73	104.17	-10.76	786.70	110.33	-5.48	439.02	107.22	-8.14	239.78
8	3	4.84	0.34	329.64	275.03	-16.57	9761.52	320.33	-2.82	3482.25	300.02	-8.99	2664.12
9	12	1.68	0.18	10.83	10.42	-3.82	1.70	11.32	4.50	1.30	15.54	43.48	23.89
10	5	2.05	0.27	18.39	16.69	-9.24	14.76	18.53	0.78	6.49	22.44	22.03	21.08
11	3	3.15	0.06	34.60	31.42	-9.20	30.54	47.00	35.83	163.83	56.01	61.89	470.06
12	4	3.36	0.59	102.76	87.68	-14.68	982.57	91.36	-11.09	665.10	84.78	-17.50	488.54
13	9	1.57	0.89	22.52	20.85	-7.38	27.02	20.73	-7.94	31.46	18.30	-18.72	24.09
14	16	2.38	0.55	36.67	35.14	-4.18	29.40	35.01	-4.54	27.12	33.62	-8.33	17.01
15	12	3.47	0.16	61.35	59.23	-3.46	49.03	65.40	6.60	48.99	77.94	27.04	296.83
16	17	4.09	0.23	131.08	127.37	-2.84	196.18	133.16	1.58	139.51	141.99	8.32	222.99
17	13	1.74	0.47	17.61	16.78	-4.69	7.49	16.88	-4.16	6.17	18.32	4.05	3.56
18	5	2.34	0.50	33.19	30.05	-9.46	72.92	31.03	-6.51	50.03	32.02	-3.53	16.27
19	5	3.78	0.38	119.74	108.34	-9.52	795.68	115.99	-3.13	419.24	113.60	-5.13	202.64
20	3	4.70	0.46	335.43	271.93	-18.93	11503.45	304.13	-9.33	5472.89	276.43	-17.59	5563.17
21	10	1.30	0.22	7.97	7.65	-4.01	1.19	8.24	3.40	0.77	11.33	42.15	12.72
22	6	1.82	0.17	12.10	11.36	-6.09	3.73	13.27	9.63	2.99	17.66	45.89	33.19
23	4	2.69	0.92	71.38	59.74	-16.30	553.57	60.45	-15.32	535.94	51.33	-28.09	492.23
24	5	3.46	0.41	91.06	81.80	-10.17	441.56	86.49	-5.02	239.89	85.85	-5.72	120.49
Average						8.41*	1083.50		7.43*	516.45		19.17*	499.55
$n \geq 9$						4.63*	71.39		5.09*	64.48		18.50*	133.64
$n \leq 6$						12.19*	2095.62		9.77*	968.42		19.84*	865.46

*Average of absolute % bias.

likelihood estimation (GLE), (ii) Caudill's method, (iii) empirical Bayes Gaussian likelihood estimation (EB-GLE) under a saturated mean model and $\phi = 2$, (iv) the adaptive version where ϕ is estimated by the weighted least squares method described in section 2.5, (v) EB-GLE under the mean model selected in Table 2.3 and $\phi = 2$, (vi) the Bayes analogue of (iii) using various mixing distributions for the γ_i, and finally (vii) the Bayes analogue of (v).

The first comparison that we would like to make is between the group-

(b) Confidence interval

	GLE		EB-GLE						Caudill			
	CI1*		CI1*		CI1-BC*		CI2*		CI1-BC*		CI2*	
i	L	C	L	C	L	C	L	C	L	C	L	C
1	25.09	0.86	19.53	0.82	21.46	0.86	21.50	0.83	11.62	0.68	20.42	0.85
2	37.35	0.91	30.91	0.88	33.16	0.91	33.38	0.90	17.68	0.69	36.32	0.84
3	74.90	0.90	63.47	0.88	67.56	0.91	68.08	0.89	38.14	0.71	85.57	0.83
4	44.23	0.91	41.32	0.95	41.20	0.92	42.25	0.96	33.11	0.81	74.48	0.87
5	15.70	0.90	13.65	0.88	14.35	0.90	14.46	0.89	8.62	0.72	13.70	0.95
6	52.49	0.76	35.74	0.71	41.06	0.72	43.22	0.78	21.49	0.54	42.89	0.90
7	98.77	0.79	74.06	0.82	79.86	0.76	85.15	0.87	45.61	0.58	90.03	0.94
8	287.19	0.66	187.52	0.79	197.61	0.59	290.21	0.97	142.73	0.50	335.07	0.94
9	4.64	0.87	4.19	0.96	4.15	0.92	4.47	0.98	3.64	0.81	22.32	0.94
10	13.09	0.79	9.96	0.91	10.19	0.79	11.88	0.98	6.45	0.59	26.87	0.92
11	14.05	0.64	12.17	0.01	9.37	0.63	61.02	0.90	7.09	0.48	140.92	0.98
12	112.69	0.74	77.35	0.71	88.36	0.71	93.44	0.78	47.38	0.52	91.55	0.84
13	24.47	0.88	18.56	0.84	20.55	0.87	20.48	0.86	11.21	0.71	19.86	0.83
14	22.27	0.90	19.49	0.87	20.48	0.90	20.65	0.88	11.58	0.68	21.15	0.90
15	24.68	0.88	22.64	0.93	22.14	0.89	25.02	0.95	14.02	0.69	73.43	0.87
16	52.85	0.90	48.52	0.96	48.83	0.93	49.56	0.97	36.59	0.77	62.38	0.93
17	10.88	0.89	9.44	0.88	9.92	0.89	10.00	0.89	6.52	0.73	10.49	0.96
18	32.49	0.82	24.03	0.80	26.66	0.78	27.63	0.83	14.29	0.58	28.81	0.97
19	99.83	0.81	75.12	0.86	80.37	0.79	87.25	0.93	46.48	0.57	97.52	0.95
20	327.78	0.68	209.81	0.70	231.75	0.61	295.81	0.88	153.58	0.51	315.27	0.85
21	4.03	0.89	3.51	0.96	3.53	0.91	3.73	0.98	3.33	0.84	16.32	0.92
22	6.64	0.82	5.42	0.92	5.22	0.83	7.26	0.97	4.01	0.68	30.30	0.91
23	107.91	0.79	68.39	0.71	82.45	0.76	82.70	0.75	39.60	0.58	69.74	0.71
24	77.20	0.80	57.92	0.84	62.42	0.78	66.69	0.89	34.86	0.59	71.53	0.96
Average	65.47	0.82	47.20	0.82	50.94	0.81	61.08	0.90	31.65	0.65	74.87	0.90
$n \geq 9$	28.42	0.89	24.60	0.90	25.61	0.90	26.13	0.91	16.34	0.74	38.04	0.89
$n \leq 6$	102.51	0.76	69.79	0.73	76.28	0.73	96.02	0.88	46.96	0.56	111.71	0.91

*Confidence intervals given by (2.8), (2.9) and (2.10).

specific GLE (which does not pool information across different demographic groups) with the other methods, which all attempt to pool information across groups in some way. To be concrete, we compare the group-specific GLE with EB-GLE based on the saturated mean model and $\phi = 2$. We can see from Table 2.4 that the EB-GLE, which borrows strength from other groups has smaller MSE than the group-specific GLE in 21 of the 24 demographic groups for estimating P_{95}. Across all groups, the average MSE is 516.45 for EB-GLE, which is much smaller than the average MSE of 1083.50 for the group-specific GLEs. Next, we compare the 2 estimators in terms of length and coverage of the confidence interval CI1 of P_{95} given by (2.8). From Table 2.4, we see that the average length (over all simulations as well as across the 24 demographic groups) of the confidence intervals constructed using group-specific GLE is 65.47 with average coverage 0.82 (0.89 over the 12 groups with at least 9 pools of individuals, 0.76 over the 12 groups with 6 or less pools). For the sake of brevity, we will denote this coverage breakdown in the format 0.82 (0.89, 0.76) from now on. Thus there is undercoverage, and the primary reason seems to be that of insufficient number of pools in some groups (only 3 pools each in 3 groups, and 4 pools each in another 3 groups). For the confidence intervals constructed using EB-GLE, the average length is 47.2 (compared with 65.47 for group-specific GLE) and the coverage remains at 0.82 (0.90, 0.73). Looking at the results more closely, the coverage is only 0.01 for group 11, which has only 3 pools of data. This is caused by the 36% bias of EB-GLE in estimating P_{95} for this group. This is in deep contrast to the less than 10% bias in absolute value of EB-GLE in 20 of the remaining 23 groups. A plausible explanation for an occasional big bias of EB-GLE is this. Think of EB-GLE as some

kind of shrinkage estimator, it is to be expected that shrinkage will not be beneficial universally across all groups, and so it is not inconceivable that while the EB-GLE can improve the estimation in most groups, it is at the expense of poorer performance in one or a few groups. One remedy is the following. Since we are using the bootstrap to estimate the standard error of $\log\left(\hat{P}_{95}\right)$ anyway, we can use the same bootstrap samples to estimate the bias of $\log\left(\hat{P}_{95}\right)$ in estimating the true $\log\left(P_{95}\right)$, which leads us to the bias-corrected confidence interval CI1-BC given by (2.10). The results of the bias-corrected confidence intervals based on EB-GLE are also shown in Table 2.4. Compared with the non-bias corrected version, the average length is increased slightly to 50.94, and the average coverage is maintained at 0.81 (0.90, 0.73). In particular, the coverage in group 11 is improved from 0.01 to 0.63, which is comparable to the coverage of 0.64 achieved by the group-specific GLE in that group, and this is achieved by intervals shorter in length on the average (9.37 versus 14.05). Thus the bias-corrected confidence interval of P_{95} based on EB-GLE seems to perform quite well. Rather than estimating the bias, another possibility is to use the bootstrap estimate of RMSE of $\log\left(\hat{P}_{95}\right)$ instead of the bootstrap standard error to result in CI2 given by (2.9), which has superior coverage 0.90 (0.91, 0.88), but the average length (over all 24 groups) increased slightly to 61.08. Even for demographic group 11, the coverage is improved to 0.90, but this is not very useful, because the true P_{95} for this group is only 34.60, yet the average half-length of the intervals is 30.51 when RMSE is used instead of SE.

The next thing to look at is to see if EB-GLE based on, say, the saturated mean model and $\phi = 2$ is better than Caudill's estimator. Comparing these two estimators of P_{95}, EB-GLE is less biased than Caudill's estima-

tor in 21 out of the 24 demographic groups. The average (over 24 groups) absolute percent bias is 19.17% for Caudill's estimator, and only 7.43% for EB-GLE. Caudill's estimator has small variance, and as a result the associated confidence intervals are short, but due to its huge bias, the coverage 0.52 (0.49, 0.55) is very poor. This is not shown in Table 2.4 to save space. After bootstrap bias correction, we can see from Table 2.4 that the average coverage increases only slightly to 0.65 (0.74, 0.56), which is a sign that the variability of the bootstrap estimate of bias cannot be ignored. If we use RMSE instead of SE, then the resulting CI2 confidence intervals of P_{95} based on Caudill's estimator has average coverage 0.90 (0.89, 0.91), but the average length of 74.87 is longer than the average length of 61.08 for CI2 based on EB-GLE, which has similar average coverage.

Next, we compare EB-GLE with $\phi = 2$ with the adaptive version where ϕ is estimated from the data using the weighted least squares method. The mean of $\hat{\phi}$ over the 1000 simulations is 1.831 with standard error 0.125, which is compatible with the value $\phi = 2$ suggested by Figure 2.1. It is re-assuring to see from Table 2.5 that the weighted least squares method seems to do a good job in estimating ϕ, and that the adaptive EB-GLE (AEB-GLE) performs only slightly worse than EB-GLE based on $\phi = 2$. For the bias-corrected confidence intervals, the average length when ϕ is estimated is 53.47, slightly longer than the average length of 50.94 when ϕ is fixed at 2, and achieve similar coverage. We have also tried EB-GLE with a mis-specified $\phi = 1$. The results are poor and will not be shown to save space. Thus the choice of ϕ matters. As commented earlier, $\phi = 2$ is the natural choice when the exposure distributions are log-normal. If there is any doubt, we can estimate ϕ from the pooled data at hand and compute

Table 2.5: Mean, percent bias (% bias) and mean squared error (MSE) of the empirical Bayes Gaussian likelihood estimator (EB-GLE), adaptive empirical Bayes Gaussian likelihood estimator (AEB-GLE) and empirical Bayes Gaussian likelihood estimator with selected mean model (EB-GLEM) of the 95^{th} percentile P_{95} for 24 demographic groups based on 1000 simulations, together with average length (L) and coverage (C) of the 95% confidence intervals (CIs) based on three methods.

(a) Estimation

			EB-GLE			AEB-GLE			EB-GLEM		
i	n	P_{95}	Mean	% bias	MSE	Mean	% bias	MSE	Mean	% bias	MSE
1	9	26.08	23.87	-8.47	35.11	22.39	-14.14	37.16	30.91	18.52	43.22
2	12	48.34	45.41	-6.07	70.28	41.49	-14.17	102.87	46.24	-4.34	41.65
3	13	105.50	99.80	-5.40	276.56	88.69	-15.93	557.93	95.87	-9.13	211.14
4	19	123.54	127.81	3.46	122.16	116.15	-5.98	250.74	132.40	7.17	204.92
5	15	25.43	24.17	-4.95	14.36	22.52	-11.47	17.87	24.40	-4.06	10.35
6	4	46.96	41.20	-12.26	143.37	39.19	-16.55	143.67	47.79	1.76	39.05
7	5	116.73	110.33	-5.48	439.02	101.88	-12.72	568.62	118.30	1.35	224.68
8	3	329.64	320.33	-2.82	3482.25	291.05	-11.71	4554.78	342.70	3.96	2392.43
9	12	10.83	11.32	4.50	1.30	11.29	4.23	1.22	11.52	6.38	1.49
10	5	18.39	18.53	0.78	6.49	18.44	0.30	5.07	20.62	12.13	8.60
11	3	34.60	47.00	35.83	163.83	46.69	34.95	155.96	48.50	40.18	203.73
12	4	102.76	91.36	-11.09	665.10	84.71	-17.57	760.26	100.52	-2.18	331.91
13	9	22.52	20.73	-7.94	31.46	19.50	-13.40	31.30	22.84	1.43	12.23
14	16	36.67	35.01	-4.54	27.12	32.10	-12.48	40.45	35.34	-3.64	19.22
15	12	61.35	65.40	6.60	48.99	61.73	0.62	37.93	69.00	12.46	93.28
16	17	131.08	133.16	1.58	139.51	120.44	-8.12	361.79	140.91	7.50	267.50
17	13	17.61	16.88	-4.16	6.17	16.05	-8.84	6.60	17.16	-2.55	4.88
18	5	33.19	31.03	-6.51	50.03	29.71	-10.50	46.19	37.04	11.60	34.53
19	5	119.74	115.99	-3.13	419.24	107.05	-10.60	513.50	115.57	-3.48	265.11
20	3	335.43	304.13	-9.33	5472.89	274.94	-18.03	7339.06	357.55	6.59	3446.00
21	10	7.97	8.24	3.40	0.77	8.32	4.37	0.79	8.61	8.05	1.05
22	6	12.10	13.27	9.63	2.99	13.47	11.34	3.40	15.43	27.48	12.87
23	4	71.38	60.45	-15.32	535.94	56.04	-21.50	565.90	51.16	-28.32	466.98
24	5	91.06	86.49	-5.02	239.89	80.42	-11.68	294.21	96.83	6.33	260.99
Average				7.43*	516.45		12.13*	683.22		9.61*	358.24
$n \geq 9$				5.09*	64.48		9.48*	120.55		7.10*	75.91
$n \leq 6$				9.77*	968.42		14.79*	1245.89		12.11*	640.57

*Average of absolute % bias.

(b) Confidence interval

	EB-GLE				AEB-GLE				EB-GLEM					
	CI1-BC*		CI2*		CI1-BC*		CI2*		CI1*		CI1-BC*		CI2*	
i	L	C	L	C	L	C	L	C	L	C	L	C	L	C
1	21.46	0.86	21.50	0.83	20.45	0.85	23.47	0.84	19.02	0.83	15.52	0.81	40.16	0.95
2	33.16	0.91	33.38	0.90	34.72	0.92	42.68	0.91	24.27	0.91	25.32	0.84	27.85	0.93
3	67.56	0.91	68.08	0.89	81.63	0.95	102.03	0.95	42.87	0.84	46.47	0.78	56.62	0.86
4	41.20	0.92	42.25	0.96	62.66	0.97	80.98	0.98	44.26	0.91	42.48	0.92	50.58	0.95
5	14.35	0.90	14.46	0.89	13.97	0.91	17.77	0.90	11.91	0.90	12.37	0.87	12.93	0.91
6	41.06	0.72	43.22	0.78	37.68	0.69	43.87	0.75	24.90	0.93	22.89	0.53	60.65	1.00
7	79.86	0.76	85.15	0.87	80.80	0.78	96.77	0.82	61.44	0.94	59.20	0.67	93.52	1.00
8	197.61	0.59	290.21	0.97	210.04	0.61	282.02	0.84	172.87	0.91	156.01	0.52	391.67	1.00
9	4.15	0.92	4.47	0.98	3.89	0.83	4.38	0.97	4.15	0.94	3.98	0.90	4.96	0.97
10	10.19	0.79	11.88	0.98	9.20	0.75	11.64	0.98	8.75	0.91	7.65	0.69	18.90	0.99
11	9.37	0.63	61.02	0.90	9.24	0.62	58.93	0.88	11.54	0.00	7.69	0.43	101.20	0.99
12	88.36	0.71	93.44	0.78	85.20	0.70	100.33	0.73	67.78	0.88	65.40	0.63	107.26	0.98
13	20.55	0.87	20.48	0.86	19.54	0.86	21.80	0.86	14.11	0.94	13.79	0.79	19.00	0.97
14	20.48	0.90	20.65	0.88	21.37	0.92	27.68	0.91	16.51	0.90	17.14	0.85	18.09	0.92
15	22.14	0.89	25.02	0.95	25.50	0.90	30.38	0.96	24.24	0.77	21.34	0.87	47.14	0.97
16	48.83	0.93	49.56	0.97	71.31	0.97	91.97	0.98	54.96	0.95	51.96	0.94	65.01	0.97
17	9.92	0.89	10.00	0.89	9.20	0.89	11.25	0.89	8.45	0.90	8.71	0.86	9.01	0.92
18	26.66	0.78	27.63	0.83	24.49	0.75	28.35	0.81	18.90	0.92	16.61	0.64	44.01	0.99
19	80.37	0.79	87.25	0.93	81.72	0.79	98.83	0.85	56.84	0.85	57.28	0.63	85.45	0.98
20	231.75	0.61	295.81	0.88	239.52	0.62	305.95	0.75	221.06	0.95	182.36	0.56	553.53	1.00
21	3.53	0.91	3.73	0.98	3.30	0.84	3.65	0.98	3.51	0.95	3.32	0.89	4.42	0.97
22	5.22	0.83	7.26	0.97	4.87	0.78	7.47	0.93	6.10	0.28	4.79	0.82	17.92	0.91
23	82.45	0.76	82.70	0.75	77.87	0.73	86.55	0.74	26.42	0.35	33.26	0.49	65.64	0.65
24	62.42	0.78	66.69	0.89	61.48	0.78	74.16	0.82	63.76	0.95	58.75	0.80	87.61	0.99
Average	50.94	0.81	61.08	0.90	53.74	0.81	68.87	0.88	42.03	0.82	38.93	0.74	82.63	0.95
$n \geq 9$	25.61	0.90	26.13	0.91	30.63	0.90	38.17	0.93	22.35	0.89	21.87	0.86	29.65	0.94
$n \leq 6$	76.28	0.73	96.02	0.88	76.84	0.72	99.57	0.82	61.70	0.74	55.99	0.61	135.61	0.96

*Confidence intervals given by (2.8), (2.9) and (2.10).

the adaptive EB-GLE.

We now investigate what happens if a more parsimonious mean model, such as the one identified in Table 2.3 based on the 2003-04 NHANES data, which the present simulation study tries to mimic, is used to obtain the EB-GLE instead of the saturated model. As one would expect, the use of a more parsimonious mean model reduces the variance but it is at the expense of bias. It can be seen from Table 2.5 that this model-based EB-GLE of P_{95} is severely biased in 4 demographic groups, and the associated CI1 confidence intervals have low coverage in 3 of the 4 groups (0 for group 11, 0.28 for group 22, 0.35 for group 23). Bias correction improves the coverage in these groups, but lowers the coverage in other groups, and the average coverage over all 24 groups is actually made worse by bias correction. The use of RMSE instead of SE leads to good coverage, but the average length is almost doubled to 82.63. This shows that while a properly selected mean model can reduce variance, it will lead to more biased estimates, as compared to a saturated model. Furthermore, the estimation of the group-specific 95^{th} percentile $\exp(\mu_i + 1.645\sigma_i)$ of a log-normal distribution from the estimates of μ_i and σ_i is a kind of extrapolation, and the effect of the bias in estimating the μ_i and σ_i will be blown up.

Finally, in Table 2.6, we focus on the Bayes analogue B-GLE of EB-GLE under various choices (inverse-gamma, gamma, log-normal) of the mixing distribution. It can be seen that the choice does not change the results that much. A comparison of Table 2.5 with Table 2.6 suggests that EB-GLE and B-GLE are quite comparable. For the saturated model case, the quasi Bayes credible interval, like its EB counterpart, has low coverage for demographic group 11. As commented in the last paragraph, when the

Table 2.6: Mean, percent bias (% bias) and mean squared error (MSE) of the Bayesian Gaussian likelihood estimator (B-GLE) under various choices of the mixing distribution and B-GLE under a selected mean model (B-GLEM) in estimating the 95^{th} percentile P_{95} for 24 demographic groups based on 1000 simulations, together with average length (L) and coverage (C) of 95% credible intervals (CrIs).

(a) Estimation

			B-GLE									B-GLEM		
			Inverse-Gamma			Gamma			Log-normal			Inverse-Gamma		
i	n	P_{95}	Mean	% bias	MSE	Mean	% bias	MSE	Mean	% bias	MSE	Mean	% bias	MSE
1	9	26.08	23.73	-9.02	32.22	23.77	-8.85	33.02	23.75	-8.93	33.18	30.36	16.40	36.13
2	12	48.34	45.16	-6.58	65.34	44.87	-7.18	62.17	44.81	-7.31	64.81	45.63	-5.60	41.63
3	13	105.50	99.30	-5.88	293.40	99.38	-5.80	279.49	99.15	-6.02	290.54	94.04	-10.86	251.65
4	19	123.54	130.05	5.27	137.30	129.10	4.51	149.49	129.35	4.70	144.57	132.96	7.62	202.13
5	15	25.43	24.16	-4.99	12.35	24.33	-4.34	10.97	24.25	-4.66	11.47	24.20	-4.84	9.77
6	4	46.96	41.68	-11.25	124.19	42.07	-10.43	119.21	41.86	-10.86	122.52	46.06	-1.93	31.17
7	5	116.73	111.26	-4.69	387.60	112.16	-3.91	342.35	111.41	-4.55	346.98	115.20	-1.31	193.27
8	3	329.64	326.17	-1.05	3096.76	329.42	-0.07	3477.33	327.40	-0.68	3358.50	332.82	0.96	1834.22
9	12	10.83	11.61	7.16	1.66	11.51	6.31	1.77	11.53	6.43	1.65	11.59	7.00	1.60
10	5	18.39	18.86	2.59	5.48	19.01	3.38	7.06	18.91	2.83	6.59	20.02	8.91	5.64
11	3	34.60	48.38	39.81	200.03	47.65	37.71	183.32	47.86	38.31	186.77	46.15	33.39	141.17
12	4	102.76	90.48	-11.95	549.95	91.83	-10.63	538.33	91.36	-11.09	555.30	97.75	-4.88	261.43
13	9	22.52	20.14	-10.58	28.15	20.00	-11.17	28.24	20.01	-11.11	28.36	22.19	-1.44	10.45
14	16	36.67	35.01	-4.55	24.66	35.31	-3.72	22.22	35.21	-3.98	23.31	35.05	-4.44	18.26
15	12	61.35	67.27	9.65	67.87	66.03	7.62	65.09	66.39	8.21	63.14	68.18	11.12	77.22
16	17	131.08	135.40	3.29	156.84	135.78	3.58	186.24	135.49	3.36	173.09	141.85	8.21	286.09
17	13	17.61	16.79	-4.62	5.89	16.92	-3.87	5.13	16.84	-4.33	5.36	16.88	-4.11	4.77
18	5	33.19	30.62	-7.75	39.86	30.79	-7.23	39.72	30.60	-7.80	40.87	35.64	7.37	21.14
19	5	119.74	116.73	-2.51	347.65	118.70	-0.87	371.03	118.00	-1.45	364.69	111.55	-6.84	270.59
20	3	335.43	310.32	-7.49	4739.95	313.74	-6.47	5127.90	312.01	-6.98	5165.28	347.50	3.60	2470.59
21	10	7.97	8.37	5.01	0.86	8.36	4.94	1.00	8.34	4.65	0.90	8.53	7.05	0.98
22	6	12.10	13.56	12.04	3.70	13.53	11.80	4.10	13.52	11.75	3.79	15.14	25.14	11.03
23	4	71.38	59.29	-16.93	505.11	59.78	-16.25	479.72	59.68	-16.40	488.60	49.14	-31.16	553.47
24	5	91.06	86.68	-4.82	207.27	87.88	-3.49	221.68	87.36	-4.06	221.50	95.63	5.02	213.17
Average				8.31*	459.75		7.67*	489.86		7.94*	487.57		9.13	289.48
$n \geq 9$				6.38*	68.88		5.99*	70.40		6.14*	70.03		7.39	78.39
$n \leq 6$				10.24*	850.63		9.35*	909.31		9.73*	905.12		10.88	500.57

*Average of absolute % bias.

(b) Credible interval

	B-GLE						B-GLEM	
	Inverse-gamma		Gamma		Log-normal		Inverse-gamma	
i	L	C	L	C	L	C	L	C
1	12.05	0.70	11.49	0.71	11.80	0.71	15.37	0.74
2	19.30	0.76	18.59	0.76	18.94	0.76	16.86	0.82
3	40.78	0.77	39.66	0.79	40.21	0.78	28.12	0.65
4	42.23	0.90	44.46	0.91	43.45	0.90	46.96	0.86
5	9.10	0.80	9.02	0.83	9.09	0.83	8.53	0.84
6	30.55	0.83	30.11	0.83	30.49	0.83	24.95	0.97
7	69.58	0.94	70.91	0.95	70.40	0.94	61.57	0.96
8	263.35	0.97	274.05	0.96	269.67	0.96	219.02	0.97
9	4.65	0.88	4.98	0.90	4.83	0.90	4.60	0.89
10	11.46	0.96	12.17	0.96	11.88	0.96	10.73	0.94
11	37.09	0.19	41.70	0.67	39.30	0.48	23.66	0.18
12	65.69	0.84	65.77	0.86	66.35	0.85	63.47	0.95
13	10.47	0.67	9.76	0.65	10.13	0.66	9.95	0.87
14	12.83	0.81	12.71	0.82	12.85	0.82	11.62	0.84
15	27.06	0.83	28.96	0.87	28.03	0.86	25.13	0.78
16	46.09	0.92	48.83	0.90	47.53	0.90	55.72	0.85
17	6.69	0.83	6.72	0.87	6.72	0.87	6.36	0.87
18	19.53	0.89	19.47	0.88	19.54	0.86	19.03	0.93
19	72.45	0.94	75.26	0.94	74.57	0.93	50.67	0.92
20	256.96	0.93	261.74	0.92	260.81	0.92	250.04	0.97
21	3.64	0.91	3.89	0.92	3.78	0.93	3.77	0.90
22	7.43	0.87	8.17	0.90	7.86	0.90	8.17	0.57
23	46.49	0.67	44.22	0.67	45.74	0.67	17.92	0.01
24	53.95	0.94	55.52	0.93	55.18	0.93	63.62	0.94
Average	48.73	0.82	49.92	0.85	49.55	0.84	43.58	0.80
$n \geq 9$	19.58	0.82	19.92	0.83	19.78	0.83	19.42	0.83
$n \leq 6$	77.88	0.83	79.92	0.87	79.32	0.85	67.74	0.77

mean model selected in Table 2.3 is used, the bias is increased, leading to low coverage for groups 11 and 23, and to a lesser degree, also group 22. Since the computation of the quasi Bayes estimator using the software JAGS requires the use of MCMC sampling, it is computing intensive, and as a result it is not feasible to use the bootstrap to estimate the bias or mean squared error of the quasi Bayes estimator. Thus while the EB-GLE and B-GLE have similar performance, the former has the advantage that it can be improved further via bootstrap estimation of the bias or RMSE.

2.9 Discussion

We have proposed EB-GLE and its Bayesian analogue B-GLE to estimate the log-normal distribution based on pooled samples, which is easily implemented and more efficient than the estimators proposed by Caudill (2012). Our simulation study shows that EB-GLE and B-GLE perform similarly, but bootstrap resampling is only feasible for the former. We recommend the bias-corrected confidence interval CI1-BC based on EB-GLE which has good coverage property for those demographic groups with sufficient number of pools. It is to be expected that for those groups with only 3 or 4 pools, there is really not much that one can do. One could used the t instead of standard normal percentiles to construct confidence intervals. Another possibility is to use the RMSE instead of SE, but our simulation study shows that the resulting interval CI2 will sometimes be too long to be of practical value. Our study also shows that the reduction in variance which arises from the use of a more parsimonious model of the means can be offset by an increase in bias, leading to poor confidence interval coverage in a few groups with insufficient number of pools. One can make a

case for the use of the saturated mean model because pooling does not really affect the estimation of the mean that much. Thus even if we have only 5 pools within a demographic group, say, of size 8 each, the mean of the 5 pool averages is actually the average of 40 individual measurements, which may be sufficiently precise, and there may not be a need to model the mean more parsimoniously. For estimating the variance (hence also the percentiles, which are functions of the mean and variance), efficiency is lost by pooling, and it is desirable to borrow strength from other demographic groups, which we do by treating the squared coefficients of variation as random effects. Our preferred estimator is EB-GLE based on the saturated mean model with bias correction carried out in constructing confidence intervals. This recommendation is also supported by another simulation study with parameter values in the 24 demographic groups set to values inferred from Table II of Caudill (2012) for NHANES 2005-06 in Table 2.7. There is some difference between EB-GLE and B-GLE and B-GLE seems better. However, B-GLE is more computing intensive.

Besides assuming log-normal distribution for the data, nonparametric density estimation is also possible for pooled data, when no clear parametric distributions are available to adequately fit the data. The nonparametric approach to estimate cumulants/moments from pooled data has been described in section 2 of Xu and Kuk (2014). The moments can be used to determine the density function. An alternative way is to look at characteristic function. Nonparametric method can be used to estimate characteristic function of the sums, and then take the power of one over pool size K to obtain characteristic function of individual data. Based on inversion formula, density function of individual data can be derived from its characteristic

Table 2.7: Mean, percent bias (% bias) and mean squared error (MSE) of the group-specific Gaussian likelihood estimator (GLE), Caudills estimator, empirical Bayes Gaussian likelihood estimator (EB-GLE), adaptive empirical Bayes Gaussian likelihood estimator (AEB-GLE) and Bayesian Gaussian likelihood estimator (B-GLE) of the 95^{th} percentile P_{95} for 24 demographic groups of NHANES 2005-06 based on 1000 simulations, together with average length (L) and coverage (C) of the 95% confidence intervals (CIs) based on three methods and credible intervals (CrIs).

(a) Estimation of GLE, Caudill's estimator and EB-GLE

			GLE			Caudill			EB-GLE		
i	n	P_{95}	Mean	% bias	MSE	Mean	% bias	MSE	Mean	% bias	MSE
1	9	16.30	15.81	-3.00	3.02	14.64	-10.16	3.17	15.85	-2.76	2.24
2	12	31.40	30.57	-2.65	8.47	27.63	-12.02	15.24	30.56	-2.68	6.82
3	12	87.70	86.04	-1.90	44.21	84.69	-3.44	17.41	87.69	-0.02	29.50
4	15	149.80	148.22	-1.05	38.73	165.17	10.26	264.70	155.62	3.89	62.89
5	13	19.00	18.63	-1.97	2.41	17.90	-5.80	1.63	18.77	-1.20	1.78
6	6	35.40	33.11	-6.46	29.20	28.60	-19.21	49.03	32.86	-7.18	23.48
7	5	96.80	89.12	-7.93	276.33	76.37	-21.11	443.39	88.66	-8.41	212.35
8	5	245.00	226.30	-7.63	1816.78	187.73	-23.38	3445.37	224.33	-8.44	1404.18
9	11	10.10	9.85	-2.51	0.72	9.71	-3.87	0.29	9.97	-1.32	0.49
10	9	19.30	18.84	-2.40	2.02	20.35	5.44	1.49	19.68	1.98	1.24
11	4	44.00	39.93	-9.26	68.75	35.60	-19.08	76.65	40.13	-8.79	46.05
12	4	91.00	79.41	-12.74	561.61	55.45	-39.07	1304.14	76.02	-16.46	576.14
13	10	12.60	12.30	-2.39	1.02	12.45	-1.18	0.25	12.55	-0.41	0.65
14	16	22.80	22.55	-1.11	1.18	24.94	9.39	4.91	23.34	2.38	1.13
15	13	71.60	70.71	-1.24	12.16	78.36	9.44	49.92	74.21	3.64	14.87
16	17	138.50	137.05	-1.05	53.46	141.05	1.84	28.23	140.36	1.35	42.61
17	14	11.70	11.50	-1.67	0.41	12.61	7.79	0.98	11.88	1.57	0.30
18	7	28.20	26.81	-4.93	12.19	24.87	-11.79	12.43	27.01	-4.22	8.28
19	7	102.70	98.55	-4.04	108.53	97.70	-4.87	41.41	101.81	-0.87	54.61
20	5	282.00	255.33	-9.46	3475.67	181.17	-35.75	10472.92	246.11	-12.73	3558.56
21	16	8.10	8.01	-1.13	0.16	8.91	9.95	0.74	8.26	1.99	0.14
22	9	12.70	12.22	-3.78	2.09	11.25	-11.46	2.39	12.23	-3.72	1.59
23	6	50.10	47.65	-4.89	28.46	48.71	-2.78	5.29	49.73	-0.74	12.10
24	3	93.10	78.39	-15.80	726.14	59.72	-35.85	1168.09	76.23	-18.13	652.66
Average				4.62*	303.07		13.12*	725.42		4.79*	279.78

*Average of absolute % bias.

(b) Estimation of AEB-GLE and B-GLE

	AEB-GLE			B-GLE		
i	Mean	% bias	MSE	Mean	% bias	MSE
1	15.59	-4.38	2.31	15.80	-3.06	2.29
2	29.83	-4.99	9.08	30.60	-2.55	7.03
3	85.75	-2.22	54.65	87.98	0.32	27.86
4	153.35	2.37	92.40	156.52	4.49	77.14
5	18.44	-2.94	1.99	18.73	-1.44	1.74
6	32.09	-9.35	26.88	32.86	-7.17	23.42
7	86.37	-10.78	271.75	89.55	-7.49	190.83
8	217.61	-11.18	2117.68	224.71	-8.28	1375.86
9	9.85	-2.46	0.45	10.01	-0.91	0.44
10	19.51	1.08	1.07	19.77	2.42	1.20
11	39.36	-10.56	50.91	40.16	-8.72	43.75
12	73.36	-19.38	660.24	75.79	-16.71	538.62
13	12.41	-1.48	0.61	12.58	-0.12	0.67
14	23.09	1.28	1.03	23.47	2.92	1.31
15	73.20	2.23	16.70	74.46	4.00	17.68
16	137.52	-0.71	115.01	140.70	1.59	41.14
17	11.79	0.78	0.25	11.95	2.14	0.35
18	26.52	-5.96	9.53	27.27	-3.30	7.62
19	99.90	-2.72	90.58	102.34	-0.35	51.96
20	236.05	-16.29	4911.84	244.54	-13.29	3345.70
21	8.21	1.41	0.11	8.29	2.35	0.15
22	12.04	-5.20	1.61	12.31	-3.11	1.49
23	49.00	-2.20	14.47	50.13	0.06	12.47
24	73.96	-20.56	722.70	77.30	-16.97	568.76
Average		5.94*	382.24		4.74*	264.15

*Average of absolute % bias.

(c) Confidence interval of GLE, Caudill's estimator and EB-GLE

	GLE		Caudill						EB-GLE					
	CI1*		CI1*		CI1-BC*		CI2*		CI1*		CI1-BC*		CI2*	
i	L	C	L	C	L	C	L	C	L	C	L	C	L	C
1	6.26	0.87	2.31	0.28	2.61	0.56	7.26	0.81	5.42	0.86	5.62	0.86	5.88	0.87
2	10.69	0.89	3.74	0.07	4.31	0.53	15.09	0.79	9.63	0.87	9.93	0.88	10.29	0.88
3	23.50	0.87	10.18	0.74	10.73	0.56	24.50	0.92	20.51	0.91	20.73	0.87	21.30	0.93
4	22.54	0.89	19.15	0.07	17.68	0.81	56.93	0.89	20.22	0.83	19.77	0.89	25.85	0.91
5	5.58	0.87	2.24	0.54	2.42	0.55	6.25	0.87	5.00	0.88	5.10	0.87	5.23	0.88
6	18.11	0.81	5.67	0.04	7.01	0.49	22.29	0.70	14.95	0.75	16.07	0.79	17.34	0.77
7	52.70	0.77	16.53	0.05	20.75	0.49	62.82	0.68	41.70	0.70	45.34	0.74	49.69	0.73
8	138.56	0.79	45.28	0.03	58.51	0.48	178.53	0.69	110.50	0.71	120.98	0.77	132.97	0.74
9	3.11	0.88	1.44	0.78	1.51	0.65	3.09	0.93	2.73	0.89	2.78	0.87	2.86	0.90
10	4.86	0.85	2.20	0.52	2.12	0.56	6.19	0.92	3.97	0.95	3.95	0.84	4.45	0.99
11	25.34	0.73	7.84	0.11	9.64	0.46	26.77	0.66	18.78	0.66	20.66	0.69	23.57	0.70
12	77.53	0.74	19.72	0.02	31.58	0.49	99.15	0.68	61.37	0.64	73.40	0.72	79.74	0.70
13	3.77	0.88	1.69	0.89	1.73	0.61	3.44	0.97	3.21	0.91	3.26	0.88	3.36	0.93
14	3.94	0.90	2.13	0.02	1.99	0.62	7.77	0.79	3.50	0.93	3.47	0.89	3.82	0.96
15	12.51	0.88	7.44	0.04	6.92	0.67	25.43	0.82	10.89	0.90	10.68	0.87	13.22	0.93
16	26.83	0.90	16.62	0.90	16.65	0.72	25.03	0.95	24.06	0.95	24.02	0.89	24.73	0.96
17	2.34	0.88	1.41	0.26	1.32	0.67	3.77	0.84	2.04	0.95	2.03	0.89	2.16	0.98
18	11.76	0.83	4.01	0.19	4.56	0.50	12.97	0.74	9.75	0.80	10.20	0.81	10.85	0.81
19	34.53	0.81	13.92	0.67	14.69	0.53	33.99	0.91	27.49	0.87	27.94	0.80	30.41	0.94
20	204.63	0.79	57.24	0.00	87.54	0.54	302.05	0.76	169.91	0.70	194.36	0.78	210.79	0.76
21	1.48	0.89	1.11	0.16	1.03	0.79	3.06	0.85	1.32	0.96	1.31	0.92	1.40	0.98
22	4.96	0.85	1.90	0.23	2.16	0.55	5.89	0.77	4.31	0.84	4.49	0.84	4.71	0.84
23	17.12	0.81	6.44	0.81	6.62	0.48	16.33	0.95	13.04	0.88	13.20	0.78	15.04	0.97
24	79.47	0.67	20.74	0.04	31.51	0.42	86.65	0.58	57.04	0.54	70.29	0.62	79.78	0.62
Average	33.00	0.83	11.29	0.31	14.40	0.57	43.14	0.81	26.72	0.83	29.57	0.82	32.48	0.86

*Confidence intervals given by (2.8), (2.9) and (2.10).

(d) Confidence interval of AEB-GLE and credible interval of B-GLE

	AEB-GLE						B-GLE	
	CI1*		CI1-BC*		CI2*		CrI	
i	L	C	L	C	L	C	L	C
1	4.98	0.82	5.29	0.84	6.37	0.86	5.64	0.93
2	9.83	0.79	10.43	0.89	13.06	0.90	9.60	0.92
3	26.32	0.86	27.14	0.92	32.79	0.93	23.58	0.96
4	32.97	0.94	32.49	0.96	37.88	0.97	29.97	0.85
5	4.79	0.83	5.00	0.85	6.15	0.88	5.19	0.94
6	14.18	0.70	15.70	0.80	19.24	0.79	15.46	0.87
7	42.75	0.64	47.71	0.76	57.93	0.76	45.82	0.88
8	125.03	0.65	140.25	0.79	167.36	0.79	116.93	0.87
9	2.43	0.86	2.53	0.84	2.95	0.89	3.00	0.97
10	3.78	0.93	3.80	0.80	4.54	0.98	5.61	0.97
11	17.67	0.60	19.90	0.67	24.95	0.69	22.48	0.90
12	59.08	0.58	73.57	0.71	87.13	0.71	50.76	0.68
13	2.92	0.89	3.01	0.85	3.47	0.91	3.76	0.97
14	3.66	0.94	3.67	0.89	4.31	0.97	4.60	0.93
15	14.24	0.94	14.12	0.93	16.71	0.97	15.95	0.89
16	39.43	0.88	39.89	0.96	47.73	0.96	28.45	0.95
17	1.87	0.94	1.88	0.84	2.16	0.97	2.62	0.95
18	9.27	0.74	9.92	0.79	12.02	0.80	11.10	0.95
19	31.98	0.80	33.00	0.82	38.78	0.86	36.45	0.97
20	184.12	0.65	219.25	0.80	258.94	0.80	144.10	0.79
21	1.18	0.95	1.19	0.86	1.34	0.97	1.66	0.94
22	3.90	0.77	4.15	0.83	4.97	0.84	4.52	0.92
23	13.22	0.83	13.59	0.79	16.63	0.92	19.16	0.98
24	53.86	0.50	68.56	0.61	84.10	0.62	57.61	0.73
Average	29.31	0.79	33.17	0.83	39.65	0.86	27.67	0.90

*Confidence intervals given by (2.8), (2.9) and (2.10).

function.

We have assumed equal weights so far. However, our proposed method can be easily extended to the case of unequal weights. Assume that $A_{ij} = \sum_{k=1}^{K} \omega_{ijk} X_{ijk}$ is the j^{th} pool average with normalized weights ω_{ijk} in the i^{th} demographic group and $\sum_{k=1}^{K} \omega_{ijk} = 1$. To approximate the distribution by normal distribution, we can write down the expectation and variance of the weighted averages,

$$E\left[A_{ij}\right] = E[X_{ijk}] = \alpha_i,$$
$$\text{var}\left[A_{ij}\right] = \sum_{k=1}^{K} \omega_{ijk}^2 \,\text{var}\left[X_{ijk}\right] = \sum_{k=1}^{K} \omega_{ijk}^2 \beta_i^2 = \varphi_{ij}\beta_i^2,$$

where $\varphi_{ij} = \sum_{k=1}^{K} \omega_{ijk}^2$. We use $N\left(\alpha_i, \varphi_{ij}\beta_i^2\right)$ as the working distribution

for A_{ij}, and the marginal log-likelihood function after integrating out the $\gamma_i = (\beta_i/\alpha_i)^2 \sim \Gamma^{-1}(\kappa, \lambda)$ is given by

$$l(\alpha_1, \cdots, \alpha_d, \kappa, \lambda) = \sum_{i=1}^{d} \Bigg\{ 2\kappa \log \alpha_i - \left(\kappa + \frac{n_i}{2}\right) \log \left[\sum_{j=1}^{n_i} \frac{(A_{ij} - \alpha_i)^2}{\varphi_{ij}} + 2\lambda\alpha_i^2 \right] + \log \Gamma\left(\kappa + \frac{n_i}{2}\right) \Bigg\} + d\kappa \log(2\lambda) - d \log \Gamma(\kappa).$$

We can obtain the EB-GLE with unequal weights by maximizing the above marginal log-likelihood. Unfortunately, we cannot find the weights ω_{ijk} used in NHANES 2005-06 from the online link (http://wwwn.cdc.gov/nchs/nhanes/search/nhanes05_06.aspx), and so we are not able to apply the above method to the weighted pool averages collected in 2005-06.

The empirical Bayes approach is a popular approach to pool information across groups, In our particular context, we treat the group-specific exposure means α_i as fixed effects and model them by either a saturated or non-saturated model; and we treat the squared coefficients of variation γ_i as random effects that follow a common distribution. It is this common distribution or homogeneity assumption which allows us to pool information across groups. Our model shares some similarity with hierarchical or multilevel modeling. Shrinkage method is another way to combine information across groups and empirical Bayes or Bayes estimator can be considered as one kind of shrinkage estimators. Strictly speaking, what we propose in this paper are quasi Bayes and quasi EB estimators because we are treating the working Gaussian likelihoods as if they are the true likelihoods. To obtain the exact EB or Bayes estimators would typically require the EM algorithm for the former, and data augmentation for the latter, which are

computing intensive. In comparison, the proposed EB-GLE and B-GLE are simpler to compute, and our study suggests that the Gaussian likelihoods work quite well when the pool size is 8. Another way to enable pooling of information across "neighboring" groups is by some kind of nonparametric smoothing. Unfortunately in our situation, race and gender are categorical variables, and we do not have enough age groups to perform any kind of meaningful smoothing by age. The same comment would apply to quantile regression method, which in a way is the natural method to use, given that our main interest is focused on the percentiles. Furthermore, pooled data does not contain too much information about the quantiles of individual exposure due to the convergence to normal distribution effect of averaging. Thus quantile regression based on pooled data is a challenging problem that we will address in future investigation.

Estimation of the 95^{th} and other extreme percentiles is highly sensitive to the distributional assumption made. This is because the 95^{th} percentile will be given by different functions of, say, the mean and variance, depending on whether the exposure distribution is log-normal, gamma, or inverse Gaussian, just to name a few examples. But no matter what the exposure distribution is, the proposed EB-GLE method can be applied in almost the same way, and this is another advantage of our method. In the biomonitoring literature, the log-normal distribution is the popular choice, and this is backed up by our Box-Cox transformation analysis based on the 2003-04 NHANES data. It is more difficult to test distributional assumption when only pooled data are available, such as in NHANES 2005-06. This is because regardless of what the distribution of the individual exposures is, the distribution of the pool averages will tend to the normal distribution

as pool size increases by virtue of the Central Limit Theorem. One way to overcome this problem is to use historical data, For example, we can use the 2003-04 individual data to justify the log-normal assumption, and keep faith in it for the 2005-06 data. Another possible strategy is to use a mixed design to collect some individual data in addition to pooled data. Collecting individual data will boost our ability to test distributional assumption, while pooling will save cost. There is some design issue to be solved, such as finding the optimal ratio of pooled to individual data, and subsequently how to compute estimators based on a combination of pooled and individual data. We will again leave these for future investigations.

Another area where pooling is used is in genetic association study. Pooled genotype data are often reported instead of individual genotype data to save genotyping cost. Kuk et al. (2014) have also proposed a random effects formulation for certain baseline parameters to enable pooling of information across different genetic markers.

Chapter 3

Collapsed Data MLE

This chapter is organized as follows. Section 3.1 summarizes the collapsed data maximum likelihood estimator and highlights the main findings; Section 3.2 provides the details of our method; Section 3.3 considers a real data analysis and section 3.4 concludes this chapter with some discussion and extensions.

The materials presented in this chapter have been published in Kuk et al. (2013b).

3.1 Summary

In this chapter, we propose an estimation method that does not suffer from the aforementioned drawbacks of the expectation maximization (EM) algorithm (see section 1.2.4). To begin with, our method is non-iterative in nature, and the amount of computation does not increase with pool size. Finally, the number of putative haplotypes with positive probability estimates does not grow exponentially with the number of markers. In fact, the number of haplotypes that we need to deal with is no greater than the number of pools, and is often much less, especially for rare variants (RVs).

This desirable algorithm is made possible by collapsing the pool total at each marker (which can take on values $0, 1, \cdots, K = 2k$, where k is the number of individuals in each pool) to just "0" or "at least 1", as done in the literature of group testing (Dorfman, 1943; Gastwirth and Hammick, 1989). It should be pointed out, though, that the group testing literature deals mainly with estimating the prevalence of a single binary trait, whereas we are now dealing with multiple genetic variants. The haplotype frequency estimates produced by our method are in fact the maximum likelihood estimates (MLEs) based on the collapsed data.

Pirinen (2009) pointed out that it may be possible to use database information to create a list of frequently occurring haplotypes and that the EM algorithm will run much faster if the underlying unobserved haplotypes are restricted to come from this list only. We call the resulting procedure the EML algorithm. Since we are focusing on RVs, there may not be an external list of common haplotypes available from existing databases. Instead, we construct at the outset an internal list of all possible haplotypes compatible with the observed pool totals. This list is constructed only once and will not be repeated at each iteration of the EML algorithm.

We conduct a running time analysis to compare the collapsed data method with the EML algorithm. The EM algorithm without a list is much slower and will not be considered. For rare alleles, the total allele frequency at most markers and for most pools will equal to 0 or a small number like 1 or 2, and this greatly reduces the amount of computation required to construct the list of all possible underlying haplotypes. This makes the EML algorithm a worthy competitor of the collapsed data method. When applied to genotype data collected from 148 obese persons for 25 rare vari-

ants around the *MGLL* gene, and 32 rare variants around the *FAAH* gene, the collapsed data method is 50 to 200 times faster than the EML algorithm when the pool size is 2. When the pool size is 4, the collapsed data method took 0.61 and 6.45 seconds to run on an intel (R) Core (TM) 2 desktop for the cases of 25 and 32 rare variants respectively, but the EML algorithm was still running after 10 hours and as a result was aborted. The price paid for proposing such a fast and simple algorithm is possible loss of estimation efficiency. We show theoretically that there will not be much loss of efficiency for the case of rare variants. Even if the variants are not rare, the collapsed data method will still be useful for the purpose of obtaining consistent initial estimates to input to more sophisticated algorithms, and for screening a large number of possible causal variants to a more manageable set within a narrower region to be studied further using more advanced methods or molecular haplotyping.

We conclude this chapter by an application to identify rare variants associated with obesity. Using resequenced data collected from a case control study involving 148 obese persons and 150 controls, and a method called RARECOVER, Bhatia et al. (2010) identified 12 RVs associated with obesity in a $5Kbp$ window containing 25 RVs just upstream of the *MGLL* gene. We apply the collapsed data method to estimate the haplotype distribution for the 25 RVs in this window. Comparing the haplotype frequency estimates for the cases and controls, we are able to identify a much more parsimonious subset of 3 RVs than the 12 RVs selected by RARECOVER. From the set of 32 rare variants around the *FAAH* gene, we discover an interesting potential interaction between two of them. We conclude this chapter with some discussion on the effect of noise due to pooling, and the

analysis of replicated measurements with correlation.

3.2 Statistical Models and Methods

3.2.1 Collapsed data estimator

Focusing on bi-allelic loci, the two possible alleles at each locus can be represented by "1" (the minor allele) and "0". As a result, the alleles at selected loci of a chromosome can be represented by a binary haplotype vector. Since human chromosomes come in pairs, there are 2 haplotype vectors for each individual, one maternal, and one paternal. Suppose we have n pools of k individuals each so that there are $K = 2k$ haplotypes within each pool. Denote by $Y_{ij} = (Y_{1ij}, \cdots, Y_{Lij})$ the j^{th} haplotype in the i^{th} pool, where $i = 1, \cdots, n$, $j = 1, \cdots, K$, and L is the number of markers being typed. Assuming Hardy-Weinberg equilibrium, the nK haplotype vectors are independent and identically distributed with probability function

$$f(y_1, \cdots, y_L) = P\left(Y_{1ij} = y_1, \cdots, Y_{Lij} = y_L\right)$$

for every L-tuple $y = (y_1, \cdots, y_L)$ belonging to the Cartesian product $\Omega = \{0, 1\}^L$. With pooling, the observed data are the pool totals

$$T_i = \sum_{j=1}^{K} Y_{ij} = \left(\sum_{j=1}^{K} Y_{1ij}, \cdots, \sum_{j=1}^{K} Y_{Lij}\right) = \left(T_{1i}, \cdots, T_{Li}\right), \quad i = 1, \cdots, n.$$

The probability function $p\left(t_1, \cdots, t_L\right)$ of each pool total, with $(t_1, \cdots, t_L) \in \{0, 1, \cdots, K\}^L$, is given by the K-fold convolution of the haplotype probability function $f\left(y_1, \cdots, y_L\right)$ and so the likelihood based on the observed pooled data is highly intractable and not easy to maximize directly. If the

individual haplotypes were actually observed, then the population haplotype distribution function can be estimated simply by the empirical haplotype distribution. By taking conditional expectation to estimate iteratively the unobserved haplotype frequencies in the sample, the EM algorithm can be used to obtain the MLEs of $\{f(y), y \in \Omega\}$ based on the observed pool totals, but as pointed out earlier, the EM algorithm is very computing intensive and not viable if L or k is large.

Next, we will show that the MLEs of $\{f(y), y \in \Omega\}$ based on the collapsed data

$$Z_i = \left(I\left\{\sum_{j=1}^{K} Y_{1ij} \geq 1\right\}, \cdots, I\left\{\sum_{j=1}^{K} Y_{Lij} \geq 1\right\}\right) = (Z_{1i}, \cdots, Z_{Li})$$

defined via indicator functions are very easy to obtain. Note that what Z_i does is to collapse each total allele frequency to either "0" or "at least 1" as done in classical group testing.

From here on, we will call $\{Y_{ij}, i = 1, \cdots, n, j = 1, \cdots, K\}$ the complete haplotype data (usually not observed but will be used as a benchmark); $\{T_i, i = 1, \cdots, n\}$ the pooled genotype data (or individual genotype data if pool size is 1), and $\{Z_i, i = 1, \cdots, n\}$ the collapsed data. In this chapter, we refer to k as the pool size, not K.

Suppressing its dependence on the pool size k, let

$$g(z_1, \cdots, z_L) = P(Z_{1i} = z_1, \cdots, Z_{Li} = z_L)$$

with $z = (z_1, \cdots, z_L) \in \Omega = \{0, 1\}^L$ be the probability function of the collapsed data. The likelihood function based on the collapsed data $Z_1, \cdots, Z_n$

is multinomial, i.e., proportional to

$$\prod_{z\in\Omega} g(z)^{n_Z(z)},$$

where for every $z = (z_1, \cdots, z_L) \in \Omega$, $n_Z(z)$ is the number of pools with $Z_i = z$. The value of $g(z)$ which maximizes the above multinomial likelihood is simply the sample proportion

$$\hat{g}(z) = \frac{n_Z(z)}{n}. \tag{3.1}$$

We consider next the probabilities of zero pool totals at various loci. Let Λ be a subset of $\{1, 2, \cdots, L\}$ which specifies the positions of the zeros, then

$$g_0(\Lambda) = P\left(Z_{li} = 0, l \in \Lambda\right)$$

can be obtained by summing $g(z)$ over all those $z = (z_1, \cdots, z_L)$ satisfying $z_\Lambda = (z_l, l \in \Lambda) = 0$. By summing the MLE $\hat{g}(z)$ of $g(z)$ in a similar way, we can see that the MLE of $g_0(\Lambda)$ based on the collapsed data $Z_1, \cdots, Z_n$ is given by

$$\hat{g}_0(\Lambda) = \frac{n_{0Z}(\Lambda)}{n}, \tag{3.2}$$

where $n_{0Z}(\Lambda)$ is the number of pools with $Z_{li} = 0$ for $l \in \Lambda$ (i.e., with no minor alleles at the positions specified by Λ).

The quantities that we are interested in estimating are the haplotype frequencies $\{f(y), y \in \Omega\}$, and not $g(z)$ or $g_0(\Lambda)$. To find the MLE of $f(y)$, we will derive an equation expressing $f(y)$ as a function of $g_0(\Lambda)$. Since we have already obtained in (3.2) the collapsed data MLE $\hat{g}_0(\Lambda)$ of $g_0(\Lambda)$, the collapsed data MLE $\hat{f}(y)$ of $f(y)$ can be obtained by substituting $\hat{g}_0(\Lambda)$

into the equation for $f(y)$. We begin by relating $f(0,\cdots,0)$ to $g(0,\cdots,0)$ first. By definition

$$\begin{aligned} g(0,\cdots,0) &= P\left(\sum_{j=1}^{K} Y_{lij} = 0, l = 1,\cdots,L\right) \\ &= P\left(\bigcap_{j=1}^{K} \{Y_{lij} = 0, l = 1,\cdots,L\}\right) \\ &= f(0,\cdots,0)^K \end{aligned}$$

is the probability of the event that all K haplotypes in a pool are zero vectors, and so

$$f(0,\cdots,0) = g(0,\cdots,0)^{\frac{1}{K}}. \tag{3.3}$$

More generally, let Λ be a non-empty subset of $\{1,\cdots,L\}$, we have

$$\begin{aligned} g_0(\Lambda) &= P\left(\sum_{j=1}^{K} Y_{lij} = 0, l \in \Lambda\right) \\ &= P\left(\bigcap_{j=1}^{K} \{Y_{lij} = 0, l \in \Lambda\}\right) \\ &= P(Y_{lij} = 0, l \in \Lambda)^K. \end{aligned}$$

If we define

$$f_0(\Lambda) = P(Y_{lij} = 0, l \in \Lambda),$$

as the probability that the allele type is "0" at the positions specified by Λ in a single haplotype, then the equation above can be rewritten as $g_0(\Lambda) = f_0(\Lambda)^K$, or equivalently

$$f_0(\Lambda) = g_0(\Lambda)^{\frac{1}{K}}, \tag{3.4}$$

which is a generalization of (3.3). It follows from (3.2) and (3.4) that the collapsed data MLE of $f_0(\Lambda)$ is

$$\hat{f}_0(\Lambda) = \hat{g}_0(\Lambda)^{\frac{1}{K}} = \left[\frac{n_{0Z}(\Lambda)}{n}\right]^{\frac{1}{K}}. \tag{3.5}$$

Now that we know how to estimate $f_0(\Lambda)$, all the haplotype frequencies $f(y), y \in \Omega$, can be estimated too because $f(y)$ can be expressed in terms of the $f_0(\Lambda)$ as follows. Note that

$$\begin{aligned} f(y_1, \cdots, y_L) &= P(Y_{1ij} = y_1, \cdots, Y_{Lij} = y_L) \\ &= \mathrm{E}\left[\prod_{l\in\Lambda(y)}(1 - Y_{lij}) \prod_{l\in\Lambda'(y)} Y_{lij}\right] \\ &= \mathrm{E}\left[\prod_{l\in\Lambda(y)} W_{lij} \prod_{l\in\Lambda'(y)}(1 - W_{lij})\right], \end{aligned}$$

where $\Lambda(y)$ denotes the positions of the 0's in $y = (y_1, \cdots, y_L)$, $\Lambda'(y)$ is the complement of $\Lambda(y)$ which gives the positions of the 1's, and $W_{lij} = 1 - Y_{lij}$. Now $\mathrm{E}\left[\prod_{l\in\Lambda} W_{lij}\right] = f_0(\Lambda)$ by definition, and so the last expectation in the above equation can be expanded as

$$\begin{aligned} f(y) &= \mathrm{E}\left[\prod_{l\in\Lambda(y)} W_{lij} \prod_{l\in\Lambda'(y)}(1 - W_{lij})\right] \\ &= f_0(\Lambda(y)) + \sum_{r=1}^{m}(-1)^r \sum_{\substack{S\subset\Lambda'(y) \\ |S|=r}} f_0(\Lambda(y) \cup S), \end{aligned}$$

which can also be derived using the inclusion-exclusion principle, and m is the number of 1's in the haplotype vector y. Substituting (3.5) into the

above equation, we obtain

$$\begin{aligned}\hat{f}(y) =& \hat{f}_0\left(\Lambda(y)\right) + \sum_{r=1}^{m}(-1)^r \sum_{\substack{S\subset\Lambda'(y)\\|S|=r}} \hat{f}_0\left(\Lambda(y)\cup S\right)\\ =& \left[\hat{g}_0\left(\Lambda(y)\right)\right]^{\frac{1}{K}} + \sum_{r=1}^{m}(-1)^r \sum_{\substack{S\subset\Lambda'(y)\\|S|=r}} \left[\hat{g}_0\left(\Lambda(y)\cup S\right)\right]^{\frac{1}{K}}\\ =& \left[\frac{n_{0Z}\left(\Lambda(y)\right)}{n}\right]^{\frac{1}{K}} + \sum_{r=1}^{m}(-1)^r \sum_{\substack{S\subset\Lambda'(y)\\|S|=r}} \left[\frac{n_{0Z}\left(\Lambda(y)\cup S\right)}{n}\right]^{\frac{1}{K}}\end{aligned} \tag{3.6}$$

as the collapsed data MLE of $f(y)$. $\hat{f}(y)$ may be negative according to (3.6) due to calculation error, but it's rare based on my experience. To restrict the estimate $\hat{f}(y)$ within the legitimate range, one way is to truncate the estimate between zero and one.

3.2.2 Running time analysis and comparison with the EML algorithm

It can be seen from (3.6) that the collapsed data estimation method is non-iterative and the amount of computation does not depend that much on K since (3.6) depends on K only through the power $\frac{1}{K}$. The only apparent difficulty that might limit the applicability of this method is that we seemingly need to evaluate (3.6) for $2^L - 1$ choices of y which increases exponentially with L. That this is not necessary is a consequence of the following lemma.

Lemma 3.1. If $g(y) = P\left(Z = y\right) = 0$, then $f(y) = P\left(Y = y\right) = 0$.

To prove lemma 3.1, we note that if every haplotype $Y_1, \cdots, Y_K$ in a pool is equal to $y = (y_1, \cdots, y_L)$, then the vector of total allele frequencies

is $t = (t_1, \cdots, t_L) = (Ky_1, \cdots, Ky_L)$, with the consequence that $t_l = 0$ if $y_l = 0$, and $t_l = K \geq 1$ if $y_l = 1$. By definition, the collapsed data is $z_l = 0$ if $t_l = 0$, and $z_l = 1$ if $t_l \geq 1$, so it follows that $z = (z_1, \cdots, z_L) = (y_1, \cdots, y_L) = y$. Since $\{Y_1 = \cdots = Y_K = y\} \Rightarrow \{Z = y\}$, we have $0 \leq f(y)^K = P(Y_1 = \cdots = Y_K = y) \leq P(Z = y) = g(y)$. It follows that if $g(y) = 0$, then $f(y)$ must also equal 0.

Let $\hat{f}(\cdot)$ and $\hat{g}(\cdot)$ be the collapsed data MLE of $f(\cdot)$ and $g(\cdot)$ respectively. By the invariance property of MLE, the relationship between $\hat{f}(\cdot)$ and $\hat{g}(\cdot)$ is the same as that between $f(\cdot)$ and $g(\cdot)$, and so lemma 3.1 can be applied to the estimates as well. Thus we know without calculation that if $\hat{g}(y) = 0$, then $\hat{f}(y) = 0$, and this results in the following rule for finding the non-zero $\hat{f}(y)$.

Rule 3.1. Use (3.6) to evaluate $\hat{f}(y)$ only for those y with $\hat{g}(y) > 0$.

In view of (3.1), rule 3.1 means that we only need to use (3.6) to evaluate $\hat{f}(y)$ for those y which coincides with at least one of the collapsed pool totals Z_i, $i = 1, \cdots, n$. There are at most n such y's, which does not increase with L, and the number will often be much less than n due to repetition among the Z_i's. The number of times one needs to evaluate (3.6) will be especially small for the case of rare variants because most of the Z_i's will be vectors of all zeros or mostly zeros, and so the number of distinct Z_i's is much smaller than n. As a result, the collapsed data method of estimation is very fast.

In comparison, the most time consuming part of the EM algorithm is to obtain for each pool the collection of all possible underlying haplotype vectors that sum up to the observed pool total. Omitting the subscript for numbering the pools, this amounts to finding all possible combinations of K binary vectors $Y_1 = (Y_{11}, \cdots, Y_{L1}), \cdots, Y_K = (Y_{1K}, \cdots, Y_{LK})$ that add

up to the observed pool total $T = (T_1, \cdots, T_L)$. The algorithm that we use to do this is the following. For every $l = 1, \cdots, L$, since $T_l = Y_{l1} + \cdots + Y_{lK}$ is given, it means that the value at locus l must equal to 1 for T_l haplotype vectors, and equal to 0 for the remaining $K - T_l$ haplotype vectors; and there are ${}_KC_{T_l}$ ways to choose the first set of T_l vectors. Doing this for all L loci, we have to consider a total of $\prod_{l=1}^{L} {}_KC_{T_l}$ possibilities which increases very quickly with K and L, and this is for one pool only. Fortunately, if the alleles are rare, we can expect the values of the haplotype vectors to be fully resolved at most loci, and have little phase ambiguity at the other loci. In other words, we can expect most of the T_l to be 0, and the remaining T_l to be small, with the consequence that $\prod_{l=1}^{L} {}_KC_{T_l}$ is not prohibitively large. Each of the $\prod_{l=1}^{L} {}_KC_{T_l}$ ways of assigning zeros and ones will result in K haplotype vectors, and we simply take the union of all such haplotype vectors to create a list of haplotypes compatible with the total allele frequencies of a single pool. We take the union of the pool specific lists to obtain a merged list. We can save a lot of computer time by obtaining this (merged) list at the outset once and for all, and then constrain the EM algorithm to estimate the frequencies of only those haplotypes on the list. We call this the EML algorithm which is similar in spirit to the AEML algorithm proposed by Pirinen (2009), except that we do not use normal approximation (as we are working with small pool sizes), and we do not assume the existence of an external list due to insufficient database information about rare alleles.

To compare the running time of the collapsed data method and the EML algorithm, we consider data collected from the CRESCENDO cohort (http://clinicaltrials.gov/ct/show/NCT00263042) of 298 individuals at the two extreme ends of the Body Mass Index (BMI): 148 obese in-

dividuals (cases) with BMI greater than $40kg/m^2$, and 150 controls with BMI lower than $30kg/m^2$. Individual samples were re-sequenced around two genes known to be involved in endocannabinoid metabolism: *FAAH* on chromosome 1, and *MGLL* on chromosome 3. There are $31Kbp$ of re-sequenced data near the *FAAH* gene, and $157Kbp$ near the *MGLL* locus. Bhatia et al. (2010) discovered two $5Kbp$ regions enriched in rare variants located just upstream of the *FAAH* and *MGLL* genes respectively, with 32 RVs in the first region, and 25 RVs in the second region. Our running time analysis makes use of data from the cases only as the control data leads to rather trivial haplotype probability estimates concentrated almost entirely at the ancestral haplotype $(0, 0, \cdots, 0)$ of all zeros, plus a few haplotypes with only one or two 1's. Table 3.1 reports the running times of the collapsed data method and the EML algorithm for estimating the haplotype distributions of the 25 RVs in the *MGLL* region, and the 32 RVs in the *FAAH* region, when the 148 cases are grouped into pools of size 1, 2 and 4 respectively. The lengths of the EML lists of possible haplotypes are also shown. It can be seen that the collapsed data method always runs faster than the EML algorithm, and much faster when the pool size k and

Table 3.1: Running times in seconds of the collapsed data (CD) method and the EML algorithm for estimating the haplotype distributions of the 25 RVs in the *MGLL* region and the 32 RVs in the *FAAH* region when 148 obese individuals are grouped into pools of various sizes.

Pool size	*MGLL*						*FAAH*					
	$k=1$		$k=2$		$k=4$		$k=1$		$k=2$		$k=4$	
	CD	EML	CD	EML	CD	EML	CD	EML	CD	EML	CD	EML
Time*	0.33	1.14	0.36	18.71	0.61	>10 h	0.44	0.72	0.67	126.38	6.45	>10 h
Length†		67		88		136		37		125		611

*Running time.

†Length of haplotype list.

the number L of RVs increase. This is because increasing k and L does not increase the running time of the collapsed data method substantially, whereas the running time of the EML algorithm increases very rapidly. For the *MGLL* region with 25 RVs, the running time of EML is 1.14 seconds on an intel (R) Core (TM) 2 desktop when $k = 1$. The running time increases to 18.71 seconds when $k = 2$, and to more than 10 hours when $k = 4$. One explanation for this huge increase is the fact that when the 148 individuals are grouped into 37 pools of size 4 each, one of the 37 pools has a total allele count of 3 at one site, a count of 1 at five other sites, and 0 at the remaining 19 sites. Thus just for this pool alone, the EML algorithm has to go through $\prod_{l=1}^{L} {}_{K}C_{T_l} = {}_{8}C_{3} \left({}_{8}C_{1}\right)^{5} \left({}_{8}C_{0}\right)^{19} = 1,835,008$ possibilities to come up with the list of haplotypes compatible with the observed pool total. For the *FAAH* region with 32 RVs, the EML algorithm takes 126.38 seconds to run when the pool size is 2 (partly because one of the 74 pools of size 2 has a total allele count of 2 at four sites, a count of 1 at two other sites, and 0 at the remaining 26 sites, leading to $\left({}_{4}C_{2}\right)^{4} \left({}_{4}C_{1}\right)^{2} \left({}_{4}C_{0}\right)^{26} = 20,736$ possibilities). When the pool size is increased to $k = 4$, one of the pools has a total of 2 at five sites, 1 at four sites, and 0 elsewhere, which translates to $\left({}_{8}C_{2}\right)^{5} \left({}_{8}C_{1}\right)^{4} \left({}_{8}C_{0}\right)^{23} \approx 7e10$ possibilities. As a result, the EML algorithm takes more than 10 hours to run again and is aborted. Thus we have seen how $\prod_{l=1}^{L} {}_{K}C_{T_l}$ can get to become very large even though k is as small as 4 and the alleles are rare. The number $\prod_{l=1}^{L} {}_{K}C_{T_l}$ will grow even bigger if the alleles are not rare because of larger T_l values. In contrast, the total number of haplotypes that the collapsed data method has to consider is given by the number of pools with distinct pool total configurations, which is bounded above by the number of pools, and will not grow with k and L.

This is the main reason why the collapsed data method is faster than the EML algorithm, particularly when k and L are increased.

Table 3.2: Estimates of haplotype frequencies for the 25 RVs in the *MGLL* region obtained from pooled genotype data of 148 obese individuals using the collapsed data (CD) method and the EML algorithm, with standard errors in parentheses.

	$k=1$		$k=2$		$k=4$
Positions of "1"s	CD	EML	CD	EML	CD
None	0.7927 (0.0251)	0.7941 (0.0238)	0.7912 (0.0286)	0.8202 (0.0231)	0.7572 (0.0447)
1	0.0536 (0.0145)	0.0505 (0.0132)	0.0497 (0.0171)	0.0397 (0.0126)	0.0549 (0.0311)
2	0.0043 (0.0042)	0.0034 (0.0034)			
3	0.0456 (0.0134)	0.0433 (0.0123)	0.0381 (0.0153)	0.0291 (0.011)	0.0394 (0.0274)
5	0.0043 (0.0042)	0.0034 (0.0034)			
6	0.0043 (0.0042)	0.0034 (0.0034)	0.0067 (0.0067)	0.0034 (0.0034)	
9	0.0085 (0.0060)	0.0072 (0.0051)	0.0133 (0.0093)	0.0079 (0.0056)	
11	0.0043 (0.0042)	0.0034 (0.0034)			
15	0.0043 (0.0042)	0.0034 (0.0034)			
19	0.0043 (0.0042)	0.0068 (0.0048)	0.0067 (0.0067)	0.0069 (0.0048)	0.0214 (0.0212)
20	0.0043 (0.0042)	0.0068 (0.0048)	0.0067 (0.0067)	0.0069 (0.0048)	0.0214 (0.0212)
21	0.0043 (0.0042)	0.0034 (0.0034)			
22	0.0127 (0.0073)	0.0101 (0.0058)	0.0197 (0.0113)	0.0101 (0.0058)	0.0394 (0.0274)
23	0.0043 (0.0042)	0.0034 (0.0034)	0.0067 (0.0067)	0.0034 (0.0034)	0.0214 (0.0212)
24	0.0127 (0.0073)	0.0101 (0.0058)	0.0067 (0.0067)	0.004 (0.0039)	
1, 3	0.0048 (0.0055)	0.0040 (0.0049)	0.0090 (0.0091)	0.0059 (0.0063)	0.0172 (0.0214)
1, 9	0.0034 (0.0040)	0.0029 (0.0034)	0.0032 (0.0057)	0.0022 (0.0035)	0.0259 (0.0181)
1, 15			0.0056 (0.0056)	0.0034 (0.0034)	
1, 24			0.0098 (0.0078)	0.0064 (0.0049)	0.0137 (0.0136)
2, 3			0.0059 (0.0058)	0.0034 (0.0034)	0.0155 (0.0154)
3, 9					0.0155 (0.0154)
3, 11			0.0059 (0.0058)	0.0034 (0.0034)	0.0155 (0.0154)
3, 14	0.0040 (0.0040)	0.0034 (0.0034)	0.0059 (0.0058)	0.0034 (0.0034)	0.0155 (0.0154)
5, 21			0.0067 (0.0067)	0.0034 (0.0034)	
6, 7	0.0250 (0.0101)	0.0203 (0.0082)	0.0314 (0.0138)	0.0182 (0.0081)	0.0394 (0.0274)
19, 20		0.0017 (0.0017)		0.0033 (0.0035)	
1, 3, 15					0.0087 (0.0087)
3, 6, 7	0.0026 (0.0039)	0.0034 (0.0034)	0.0066 (0.0076)	0.0057 (0.0049)	0.0233 (0.0217)
5, 6, 21					0.0214 (0.0212)
6, 7, 24					0.0155 (0.0154)
6, 19, 20		0.0017 (0.0017)			
7, 19, 20		0.0017 (0.0017)			
1, 6, 7, 24	0.0039 (0.0038)	0.0034 (0.0034)			
6, 7, 19, 20	0.0041 (0.0041)	0.0017 (0.0017)	0.0057 (0.0056)	0.0033 (0.0034)	
1, 3, 6, 7, 24			0.0041 (0.0041)	0.0032 (0.0034)	0.0092 (0.0069)
1, 6, 7, 19, 20					0.0077 (0.0089)
1, 12, 13, 22, 25	0.0039 (0.0039)	0.0034 (0.0034)	0.0053 (0.0053)	0.0034 (0.0034)	
1, 12, 13, 22, 24, 25					0.0094 (0.0094)
Sum of other haplotype probabilities		1.44e-13		4.39e-17	

Table 3.2 reports the haplotype frequency estimates for the 25 RVs in the *MGLL* region obtained using the collapsed data method and the EML algorithm, and Table 3.3 does the same for the 32 RVs in the *FAAH* region.

Table 3.3: Estimates of haplotype frequencies for the 32 RVs in the *FAAH* region obtained from pooled genotype data of 148 obese individuals using the collapsed data (CD) method and the EML algorithm, with standard errors in parentheses.

	$k=1$		$k=2$		$k=4$
Positions of "1"s	CD	EML	CD	EML	CD
None	0.7623 (0.0266)	0.7065 (0.0267)	0.7467 (0.0323)	0.7351 (0.0260)	0.7787 (0.0405)
1	0.0044 (0.0044)	0.0034 (0.0034)			
3	0.0044 (0.0044)	0.0034 (0.0034)			
5	0.0044 (0.0044)	0.0034 (0.0034)	0.0080 (0.0080)	0.0034 (0.0034)	
7	0.0044 (0.0044)	0.0068 (0.0048)	0.0080 (0.0080)	0.0073 (0.0051)	
9	0.0044 (0.0044)	0.0034 (0.0034)			
10	0.0132 (0.0076)	0.0101 (0.0058)	0.0232 (0.0133)	0.0101 (0.0058)	
11	0.0044 (0.0044)	0.0034 (0.0034)	0.0080 (0.0080)	0.0034 (0.0034)	0.0179 (0.0178)
14	0.0044 (0.0044)	0.0034 (0.0034)	0.0080 (0.0080)	0.0034 (0.0034)	
17	0.0044 (0.0044)	0.0034 (0.0034)	0.0080 (0.0080)	0.0034 (0.0034)	
20	0.0044 (0.0044)	0.0034 (0.0034)			
21	0.0261 (0.0105)	0.0263 (0.0098)	0.0305 (0.0150)	0.0220 (0.0089)	0.0334 (0.0233)
22	0.0088 (0.0062)	0.0068 (0.0048)	0.0080 (0.0080)	0.0034 (0.0034)	0.0179 (0.0178)
24	0.0132 (0.0076)	0.0304 (0.0100)	0.0232 (0.0133)	0.0300 (0.0105)	0.0179 (0.0178)
25	0.0088 (0.0062)	0.0135 (0.0067)	0.0157 (0.011)	0.0135 (0.0067)	0.0179 (0.0178)
26	0.0044 (0.0044)	0.0034 (0.0034)			
28	0.0597 (0.0155)	0.0670 (0.0149)	0.0705 (0.0215)	0.0598 (0.0141)	0.0471 (0.0267)
30	0.0044 (0.0044)	0.0035 (0.0035)			
31	0.0044 (0.0044)	0.0034 (0.0034)	0.0080 (0.0080)	0.0034 (0.0034)	
32	0.0044 (0.0044)	0.0037 (0.0037)			
2, 25	0.0044 (0.0044)	0.0034 (0.0034)			
3, 24				0.0034 (0.0034)	
5, 28					0.0122 (0.0122)
7, 24	0.0297 (0.0111)	0.0507 (0.0127)	0.0330 (0.0149)	0.0482 (0.0127)	0.0155 (0.0154)
7, 25					0.0155 (0.0154)
10, 21					0.0137 (0.0136)
12, 13	0.0044 (0.0044)	0.0034 (0.0034)	0.0080 (0.0080)	0.0034 (0.0034)	0.0179 (0.0178)
21, 23	0.0043 (0.0043)	0.0034 (0.0034)			
21, 28	0.0021 (0.0041)	9e-04 (0.0036)			
21, 30	0.0041 (0.0042)	0.0032 (0.0034)			
21, 32			0.0071 (0.0071)	0.0034 (0.0034)	
22, 30	0.0043 (0.0043)	0.0033 (0.0034)	0.0077 (0.0077)	0.0037 (0.0037)	
24, 28	0.0031 (0.0041)	0.0034 (0.0034)	5e-04 (0.0067)	0.0010 (0.0036)	0.0054 (0.0129)
26, 28			0.0061 (0.0061)	0.0034 (0.0034)	
28, 30			0.0061 (0.0061)	0.0034 (0.0034)	
28, 31					0.0122 (0.0122)
28, 32	0.0038 (0.0041)	0.0030 (0.0034)	0.0061 (0.0061)	0.0034 (0.0034)	
1, 21, 23			0.0071 (0.0071)	0.0034 (0.0034)	
2, 20, 25			0.0075 (0.0075)	0.0034 (0.0034)	
3, 7, 24			0.0063 (0.0062)		0.0137 (0.0136)
4, 7, 24	0.0042 (0.0042)	0.0034 (0.0034)	0.0063 (0.0062)	0.0034 (0.0034)	0.0137 (0.0136)
7, 10, 24					0.0137 (0.0136)
7, 17, 24					0.0137 (0.0136)
7, 21, 24			0.0028 (0.0061)	0.0019 (0.0035)	
7, 22, 24	0.0038 (0.0041)	0.0034 (0.0034)			
7, 24, 28					0.0128 (0.0141)
7, 24, 30	0.0040 (0.0041)	0.0034 (0.0034)	0.0063 (0.0062)	0.0034 (0.0034)	
14, 21, 32					0.0137 (0.0136)
21, 22, 30			0.0059 (0.0068)	0.0030 (0.0034)	
22, 28, 30					0.0111 (0.0110)
24, 28, 32					0.0102 (0.0101)
26, 28, 30					0.0122 (0.0122)
1, 21, 23, 28					0.0102 (0.0101)
7, 21, 24, 28					0.0067 (0.0077)
7, 21, 24, 30					0.0111 (0.0110)
7, 22, 24, 28			0.0053 (0.0049)	0.0034 (0.0034)	
8, 27, 28, 29		0.0034 (0.0034)		0.0034 (0.0034)	
21, 22, 28, 30					0.0070 (0.0090)
7, 10, 22, 24, 28					0.0082 (0.0073)
8, 26, 27, 28, 29	0.0041 (0.0041)	0.0034 (0.0034)			
8, 9, 26, 27, 28, 29			0.0060 (0.0060)	0.0034 (0.0034)	
2, 8, 9, 20, 25, 26, 27, 28, 29					0.0111 (0.0110)
Sum of other haplotype probabilities		1.30e-24		1.12e-06	

We only report the EML estimates for pool sizes 1 and 2 as it is too time consuming to run the EML algorithm when $k = 4$ (see Table 3.1). For the EML method, haplotypes with estimated probabilities less than 10^{-5} are not listed out separately, but the total probability of all these omitted haplotypes is reported and is found to be negligible. It can be seen that the collapsed data estimates and the EML estimates are quite close to each other. When the pool size is increased, we can see that for the collapsed data method, some haplotypes with a few "1"s drop out from the list of haplotypes with positive probability estimates, and their places are taken by a few haplotypes with more "1"s. This is not surprising in view of rule 3.1 as increasing the pool size will lead to non-zero allele frequencies at more loci. However, the haplotypes being swapped all have small probabilities and do not really change the overall picture of the haplotype distribution. It is also reassuring to see that the collapsed data method captures almost the same haplotypes as the EML method which is based on a much longer list of possible haplotypes. Comparing the columns of Table 3.2 for the same k, we can see that the EML algorithm produces just three more haplotypes, each with an estimated frequency of 0.0017, than the collapsed data estimation method when $k = 1$; and only one more haplotype with an estimated frequency of 0.0033 when $k = 2$. A similar pattern emerges from Table 3.3 for the case of 32 RVs. This suggests that the phenomenon of haplotypes with very few "1"s giving way to haplotypes with more "1"s as pool size increases has more to do with pooling itself than the further collapsing of pooled data and the subsequent use of rule 3.1. In any case, we believe that the problem has been exaggerated in Tables 3.2 and 3.3 due to insufficient number of pools. Because we fix the number of individuals

at 148, there are only 37 pools of size 4. If we have instead 148 pools of 4 individuals each, which would cost about the same to genotype as 148 individuals, then there will be less chance for the collapsed data method to miss important haplotypes. To be concrete, let us consider the haplotype $y = (0, \cdots, 0, 1, 0)$ with a solitary "1" at position 24 for the *MGLL* data. Suppose we assume that $f(y) = 0.013$ and $f(0, \cdots, 0) = 0.8$, which seem reasonable based on the estimates reported in Table 3.2. For a pool of k individuals ($2k$ haplotypes), let T be the sum of the $2k$ haplotypes, and Z the result of collapsing T component-wise according to whether the total allele count at each site is "0" or "at least 1". If one of the $2k$ haplotypes is $y = (0, \cdots, 0, 1, 0)$, and the rest are all ancestral (i.e., the most common haplotype of all zeros), which occurs with probability $2k(0.013)(0.8)^{2k-1}$, then $T = y + (2k-1)0 = y = Z$. Since this is just one but not the only way to get $Z = y$, we can conclude that $P(Z = y) \geq 2k(0.013)(0.8)^{2k-1}$. It follows that if the pool size is $k = 2$, we have $P(Z = y) \geq 4(0.013)(0.8)^3 = 0.0266$. With 74 pools of size 2 each, the probability that $Z \neq y$ for all 74 pools (which will cause the collapsed data method to miss this haplotype according to rule 3.1) is less than or equal to $(1 - 0.0266)^{74} = 0.136$. Thus the probability that the collapsed data method based on 74 pools of size 2 each will not miss $y = (0, \cdots, 0, 1, 0)$ is at least 0.864, which is indeed the case in Table 3.2. When the pool size is $k = 4$, $P(Z = y) \geq 8(0.013)(0.8)^7 = 0.0218$. With only 37 pools of size 4, the probability that y is missed by the collapsed data method is less than or equal to $(1-0.0218)^{37} = 0.442$, and so the probability that $y = (0, \cdots, 0, 1, 0)$ is not missed is at least 0.558. The fact that the lower bound is only 0.558 explains why $y = (0, \cdots, 0, 1, 0)$ disappears from the list in Table 3.2 when $k = 4$. Have we had 148 pools

of size 4 each, the probability of missing $y = (0, \cdots, 0, 1, 0)$ is smaller than or equal to $(1 - 0.0218)^{148} = 0.038$, and so the probability that the collapsed data method will not miss y is at least 0.962. The above argument can be generalized to show that for every y with $f(y) > 0$ and for every pool size, we can find the number of pools necessary to ensure that the collapsed data method will not miss y with probability larger than some desired value, say, 0.95. Thus increasing the number of pools will alleviate the problem of missing haplotypes.

3.2.3 Variance and efficiency formulae

The collapsed data method is very fast to run; the only concern is that there might be some loss of estimation efficiency. To investigate this, we will derive some variance and efficiency formulae, and show that the collapsed data method is well suited to estimating certain union type probabilities when the variants are rare. We start with $f(0, \cdots, 0)$, the probability that all the alleles are zeros. As $\hat{g}(0, \cdots, 0)$ is just the sample proportion of pools with zero sums,

$$\text{var}\left[\hat{g}(0, \cdots, 0)\right] = \frac{g(0, \cdots, 0)\left[1 - g(0, \cdots, 0)\right]}{n}.$$

Now $\hat{f}(0, \cdots, 0) = \hat{g}(0, \cdots, 0)^{\frac{1}{K}}$ according to (3.3), and so we can use the delta method to obtain

$$\begin{aligned}\text{var}\left[\hat{f}(0, \cdots, 0)\right] &= \left[\frac{1}{K} g(0, \cdots, 0)^{\frac{1}{K}-1}\right]^2 \frac{g(0, \cdots, 0)\left[1 - g(0, \cdots, 0)\right]}{n} \\ &= \frac{f(0, \cdots, 0)^{2-K}\left[1 - f(0, \cdots, 0)^K\right]}{nK^2}\end{aligned} \tag{3.7}$$

as the asymptotic variance of $\hat{f}(0,\cdots,0)$ when the number of pools n increases. If complete haplotype data were available (i.e. if all $2nk$ haplotypes in the n pools were known), the complete data MLE of $f(0,\cdots,0)$ is simply $\hat{f}_C(0,\cdots,0) = n_Y(0)/(2nk)$, where $n_Y(0)$ is the number of haplotypes of ancestral type (i.e., consisting of all zeros) in the sample of $2nk$ haplotypes. It has variance

$$\text{var}\left[\hat{f}_C(0,\cdots,0)\right] = \frac{f(0,\cdots,0)\left[1-f(0,\cdots,0)\right]}{2nk}.$$

As n increases, the asymptotic efficiency of $\hat{f}(0,\cdots,0)$ relative to $\hat{f}_C(0,\cdots,0)$ is

$$\begin{aligned}\text{ARE}\left[\hat{f}(0,\cdots,0),\hat{f}_C(0,\cdots,0)\right] &= \frac{\text{var}\left[\hat{f}_C(0,\cdots,0)\right]}{\text{var}\left[\hat{f}(0,\cdots,0)\right]} \\ &= Kf(0,\cdots,0)^{K-1}\left[\frac{1-f(0,\cdots,0)}{1-f(0,\cdots,0)^K}\right] \\ &= \frac{Kf(0,\cdots,0)^{K-1}}{1+f(0,\cdots,0)+\cdots+f(0,\cdots,0)^{K-1}}.\end{aligned} \tag{3.8}$$

If all the variants are rare, $f(0,\cdots,0)$ will be close to 1, and $\text{ARE}\left[\hat{f}(0,\cdots,0),\hat{f}_C(0,\cdots,0)\right]$ will also be close to 1. Often we will be dealing with a large number of markers, and a proper framework to handle this is to assume that $f(0,\cdots,0)$ tends to a limit c as L increases. This is reasonable because even though each variant is rare, when there is enough of them, there is a non-negligible probability that at least 1 of the rare variants will occur. Mathematically, if the L variants occur independently with a small probability a_L/L each, and $\lim_{L\to\infty} a_L = a > 0$, then $\lim_{L\to\infty} f(0,\cdots,0) = \lim_{L\to\infty}\left(1-\frac{a_L}{L}\right)^L = e^{-a} = c < 1$, which is just the Poisson

law for rare events, and the limiting ARE becomes

$$\frac{Kc^{K-1}}{1+c+\cdots+c^{K-1}} < 1.$$

If the variants are so rare that the probability of occurrence of each variant is a_L/L, with $\lim_{L\to\infty} a_L = a = 0$, then $c = 1$, and $\hat{f}(0,\cdots,0)$ becomes fully efficient.

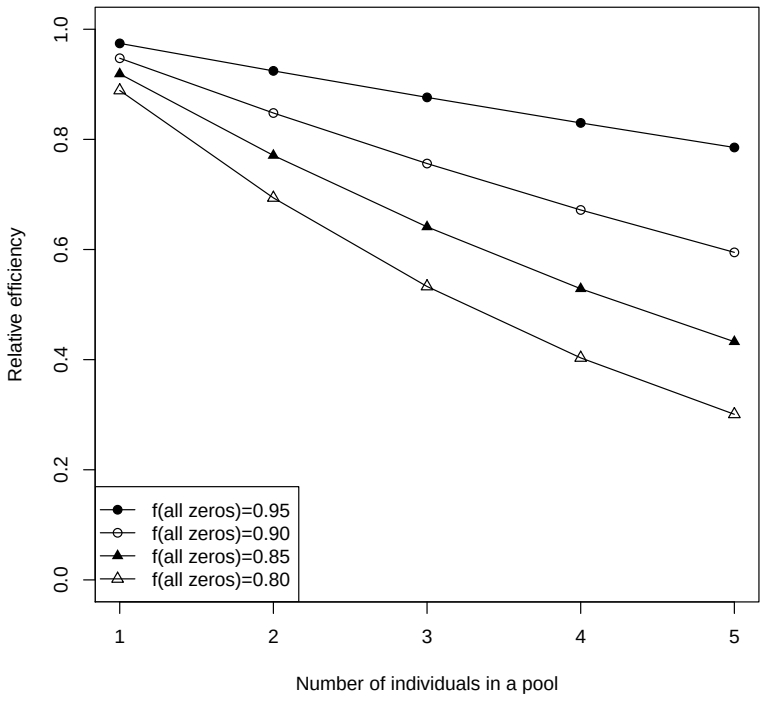

Figure 3.1: Asymptotic relative efficiency of the collapsed data MLE versus the complete data MLE of the haplotype frequency of all zeros for various choices of the true frequency.

To see it graphically, we plot in Figure 3.1 the asymptotic relative efficiency given by (3.8) as a function of pool size $k = K/2$ for 4 choices of $f(0,\cdots,0)$, 0.8, 0.85, 0.9 and 0.95. We can see that the ARE of $\hat{f}(0,\cdots,0)$ relative to $\hat{f}_C(0,\cdots,0)$ depends on how large $f(0,\cdots,0)$ is (i.e., how rare

the alleles are) and the pool size. There is not much loss in efficiency in using $\hat{f}(0,\cdots,0)$ instead of $\hat{f}_C(0,\cdots,0)$ for pools of size 1 as the AREs are all very close to 1. When the pool size is 2, the ARE ranges from 69% to 92% as $f(0,\cdots,0)$ increases from 0.8 to 0.95. If the alleles are so rare that $f(0,\cdots,0) = 0.95$, the ARE lies above 80% even for pool size 5. Note that the ARE that we compute using (3.8) is conservative in 2 ways. First, we are benchmarking $\hat{f}(0,\cdots,0)$ against an estimator $\hat{f}_C(0,\cdots,0)$ which cannot be computed from the observed pooled genotype data as it requires complete haplotype information. Secondly, the ARE formula (3.8) does not take costs into consideration. Suppose it is M times more expensive to obtain the haplotype information of an individual than to genotype the pooled sample of a group of individuals, then for the same cost of genotyping n pools of k individuals each, we can only obtain haplotype information for $n_I = n/M$ individuals. Thus a fairer comparison is to compare var $\left[\hat{f}(0,\cdots,0)\right]$ with $f(0,\cdots,0)\left[1-f(0,\cdots,0)\right]/(2n_I)$, and so we should multiply (3.8) by Mk. Even under the conservative assumption that $M = 1$ (i.e., cost of phasing = cost of genotyping), we should multiply (3.8) by k, the pool size.

The variance of the EM estimates (based on pooled genotyped data) should be in between that of $\hat{f}(0,\cdots,0)$ (obtained by collapsing the pooled genotype data) and $\hat{f}_C(0,\cdots,0)$ (based on complete haplotype data), but the EM algorithm is a lot more computing intensive than the collapsed data method. Thus if it can be demonstrated that the simpler estimator $\hat{f}$ loses little efficiency relative to the gold standard $\hat{f}_C$, then there is no need to compute the EM estimate.

We consider now more generally the relative efficiency for estimating

$f_0(\Lambda) = P(Y_{lij} = 0, l \in \Lambda)$ for non-empty subsets Λ of $\{1, \cdots, L\}$. Note that $f(0, \cdots, 0)$ is just $f_0(\Lambda)$ when $\Lambda = \{1, \cdots, L\}$. Using derivations similar to those used to derive (3.8), we obtain

$$\text{ARE}\left[\hat{f}_0(\Lambda), \hat{f}_{0C}(\Lambda)\right] = \frac{K f_0(\Lambda)^{K-1}}{1 + f_0(\Lambda) + \cdots + f_0(\Lambda)^{K-1}} \tag{3.9}$$

as the generalization of (3.8), where $\hat{f}_{0C}(\Lambda)$ is the complete data MLE of $f_0(\Lambda)$. By definition, $f_0(\Lambda) \geq f(0, \cdots, 0)$ and so

$$\begin{aligned}
\text{ARE}\left[\hat{f}_0(\Lambda), \hat{f}_{0C}(\Lambda)\right] &= \frac{K f_0(\Lambda)^{K-1}}{1 + f_0(\Lambda) + \cdots + f_0(\Lambda)^{K-1}} \\
&\geq \frac{K f(0, \cdots, 0)^{K-1}}{1 + f(0, \cdots, 0) + \cdots + f(0, \cdots, 0)^{K-1}} \\
&= \text{ARE}\left[\hat{f}(0, \cdots, 0), \hat{f}_C(0, \cdots, 0)\right].
\end{aligned}$$

The implication of this result is that relative to the complete data MLE, the collapsed data MLE will do even better in estimating the probability of a subset of zeros rather than the probability of all zeros.

Summarizing our results so far, the proposed method is well suited to estimating probabilities of the type $f_0(\Lambda) = P(Y_{lij} = 0, l \in \Lambda)$, and with little loss of efficiency if the variants are rare and the pool size is not too large. This serves our purpose well because probabilities like $f_0(\Lambda) = P(Y_{lij} = 0, l \in \Lambda)$ play a special role in the study of rare genetic variants. As the frequency of occurrence of each rare variant is very low, some kind of collapsing is needed. Bhatia et al. (2010) defined the union of a set of variants to take on the value 1 when at least one of the variants occurs, and 0 if none of the variants occurs. If $\Lambda \subset \{1, \cdots, L\}$ indexes the variants of which we are taking the union, then the probability that the union

variant occurs is $p_U(\Lambda) = 1 - f_0(\Lambda)$ which of course can be estimated by $\hat{p}_U(\Lambda) = 1 - \hat{f}_0(\Lambda)$, where $\hat{f}_0(\Lambda)$ is given by (3.5). Analogous to (3.7), we can obtain

$$\text{var}\left[\hat{p}_U(\Lambda)\right] = \text{var}\left[\hat{f}_0(\Lambda)\right] = \frac{\hat{f}_0(\Lambda)^{2-K}\left[1 - \hat{f}_0(\Lambda)^K\right]}{nK^2}$$

from the binomial variance of $\hat{g}_0(\Lambda)$ using delta method. The asymptotic relative efficiency of $\hat{p}_U(\Lambda) = 1 - \hat{f}_0(\Lambda)$ is the same as that of $\hat{f}_0(\Lambda)$ and so is also given by (3.9).

We end this section by discussing variance estimation for the haplotype frequency estimates $\hat{f}(y), y \in \{0,1\}^L$. Since it follows from (3.6) that every $\hat{f}(y)$ can be expressed as a linear combination of the probabilities $\hat{f}_0(\Lambda)$, we can obtain the variance of $\hat{f}(y)$ from the covariance matrix of $\hat{f}_0(\Lambda), \Lambda \subset \{1, \cdots, L\}$. To obtain the covariance matrix of $\hat{f}_0(\Lambda)$, we use variance formula like the one above and the fact that

$$\begin{aligned}
&\text{cov}\left[\hat{f}_0(\Lambda_1), \hat{f}_0(\Lambda_2)\right] \\
=&\left[\frac{1}{K} g_0(\Lambda_1)^{\frac{1}{K}-1}\right]\left[\frac{1}{K} g_0(\Lambda_2)^{\frac{1}{K}-1}\right]\frac{g_0(\Lambda_1 \cup \Lambda_2) - g_0(\Lambda_1) g_0(\Lambda_2)}{n} \\
=&\frac{f_0(\Lambda_1)^{1-K} f_0(\Lambda_2)^{1-K}\left[f_0(\Lambda_1 \cup \Lambda_2)^K - f_0(\Lambda_1)^K f_0(\Lambda_2)^K\right]}{nK^2}.
\end{aligned}$$

3.3 An Analysis of Rare Variants Associated with Obesity

As many common diseases cannot be explained by common variants, there is a theory that they might be caused by multiple rare variants. The lack of rare variant data had hampered the investigation of this "common disease, rare variant" hypothesis, but recent advances in technology have made it

possible to re-sequence large stretches of a genome in a cost effective way. Thus the time is ripe for rare variant analysis, and as a result there is a huge surge of papers on this topic. The analysis of rare variants presents many new challenges. Most existing methods of data analysis are not designed with rare attributes in mind and their naive applications will lead to imprecise estimates and tests of low power. As a result, various strategies to handle rare variant data have been proposed, including collapsing (Li and Leal, 2008), weighting (Madsen and Browning, 2009), thresholding (Price et al., 2010), C-alpha test (Neale et al., 2011) based on comparing the expected variance with the actual variance of the distribution of allele frequencies, score-type tests (Lin and Tang, 2011) and the sequence kernel association test (Wu et al., 2011). All the above procedures are concerned with testing the overall significance of a collection of variants rather than variant selection, and are not developed with pooled data in mind which, as noted by Kim et al. (2010), could be important for rare variants. Bhatia et al. (2010) proposed a covering method called "RARECOVER" for selecting rare variants associated with common diseases. RARECOVER is basically a step-up greedy procedure whereby at each step, the variant which maximizes the Pearson's chi square statistic upon taking union with the variants selected so far is added to the existing set if the increase in Pearson's statistic exceeds a certain threshold. Applying RARECOVER to the case control data described earlier in section 3.2.2, Bhatia et al. (2010) identified 12 RVs out of the 25RVs in a $5Kbp$ window just upstream of the *MGLL* gene as potential causal variants of obesity.

To provide an alternative analysis, we use the collapsed data method to compute the haplotype frequency estimates separately for the 148 cases and

150 controls, and the results are shown in the left panel of Table 3.4. Due to limited information, it is difficult to establish significance based on single variant analyses, and it is important to borrow strength across different rare variants. It is hoped that we can select variants with low individual minor allele frequency in such a way that their union has a higher frequency of occurrence of say, 10 to 20%. As argued earlier, the collapsed data method is particularly well suited to the estimation of union probabilities with high efficiency given by (3.9) for the case of rare variants and small pool size. To identify which RVs to include in the union, we compare the haplotype frequency estimates reported in the left panel of Table 3.4 for the cases and controls. In view of our earlier comment that it is the insufficient power of single variant analysis which necessitates the formation of union variants, we will use a more liberal level of 10% significance to perform the initial screening of haplotypes. The three haplotypes depicted in bold type in the left panel of Table 3.4 are those with case frequency significantly higher than the control frequency at the 10% level (i.e., standardized difference greater than 1.28). From the haplotypes with only one "1", we pick out RV3 (chr3_129031107) and RV9 (chr3_129031864) as candidate RVs that are associated with obesity. From the haplotypes with two "1"s, we pick out RV6 (chr3_129031590) and RV7 (chr3_129031591). For these 4 RVs, their union probability is estimated to be .0958 (SE = .0176) for the cases and .027 (SE = .0094) for the controls to result in a standardized difference of 3.45. A closer examination of Table 3.4 reveals that there is some redundancy between RV6 and RV7 as these 2 RVs tend to occur together always. Thus it may be sufficient to include only one of them in the union variant. Between the two, RV6 is slightly more significant (marginal frequency of 10

for cases versus 3 for controls, as opposed to 9 and 3). If we leave out RV7, the union probability of the remaining RV3, RV6 and RV9 is estimated to be .0958 (SE = .0176) for the cases and .0236 (SE = .0088) for the controls to result in a larger standardized difference of 3.67. If we replace RV6 by RV7 in the union, the union probability is estimated to be .0921 (SE = .0172) for the cases and .027 (SE = .0094) for the controls with a smaller standardized difference of 3.32. Since the probability of haplotype is small, we also calculate the relative risk. Z scores for the logarithm of relative risk were 3.22 for the four rare variants (RV3, RV6, RV7 and RV9), 3.37 if we leave out RV7 and 3.31 if we leave out RV6. Similar results are obtained using both difference and relative risk. In view of the above discussion, we select RV3, RV6 and RV9 as potential causal RVs out of the total of 25 RVs. This is a more parsimonious selection than the 12 RVs picked by RARECOVER.

Haplotype based inference can provide more information than single marker analyses. For example, the high linkage disequilibrium between RVs 6 and 7 for the *MGLL* data mentioned above was discovered by inspecting the left panel of Table 3.4. To give another example, we look at the right panel of Table 3.4, which shows the collapsed data estimates of the haplotype frequencies in the *FAAH* region. It can be seen that RVs 7 and 24 (depicted in bold type) occur together a lot. Based on single locus allele frequencies, RV24 is not associated with obesity (total allele count is 28 for cases, 24 for controls, with a standardized difference of 0.63), whereas RV7 is marginally significant (20 versus 11 for a standardized difference of 1.7). From the right panel of Table 3.4, we see that the estimated frequencies that only RV7 occurs (out of the 32 RVs in the *FAAH* region) are about

Table 3.4: Estimates of haplotype frequencies and probabilities of various variant combinations for the 25 RVs in the *MGLL* region and the 32 RVs in the *FAAH* region obtained by collapsing data from 148 cases and 150 controls, with $k = 1$ and standard errors in parentheses.

MGLL			*FAAH*		
Positions of "1"s	Cases	Controls	Positions of "1"s	Cases	Controls
None	0.7927 (0.0251)	0.8981 (0.0180)	None	0.7623 (0.0266)	0.8327 (0.0226)
1	0.0536 (0.0145)	0.0364 (0.0113)	1	0.0044 (0.0044)	
2	0.0043 (0.0042)		3	0.0044 (0.0044)	
3	0.0456 (0.0134)	0.0147 (0.0073)	5	0.0044 (0.0044)	
4		0.0037 (0.0037)	6		0.0040 (0.0040)
5	0.0043 (0.0042)		7	0.0044 (0.0044)	0.0040 (0.0040)
6	0.0043 (0.0042)		9	0.0044 (0.0044)	
7			10	0.0132 (0.0076)	0.0119 (0.0068)
8		0.0074 (0.0052)	11	0.0044 (0.0044)	0.0159 (0.0079)
9	0.0085 (0.006)		14	0.0044 (0.0044)	
10		0.0037 (0.0037)	15		0.0040 (0.0040)
11	0.0043 (0.0042)		17	0.0044 (0.0044)	
15	0.0043 (0.0042)		18		0.0040 (0.0040)
19	0.0043 (0.0042)		19		0.0040 (0.0040)
20	0.0043 (0.0042)		20	0.0044 (0.0044)	
21	0.0043 (0.0042)		21	0.0261 (0.0105)	0.0198 (0.0088)
22	0.0127 (0.0073)	0.0037 (0.0037)	22	0.0088 (0.0062)	
23	0.0043 (0.0042)		24	0.0132 (0.0076)	0.0314 (0.0109)
24	0.0127 (0.0073)	0.0111 (0.0064)	25	0.0088 (0.0062)	0.0198 (0.0088)
1, 3	0.0048 (0.0055)		26	0.0044 (0.0044)	
1, 9	0.0034 (0.0040)		28	0.0597 (0.0155)	0.0353 (0.0116)
1, 16		0.0036 (0.0036)	30	0.0044 (0.0044)	
1, 17		0.0036 (0.0036)	31	0.0044 (0.0044)	
1, 18		0.0036 (0.0036)	32	0.0044 (0.0044)	
3, 14	0.0040 (0.0040)	0.0036 (0.0036)	2, 25	0.0044 (0.0044)	
6, 7	0.0250 (0.0101)	0.0037 (0.0037)	6, 16		0.0040 (0.0040)
3, 6, 7	0.0026 (0.0039)	0.0036 (0.0036)	**7, 24**	0.0297 (0.0111)	0.0188 (0.0084)
1, 6, 7, 24	0.0039 (0.0038)		12, 13	0.0044 (0.0044)	
6, 7, 19, 20	0.0041 (0.0041)		21, 23	0.0043 (0.0043)	
1, 12, 13, 22, 25	0.0039 (0.0039)		21, 30	0.0041 (0.0042)	
			21, 28	0.0021 (0.0041)	0.0030 (0.0038)
			22, 30	0.0043 (0.0043)	
			24, 28	0.0031 (0.0041)	
			28, 32	0.0038 (0.0041)	
			4, **7, 24**	0.0042 (0.0042)	
			7, 24, 30	0.0040 (0.0041)	
			7, 22, **24**	0.0038 (0.0041)	
			8, 26, 27, 28, 29	0.0041 (0.0041)	
Variant combinations			Variant combinations		
3 or 6 or 7 or 9	0.0958 (0.0176)	0.0270 (0.0094)	7 but not 24	0.0044 (0.0044)	0.0040 (0.0040)
3 or 6 or 9	0.0958 (0.0176)	0.0236 (0.0088)	24 but not 7	0.0163 (0.0081)	0.0314 (0.0109)
3 or 7 or 9	0.0921 (0.0172)	0.0270 (0.0094)	7 and 24	0.0416 (0.0129)	0.0188 (0.0084)

the same for cases and controls, .0044 versus .04, and the standardized difference is .07. Interestingly, the estimated frequency of RV24 occurring only is higher among the controls than cases (.0314 versus .0132), and the standardized difference is -1.37 which seems to suggest weakly that RV24 alone may be a protective variant. The estimated frequency of both RVs 7 and 24 occurring (and nothing else) is .0297 which is higher than the corresponding frequency of .0188 for the controls, even though the standardized difference is only .78. These findings suggest that it may be worthwhile to study the association between obesity and the following 2-variant combinations: "RV7 but not RV24", "RV24 but not RV7" and "both RV7 and RV24". The frequencies of occurrences of these combinations can be obtained separately for the cases and controls by summing the frequencies of all those haplotypes satisfying the given conditions, and standard errors of these sums can be obtained from the variances and covariances of the haplotype frequency estimates. This is done at the lower right hand corner of Table 3.4. We can see that "RV7 but not RV24" is not associated with obesity (standardized difference is .07). For the combination "RV24 but not RV7", the standardized difference is -1.11. The negative difference suggests protectiveness, although its magnitude is not large enough to be statistically significant (2-sided p value is .267). Finally, for "both RV7 and RV24", the standardized difference is 1.48 with a 1-sided p value of .07. We use a 1-sided p value here since we are primarily interested in finding variants which cause obesity. Thus our analysis based on haplotype frequency estimates suggests that there is potentially an interesting interaction effect between RV7 and RV24 on obesity. Even though the evidence is not conclusive, there is a hint that both variants have to occur to cause

obesity. Further investigation is required to verify or disprove the above conjectures. With 32 RVs in the *FAAH* region, there are 496 pairs of RVs to explore for possible interaction. By looking closely at the 40 or so *FAAH* haplotypes with positive estimated probabilities listed in Table 3.4, we are able to flag out RVs 7 and 24 as a potentially interesting pair. Lin and Zeng (2006) also noted that it is the sparseness of the estimated haplotype frequency distribution which makes "haplotyping an effective data-reduction strategy".

3.4 Discussion and Extensions

We have proposed a very fast method to compute haplotype frequency estimates from individual or pooled genotype data which is non-iterative in nature, feasible for all pool size and number of markers, and is significantly faster than the EML algorithm. This is made possible by collapsing the total allele counts to "0" and "≥ 1". Efficiency calculation suggests that the method is well suited to the estimation of union probabilities with little loss of efficiency if the variants are rare and the pool size is not too large. We conclude that the proposed method is adequate and useful for the purpose of screening out rare variants to form union variants, and highly reliable for estimating union probabilities. Even if the variants are not rare, the proposed method can be used to provide quick initial estimates of the haplotype frequencies. It is also valuable as a quick screening tool to filter a large number of variants down to a more manageable set for further study.

In obtaining the results for Tables 3.2, 3.3 and 3.4, we have assumed that we can observe the total allele counts exactly when K haplotypes are pooled together. In practice, pooling will add noise to the data. One way to

simulate this noise is to use the binomial model proposed by Quade et al. (2005). To be specific, suppose the pool size is k, and the true total allele count at a particular locus is t (i.e., in the pool of $K = 2k$ haplotypes, t of them have the allele "1" at that locus, and the remaining $2k - t$ haplotypes have the allele "0"). We assume that the genetic materials of the k individuals are mixed together in equal proportions and a total of KD readings are to be taken from the pooled sample so that there are D reads per haplotype on the average, which is known as the sequencing depth. This results in an amplified count $A \sim \text{Binomial}\,(KD, p = t/K)$, and we divide A by D, and round it to the nearest integer to obtain the observed count y. If A/D is midway between two integers, we pick one of them as y with equal probability. Note that for a fixed K, $A/D \to t$ in probability as the sequencing depth D increases, and so the amount of noise decreases with D. In practice, we often fix the total number of reads KD, and so the depth will decrease with pool size, which serves to explain why increasing the pool size can lead to noisier data. To investigate the effect of noise induced by pooling on the collapsed data estimator, we take the *MGLL* data as the "true" data, and we add noise at every locus independently using the above binomial model with pool size 1, 2 and 4, and $KD = 64$. This is done 200 times for each pool size. For each set of "noisy" data generated, we use the collapsed data method to estimate the haplotype frequencies. The average estimates over the 200 samples and their standard errors are reported in Table 3.5. It can be seen that the averages of the "noisy" data estimates are quite close to the collapsed data estimates based on the true allele count data without noise. The standard error increases with pool size, but this is to be expected because we fix the total number of reads at 64, and so the

sequencing depth decreases from $D = 32$ when the pool size is 1 to $D = 8$ when the pool size is 4. But even when the pool size is 4, there is good agreement between the estimates with and without noise added.

Table 3.5: Collapsed data estimates of haplotype frequencies for the 25 RVs in the *MGLL* region with and without "noise" added to the pooled genotype data of 148 obese individuals, with standard errors in parentheses.

	$k=1$		$k=2$		$k=4$	
Positions of "1"s	No noise	Noise added	No noise	Noise added	No noise	Noise added
None	0.7927	0.7927 (0)	0.7912	0.7929 (0.0037)	0.7572	0.7718 (0.0150)
1	0.0536	0.0536 (0)	0.0497	0.0493 (0.0025)	0.0549	0.0552 (0.0129)
2	0.0043	0.0043 (0)				
3	0.0456	0.0456 (0)	0.0381	0.0377 (0.0025)	0.0394	0.0401 (0.0134)
5	0.0043	0.0043 (0)				
6	0.0043	0.0043 (0)	0.0067	0.0070 (0.0020)		
9	0.0085	0.0085 (0)	0.0133	0.0134 (0.0012)		
11	0.0043	0.0043 (0)				
15	0.0043	0.0043 (0)				
19	0.0043	0.0043 (0)	0.0067	0.0067 (0.0001)	0.0214	0.0191 (0.0023)
20	0.0043	0.0043 (0)	0.0067	0.0067 (0.0001)	0.0214	0.0191 (0.0023)
21	0.0043	0.0043 (0)				
22	0.0127	0.0127 (0)	0.0197	0.0194 (0.0012)	0.0394	0.0337 (0.0089)
23	0.0043	0.0043 (0)	0.0067	0.0066 (0.0009)	0.0214	0.0180 (0.0054)
24	0.0127	0.0127 (0)	0.0067	0.0067 (0.0011)		
1, 3	0.0048	0.0048 (0)	0.0090	0.0086 (0.0017)	0.0172	0.0120 (0.0083)
1, 9	0.0034	0.0034 (0)	0.0032	0.0032 (0.0006)	0.0259	0.0187 (0.0078)
1, 15			0.0056	0.0055 (0.0008)		
1, 24			0.0098	0.0096 (0.0011)	0.0137	0.0098 (0.0050)
2, 3			0.0059	0.0056 (0.0011)	0.0155	0.0109 (0.0059)
3, 9					0.0155	0.0110 (0.0053)
3, 11			0.0059	0.0057 (0.0008)	0.0155	0.0117 (0.0053)
3, 14	0.0040	0.0040 (0)	0.0059	0.0058 (0.0004)	0.0155	0.0116 (0.0054)
5, 21			0.0067	0.0064 (0.0013)		
6, 7	0.0250	0.0250 (0)	0.0314	0.0303 (0.0023)	0.0394	0.0329 (0.0121)
1, 3, 15					0.0087	0.0061 (0.0036)
3, 6, 7	0.0026	0.0026 (0)	0.0066	0.0064 (0.0013)	0.0233	0.0151 (0.0086)
5, 6, 21					0.0214	0.0140 (0.0084)
6, 7, 24					0.0155	0.0099 (0.0064)
1, 6, 7, 24	0.0039	0.0039 (0)				
6, 7, 19, 20	0.0041	0.0041 (0)	0.0057	0.0055 (0.0007)		
1, 3, 6, 7, 24			0.0041	0.0039 (0.0010)	0.0092	0.0051 (0.0037)
1, 6, 7, 19, 20					0.0077	0.0061 (0.0028)
1, 12, 13, 22, 25	0.0039	0.0039 (0)	0.0053	0.0050 (0.0012)		
1, 12, 13, 22, 24, 25					0.0094	0.0057 (0.0043)

A good experimental design may include replicates obtained, for example, by performing multiple assaying of the specimens from each pool of individuals. Let $T_i^{(r)} = \left(T_{1i}^{(r)}, \cdots, T_{Li}^{(r)}\right)$ be the total allele counts observed at the L loci for the r^{th} replicate of the i^{th} pool, $i = 1, \cdots, n$, $r = 1, \cdots, R$; and $Z_i^{(r)} = \left(Z_{1i}^{(r)}, \cdots, Z_{Li}^{(r)}\right) = \left(I\left\{T_{1i}^{(r)} \geq 1\right\}, \cdots, I\left\{T_{Li}^{(r)} \geq 1\right\}\right)$ the data

that result from collapsing each total count to "0" or "at least 1". The collapsed data estimation method can be extended easily to handle replicated data by noting the following. We can see from (3.6) that to estimate the haplotype probability function $f(y_1, \cdots, y_L)$, it suffices to know how to estimate the probability $g_0(\Lambda)$ that the collapsed counts are zeros at a subset of loci for all possible choices of the subset Λ. Without replicates, the collapsed data estimator of $g_0(\Lambda)$ is

$$\hat{g}_0(\Lambda) = \frac{\sum_{i=1}^{n} I\{Z_{li} = 0 \text{ for } l \text{ in } \Lambda\}}{n} = \frac{n_{0Z}(\Lambda)}{n}, \tag{3.10}$$

where $n_{0Z}(\Lambda) = \sum_{i=1}^{n} I\{Z_{li} = 0 \text{ for } l \text{ in } \Lambda\}$ is simply counting the number of pools with zero allele counts at the sites specified by Λ, and we can estimate the variance of $\hat{g}_0(\Lambda)$ by the binomial variance formula

$$\widehat{\text{var}}[\hat{g}_0(\Lambda)] = \frac{\hat{g}_0(\Lambda)[1 - \hat{g}_0(\Lambda)]}{n}.$$

When there are replicates, the obvious extension of (3.10) is

$$\hat{g}_0(\Lambda) = \frac{\sum_{i=1}^{n}\sum_{r=1}^{R} I\left\{Z_{li}^{(r)} = 0 \text{ for } l \text{ in } \Lambda\right\}}{nR} = \frac{\sum_{i=1}^{n} P_i}{n} = \bar{P}, \tag{3.11}$$

where $P_i = R^{-1}\sum_{r=1}^{R} I\left\{Z_{li}^{(r)} = 0 \text{ for } l \text{ in } \Lambda\right\}$ is the proportion of replicates in the i^{th} pool with zero allele counts at the sites specified by Λ. The binomial variance formula is no longer applicable because the replicated measurements from the same pool are likely to be positively correlated. Since $\hat{g}_0(\Lambda)$ defined by (3.11) is just the sample mean $\bar{P}$ of the P_i from independent pools of individuals, we can use the following variance estimate

instead

$$\widehat{\text{var}}\left[\hat{g}_0\big(\Lambda\big)\right] = \widehat{\text{var}}\left[\bar{P}\right] = \frac{s_P^2}{n}, \tag{3.12}$$

where s_P^2 is the sample variance of $P_1, \cdots, P_n$. The ratio of $n^{-1}s_P^2$ to the binomial variance estimate $\hat{g}_0(\Lambda)\left[1 - \hat{g}_0(\Lambda)\right]/(nR)$ can be viewed as an over-dispersion or variance inflation factor. Finally, we can substitute (3.11) into (3.6) to obtain estimates of $f(y_1, \cdots, y_L)$, and we can obtain variance estimates using (3.12) and the delta method.

Chapter 4

EM with an Internal List

This chapter is organized as follows. Section 4.1 highlights the main findings of our method; Section 4.2 describes the details of our expectation maximization (EM) algorithm restricted within an internal list based on collapsed data; Section 4.3 considers a real data analysis and a simulation study; Section 4.4 concludes this chapter with some discussion.

The materials presented in this chapter have been published in Kuk et al. (2013a).

4.1 Summary

Gasbarra et al. (2011) advocate the use of database information to create a list of frequently occurring haplotypes. We do not assume the existence of an external list for two reasons. First, database information for rare alleles is currently still lacking. Secondly, an EM type algorithm restricted to a list is sensitive to the correct choice and completeness of the external list used. Instead, we use the data on hand to construct an internal list.

Motivated by the collapsed data estimation method developed by Kuk et al. (2013b) which only keeps track of whether an allele count is "0" or

"≥ 1", we propose a collapsed data (CD) list of possible haplotypes. It will be shown in section 4.2 that for rare genetic variants, the CD list has inflated probabilities of capturing the true underlying haplotypes. To improve coverage, we augment the CD list by adding those haplotypes with only one "1" (i.e., only one rare variant occurs) to result in an augmented CD (ACD) list. The EM algorithm restricted to the ACD list still does not perform satisfactorily in our simulation studies, apparently due to the inclusion of too many false haplotypes. In response, we propose an ATCD (augmented and trimmed CD) list where those haplotypes with estimated frequencies lower than a threshold at each iteration of the algorithm are removed from the list. We propose a method to select the threshold by benchmarking the resulting EM estimate of the frequency of the ancestral haplotype of all zeros (i.e., no variant occurs) with the corresponding estimate obtained using the collapsed data method of Kuk et al. (2013b).

To assess the performance of the various estimators, we simulate genotype data resembling those collected for the 148 obese individuals in the CRESCENDO cohort study http://clinicaltrials.gov/ct/show/NCT00263042, at 25 loci near the *MGLL* gene on chromosome 3, and 32 loci near the *FAAH* gene on chromosome 1. The EM estimates based on the CD list and the ACD list do not perform well in the simulation study. In particular, they over-estimate the haplotype frequency of the ancestral haplotype of all zeros. The EM estimates based on the ATCD list, on the other hand, perform very well. In the two scenarios involving 25 and 32 loci, the EM-ATCDL estimates outperform the EM estimates based on other lists as well as the collapsed data maximum likelihood estimates (MLE). We conclude that the augmented and trimmed CD list is a useful list for the

EM algorithm to base upon in estimating the haplotype distributions of rare variants.

4.2 Statistical Models and Methods

4.2.1 Collapsed data list

Focusing on bi-allelic loci, the two possible alleles at each locus can be represented by "1" (the minor or variant allele) and "0" (the major allele). As a result, the alleles at selected loci of a chromosome can be represented by a binary haplotype vector. Since human chromosomes come in pairs, there are 2 haplotype vectors for each individual, one maternal, and one paternal. Suppose we have n pools of k individuals each so that there are $K = 2k$ haplotypes within each pool. Denote by $Y_{ij} = (Y_{1ij}, \cdots, Y_{Lij})$ the j^{th} haplotype in the i^{th} pool, where $i = 1, \cdots, n$, $j = 1, \cdots, K$, and L is the number of loci to be genotyped. Assuming Hardy-Weinberg equilibrium, the nK haplotype vectors are independent and identically distributed with probability function

$$f(y_1, \cdots, y_L) = P\left(Y_{1ij} = y_1, \cdots, Y_{Lij} = y_L\right)$$

for every L-tuple $y = (y_1, \cdots, y_L)$ belonging to the Cartesian product $\Omega = \{0, 1\}^L$. With pooling, the observed data are the pool totals

$$T_i = \sum_{j=1}^{K} Y_{ij} = \left(\sum_{j=1}^{K} Y_{1ij}, \cdots, \sum_{j=1}^{K} Y_{Lij}\right) = \left(T_{1i}, \cdots, T_{Li}\right), \ i = 1, \cdots, n.$$

The probability function $p\left(t_1, \cdots, t_L\right)$ of each pool total is given by the K-fold convolution of the haplotype probability function $f(y_1, \cdots, y_L)$ and

so the likelihood based on the observed pooled data is highly intractable and not easy to maximize directly.

Kuk et al. (2013b) defined the collapsed data via indicator functions as

$$Z_i = \left(I\left\{\sum_{j=1}^{K} Y_{1ij} \geq 1\right\}, \cdots, I\left\{\sum_{j=1}^{K} Y_{Lij} \geq 1\right\}\right) = (Z_{1i}, \cdots, Z_{Li}) .$$

Note that what Z_i does is to collapse each total allele frequency to either "0" (coded as 0) or "at least 1" (coded as 1) as done in classical group testing (Dorfman, 1943). From here on, we will call $\{Y_{ij}, i = 1, \cdots, n, j = 1, \cdots, K\}$ the complete haplotype data (usually not observed); $\{T_i, i = 1, \cdots, n\}$ the pooled genotype data (reduces to individual genotype data if the pool size is 1), and $\{Z_i, i = 1, \cdots, n\}$ the collapsed data. We refer to k as the pool size, not K.

4.2.2 EM with an internal list

Since the collapsed data is a reduction of the pooled data, the collapsed data MLE is less efficient than the pooled data MLE. Kuk et al. (2013b) showed that the loss of estimation efficiency due to the collapsing of pooled data is not large for rare variants and small pool size. However, if the pool size is moderate or large, which is recommended from the cost saving point of view, an estimator based on the original pooled data without collapsing can be substantially more efficient than the collapsed data MLE. This is why we want to modify the EM algorithm for finding the pooled data MLE to make it computationally feasible.

If the individual haplotypes $Y_{ij},\ i = 1, \cdots, n,\ j = 1, \cdots, K$, were actually observed, then the population haplotype distribution function can be

estimated simply by the empirical haplotype distribution. In other words, the so-called complete data MLE of $f(y),\ y \in \Omega$, is

$$\hat{f}_C(y) = \frac{m(y)}{nK}, \tag{4.1}$$

where $m(y) = \sum_{i=1}^{n} \sum_{j=1}^{K} I\left(Y_{ij} = y\right)$ is the number of times y appears in Y_{ij}. The E-step of the EM algorithm involves taking conditional expectation of $m(y)$ given the observed data and current estimates $\hat{f}^{(t)}(y),\ y \in \Omega$, to get

$$\begin{aligned}
\hat{m}^{(t)}(y) &= \mathrm{E}\left[m(y)|T_1 = t_1, \cdots, T_n = t_n\right] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{K} P\left(Y_{ij} = y|T_i = t_i\right) \\
&= \sum_{i=1}^{n} KP\left(Y_{i1} = y|T_i = t_i\right),
\end{aligned}$$

where

$$\begin{aligned}
P\left(Y_{i1} = y|T_i = t_i\right) &= \frac{P\left(Y_{i1} = y, T_i = t_i\right)}{P\left(T_i = t_i\right)} \\
&= \frac{\sum\limits_{\substack{y_2 \in \Omega, \cdots, y_K \in \Omega \\ y + y_2 + \cdots + y_K = t_i}} \left[\hat{f}^{(t)}(y) \prod_{j=2}^{K} \hat{f}^{(t)}(y_j)\right]}{\sum\limits_{\substack{y_1 \in \Omega, \cdots, y_K \in \Omega \\ y_1 + \cdots + y_K = t_i}} \left[\prod_{j=1}^{K} \hat{f}^{(t)}(y_j)\right]}
\end{aligned} \tag{4.2}$$

Since the complete data multinomial likelihood belongs to the exponential family, the M-step can be carried out analytically to yield the updating formula

$$\hat{f}^{(t+1)}(y) = \frac{\hat{m}^{(t)}(y)}{nK} \tag{4.3}$$

which is just (4.1) with $m(y)$ replaced by the imputed value $\hat{m}^{(t)}(y)$.

The E-step of the EM algorithm is very time consuming. As one can see from (4.2), it involves finding all possible underlying haplotype vectors that sum up to the observed pool total. The combinatorial problem is greatly reduced if we can restrict the possible haplotypes to come from a relatively short list.

Let $R \subset \Omega$ be a reduced list of possible haplotypes obtained by whatever method. The generic EM with a list algorithm operates in the same way as the EM algorithm described above except that the updating formula (4.3) is only applied to $y \in R \subset \Omega$, and Ω is replaced by R under the summation symbols in Equation (4.2).

Kuk et al. (2013b) described a combinatorial method to arrive at a reduced list R, but the resulting EML algorithm is still very time consuming. As can be seen from Table 4.1, the EM with a list (EML) method (Kuk et al., 2013b) is not feasible for pool size larger than 2. Thus there is a need for alternative methods to arrive at a reduced list. Motivated by the fact that the collapsed data MLE $\hat{f}_{CD}(y) > 0$ only for "those haplotypes y which coincide with at least one of the collapsed data vectors $Z_i, i = 1, \cdots, n$, in the sample", it seems sensible to apply the EM algorithm with haplotypes

Table 4.1: Running times of EM algorithms based on different lists

	MGLL			*FAAH*		
	EML	EM-CDL	EM-ATCDL	EML	EM-CDL	EM-ATCDL
$k = 1$	1.14	0.08	3.68	0.72	0.13	4.57
$k = 2$	18.71	0.10	7.05	126.38	0.17	6.78
$k = 4$	> 10 h	0.23	7.39	> 10 h	0.13	27.93

Running times in seconds for EML (EM with combinatorially determined list), EM-CDL (EM with CD list) and EM-ATCDL (EM with augmented and trimmed CD list with adaptive threshold) for estimating the haplotype distributions of the 25 rare variants in the *MGLL* region and the 32 rare variants in the *FAAH* region when 148 obese individuals are grouped into pools of various sizes.

restricted to this list, which we call the CD list. Let y be a non-ancestral haplotype (i.e., $y \neq \mathbf{0}$, the vector of all zeros) with frequency $f(y) > 0$, the probability that it is captured in a list of n randomly sampled haplotypes is $1-[1-f(y)]^n$ ($\approx 1-e^{-nf(y)}$ if $f(y)$ is small and n is large), whereas the probability that it is captured by the CD list constructed from n pools of k individuals each is $1-[1-g(y)]^n \approx 1-e^{-ng(y)}$. Thus if $g(y) > f(y)$, the probability that y is captured by the CD list is higher than the probability that it is captured by direct sampling of haplotypes (not to mention the extra cost incurred in resolving the phase ambiguity to sample the haplotypes directly), and by increasing the number of pools n, we can make the capture probability arbitrarily large. For example, if we want the CD list to capture y with probability at least $1-\varepsilon$, all we have to do is to solve $1-e^{-ng(y)} \geq 1-\varepsilon$ (after Poisson approximation) to get $n \geq \frac{-\log(\varepsilon)}{g(y)}$. A sufficient condition for $g(y)$ to be greater than $f(y)$ is given below.

Lemma 4.1. Let y be non-ancestral, $g(y) > f(y)$ if $f(\mathbf{0}) > \left(\frac{1}{2k}\right)^{\frac{1}{2k-1}}$, where k is the number of individuals in each pool.

Proof. A sufficient condition for $Z = y$ is that one of the $2k$ haplotype vectors $Y_1, \cdots, Y_{2k}$ in a pool of k individuals is equal to $y = (y_1, \cdots, y_L)$, and the other $2k-1$ haplotype vectors are all zero vectors. Thus $g(y) \geq 2kf(y)f(\mathbf{0})^{2k-1}$, and the lemma follows. □

The values of $\left(\frac{1}{2k}\right)^{\frac{1}{2k-1}}$ for various choices of the pool size k are given in Table 4.2. Thus if the alleles are sufficiently rare in the sense that $f(\mathbf{0})$ is larger than the threshold given in Table 4.2, then there is a better chance of capturing each non-ancestral haplotype with $f(y) > 0$ by the CD list than by direct sampling of haplotypes. This is achieved by re-distributing

Table 4.2: Sufficient conditions for non-ancestral haplotype frequencies to be increased by collapsing data

Lower threshold of $f(\mathbf{0})$				
$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
0.5000	0.6300	0.6988	0.7430	0.7743

Sufficient conditions for collapsed data frequencies $\{g(y), y \neq \mathbf{0}\}$ to be greater than haplotype frequencies $\{f(y), y \neq \mathbf{0}\}$ for various choices of pool size k.

the probability of the ancestral haplotype in the process of pooling and collapsing. In other words, the reason why it is possible to have $g(y) > f(y)$ for non-ancestral y is because $g(\mathbf{0}) = f(\mathbf{0})^{2k} < f(\mathbf{0})$. We cannot have $g(y) > f(y)$ for all haplotypes y because both $g(y)$ and $f(y)$ must sum to 1. Table 4.3 shows how the probabilities are being re-distributed for a 25-loci case. The true haplotype distribution $f(y)$ is listed in column 1 of Table 4.3, whereas the distributions $g(y)$ of the collapsed data for various pool sizes are given in the subsequent columns. For non-ancestral y, we can see from Table 4.3 that $g(y) > f(y)$, and more so when the pool size is increased (up to a point), which is good news for the CD list. For example, $f(1, 0, \cdots, 0) = 0.0509$, whereas $g(1, 0, \cdots, 0) = 0.0839$ when $k = 1$, and continues to increase to 0.1143 and 0.1169 when the pool size is increased to 2 and 3. We are particularly interested in the capability of the CD list in capturing haplotypes with multiple 1's. For the last haplotype listed in Table 4.3 (which contains five 1's), $f(y) = 0.0034$, but $g(y)$ is 0.0097 when the pool size is 3. Thus if we have $n = 200$ pools (which is one setting of our simulation study) of $k = 3$ individuals each, the probability that this haplotype is captured by the CD list is $0.8577 = 1 - (1 - 0.0097)^{200} \approx 1 - e^{-200(0.0097)} = 0.8563$. But $g(y)$ will also assign positive probabilities to some haplotypes y even though $f(y) = 0$ since $\sum_{y:f(y)>0} g(y) < 1$, which

is why we propose to trim the CD list. To see how $g(y)$ can be positive even though $f(y) = 0$, consider the following case with just 2 loci. Suppose $f(1,0) > 0$, $f(0,1) > 0$, but $f(1,1) = 0$. By pooling k individuals together, it is obviously possible to have total allele counts $T_1 \geq 1,\ T_2 \geq 1$ at both loci, and hence $(Z_1, Z_2) = (1,1)$, which means that $(1,1)$ will appear on the CD list even though $f(1,1) = 0$.

Table 4.3: Induced collapsed data frequencies

Haplotype y	$f(y)$	$g(y)$			
Positions of '1's	TRUE	$k=1$	$k=2$	$k=3$	$k=4$
None	0.7995	0.6392	0.4085	0.2611	0.1669
1	0.0509	0.0839	0.1143	0.1169	0.1065
2	0.0034	0.0055	0.0070	0.0067	0.0058
3	0.0436	0.0716	0.0967	0.0980	0.0883
5	0.0034	0.0055	0.0070	0.0067	0.0058
6	0.0034	0.0055	0.0070	0.0067	0.0058
9	0.0073	0.0117	0.0151	0.0146	0.0125
11	0.0034	0.0055	0.0070	0.0067	0.0058
15	0.0034	0.0055	0.0070	0.0067	0.0058
19	0.0068	0.0109	0.0141	0.0136	0.0117
20	0.0068	0.0109	0.0141	0.0136	0.0117
21	0.0034	0.0055	0.0070	0.0067	0.0058
22	0.0102	0.0164	0.0213	0.0206	0.0178
23	0.0034	0.0055	0.0070	0.0067	0.0058
24	0.0102	0.0164	0.0213	0.0206	0.0178
1, 3	0.0040	0.0117	0.0307	0.0482	0.0610
1, 9	0.0029	0.0058	0.0105	0.0135	0.0148
3, 14	0.0034	0.0057	0.0082	0.0088	0.0084
6, 7	0.0204	0.0332	0.0439	0.0435	0.0384
3, 6, 7	0.0034	0.0077	0.0164	0.0231	0.0271
1, 6, 7, 24	0.0034	0.0060	0.0097	0.0119	0.0132
1, 12, 13, 22, 25	0.0034	0.0059	0.0087	0.0097	0.0096
Sum of haplotype probabilities	1.0000	0.9751	0.8822	0.7650	0.6462

Haplotype frequencies $f(y)$ for a 25-loci case and the induced collapsed data frequencies $g(y)$ for various pool sizes k.

The CD list misses some haplotypes with $f(y) > 0$, while some other haplotypes with $f(y) = 0$ are erroneously included. This suggests that the

CD list needs to be augmented as well as trimmed. Since we are focusing on rare variants, we augment the CD list by adding all those vectors with only one "1" to the list if they are not already there. Thus we are adding at most L haplotypes to the CD list. Beginning the EM iteration with the augmented CD list, we remove a haplotype from the list if its estimated frequency at the current iteration of the EM algorithm is less than a threshold. The way we select the threshold (typically over a grid) is to choose the one that results in an estimate of the ancestral haplotype frequency $f(\mathbf{0})$ closest to the collapsed data MLE $\hat{f}_{CD}(\mathbf{0})$, which should be a reasonable benchmark.

4.3 Results

To identify rare genetic variants associated with obesity, investigators of the CRESCENDO cohort study obtained re-sequenced data for 148 obese persons and 150 controls around two genes known to be involved in endocannabinoid metabolism: *FAAH* on chromosome 1, and *MGLL* on chromosome 3. There are 31Kbp of re-sequenced data near the *FAAH* gene, and 157Kbp near the *MGLL* locus. Bhatia et al. (2010) discovered two 5Kbp regions enriched in rare variants (RVs) located just upstream of the *FAAH* and *MGLL* genes respectively, with 32 RVs in the first region, and 25 RVs in the second region. To estimate the underlying haplotype distributions, we apply the algorithms proposed in this chapter, as well as the EM with a list (EML) method described in Kuk et al. (2013b), where the list is determined combinatorially. The collapsed data maximum likelihood estimates (CDMLE) are also computed. To save space, we only report the estimates based on the obese individuals, which is the more interesting case, as there

are very few mutations among the control subjects. Table 4.4 reports the CDMLE's, as well as the estimates obtained using EML, EM-CDL (EM with CD list) and EM-ATCDL (EM with augmented and trimmed CD list) algorithms for the 25 loci case. The estimates on the left panel ($k = 1$) are based on individual genotype data, whereas the right panel ($k = 2$) estimates are based on pooled genotype data that result from grouping the 148 obese individuals into 74 pools of size 2 each. Obviously, the estimates based on 148 pools of size 1 (i.e., individual genotype data) should be more reliable than those based on 74 pools of size 2, and so we should use the estimates on the left panel of Table 4.4 as the benchmark. It is interesting to note that as the pool size k increases to 2, the CDMLE, EML and EM-CDL estimates remove some haplotypes that are assigned probabilities in the $k = 1$ case, and in their place, some other haplotypes not presented in the $k = 1$ case are assigned probabilities in the $k = 2$ case. We will see later in the Methods section that it is an inherent property of the CD list to include extraneous false haplotypes as pool size increases. By augmenting and trimming the CD list in the proposed way, the EM-ATCDL estimates based on $k = 1$ and 2 are much more comparable with similar support, which is desirable.

Table 4.1 reports the running times of various algorithms. It can be seen that the EML algorithm takes longer to run than EM-CDL and EM-ATCDL, and is computationally prohibitive (takes longer than 10 hours on an Intel (R) Core (TM) 2 desktop) when the pool size is $k = 4$ in both the 25 and 32 loci cases. Both EM-CDL and EM-ATCDL remain computationally feasible when $k = 4$. Understandably, EM-CDL is a bit faster to run as no augmentation and trimming is involved.

Table 4.4: Haplotype frequency estimates in the *MGLL* region using data from 148 obese individuals

	$k = 1,\ n = 148$				$k = 2,\ n = 74$			
Position of '1's	CDMLE	EML	EM-CDL	EM-ATCDL	CDMLE	EML	EM-CDL	EM-ATCDL
None	0.7927	0.7941	0.7995	0.7984	0.7912	0.8202	0.8169	0.7898
1	0.0536	0.0505	0.0509	0.0544	0.0497	0.0397	0.0398	0.0494
2	0.0043	0.0034	0.0034	0.0034				0.0034
3	0.0456	0.0433	0.0436	0.0441	0.0381	0.0291	0.0291	0.0440
5	0.0043	0.0034	0.0034	0.0034				0.0034
6	0.0043	0.0034	0.0034	0.0034	0.0067	0.0034	0.0034	0.0034
9	0.0085	0.0072	0.0073	0.0103	0.0133	0.0079	0.0079	0.0101
11	0.0043	0.0034	0.0034	0.0034				0.0034
14								0.0034
15	0.0043	0.0034	0.0034	0.0034				0.0034
19	0.0043	0.0068	0.0068	0.0068	0.0067	0.0069	0.0101	0.0101
20	0.0043	0.0068	0.0068	0.0068	0.0067	0.0069	0.0101	0.0101
21	0.0043	0.0034	0.0034	0.0034				0.0034
22	0.0127	0.0101	0.0102	0.0103	0.0197	0.0101	0.0101	0.0101
23	0.0043	0.0034	0.0034	0.0034	0.0067	0.0034	0.0034	0.0034
24	0.0127	0.0101	0.0102	0.0103	0.0067	0.0040	0.0040	0.0040
1, 3	0.0048	0.0040	0.0040	0.0038	0.0090	0.0059	0.0059	0.0020
1, 9	0.0034	0.0029	0.0029		0.0032	0.0022	0.0022	
1, 15					0.0056	0.0034	0.0034	
1, 24					0.0098	0.0064	0.0064	0.0095
2, 3					0.0059	0.0034	0.0034	
3, 14	0.0040	0.0034	0.0034		0.0059	0.0034	0.0034	
3, 11					0.0059	0.0034	0.0034	
5, 21					0.0067	0.0034	0.0034	
6, 7	0.0250	0.0203	0.0204	0.0205	0.0314	0.0182	0.0181	0.0189
19, 20		0.0017				0.0033		
3, 6, 7	0.0026	0.0034	0.0034	0.0034	0.0066	0.0057	0.0057	0.0081
6, 19, 20		0.0017						
7, 19, 20		0.0017						
1, 6, 7, 24	0.0039	0.0034	0.0034	0.0034				
6, 7, 19, 20	0.0041	0.0017			0.0057	0.0033	0.0034	0.0034
1, 3, 6, 7, 24					0.0041	0.0032	0.0032	
1, 12, 13, 22, 25	0.0039	0.0034	0.0034	0.0034	0.0053	0.0034	0.0034	0.0034

Estimates of haplotype frequencies for the 25 rare variants in the *MGLL* region obtained by CDMLE (collapsed data maximum likelihood estimation), EML (EM with combinatorially determined list), EM-CDL (EM with CD list) and EM-ATCDL (EM with augmented and trimmed CD list with adaptive threshold) based on $n = 148/k$ pools of k individuals each.

To facilitate comparison of estimators in situations similar to those under which the original data were collected, we simulate haplotype data from the *MGLL* region (25 loci) and *FAAH* region (32 loci) according to the haplotype distributions listed as "true" in Tables 4.5 and 4.6. These distributions are actually the haplotype distributions estimated using EM-CDL from the individual genotype data of the 148 cases of the CRESCENDO cohort study, but we will treat them as the true distributions in our simulation study. Thus there are only 22 possible haplotypes for the 25 loci case,

Table 4.5: Average estimates of haplotype frequencies for a 25 loci case

Position of '1'	TRUE	$k=2$					$k=4$			
		CDMLE	EML	EM-CDL	EM-ATCDL	EM-TCDL	CDMLE	EM-CDL	EM-ATCDL	EM-TCDL
(a) $n = 100$										
None	0.7995	0.7973	0.8279	0.8283	0.8003	0.8067	0.7961	0.8535	0.7957	0.8119
		(0.0232)	(0.0169)	(0.0170)	(0.0215)	(0.0192)	(0.0204)	(0.0118)	(0.0179)	(0.0144)
1	0.0509	0.0508	0.0412	0.0412	0.0477	0.0477	0.0502	0.0344	0.0457	0.0494
		(0.0155)	(0.0115)	(0.0115)	(0.0119)	(0.0112)	(0.0152)	(0.0083)	(0.0093)	(0.0086)
2	0.0034	0.0036	0.0019	0.0019	0.0031	0.0024	0.0032	0.0008	0.0028	0.0020
		(0.0040)	(0.0022)	(0.0022)	(0.0025)	(0.0027)	(0.0044)	(0.0012)	(0.0023)	(0.0029)
3	0.0436	0.0441	0.0353	0.0353	0.0425	0.0408	0.0435	0.0277	0.0396	0.0411
		(0.0139)	(0.0104)	(0.0104)	(0.0112)	(0.0107)	(0.0144)	(0.0077)	(0.0097)	(0.0080)
5	0.0034	0.0035	0.0019	0.0019	0.0031	0.0027	0.0031	0.0008	0.0028	0.0015
		(0.0039)	(0.0021)	(0.0021)	(0.0028)	(0.0029)	(0.0043)	(0.0011)	(0.0017)	(0.0024)
6	0.0034	0.0027	0.0017	0.0016	0.0029	0.0022	0.0038	0.0013	0.0030	0.0019
		(0.0035)	(0.0021)	(0.0021)	(0.0028)	(0.0030)	(0.0050)	(0.0017)	(0.0022)	(0.0026)
9	0.0073	0.0092	0.0056	0.0056	0.0085	0.0087	0.0073	0.0029	0.0074	0.0079
		(0.0065)	(0.0039)	(0.0039)	(0.0046)	(0.0052)	(0.0066)	(0.0026)	(0.0045)	(0.0062)
11	0.0034	0.0041	0.0022	0.0022	0.0036	0.0029	0.0032	0.0008	0.0027	0.0016
		(0.0048)	(0.0025)	(0.0025)	(0.0029)	(0.0031)	(0.0046)	(0.0013)	(0.0021)	(0.0025)
15	0.0034	0.0039	0.0021	0.0021	0.0032	0.0026	0.0034	0.0009	0.0029	0.0017
		(0.0052)	(0.0028)	(0.0028)	(0.0032)	(0.0033)	(0.0050)	(0.0013)	(0.0021)	(0.0025)
19	0.0068	0.0069	0.0039	0.0039	0.0061	0.0055	0.0075	0.0022	0.0058	0.0051
		(0.0056)	(0.0031)	(0.0031)	(0.0040)	(0.0044)	(0.0066)	(0.0020)	(0.0031)	(0.0042)
20	0.0068	0.0073	0.0041	0.0041	0.0061	0.0058	0.0080	0.0023	0.0057	0.0051
		(0.0060)	(0.0035)	(0.0035)	(0.0042)	(0.0045)	(0.0065)	(0.0020)	(0.0028)	(0.0040)
21	0.0034	0.0038	0.0020	0.0020	0.0032	0.0029	0.0041	0.0011	0.0033	0.0022
		(0.0041)	(0.0022)	(0.0022)	(0.0027)	(0.0031)	(0.0051)	(0.0014)	(0.0023)	(0.0028)
22	0.0102	0.0117	0.0070	0.0070	0.0095	0.0099	0.0110	0.0043	0.0096	0.0091
		(0.0075)	(0.0047)	(0.0047)	(0.0053)	(0.0054)	(0.0085)	(0.0033)	(0.0046)	(0.0056)
23	0.0034	0.0032	0.0018	0.0018	0.0028	0.0024	0.0038	0.0010	0.0029	0.0019
		(0.0039)	(0.0023)	(0.0023)	(0.0028)	(0.0030)	(0.0045)	(0.0012)	(0.0021)	(0.0026)
24	0.0102	0.0096	0.0060	0.0060	0.0095	0.0098	0.0114	0.0043	0.0098	0.0118
		(0.0065)	(0.0041)	(0.0041)	(0.0051)	(0.0057)	(0.0076)	(0.0028)	(0.0047)	(0.0057)
1, 3	0.0040	0.0048	0.0039	0.0039	0.0045	0.0043	0.0071	0.0049	0.0061	0.0052
		(0.0049)	(0.0037)	(0.0037)	(0.0041)	(0.0040)	(0.0070)	(0.0040)	(0.0046)	(0.0048)
1, 9	0.0029	0.0030	0.0021	0.0021	0.0023	0.0018	0.0051	0.0023	0.0028	0.0017
		(0.0035)	(0.0024)	(0.0024)	(0.0033)	(0.0033)	(0.0055)	(0.0021)	(0.0031)	(0.0032)
6, 7	0.0204	0.0210	0.0150	0.0148	0.0215	0.0195	0.0203	0.0104	0.0210	0.0219
		(0.0098)	(0.0072)	(0.0072)	(0.0075)	(0.0077)	(0.0105)	(0.0043)	(0.0068)	(0.0060)
3, 14	0.0034	0.0035	0.0021	0.0021	0.0020	0.0028	0.0031	0.0012	0.0015	0.0021
		(0.0037)	(0.0022)	(0.0022)	(0.0032)	(0.0031)	(0.0038)	(0.0015)	(0.0024)	(0.0026)
3, 6, 7	0.0034	0.0037	0.0028	0.0029	0.0025	0.0030	0.0047	0.0028	0.0022	0.0016
		(0.0047)	(0.0036)	(0.0036)	(0.0044)	(0.0047)	(0.0053)	(0.0025)	(0.0033)	(0.0035)
1, 6, 7, 24	0.0034	0.0031	0.0021	0.0021	0.0006	0.0006	0.0036	0.0018	0.0006	0.0000
		(0.0033)	(0.0023)	(0.0023)	(0.0023)	(0.0021)	(0.0037)	(0.0018)	(0.0017)	(0.0004)
1, 12, 13, 22, 25	0.0034	0.0037	0.0024	0.0024	0.0026	0.0029	0.0038	0.0016	0.0005	0.0025
		(0.0038)	(0.0025)	(0.0025)	(0.0031)	(0.0029)	(0.0035)	(0.0015)	(0.0018)	(0.0027)
Sum of remaining haplotype probabilities		0.0376	0.0248	0.0248	0.0117	0.0124	0.0893	0.0366	0.0255	0.0109
Sum of probabilities of missed haplotypes		0.0247	0.0241	0.0244	0.0218	0.0286	0.0296	0.0285	0.0179	0.0367
Sum of squared errors		0.00166	0.00186	0.00189	0.00110	0.00106	0.00201	0.00415	0.00091	0.00089
Length of list		26.77	116.28	26.77	19.06	18.25	45.23	45.23	25.38	15.62
SD of length		(3.26)	(81.30)	(3.26)	(3.03)	(2.18)	(3.94)	(3.94)	(4.78)	(2.40)

(b) $n = 200$

None	0.7995	0.7979	0.8248	0.8250	0.7981	0.8009	0.7990	0.8451	0.7970	0.8040
		(0.0150)	(0.0117)	(0.0117)	(0.0148)	(0.0132)	(0.0154)	(0.0086)	(0.0133)	(0.0125)
1	0.0509	0.0514	0.0433	0.0433	0.0492	0.0503	0.0502	0.0387	0.0462	0.0507
		(0.0103)	(0.0082)	(0.0082)	(0.0089)	(0.0088)	(0.0121)	(0.0069)	(0.0077)	(0.0080)
2	0.0034	0.0035	0.0020	0.0020	0.0033	0.0031	0.0037	0.0011	0.0030	0.0024
		(0.0032)	(0.0018)	(0.0018)	(0.0024)	(0.0026)	(0.0035)	(0.0011)	(0.0013)	(0.0021)
3	0.0436	0.0430	0.0362	0.0362	0.0435	0.0426	0.0441	0.0322	0.0406	0.0426
		(0.0092)	(0.0075)	(0.0075)	(0.0082)	(0.0074)	(0.0105)	(0.0054)	(0.0071)	(0.0065)
5	0.0034	0.0033	0.0018	0.0018	0.0032	0.0028	0.0034	0.0011	0.0030	0.0023
		(0.0028)	(0.0016)	(0.0016)	(0.0020)	(0.0023)	(0.0035)	(0.0011)	(0.0013)	(0.0021)
6	0.0034	0.0038	0.0023	0.0023	0.0033	0.0031	0.0033	0.0014	0.0028	0.0022
		(0.0031)	(0.0019)	(0.0020)	(0.0021)	(0.0025)	(0.0035)	(0.0013)	(0.0014)	(0.0019)
9	0.0073	0.0080	0.0054	0.0054	0.0079	0.0088	0.0081	0.0036	0.0070	0.0092
		(0.0038)	(0.0026)	(0.0026)	(0.0032)	(0.0037)	(0.0043)	(0.0019)	(0.0026)	(0.0031)
11	0.0034	0.0032	0.0018	0.0018	0.0030	0.0027	0.0032	0.0010	0.0029	0.0023
		(0.0026)	(0.0016)	(0.0016)	(0.0022)	(0.0024)	(0.0031)	(0.0011)	(0.0015)	(0.0021)
15	0.0034	0.0035	0.0019	0.0019	0.0030	0.0028	0.0038	0.0012	0.0031	0.0028
		(0.0033)	(0.0018)	(0.0018)	(0.0023)	(0.0025)	(0.0031)	(0.0010)	(0.0015)	(0.0021)
19	0.0068	0.0063	0.0039	0.0039	0.0062	0.0061	0.0066	0.0026	0.0056	0.0057
		(0.0039)	(0.0025)	(0.0025)	(0.0029)	(0.0034)	(0.0041)	(0.0015)	(0.0018)	(0.0026)
20	0.0068	0.0068	0.0042	0.0042	0.0062	0.0063	0.0063	0.0026	0.0054	0.0059
		(0.0039)	(0.0025)	(0.0025)	(0.0028)	(0.0030)	(0.0038)	(0.0015)	(0.0022)	(0.0026)
21	0.0034	0.0038	0.0022	0.0022	0.0035	0.0032	0.0037	0.0012	0.0031	0.0025
		(0.0035)	(0.0020)	(0.0020)	(0.0023)	(0.0026)	(0.0034)	(0.0011)	(0.0015)	(0.0021)
22	0.0102	0.0105	0.0071	0.0071	0.0097	0.0103	0.0112	0.0052	0.0086	0.0098
		(0.0058)	(0.0037)	(0.0037)	(0.0041)	(0.0041)	(0.0052)	(0.0022)	(0.0030)	(0.0029)
23	0.0034	0.0039	0.0022	0.0022	0.0031	0.0030	0.0035	0.0011	0.0030	0.0025
		(0.0027)	(0.0015)	(0.0016)	(0.0021)	(0.0022)	(0.0030)	(0.0010)	(0.0015)	(0.0021)
24	0.0102	0.0105	0.0069	0.0069	0.0107	0.0114	0.0106	0.0049	0.0088	0.0115
		(0.0050)	(0.0033)	(0.0033)	(0.0043)	(0.0042)	(0.0059)	(0.0023)	(0.0036)	(0.0043)
1, 3	0.0040	0.0046	0.0037	0.0037	0.0041	0.0040	0.0052	0.0041	0.0049	0.0047
		(0.0041)	(0.0032)	(0.0032)	(0.0033)	(0.0034)	(0.0048)	(0.0029)	(0.0030)	(0.0031)
1, 9	0.0029	0.0036	0.0026	0.0026	0.0030	0.0022	0.0030	0.0018	0.0027	0.0010
		(0.0027)	(0.0019)	(0.0019)	(0.0026)	(0.0028)	(0.0031)	(0.0015)	(0.0022)	(0.0020)
6, 7	0.0204	0.0191	0.0148	0.0146	0.0215	0.0199	0.0206	0.0124	0.0212	0.0223
		(0.0067)	(0.0048)	(0.0047)	(0.0056)	(0.0054)	(0.0073)	(0.0037)	(0.0048)	(0.0046)
3, 14	0.0034	0.0030	0.0019	0.0019	0.0020	0.0027	0.0036	0.0015	0.0025	0.0029
		(0.0029)	(0.0018)	(0.0018)	(0.0026)	(0.0025)	(0.0026)	(0.0011)	(0.0020)	(0.0018)
3, 6, 7	0.0034	0.0039	0.0029	0.0029	0.0021	0.0027	0.0038	0.0025	0.0024	0.0019
		(0.0035)	(0.0025)	(0.0025)	(0.0029)	(0.0031)	(0.0041)	(0.0022)	(0.0028)	(0.0029)
1, 6, 7, 24	0.0034	0.0039	0.0028	0.0028	0.0007	0.0006	0.0039	0.0020	0.0002	0.0000
		(0.0025)	(0.0019)	(0.0019)	(0.0019)	(0.0018)	(0.0026)	(0.0014)	(0.0009)	(0.0002)
1, 12, 13, 22, 25	0.0034	0.0035	0.0025	0.0024	0.0027	0.0032	0.0033	0.0017	0.0004	0.0031
		(0.0022)	(0.0016)	(0.0017)	(0.0023)	(0.0023)	(0.0026)	(0.0013)	(0.0014)	(0.0020)
Sum of remaining haplotype probabilities		0.0340	0.0227	0.0227	0.0102	0.0074	0.0703	0.0310	0.0255	0.0076
Sum of probabilities of missed haplotypes		0.0103	0.0097	0.0100	0.0132	0.0173	0.0131	0.0118	0.0111	0.0200
Sum of squared errors		0.00077	0.00125	0.00126	0.00059	0.00054	0.00101	0.00281	0.00055	0.00048
Length of list		39.65	152.30	39.65	22.08	19.63	71.24	71.24	29.29	19.81
SD of length		(3.90)	(71.65)	(3.90)	(4.20)	(3.70)	(5.02)	(5.02)	(5.82)	(5.22)

Average estimates of haplotype frequencies for a 25 loci case based on 100 simulations of n pools of k individuals each using CDMLE (collapsed data MLE), EML (EM with combinatorially determined list), EM-CDL (EM with CD list), EM-ATCDL (augmented and trimmed CD list) and EM-TCDL (CD list with trimming and no augmentation), with standard errors in parentheses.

Table 4.6: Average estimates of haplotype frequencies for a 32 loci case

Position of '1'	TRUE	$k=2$ CDMLE	$k=2$ EML	$k=2$ EM-CDL	$k=2$ EM-ATCDL	$k=2$ EM-TCDL	$k=4$ CDMLE	$k=4$ EM-CDL	$k=4$ EM-ATCDL	$k=4$ EM-TCDL
(a) $n = 100$										
None	0.7995	0.7979	0.8248	0.8250	0.7981	0.8009	0.7990	0.8451	0.7970	0.8040
		(0.0150)	(0.0117)	(0.0117)	(0.0148)	(0.0132)	(0.0154)	(0.0086)	(0.0133)	(0.0125)
1	0.0509	0.0514	0.0433	0.0433	0.0492	0.0503	0.0502	0.0387	0.0462	0.0507
		(0.0103)	(0.0082)	(0.0082)	(0.0089)	(0.0088)	(0.0121)	(0.0069)	(0.0077)	(0.0080)
2	0.0034	0.0035	0.0020	0.0020	0.0033	0.0031	0.0037	0.0011	0.0030	0.0024
		(0.0032)	(0.0018)	(0.0018)	(0.0024)	(0.0026)	(0.0035)	(0.0011)	(0.0013)	(0.0021)
3	0.0436	0.0430	0.0362	0.0362	0.0435	0.0426	0.0441	0.0322	0.0406	0.0426
		(0.0092)	(0.0075)	(0.0075)	(0.0082)	(0.0074)	(0.0105)	(0.0054)	(0.0071)	(0.0065)
5	0.0034	0.0033	0.0018	0.0018	0.0032	0.0028	0.0034	0.0011	0.0030	0.0023
		(0.0028)	(0.0016)	(0.0016)	(0.0020)	(0.0023)	(0.0035)	(0.0011)	(0.0013)	(0.0021)
6	0.0034	0.0038	0.0023	0.0023	0.0033	0.0031	0.0033	0.0014	0.0028	0.0022
		(0.0031)	(0.0019)	(0.0020)	(0.0021)	(0.0025)	(0.0035)	(0.0013)	(0.0014)	(0.0019)
9	0.0073	0.0080	0.0054	0.0054	0.0079	0.0088	0.0081	0.0036	0.0070	0.0092
		(0.0038)	(0.0026)	(0.0026)	(0.0032)	(0.0037)	(0.0043)	(0.0019)	(0.0026)	(0.0031)
11	0.0034	0.0032	0.0018	0.0018	0.0030	0.0027	0.0032	0.0010	0.0029	0.0023
		(0.0026)	(0.0016)	(0.0016)	(0.0022)	(0.0024)	(0.0031)	(0.0011)	(0.0015)	(0.0021)
15	0.0034	0.0035	0.0019	0.0019	0.0030	0.0028	0.0038	0.0012	0.0031	0.0028
		(0.0033)	(0.0018)	(0.0018)	(0.0023)	(0.0025)	(0.0031)	(0.0010)	(0.0015)	(0.0021)
19	0.0068	0.0063	0.0039	0.0039	0.0062	0.0061	0.0066	0.0026	0.0056	0.0057
		(0.0039)	(0.0025)	(0.0025)	(0.0029)	(0.0034)	(0.0041)	(0.0015)	(0.0018)	(0.0026)
20	0.0068	0.0068	0.0042	0.0042	0.0062	0.0063	0.0063	0.0026	0.0054	0.0059
		(0.0039)	(0.0025)	(0.0025)	(0.0028)	(0.0030)	(0.0038)	(0.0015)	(0.0022)	(0.0026)
21	0.0034	0.0038	0.0022	0.0022	0.0035	0.0032	0.0037	0.0012	0.0031	0.0025
		(0.0035)	(0.0020)	(0.0020)	(0.0023)	(0.0026)	(0.0034)	(0.0011)	(0.0015)	(0.0021)
22	0.0102	0.0105	0.0071	0.0071	0.0097	0.0103	0.0112	0.0052	0.0086	0.0098
		(0.0058)	(0.0037)	(0.0037)	(0.0041)	(0.0041)	(0.0052)	(0.0022)	(0.0030)	(0.0029)
23	0.0034	0.0039	0.0022	0.0022	0.0031	0.0030	0.0035	0.0011	0.0030	0.0025
		(0.0027)	(0.0015)	(0.0016)	(0.0021)	(0.0022)	(0.0030)	(0.0010)	(0.0015)	(0.0021)
24	0.0102	0.0105	0.0069	0.0069	0.0107	0.0114	0.0106	0.0049	0.0088	0.0115
		(0.0050)	(0.0033)	(0.0033)	(0.0043)	(0.0042)	(0.0059)	(0.0023)	(0.0036)	(0.0043)
1, 3	0.0040	0.0046	0.0037	0.0037	0.0041	0.0040	0.0052	0.0041	0.0049	0.0047
		(0.0041)	(0.0032)	(0.0032)	(0.0033)	(0.0034)	(0.0048)	(0.0029)	(0.0030)	(0.0031)
1, 9	0.0029	0.0036	0.0026	0.0026	0.0030	0.0022	0.0030	0.0018	0.0027	0.0010
		(0.0027)	(0.0019)	(0.0019)	(0.0026)	(0.0028)	(0.0031)	(0.0015)	(0.0022)	(0.0020)
6, 7	0.0204	0.0191	0.0148	0.0146	0.0215	0.0199	0.0206	0.0124	0.0212	0.0223
		(0.0067)	(0.0048)	(0.0047)	(0.0056)	(0.0054)	(0.0073)	(0.0037)	(0.0048)	(0.0046)
3, 14	0.0034	0.0030	0.0019	0.0019	0.0020	0.0027	0.0036	0.0015	0.0025	0.0029
		(0.0029)	(0.0018)	(0.0018)	(0.0026)	(0.0025)	(0.0026)	(0.0011)	(0.0020)	(0.0018)
3, 6, 7	0.0034	0.0039	0.0029	0.0029	0.0021	0.0027	0.0038	0.0025	0.0024	0.0019
		(0.0035)	(0.0025)	(0.0025)	(0.0029)	(0.0031)	(0.0041)	(0.0022)	(0.0028)	(0.0029)
1, 6, 7, 24	0.0034	0.0039	0.0028	0.0028	0.0007	0.0006	0.0039	0.0020	0.0002	0.0000
		(0.0025)	(0.0019)	(0.0019)	(0.0019)	(0.0018)	(0.0026)	(0.0014)	(0.0009)	(0.0002)
1, 12, 13, 22, 25	0.0034	0.0035	0.0025	0.0024	0.0027	0.0032	0.0033	0.0017	0.0004	0.0031
		(0.0022)	(0.0016)	(0.0017)	(0.0023)	(0.0023)	(0.0026)	(0.0013)	(0.0014)	(0.0020)
Sum of remaining haplotype probabilities		0.0340	0.0227	0.0227	0.0102	0.0074	0.0703	0.0310	0.0255	0.0076
Sum of probabilities of missed haplotypes		0.0103	0.0097	0.0100	0.0132	0.0173	0.0131	0.0118	0.0111	0.0200
Sum of squared errors		0.00077	0.00125	0.00126	0.00059	0.00054	0.00101	0.00281	0.00055	0.00048
Length of list		39.65	152.30	39.65	22.08	19.63	71.24	71.24	29.29	19.81
SD of length		(3.90)	(71.65)	(3.90)	(4.20)	(3.70)	(5.02)	(5.02)	(5.82)	(5.22)

(b) $n = 200$

None	0.7113	0.7108	0.7674	0.7677	0.7115	0.7219	0.7110	0.8075	0.7109	0.7374
		(0.0202)	(0.0132)	(0.0132)	(0.0197)	(0.0161)	(0.0232)	(0.0119)	(0.0192)	(0.0147)
1	0.0034	0.0031	0.0013	0.0013	0.0032	0.0022	0.0034	0.0005	0.0031	0.0012
		(0.0032)	(0.0014)	(0.0014)	(0.0022)	(0.0025)	(0.0054)	(0.0007)	(0.0015)	(0.0018)
3	0.0034	0.0039	0.0016	0.0016	0.0032	0.0028	0.0036	0.0005	0.0030	0.0017
		(0.0034)	(0.0014)	(0.0014)	(0.0022)	(0.0025)	(0.0047)	(0.0007)	(0.0014)	(0.0022)
5	0.0034	0.0037	0.0016	0.0016	0.0035	0.0029	0.0036	0.0006	0.0032	0.0017
		(0.0035)	(0.0015)	(0.0015)	(0.0022)	(0.0027)	(0.0049)	(0.0007)	(0.0013)	(0.0022)
7	0.0068	0.0066	0.0038	0.0041	0.0061	0.0064	0.0065	0.0027	0.0058	0.0061
		(0.0048)	(0.0024)	(0.0028)	(0.0030)	(0.0040)	(0.0070)	(0.0024)	(0.0024)	(0.0053)
9	0.0034	0.0037	0.0015	0.0015	0.0034	0.0029	0.0038	0.0006	0.0031	0.0016
		(0.0038)	(0.0016)	(0.0016)	(0.0024)	(0.0029)	(0.0049)	(0.0008)	(0.0015)	(0.0022)
10	0.0102	0.0098	0.0050	0.0050	0.0087	0.0093	0.0103	0.0026	0.0088	0.0085
		(0.0059)	(0.0029)	(0.0030)	(0.0035)	(0.0035)	(0.0072)	(0.0017)	(0.0026)	(0.0042)
11	0.0034	0.0038	0.0016	0.0016	0.0032	0.0029	0.0027	0.0004	0.0031	0.0010
		(0.0030)	(0.0013)	(0.0013)	(0.0023)	(0.0026)	(0.0044)	(0.0007)	(0.0014)	(0.0019)
14	0.0034	0.0034	0.0014	0.0014	0.0031	0.0023	0.0030	0.0004	0.0029	0.0013
		(0.0037)	(0.0016)	(0.0016)	(0.0022)	(0.0027)	(0.0048)	(0.0007)	(0.0015)	(0.0021)
17	0.0034	0.0033	0.0014	0.0014	0.0029	0.0023	0.0031	0.0005	0.0028	0.0014
		(0.0035)	(0.0014)	(0.0014)	(0.0021)	(0.0024)	(0.0047)	(0.0008)	(0.0015)	(0.0021)
20	0.0034	0.0037	0.0015	0.0015	0.0030	0.0025	0.0028	0.0004	0.0029	0.0014
		(0.0035)	(0.0014)	(0.0014)	(0.0019)	(0.0025)	(0.0042)	(0.0006)	(0.0014)	(0.0021)
21	0.0264	0.0251	0.0164	0.0164	0.0252	0.0252	0.0266	0.0117	0.0257	0.0260
		(0.0089)	(0.0051)	(0.0051)	(0.0064)	(0.0063)	(0.0135)	(0.0040)	(0.0055)	(0.0056)
22	0.0068	0.0074	0.0040	0.0040	0.0095	0.0099	0.0067	0.0016	0.0089	0.0074
		(0.0055)	(0.0029)	(0.0029)	(0.0040)	(0.0051)	(0.0067)	(0.0017)	(0.0038)	(0.0062)
24	0.0306	0.0307	0.0223	0.0234	0.0289	0.0295	0.0297	0.0194	0.0273	0.0291
		(0.0096)	(0.0066)	(0.0066)	(0.0070)	(0.0074)	(0.0146)	(0.0049)	(0.0058)	(0.0061)
25	0.0136	0.0138	0.0075	0.0075	0.0129	0.0129	0.0155	0.0047	0.0139	0.0125
		(0.0072)	(0.0036)	(0.0036)	(0.0040)	(0.0036)	(0.0100)	(0.0029)	(0.0042)	(0.0051)
26	0.0034	0.0033	0.0013	0.0013	0.0029	0.0023	0.0037	0.0005	0.0031	0.0014
		(0.0035)	(0.0014)	(0.0014)	(0.0024)	(0.0026)	(0.0052)	(0.0007)	(0.0014)	(0.0021)
28	0.0675	0.0661	0.0487	0.0488	0.0615	0.0640	0.0668	0.0390	0.0614	0.0629
		(0.0136)	(0.0089)	(0.0089)	(0.0107)	(0.0105)	(0.0162)	(0.0075)	(0.0090)	(0.0086)
30	0.0036	0.0032	0.0017	0.0017	0.0070	0.0064	0.0043	0.0009	0.0077	0.0060
		(0.0029)	(0.0015)	(0.0015)	(0.0039)	(0.0056)	(0.0053)	(0.0011)	(0.0040)	(0.0067)
31	0.0034	0.0037	0.0016	0.0016	0.0031	0.0025	0.0031	0.0004	0.0030	0.0014
		(0.0037)	(0.0016)	(0.0016)	(0.0021)	(0.0025)	(0.0044)	(0.0006)	(0.0013)	(0.0020)
32	0.0038	0.0034	0.0016	0.0016	0.0045	0.0041	0.0041	0.0007	0.0042	0.0029
		(0.0032)	(0.0015)	(0.0016)	(0.0029)	(0.0037)	(0.0060)	(0.0010)	(0.0021)	(0.0038)
2, 25	0.0034	0.0034	0.0015	0.0015	0.0021	0.0026	0.0036	0.0006	0.0013	0.0018
		(0.0033)	(0.0015)	(0.0015)	(0.0029)	(0.0028)	(0.0048)	(0.0008)	(0.0020)	(0.0023)
7, 24	0.0510	0.0507	0.0403	0.0386	0.0525	0.0530	0.0523	0.0319	0.0524	0.0519
		(0.0120)	(0.0079)	(0.0078)	(0.0094)	(0.0084)	(0.0139)	(0.0063)	(0.0089)	(0.0064)
12, 13	0.0034	0.0031	0.0013	0.0013	0.0016	0.0023	0.0029	0.0005	0.0011	0.0014
		(0.0033)	(0.0015)	(0.0015)	(0.0025)	(0.0026)	(0.0049)	(0.0009)	(0.0020)	(0.0022)
21, 23	0.0034	0.0031	0.0015	0.0015	0.0018	0.0023	0.0035	0.0007	0.0016	0.0019
		(0.0035)	(0.0016)	(0.0016)	(0.0027)	(0.0026)	(0.0042)	(0.0009)	(0.0023)	(0.0023)
21, 28	0.0009	0.0031	0.0021	0.0021	0.0022	0.0019	0.0035	0.0016	0.0015	0.0010
		(0.0038)	(0.0022)	(0.0022)	(0.0027)	(0.0030)	(0.0050)	(0.0017)	(0.0023)	(0.0027)
21, 30	0.0033	0.0035	0.0018	0.0018	0.0022	0.0018	0.0035	0.0008	0.0016	0.0012
		(0.0033)	(0.0016)	(0.0016)	(0.0027)	(0.0027)	(0.0044)	(0.0010)	(0.0024)	(0.0023)
22, 30	0.0034	0.0029	0.0014	0.0013	0.0017	0.0017	0.0037	0.0007	0.0014	0.0014
		(0.0031)	(0.0015)	(0.0015)	(0.0024)	(0.0027)	(0.0051)	(0.0009)	(0.0021)	(0.0022)
24, 28	0.0034	0.0042	0.0034	0.0033	0.0036	0.0035	0.0064	0.0036	0.0037	0.0040
		(0.0048)	(0.0030)	(0.0030)	(0.0032)	(0.0034)	(0.0072)	(0.0028)	(0.0029)	(0.0036)
28, 32	0.0030	0.0034	0.0019	0.0019	0.0022	0.0019	0.0033	0.0011	0.0018	0.0018
		(0.0033)	(0.0018)	(0.0019)	(0.0028)	(0.0028)	(0.0037)	(0.0011)	(0.0021)	(0.0026)
4, 7, 24	0.0034	0.0028	0.0015	0.0016	0.0013	0.0024	0.0028	0.0008	0.0008	0.0019
		(0.0025)	(0.0014)	(0.0014)	(0.0022)	(0.0022)	(0.0027)	(0.0008)	(0.0016)	(0.0019)

7, 22, 24	0.0034	0.0042	0.0025	0.0026	0.0019	0.0011	0.0029	0.0015	0.0018	0.0012
		(0.0036)	(0.0022)	(0.0021)	(0.0028)	(0.0024)	(0.0034)	(0.0013)	(0.0024)	(0.0025)
7, 24, 30	0.0034	0.0034	0.0023	0.0023	0.0019	0.0018	0.0033	0.0014	0.0018	0.0018
		(0.0032)	(0.0022)	(0.0022)	(0.0031)	(0.0031)	(0.0035)	(0.0013)	(0.0024)	(0.0029)
Sum of remaining haplotype probabilities		0.0828	0.0454	0.0453	0.0178	0.0083	0.2194	0.0592	0.0243	0.0160
Sum of probabilities of missed haplotypes		0.0278	0.0265	0.0271	0.0263	0.0384	0.0474	0.0462	0.0222	0.0553
Sum of squared errors		0.00161	0.00451	0.00456	0.00103	0.00100	0.00334	0.01152	0.00092	0.00152
Length of list		62.39	159.68	62.39	30.60	23.57	106.91	106.91	36.11	21.92
SD of length		(4.88)	(27.84)	(4.88)	(6.04)	(4.22)	(6.41)	(6.41)	(7.12)	(5.11)

Average estimates of haplotype frequencies for a 32 loci case based on 100 simulations of n pools of k individuals each using CDMLE (collapsed data MLE), EML (EM with combinatorially determined list), EM-CDL (EM with CD list), EM-ATCDL (augmented and trimmed CD list) and EM-TCDL (CD list with trimming and no augmentation), with standard errors in parentheses.

and 32 haplotypes for the 32 loci case. After generating the haplotypes, we form n pools of $2k$ haplotypes each ($n = 100, 200$; $k = 1, 2, 3, 4$) and the resulting pooled genotype data will be treated as the observed data to be used to construct estimates. The results reported in Tables 3 and 4 are based on 100 simulations. The gold standard that we use is the EM-PL estimator, which assumes knowledge of the perfect list (i.e, knowing exactly which $f(y) > 0$). Because the perfect list is used, the EM algorithm in this case will yield the MLE based on the pooled genotype data. We will not have such knowledge in reality and so our real interest is in comparing the performance of the following estimators: CDMLE (collapsed data MLE), EML (EM with combinatorially determined list), EM-CDL (CD list), EM-ACDL (augmented CD list), EM-ATCDL (augmented and trimmed CD list), and EM-TCDL (CD list with trimming and no augmentation). For removing haplotypes from both the ATCD and TCD lists, we try threshold values from 0.0001 to 0.002 in steps of 0.0001, and select the threshold to yield an estimate of $f(\mathbf{0})$ as close to $\hat{f}_{CD}(\mathbf{0})$ as possible. Based on the study of Kuk et al. (2013b), $\hat{f}_{CD}(\mathbf{0})$ seems to be a reasonable benchmark to use. In fact,

we can see from Tables 4.5 and 4.6 that the average of $\hat{f}_{CD}(\mathbf{0})$ (over 100 simulations) is always close to the average of the gold standard $\hat{f}_{EM-PL}(\mathbf{0})$, and this lends further support to the use of $\hat{f}_{CD}(\mathbf{0})$ as a benchmark. We have simulated data for $k = 1, 2, 3, 4$. As the EML algorithm takes too long to run (see Table 4.1), we compute the EML estimates for $k = 1$ and 2 only. To save space, we only report the results of $k = 2$ and 4 in Tables 4.5 and 4.6. The results for EM-CDL and EM-ACDL are close, and so we table the results of EM-CDL only. In order not to make the tables unduly long, we table only the averages of $\hat{f}(y)$ for those y with $f(y) > 0$, together with the sum of $\hat{f}(y)$ over the remaining y's, as well as the averages over simulations of the sum of squared errors $\sum_{y\in\Omega}\left[\hat{f}(y) - f(y)\right]^2$, $\Omega = \{0,1\}^L$, for the various estimators of $f(y)$. To supplement Tables 4.5 and 4.6, we plot the simulated averages of the sum of squared errors against pool size k for all 7 estimators, including EM-PL.

It can be seen from Tables 4.5 and 4.6 that EM-CDL overestimated the frequency $f(\mathbf{0})$ of the ancestral haplotype quite severely, and it has the largest sum of squared error among all the estimators. The performance of EML is very similar to that of EM-CDL (both unsatisfactory) but the computational cost is much higher. It suffers from assigning small probabilities to too many false haplotypes. For example, for the 25 loci case with $n = 100,\ k = 2$, the EML list on the average contains 116 haplotypes even though the true distribution is concentrated on 22 haplotypes. The total probability that the EML estimator attaches to haplotypes outside of the true 22 is only 0.0248 on the average. This foretells the need for trimming which is a point we will come back to later.

Augmenting the CD list did not help much as the results for EM-ACDL

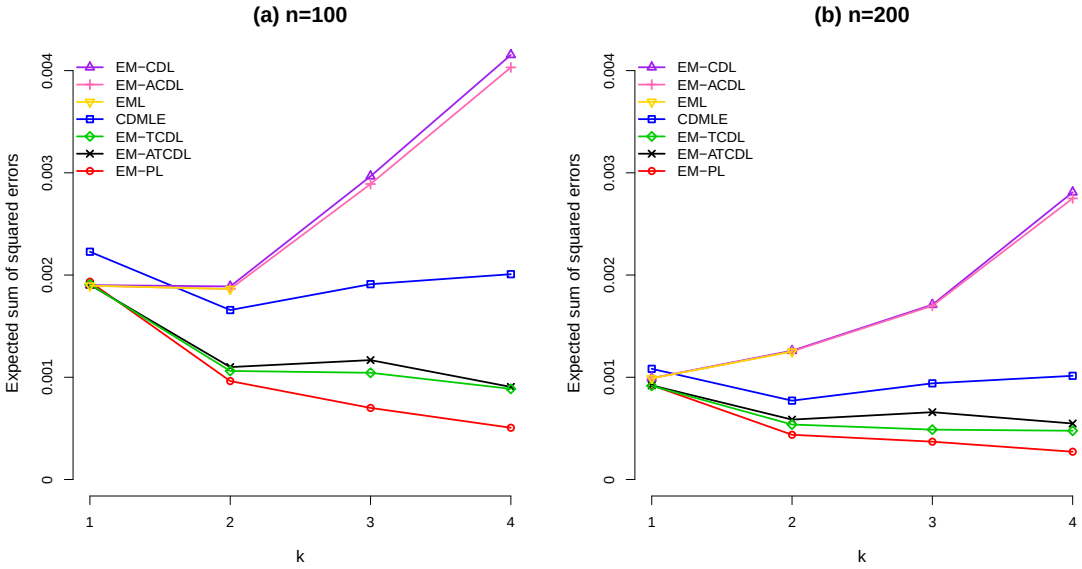

Figure 4.1: Expected sum of squared errors of various haplotype frequency estimators for a 25 loci case. Expected sum of squared errors of various haplotype frequency estimators (EM-CDL: EM with CD list; EM-ACDL: augmented CD list; EML: EM with combinatorially determined list; CDMLE: collapsed data MLE; EM-TCDL: CD list with trimming and no augmentation; EM-ATCDL: augmented and trimmed CD list; EM-PL: EM with perfect list) based on 100 simulations of n pools of k individuals each when the true haplotype distribution over 25 loci is as given in Table 4.5.

are almost the same as that of EM-CDL when $n = 200$, and only slightly better when $n = 100$ (not shown in the tables, but we can see this from Figures 4.1 and 4.2). Trimming in addition to augmenting the CD list improved things a lot, as demonstrated by the good results of EM-ATCDL in both Tables 3 and 4. From Figures 4.1 and 4.2, we can see that EM-ATCDL is clearly the best estimator among those considered, other than the perfect list estimator which is not a legitimate estimator. Since augmenting alone did not improve results much, but trimming in addition to augmenting did, we were curious to see whether trimming alone would work or not. As expected, we can see from Tables 3 and 4 that the TCD list (trimming without augmentation) is on the average shorter than the

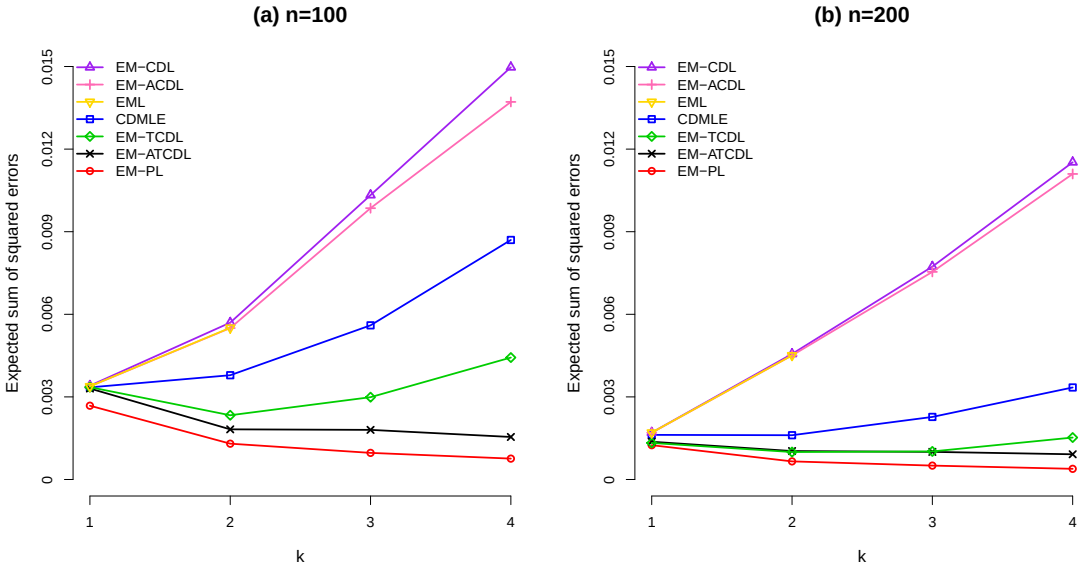

Figure 4.2: Expected sum of squared errors of various haplotype frequency estimators for a 32 loci case. Expected sum of squared errors of various haplotype frequency estimators (EM-CDL: EM with CD list; EM-ACDL: augmented CD list; EML: EM with combinatorially determined list; CDMLE: collapsed data MLE; EM-TCDL: CD list with trimming and no augmentation; EM-ATCDL: augmented and trimmed CD list; EM-PL: EM with perfect list) based on 100 simulations of n pools of k individuals each when the true haplotype distribution over 32 loci is as given in Table 4.6.

ATCD list. Consequently, the TCD list will miss more true haplotypes, and the sum of probabilities of the missed haplotypes is higher for EM-TCDL than for EM-ATCDL, and more so for the 32 loci case and when the number of pools is 100 rather than 200. In particular, the sum of probabilities of the missed haplotypes for the 32 loci case with $k = 4$ is 0.0798 (after averaging over simulations) when $n = 100$, and improves slightly to 0.0553 when $n = 200$. The corresponding figures for EM-ATCDL are 0.0328 and 0.0222. In terms of sum of squared errors, EM-TCDL is also inferior to EM-ATCDL for the 32 loci case, particularly when $n = 100$.

The collapsed data MLE advocated by Kuk et al. (2013b) behaves very similarly to the gold standard EM-PL estimator in terms of bias or expected

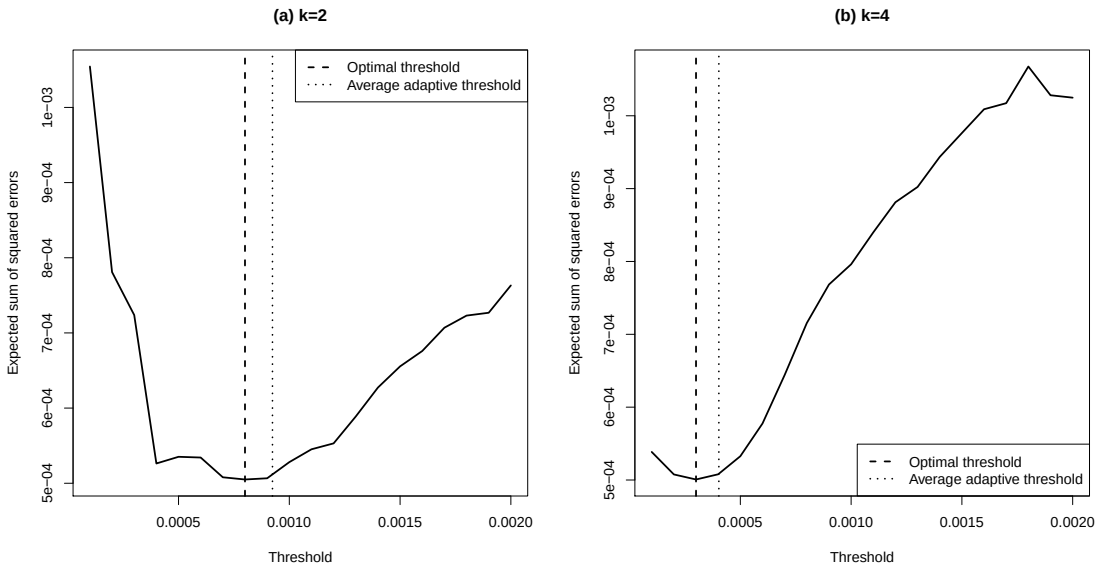

Figure 4.3: Expected sum of squared errors of the EM-ATCDL estimator with fixed threshold (25 loci case). Expected sum of squared errors of the EM-ATCDL estimator for various choices of the threshold (Optimal threshold: the threshold obtained by minimizing the averaged sum of squared errors; Average adaptive threshold: adaptively chosen thresholds obtained by minimizing the distance between $\hat{f}(\mathbf{0})$ and $f(\mathbf{0})$ over the grid 0.0001 to 0.002 in steps of 0.0001) based on 100 simulations of n pools of k individuals each when the true haplotype distribution over 25 loci is as given in Table 4.5.

value, but it suffers from having a larger variance, especially for larger pool size. In contrast, the EM-CDL estimates have small variance but large bias. By benchmarking against CDMLE, the EM-ATCDL estimates have smaller bias than EM-CDL and smaller variance than CDMLE. The main advantage of the collapsed data MLE is its simplicity and small bias. As shown by Kuk et al. (2013b), the loss in efficiency due to collapsing the pooled genotype data locus-wise to just "0" and "≥ 1" is not large for small pool size (especially when $k = 1$ which corresponds to individual genotype data) and rare alleles, but it is better to use EM-ATCDL if $k \geq 2$.

To further see if our benchmarking method of determining the threshold

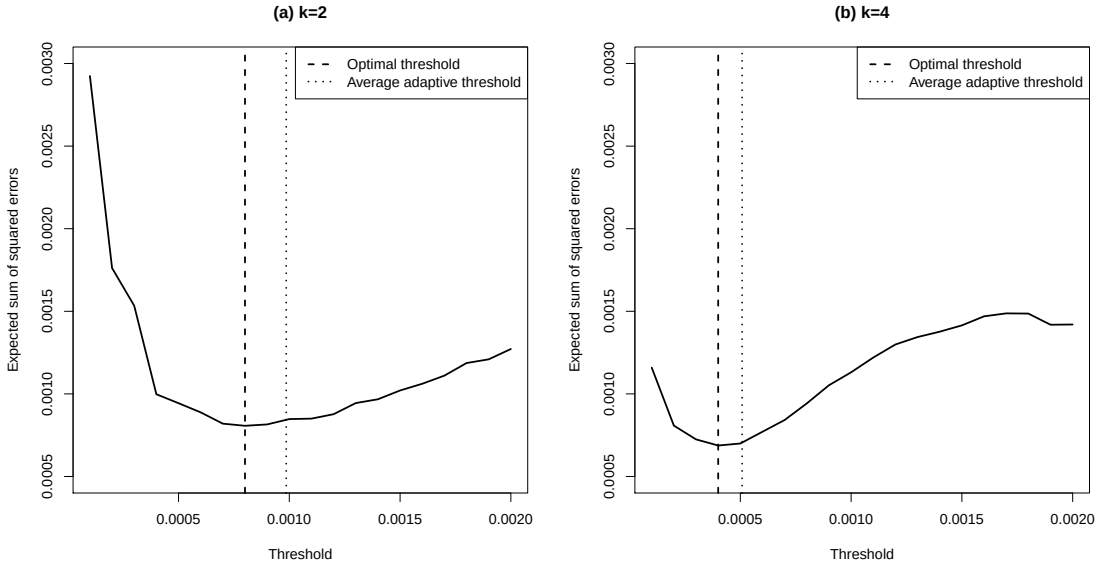

Figure 4.4: Expected sum of squared errors of the EM-ATCDL estimator with fixed threshold (32 loci case). Expected sum of squared errors of the EM-ATCDL estimator for various choices of the threshold (Optimal threshold: the threshold obtained by minimizing the averaged sum of squared errors; Average adaptive threshold: adaptively chosen thresholds obtained by minimizing the distance between $\hat{f}(\mathbf{0})$ and $f(\mathbf{0})$ over the grid 0.0001 to 0.002 in steps of 0.0001) based on 100 simulations of n pools of k individuals each when the true haplotype distribution over 32 loci is as given in Table 4.6.

for the removal of haplotypes is reasonable or not, we also compute the EM-ATCDL estimates based on fixed threshold in our simulation study to find out which threshold is "optimal". Figures 4.3 and 4.4 depict the averages of the sum of squared errors $\sum_{y\in\Omega}\left[\hat{f}(y)-f(y)\right]^2$, $\Omega=\{0,1\}^L$, over 100 simulations for the EM-ATCDL estimates $\hat{f}(y)$ as a function of the threshold value. The position of the "optimal" threshold which minimizes that averaged sum of squared errors is depicted by the vertical dashed line, whereas the average of the adaptively chosen thresholds (obtained by minimizing the distance between $\hat{f}(\mathbf{0})$ and $f(\mathbf{0})$ over the grid 0.0001 to 0.002 in steps of 0.0001) is depicted by the dotted vertical line. It can be

seen that the averages of the adaptively chosen thresholds are quite close to the "optimal" thresholds which lends support to the proposed adaptive method.

4.4 Discussion

The EM algorithm for estimating haplotype frequencies from pooled genotype data is computationally not feasible when the number of loci and/or the pool size is large due to the combinatorial challenge of finding all possible haplotypes that are compatible with the observed pool tools. Gasbarra et al. (2011) raised the possibility of using database information to form a list of frequently occurring haplotypes, and by restricting attention to only those haplotypes on such a list, Pirinen (2009) made the EM algorithm much more viable. The success of the EM with a list method is, however, dependent on the correctness of the list used. In the absence of an external list of possible haplotypes, especially for rare alleles for which there is not a lot of database information, and to protect against using the wrong list, we look at the feasibility of using the data at hand to create an internal list of possible haplotypes to be fed into the EM algorithm. Motivated by the collapsed data method studied by Kuk et al. (2013b), we propose a CD list with amplified haplotype frequencies. This alone does not work well but with appropriate augmentation and trimming, the resulting EM-ATCDL algorithm performs very well in our simulation study. It should be pointed out that even though the ATCD list originates from the CD list which is based on collapsed data: a further reduction of pooled genotype data, the EM-ATCDL estimates themselves are computed using the pooled data, which explains why they are better than the collapsed data MLEs.

The simulation results also suggest that augmenting the collapsed data list alone, or trimming the list alone, is not good enough, and it is necessary to do both. The average lengths of the various lists are also shown in Tables 3 and 4. We can see that the average length of the ATCD list ranges from 20 ($k = 1, n = 100$) to 30 ($k = 4, n = 200$) for the 25 loci case, and from 28 ($k = 1, n = 100$) to 36 ($k = 4, n = 200$) for the 32 loci case. Without using a list, there are $2^{25} \approx$ 3e7 and $2^{32} \approx$ 4e9 possible haplotypes. Thus by using the ATCD list, we can restrict our attention to only 20 to 40 haplotypes, hence the huge savings in running time. It can also be seen from Tables 3 and 4 that making a list longer does not guarantee better results, as the EML and CD lists are much longer than the ATCD list but the resulting estimates are much worse. What seems important is to add the right haplotypes and remove unnecessary ones. If an imperfect external list exists, then a sensible hybrid method is to combine it with the collapsed data list to form a union list which can be further augmented and trimmed using the techniques described in this chapter.

Currently we are only adding haplotypes with a single "1" to the list, which seems reasonable for the study of rare variants, but one can conceivably also add haplotypes with two 1's to the list. This will increase the number of possibilities substantially during the first iteration of the EM algorithm, but most of these haplotypes will be removed after one iteration.

The signs are promising that the use of the ATCD list can push the limit of the EM algorithm in terms of the number of loci and pool size that it can handle. This method is particularly well suited for estimating the haplotype distributions of rare variants which are of substantial current

interest. Note that our method does not require sampling, and is shown in simulation study to work for case of 32 loci and pool size 4, which is beyond the scope of most sampling-based methods, MCMC or deterministic. To calculate the standard errors of parameter estimates, Louis's formula is one choice. However, since we add and remove haplotypes from the list in the implementation of EM algorithm, the final estimates may not be MLEs. So we suggest using bootstrap to compute the standard errors.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

5.1.1 Human biomonitoring

Chapter 2 considers the human biomonitoring of exposure to environmental chemicals, which has become increasingly important. Individual monitoring is not viable due to low individual exposure level or insufficient volume of materials and the prohibitive cost of taking measurements from large number of subjects. Pooling of samples is an efficient and cost effective way to collect data. Estimation is, however, complicated as individual values within each pool are not observed but are only known up to their average or weighted average. The distribution of such averages is intractable when the individual measurements are log-normally distributed, which is a common assumption. We propose to replace the intractable distribution of the pool averages by a Gaussian likelihood to obtain parameter estimates. If the pool size is large, this method produces statistically efficient estimates, but regardless of pool size, the method yields consistent estimates as the number of pools increases. An empirical Bayes (EB) Gaussian likelihood approach,

as well as its Bayesian analogue, are developed to pool information from various demographic groups by using a mixed effect formulation. We also discuss methods to estimate the underlying mean-variance relationship, and to select a good model for the means, which can be incorporated into the proposed EB or Bayes framework. By borrowing strength across groups, the EB estimator is more efficient than the individual group-specific estimator. Simulation results show that the EB Gaussian likelihood estimates outperform a previous method proposed for the National Health and Nutrition Examination Surveys with much smaller bias and better coverage in interval estimation, especially after correction of bias.

5.1.2 Haplotype frequency estimation

Chapter 3 considers the estimation of haplotype frequencies based on pooled genotype in the case of rare variants. Haplotype information could lead to more powerful tests of genetic association than single locus analyses. Due to phase ambiguity, the estimation of haplotype frequencies from genotype data is non-trivial. The challenge is compounded when individuals are pooled together to save costs or to increase sample size which is crucial in the study of rare variants. Existing expectation maximization (EM) type algorithms are slow and cannot cope with large pool size or long haplotypes. We show that by collapsing the total allele frequencies of each pool suitably, the maximum likelihood estimates (MLEs) of haplotype frequencies based on the collapsed data can be calculated very quickly regardless of pool size and haplotype length. A running time analysis is provided to demonstrate the considerable savings in time that the collapsed data method can bring. The method is particularly well suited to estimating certain union probabilities

useful in the study of rare variants. Theoretical and empirical evidence are given to suggest that the proposed estimation method will not suffer much loss in efficiency if the variants are rare. The method is used to analyze re-sequencing data collected from a case control study involving 148 obese persons and 150 controls. Focusing on a region containing 25 rare variants around the *MGLL* gene, our method selects 3 rare variants as potentially causal. This is more parsimonious than the 12 variants selected by a recently proposed covering method. From another set of 32 rare variants around the *FAAH* gene, we discover an interesting potential interaction between two of them.

In chapter 4, for more efficient estimates, we propose a way to construct a data-based list of possible haplotypes to be used in conjunction with the EM algorithm to make it more feasible computationally. By viewing the pooled genotype data as incomplete data, the EM algorithm is the natural algorithm to use, but it is computationally intensive. A recent proposal to reduce the computational burden is to make use of database information to form a list of frequently occurring haplotypes, and to restrict the haplotypes to come from this list only in implementing the EM algorithm. There is, however, the danger of using an incorrect list, and there may not be enough database information to form a list externally in some applications. We investigate the possibility of creating an internal list from the data at hand. One way to form such a list is to collapse the observed total allele frequencies to "zero" or "at least one", which is shown to have the desirable effect of amplifying the haplotype frequencies. To improve coverage, we propose ways to add and remove haplotypes from the list, and a benchmarking method to determine the frequency threshold for remov-

ing haplotypes. Simulation results show that the EM estimates based on a suitably augmented and trimmed collapsed data list (ATCDL) perform satisfactorily. In two scenarios involving 25 and 32 loci respectively, the EM-ATCDL estimates outperform the EM estimates based on other lists as well as the collapsed data MLEs. The proposed augmented and trimmed CD list is a useful list for the EM algorithm to base upon in estimating the haplotype distributions of rare variants. It can handle more markers and larger pool size than existing methods, and the resulting EM-ATCDL estimates are more efficient than the EM estimates based on other lists.

We have proposed two methods for genetic studies, collapsed data MLE and a modified EM with an internal list. Running time analysis was reported in Tables 3.1 and 4.1. Collapsed data MLE runs very fast regardless of pool size and haplotype length, so it can be applied in large scale datasets. The modified EM algorithm also runs fast and can push the limit of the EM algorithm in terms of the number of loci and pool size that it can handle, which may be applicable in moderate or large datasets.

5.2 Ongoing and Future Work

5.2.1 Human biomonitoring

Pooling of samples is an efficient and cost effective way to collect data since the number of chemical measurements required can be substantially reduced. However, the individual values within each pool are not observed, and only their averages or weighted averages are measured. Previously, pooled samples were analyzed based on the strata of age group, gender and race, and only pooled level data were used in the model. However,

demographic data (e.g. age) are available for each individual within the pool and can be incorporated in the model to further smooth across demographic groups. Model selection based on pooled data is an interesting question and Gaussian version of BIC needs further investigation if it is a valid method. In general, hypothesis testing is an important statistical problem in statistical inference. In the future work, we can also consider if it is possible to perform Gaussian likelihood based score test. We discuss below some other possible future research on this topic.

• **Integrating individual and pooled level data**. Assume the unweighed average $A_{ij} = \sum_{k=1}^{K} X_{ijk}/K$ is recorded for the j^{th} pool in the i^{th} group containing K individuals with $Y_{ijk} = \log X_{ijk} \sim N(\mu_i, \sigma_i^2)$, where $i = 1, \cdots, d,\ j = 1, \cdots, n_i,\ k = 1, \cdots, K$. More information (e.g. age) are available for each individual X_{ijk}, denoted by a vector $U_{ijk} = (U_{ijk1}, \cdots, U_{ijkM})'$. Previous estimator was proposed by categorizing U_{ijk} into groups with similar demographic characteristics and then only group level information were retained and used. Intuitively, a better estimator can be investigated by incorporating the individual level data U_{ijk} in the model.

For simplicity, we consider one pool here. Suppressing the dependence on the group i and pool j, the model after incorporating individual level data is written as

$$A = \frac{\sum_{k=1}^{K} X_k}{K}, \quad \log X_k \sim N(\zeta' U_k, \sigma^2)$$

where A and $U_k = (U_{k1}, \cdots, U_{kM})'$ are observed; $X_k,\ k = 1, \cdots, K$, are latent variables; $\zeta = (\zeta_1, \cdots, \zeta_M)'$ and σ^2 are the parameters. The probability density of A is given by the K-fold convolution of log-normal densities,

which is highly intractable and not easy to maximize directly. Monte Carlo EM (MCEM) algorithm and Gaussian likelihood estimation can be applied, by using similar techniques as described in the chapter 2

When implementing MCEM, at the E-step, we take conditional expectation of the complete data log-likelihood function $l(\zeta, \sigma^2; \log X_1, \cdots, \log X_K)$ given the observed data A and $U = (U_1, \cdots, U_K)'$, and the current estimates $\hat{\zeta}_{(r)}$ and $\hat{\sigma}^2_{(r)}$ obtained from the r^{th} iteration to get

$$\begin{aligned} Q(\zeta, \sigma^2; \hat{\zeta}_{(r)}, \hat{\sigma}^2_{(r)}) =& \mathrm{E}\left[l(\zeta, \sigma^2; \log X_1, \cdots, \log X_K) | A, U; \hat{\zeta}_{(r)}, \hat{\sigma}^2_{(r)} \right] \\ =& -\frac{K}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\,\mathrm{E}\left[\sum_{k=1}^{K} (\log X_k - \zeta' U_k)^2 | A, U; \hat{\zeta}_{(r)}, \hat{\sigma}^2_{(r)} \right] \end{aligned}$$

up to an additive constant. At the M-step, we maximize $Q(\zeta, \sigma^2; \hat{\zeta}_{(r)}, \hat{\sigma}^2_{(r)})$ with respect to ζ and σ^2 to obtain the updated estimates

$$\begin{aligned} \hat{\zeta}_{(r+1)} =& (U^T U)^{-1}\,\mathrm{E}\left[U^T lX | A, U; \hat{\zeta}_{(r)}, \hat{\sigma}^2_{(r)} \right] \\ \hat{\sigma}^2_{(r+1)} =& \frac{\mathrm{E}\left[lX^T lX | A, U; \hat{\zeta}_{(r)}, \hat{\sigma}^2_{(r)} \right] - \hat{\zeta}^T_{(r+1)} U^T U \hat{\zeta}_{(r+1)}}{K} \end{aligned}$$

where $lX = (\log X_1, \cdots, \log X_K)'$. The conditional expectations

$$\begin{aligned} &\mathrm{E}\left[\sum_{k=1}^{K} (U_{km} \log X_k) \,\middle|\, A, U; \hat{\zeta}_{(r)}, \hat{\sigma}^2_{(r)} \right], \ m = 1, \cdots, M, \\ &\mathrm{E}\left[\sum_{k=1}^{K} (\log X_k)^2 \,\middle|\, A, U; \hat{\zeta}_{(r)}, \hat{\sigma}^2_{(r)} \right] \end{aligned}$$

required in the E-step of the EM algorithm have no closed-form expressions. We can consider using MCEM by approximating the above conditional expectations.

Gaussian likelihood estimator is an alternative method to use. According to the Central Limit Theorem, the pool average is approximately normally distributed if the pool size K is large with mean and variance, given by

$$\begin{aligned} \mathrm{E}[A] =& \frac{\sum_{k=1}^{K} \exp\left(\zeta^T U_k + \sigma^2/2\right)}{K} \\ \mathrm{var}[A] =& \frac{\sum_{k=1}^{K} \exp\left(2\zeta^T U_k + \sigma^2\right)\left[\exp\left(\sigma^2\right) - 1\right]}{K^2} \end{aligned}$$

If only pooled level information are used, $U_k,\ k = 1, \cdots, K$ are constant for all the individuals within the same pool. This is the case that has been discussed in the chapter 2.

5.2.2 Haplotype frequency estimation

For non-rare alleles, haplotype distributions cannot be estimated well from pooled data. The asymptotic efficiency of pooled data estimator is reduced by a factor equal to the pool size whenever the order of the cumulant to be estimated is increased by one (Kuk et al., 2010), and hence it may be appropriate to use pooled data to estimate only the low order of haplotype frequencies, e.g. the first and second order of marginal frequencies. A sensible strategy is to collect individual as well as pooled genotype data. In addition, it is interesting to see if our collapsed data MLE can be extended for family-based data where independence assumption is no longer valid. One possibility is to use a random effects formulation. We discuss below some other ongoing and possible future research on how to integrate these two data.

- **Combining individual and pooled genotype data**. A calibra-

tion type estimator based on the combined data is proposed which is more efficient than the estimator based on individual data alone. In order to take use of both individual and pooled genotype data, we propose adjusting the individual data estimators by using the first and/or second order of marginal frequencies estimated from pooled data. Denote by $\mathbf{f_0} = (f_0(1), \cdots, f_0(\Lambda_i), \cdots)^T$ the vector of the first and/or second order of marginal frequencies with "0" at positions Λ_i, where Λ_i is a non-empty subset of $\{1, \cdots, L\}$. We consider adjusting the individual data estimator of f in the following form:

$$\hat{f}_B = \hat{f}_{\text{idv}} + B^T \left(\hat{\mathbf{f}}_{\text{pol}} - \hat{\mathbf{f}}_{\text{idv}} \right).$$

where $\hat{f}_B$ is the adjusted estimator and $\hat{f}_{\text{idv}}$ is the individual data estimator of f; $\hat{\mathbf{f}}_{\text{idv}}$ and $\hat{\mathbf{f}}_{\text{pol}}$ are the individual and pooled data estimators of $\mathbf{f_0}$ respectively. The variance of the adjusted estimator $\hat{f}_B$ is given by

$$\text{var}\left[\hat{f}_B\right] = \text{var}\left[\hat{f}_{\text{idv}}\right] + B^T \,\text{cov}\left[\hat{\mathbf{f}}_{\text{pol}} - \hat{\mathbf{f}}_{\text{idv}}\right] B + 2\,\text{cov}\left[\hat{f}_{\text{idv}}, \hat{\mathbf{f}}_{\text{pol}} - \hat{\mathbf{f}}_{\text{idv}}\right] B.$$

We can choose B to minimize the above variance. Taking the first partial derivatives with respect to B yields

$$\frac{\partial\, \text{var}\left[\hat{f}_B\right]}{\partial B} = 2\,\text{cov}\left[\hat{\mathbf{f}}_{\text{pol}} - \hat{\mathbf{f}}_{\text{idv}}\right] B + 2\,\text{cov}\left[\hat{f}_{\text{idv}}, \hat{\mathbf{f}}_{\text{pol}} - \hat{\mathbf{f}}_{\text{idv}}\right]^T.$$

Let $\frac{\partial\, \text{var}\left[\hat{f}_B\right]}{\partial B} = 0$, then we can obtain the optimal B^* which minimizes the variance of the adjusted estimator. Since it is always easy to have a full

rank cov $\left[\hat{\mathbf{f}}_{\text{pol}} - \hat{\mathbf{f}}_{\text{idv}}\right]$, the optimal B^* is then given by

$$B^* = -\operatorname{cov}\left[\hat{\mathbf{f}}_{\text{pol}} - \hat{\mathbf{f}}_{\text{idv}}\right]^{-1} \operatorname{cov}\left[\hat{f}_{\text{idv}}, \hat{\mathbf{f}}_{\text{pol}} - \hat{\mathbf{f}}_{\text{idv}}\right]^T. \tag{5.1}$$

The variance of the adjusted estimator using the above optimal B^* is given by

$$\operatorname{var}\left[\hat{f}_{B^*}\right] = \operatorname{var}\left[\hat{f}_{\text{idv}}\right]\left(1 - R^2\right), \tag{5.2}$$

where

$$R^2 = \operatorname{cov}\left[\hat{f}_{\text{idv}}, \hat{\mathbf{f}}_{\text{pol}} - \hat{\mathbf{f}}_{\text{idv}}\right] \operatorname{cov}\left[\hat{\mathbf{f}}_{\text{pol}} - \hat{\mathbf{f}}_{\text{idv}}\right]^{-1} \operatorname{var}\left[\hat{f}_{\text{idv}}\right]^{-1} \operatorname{cov}\left[\hat{f}_{\text{idv}}, \hat{\mathbf{f}}_{\text{pol}} - \hat{\mathbf{f}}_{\text{idv}}\right]^T$$

is the multiple correlation. According to (5.2), the adjusted estimator $\hat{f}_{B^*}$ has smaller variance than the individual data estimator $\hat{f}_{\text{idv}}$. If all the haplotype frequency estimators are unbiased, the adjusted frequency $\hat{h}_{B^*}$ should also be unbiased. So we may expect the adjusted estimator $\hat{f}_{B^*}$ would perform better than the individual data estimator $\hat{f}_{\text{idv}}$.

• **Optimal combination ratio**. Given the same cost of genotyping for individual and pooled data, the total number of genotyping is fixed (i.e. $n = n_I + n_P$), and this brings up the question that how to assign samples in order to obtain efficient estimators. We can further investigate (5.2) to find an optimal ratio between the numbers of individual and pooled genotype data at a fixed cost of genotyping. The individual data MLE of haplotype frequency $\hat{\mathbf{f}}_I$ can be estimated through EM algorithm. $\hat{f}_{\text{idv}}$ and $\hat{\mathbf{f}}_{\text{idv}}$ can be written as a linear combination of $\hat{\mathbf{h}}_I$,

$$\hat{f}_{\text{idv}} = I_h^T \hat{\mathbf{f}}_I, \qquad \hat{\mathbf{f}}_{\text{idv}} = J^T \hat{\mathbf{f}}_I,$$

where I_h is a vector with all zeros but one "1", indicating the position of $\hat{f}_{\text{idv}}$ in $\hat{\mathbf{f}}_I$; and J is a matrix with each column specifying which haplotypes are compatible with the corresponding marginal haplotype. For example, $f_0(\Lambda) = P(Y_l = 0, l \in \Lambda) = \sum f(y_l = 0, l \in \Lambda)$, where $f_0(\Lambda)$ is the marginal frequency with "0" at positions Λ, and $f(y_l = 0, l \in \Lambda)$ is the frequency of haplotype with zeros at positions Λ. So $\left\{\text{var}\left[\hat{f}_{\text{idv}}\right] R^2\right\}$ in (5.2) can be calculated as

$$\begin{aligned}
& \text{cov}\left[\hat{f}_{\text{idv}}, \hat{\mathbf{f}}_{\text{pol}} - \hat{\mathbf{f}}_{\text{idv}}\right] \text{cov}\left[\hat{\mathbf{f}}_{\text{pol}} - \hat{\mathbf{f}}_{\text{idv}}\right]^{-1} \text{cov}\left[\hat{f}_{\text{idv}}, \hat{\mathbf{f}}_{\text{pol}} - \hat{\mathbf{f}}_{\text{idv}}\right]^T \\
= & \text{cov}\left[I_h^T \hat{\mathbf{f}}_I, \hat{\mathbf{f}}_{\text{pol}} - J^T \hat{\mathbf{f}}_I\right] \text{cov}\left[\hat{\mathbf{f}}_{\text{pol}} - J^T \hat{\mathbf{f}}_I\right]^{-1} \text{cov}\left[I_h^T \hat{\mathbf{f}}_I, \hat{\mathbf{f}}_{\text{pol}} - J^T \hat{\mathbf{f}}_I\right]^T \\
= & \text{Cov}[I_h^T \hat{\mathbf{f}}_I, -J^T \hat{\mathbf{f}}_I] \left\{\text{cov}\left[\hat{\mathbf{f}}_{\text{pol}}\right] + \text{cov}\left[J^T \hat{\mathbf{f}}_I\right]\right\}^{-1} \text{cov}\left[I_h^T \hat{\mathbf{f}}_I, -J^T \hat{\mathbf{f}}_I\right]^T \\
= & I_h^T \text{cov}\left[\hat{\mathbf{f}}_I\right] J \left\{\text{cov}\left[\hat{\mathbf{f}}_{\text{pol}}\right] + J^T \text{cov}\left[\hat{\mathbf{f}}_I\right] J\right\}^{-1} J^T \text{cov}\left[\hat{\mathbf{f}}_I\right]^T I_h,
\end{aligned}$$

let $C_I = \text{cov}\left[\hat{\mathbf{f}}_I\right]$ and $C_P = \text{cov}\left[\hat{\mathbf{f}}_{\text{pol}}\right]$; the above function can be written as

$$\begin{aligned}
& I_h^T \text{cov}\left[\hat{\mathbf{f}}_I\right] J \left\{\text{cov}\left[\hat{\mathbf{f}}_{\text{pol}}\right] + J^T \text{cov}\left[\hat{\mathbf{f}}_I\right] J\right\}^{-1} J^T \text{cov}\left[\hat{\mathbf{f}}_I\right]^T I_h \\
= & I_h^T C_I J \left(C_P + J^T C_I J\right)^{-1} J^T C_I I_h \\
= & I_h^T C_I J C_P^{-1} \left(\mathbf{I} + J^T C_I J C_P^{-1}\right)^{-1} J^T C_I I_h \\
= & I_h^T C_I J C_P^{-1} J^T \left(\mathbf{I} + C_I J C_P^{-1} J^T\right)^{-1} C_I I_h \\
= & I_h^T C_I J C_P^{-1} J^T C_I \left(\mathbf{I} + J C_P^{-1} J^T C_I\right)^{-1} I_h. \qquad (5.3)
\end{aligned}$$

Substituting (5.3) into the variance formula of $\hat{f}_{B^*}$ in (5.2), then we have

$$\begin{aligned}
\text{Var}[\hat{h}_{B^*}] =& I_h^T C_I I_h - I_h^T C_I J C_P^{-1} J^T C_I \left(\mathbf{I} + J C_P^{-1} J^T C_I\right)^{-1} I_h \\
=& I_h^T C_I \left[\mathbf{I} - J C_P^{-1} J^T C_I \left(\mathbf{I} + J C_P^{-1} J^T C_I\right)^{-1}\right] I_h
\end{aligned}$$

$$
\begin{aligned}
&= I_h^T C_I \left(\mathbf{I} + JC_P^{-1}J^T C_I\right)^{-1} I_h \\
&= I_h^T \left(C_I^{-1} + JC_P^{-1}J^T\right)^{-1} I_h
\end{aligned} \tag{5.4}
$$

which implicitly involves n_I, n_P and haplotype frequencies. Since $C_P = \frac{1}{n_P}\begin{pmatrix} f_0(1)\left[1-f_0(1)\right] & f_0(1,2)-f_0(1)f_0(2) & \cdots \\ \vdots & & \ddots \end{pmatrix}$, define $JC_P^{-1}J^T = n_P\mathbf{F}$. For the individual data, we have $C_I = O(1/n_I)$. When n_I is large, C_I can be approximated by $\frac{\mathbf{Q}}{n_I}$. So the above function (5.4) can be approximated by

$$
\text{var}\left[\hat{f}_{B^*}\right] \approx I_h^T \left(n_I\mathbf{Q}^{-1} + n_P\mathbf{F}\right)^{-1} I_h, \tag{5.5}
$$

which is a trade-off between n_I and n_P. A further look at (5.2) can give us some explanation. In (5.2), the variance of the adjusted estimator using the optimal B^* is a multiplication between var $\left[\hat{f}_{\text{idv}}\right]$ and $(1-R^2)$. So the decrease in var $\left[\hat{f}_{B^*}\right]$ can be contributed by a decrease in var $\left[\hat{f}_{\text{idv}}\right]$ or an increase in R^2. Note that the variance of the individual data MLE,

$$
\text{var}\left[\hat{f}_{\text{idv}}\right] = I_h^T C_I I_h = O\left(\frac{1}{n_I}\right)
$$

which will decrease as the number of individual data n_I increases at fixed n_P. Based on (5.3), R^2 can be calculated as

$$
R^2 = \frac{I_h^T C_I JC_P^{-1}J^T C_I \left(\mathbf{I} + JC_P^{-1}J^T C_I\right)^{-1} I_h}{I_h^T C_I I_h} \approx \frac{I_h^T \mathbf{QFQ}\left(\frac{n_I}{n_P}\mathbf{I} + \mathbf{FQ}\right)^{-1} I_h}{I_h^T \mathbf{Q} I_h} \tag{5.6}
$$

According to (5.6), R^2 will increase to 1 as the number of pooled data n_P increases at fixed n_I. So the increase in either n_I and n_P can lead to an decrease in var $\left[\hat{f}_{B^*}\right]$. An optimal combination ratio between n_I and n_P

may be obtained based on (5.5).

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723.

Angerer, J., Ewers, U., and Wilhelm, M. (2007). Human biomonitoring: state of the art. *International journal of hygiene and environmental health*, 210(3):201–228.

Bates, M. N., Buckland, S. J., Garrett, N., Caudill, S. P., and Ellis, H. (2005). Methodological aspects of a national population-based study of persistent organochlorine compounds in serum. *Chemosphere*, 58(7):943–951.

Bates, M. N., Buckland, S. J., Garrett, N., Ellis, H., Needham, L. L., Patterson Jr, D. G., Turner, W. E., and Russell, D. G. (2004). Persistent organochlorines in the serum of the non-occupationally exposed new zealand population. *Chemosphere*, 54(10):1431–1443.

Bhatia, G., Bansal, V., Harismendy, O., Schork, N. J., Topol, E. J., Frazer, K., and Bafna, V. (2010). A covering method for detecting genetic associations between rare variants and common phenotypes. *PLoS computational biology*, 6(10):e1000954.

Bignert, A., Göthberg, A., Jensen, S., Litzén, K., Odsjö, T., Olsson, M.,

and Reutergårdh, L. (1993). The need for adequate biological sampling in ecotoxicological investigations: a retrospective study of twenty years pollution monitoring. *Science of the Total Environment*, 128(2):121–139.

Caudill, S. P. (2010). Characterizing populations of individuals using pooled samples. *Journal of Exposure Science and Environmental Epidemiology*, 20(1):29–37.

Caudill, S. P. (2011). Important issues related to using pooled samples for environmental chemical biomonitoring. *Statistics in Medicine*, 30(5):515–521.

Caudill, S. P. (2012). Use of pooled samples from the national health and nutrition examination survey. *Statistics in medicine*, 31(27):3269–3277.

Caudill, S. P., Turner, W. E., and Patterson Jr, D. G. (2007a). Geometric mean estimation from pooled samples. *Chemosphere*, 69(3):371–380.

Caudill, S. P., Wong, L.-Y., Turner, W. E., Lee, R., Henderson, A., and Patterson Jr, D. G. (2007b). Percentile estimation using variable censored data. *Chemosphere*, 68(1):169–180.

Clark, A. G. (2004). The role of haplotypes in candidate gene studies. *Genetic epidemiology*, 27(4):321–333.

Crowder, M. (1985). Gaussian estimation for correlated binomial data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 229–237.

Crowder, M. (2001). On repeated measures analysis with misspecified covariance structure. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(1):55–62.

Dempster, A. P., Laird, N. M., Rubin, D. B., et al. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal statistical Society*, 39(1):1–38.

Dorfman, R. (1943). The detection of defective members of large populations. *The Annals of Mathematical Statistics*, 14(4):436–440.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The annals of Statistics*, pages 1–26.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., and Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, 11(6):446–450.

Erik, S. (2004). Biomonitoring: Pollution gets personal. *Science*, 304(5679):1892–4.

Excoffier, L. and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular biology and evolution*, 12(5):921–927.

Gasbarra, D., Kulathinal, S., Pirinen, M., and Sillanpaa, M. J. (2011). Estimating haplotype frequencies by combining data from large dna pools with database information. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 8(1):36–44.

Gastwirth, J. L. and Hammick, P. A. (1989). Estimation of the prevalence of a rare disease, preserving the anonymity of the subjects by group testing: Application to estimating the prevalence of aids antibodies in blood donors. *Journal of statistical planning and inference*, 22(1):15–27.

Gideon, S. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Gosset, W. (1927). Errors of routine analysis. *Biometrika*, 19(1-2):151–64.

Halperin, E. and Karp, R. M. (2004). Perfect phylogeny and haplotype assignment. In *Proceedings of the eighth annual international conference on Resaerch in computational molecular biology*, pages 10–19. ACM.

Homer, N., Tembe, W. D., Szelinger, S., Redman, M., Stephan, D. A., Pearson, J. V., Nelson, S. F., and Craig, D. (2008). Multimarker analysis and imputation of multiple platform pooling-based genome-wide association studies. *Bioinformatics*, 24(17):1896–1902.

Iliadis, A., Anastassiou, D., and Wang, X. (2012). Fast and accurate haplotype frequency estimation for large haplotype vectors from pooled dna data. *BMC genetics*, 13(1):94.

Ito, T., Chiku, S., Inoue, E., Tomita, M., Morisaki, T., Morisaki, H., and Kamatani, N. (2003). Estimation of haplotype frequencies, linkage-disequilibrium measures, and combination of haplotype copies in each pool by use of pooled dna data. *The American Journal of Human Genetics*, 72(2):384–398.

Kehoe, R., Thamann, F., and Cholak, J. (1933). Lead absorption and excretion in certain lead trades. *J. Indust. Hyg*, 15:306–319.

Kim, S. Y., Li, Y., Guo, Y., Li, R., Holmkvist, J., Hansen, T., Pedersen, O., Wang, J., and Nielsen, R. (2010). Design of association studies with pooled or un-pooled next-generation sequencing data. *Genetic epidemiology*, 34(5):479–491.

Kingman, J. F. C. (1982). The coalescent. *Stochastic processes and their applications*, 13(3):235–248.

Kirkpatrick, B., Armendariz, C. S., Karp, R. M., and Halperin, E. (2007). Haplopool: improving haplotype frequency estimation through dna pools and phylogenetic modeling. *Bioinformatics*, 23(22):3048–3055.

Kuk, A. Y., Li, X., and Xu, J. (2013a). An em algorithm based on an internal list for estimating haplotype distributions of rare variants from pooled genotype data. *BMC genetics*, 14(1):1–17.

Kuk, A. Y., Li, X., and Xu, J. (2013b). A fast collapsed data method for estimating haplotype frequencies from pooled genotype data with applications to the study of rare variants. *Statistics in medicine*, 32(8):1343–1360.

Kuk, A. Y., Nott, D. J., and Yang, Y. (2014). A stepwise likelihood ratio test procedure for rare variant selection in case–control studies. *Journal of human genetics*.

Kuk, A. Y., Xu, J., and Yang, Y. (2010). A study of the efficiency of pooling in haplotype estimation. *Bioinformatics*, 26:2556–2563.

Kuk, A. Y., Zhang, H., and Yang, Y. (2009). Computationally feasible estimation of haplotype frequencies from pooled dna with and without hardy–weinberg equilibrium. *Bioinformatics*, 25(3):379–386.

Li, B. and Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3):311–321.

Liang, W. E., Thomas, D. C., and Conti, D. V. (2012). Analysis and optimal design for association studies using next-generation sequencing with case-control pools. *Genetic epidemiology*, 36(8):870–881.

Lin, D. and Zeng, D. (2006). Likelihood-based inference on haplotype effects in genetic association studies. *Journal of the American Statistical Association*, 101(473):89–104.

Lin, D.-Y. and Tang, Z.-Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. *The American Journal of Human Genetics*, 89(3):354–367.

Macgregor, S., Zhao, Z. Z., Henders, A., Martin, N. G., Montgomery, G. W., and Visscher, P. M. (2008). Highly cost-efficient genome-wide association studies using dna pools and dense snp arrays. *Nucleic acids research*, 36(6):e35.

Madsen, B. E. and Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS genetics*, 5(2):e1000384.

Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z. S., Munro, H. M., Abecasis, G. R., et al. (2006). A comparison of phasing algorithms for trios and unrelated individuals. *The American Journal of Human Genetics*, 78(3):437–450.

Mardis, E. R. (2008). Next-generation dna sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, 9:387–402.

Meaburn, E., Butcher, L. M., Schalkwyk, L. C., and Plomin, R. (2006).

Genotyping pooled dna using 100k snp microarrays: a step towards genomewide association scans. *Nucleic acids research*, 34(4):e28–e28.

Morris, R. W. and Kaplan, N. L. (2002). On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genetic epidemiology*, 23(3):221–233.

Muers, M. (2010). Genomics: No half measures for haplotypes. *Nature Reviews Genetics*, 12(2):77–77.

Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S. M., Roeder, K., and Daly, M. J. (2011). Testing for an unusual distribution of rare variants. *PLoS genetics*, 7(3):e1001322.

Niu, T. (2004). Algorithms for inferring haplotypes. *Genetic epidemiology*, 27(4):334–347.

Niu, T., Qin, Z. S., Xu, X., and Liu, J. S. (2002). Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *The American Journal of Human Genetics*, 70(1):157–169.

Norton, N., Williams, N. M., O'Donovan, M. C., and Owen, M. J. (2004). Dna pooling as a tool for large-scale association studies in complex traits. *Annals of medicine*, 36(2):146–152.

Odeh, R. E. and Owen, D. B. (1980). *Tables for normal tolerance limits, sampling plans, and screening*. M. Dekker.

Pirinen, M. (2009). Estimating population haplotype frequencies from pooled snp data using incomplete database information. *Bioinformatics*, 25(24):3296–3302.

Pirinen, M., Kulathinal, S., Gasbarra, D., and SILLANPÄÄ, M. J. (2008). Estimating population haplotype frequencies from pooled dna samples using phase algorithm. *Genetics research*, 90(06):509–524.

Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). March*, pages 20–22.

Price, A. L., Kryukov, G. V., de Bakker, P. I., Purcell, S. M., Staples, J., Wei, L.-J., and Sunyaev, S. R. (2010). Pooled association tests for rare variants in exon-resequencing studies. *The American Journal of Human Genetics*, 86(6):832–838.

Quade, S. R., Elston, R. C., and Goddard, K. A. (2005). Estimating haplotype frequencies in pooled dna samples when there is genotyping error. *BMC genetics*, 6(1):25.

Roach, J. C., Glusman, G., Hubley, R., Montsaroff, S. Z., Holloway, A. K., Mauldin, D. E., Srivastava, D., Garg, V., Pollard, K. S., Galas, D. J., et al. (2011). Chromosomal haplotypes by genetic phasing of human families. *The American Journal of Human Genetics*, 89(3):382–397.

Schaid, D. J. (2004). Evaluating associations of haplotypes with traits. *Genetic epidemiology*, 27(4):348–364.

Sexton, K., Needham, L. L., and Pirkle, J. L. (2004). Human biomonitoring of environmental chemicals. *American Scientist*, 92(1):38–45.

Sham, P., Bader, J. S., Craig, I., O'Donovan, M., and Owen, M. (2002).

Dna pooling: a tool for large-scale association studies. *Nature Reviews Genetics*, 3(11):862–871.

Stephens, M. and Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *The American Journal of Human Genetics*, 76(3):449–462.

Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J., and Schork, N. J. (2011). The importance of phase information for human genomics. *Nature Reviews Genetics*, 12(3):215–223.

Thornton, J. W., McCally, M., and Houlihan, J. (2002). Biomonitoring of industrial pollutants: health and policy implications of the chemical body burden. *Public Health Reports*, 117(4):315.

Wei, G. C. G. and Tanner, M. A. (1990). A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85:699–704.

Whittle, P. (1962). Gaussian estimation in stationary time series. *Bulletin of the International Statistical Institute*, 39:105–129.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93.

Xu, J. and Kuk, A. Y. (2014). On pooling of data and its relative efficiency. *International statistical review*, to appear.

Yang, H.-C., Pan, C.-C., Lin, C.-Y., and Fann, C. S. (2006). Pda: pooled dna analyzer. *BMC bioinformatics*, 7(1):233.

Yant, W., Schrenk, H., Sayers, R., Howarth, A., and Reinhart, W. (1936). Urine sulfate determination as a measure of benzene exposure. *J. Ind. Hyg. Toxicol*, 18:69.

Zhang, H., Yang, H.-C., and Yang, Y. (2008). Poool: an efficient method for estimating haplotype frequencies from large dna pools. *Bioinformatics*, 24(17):1942–1948.