Genetic Epidemiology 35 : 291–302 (2011)

# Genetic Variance Components Estimation for Binary Traits Using Multiple Related Individuals

**Charalampos Papachristou,[1,2]* Carole Ober,[2] and Mark Abney[2]**

[1]*Department of Mathematics, Physics, and Statistics, University of the Sciences, Philadelphia, Pennsylvania*
[2]*Department of Human Genetics, University of Chicago, Chicago, Illinois*

Understanding and modeling genetic or nongenetic factors that influence susceptibility to complex traits has been the focus of many genetic studies. Large pedigrees with known complex structure may be advantageous in epidemiological studies since they can significantly increase the number of factors whose influence on the trait can be estimated. We propose a likelihood approach, developed in the context of generalized linear mixed models, for modeling dichotomous traits based on data from hundreds of individuals all of whom are potentially correlated through either a known pedigree or an estimated covariance matrix. Our approach is based on a hierarchical model where we first assess the probability of each individual having the trait and then formulate a likelihood assuming conditional independence of individuals. The advantage of our formulation is that it easily incorporates information from pertinent covariates as fixed effects and at the same time takes into account the correlation between individuals that share genetic background or other random effects. The high dimensionality of the integration involved in the likelihood prohibits exact computations. Instead, an automated Monte Carlo expectation maximization algorithm is employed for obtaining the maximum likelihood estimates of the model parameters. Through a simulation study we demonstrate that our method can provide reliable estimates of the model parameters when the sample size is close to 500. Implementation of our method to data from a pedigree of 491 Hutterites evaluated for Type 2 diabetes (T2D) reveal evidence of a strong genetic component to T2D risk, particularly for younger and leaner cases. *Genet. Epidemiol.* 35:291–302, 2011. 

**Key words: binary trait; genetic variance components; GLMMs; MCEM; diabetes; complex pedigrees**

Contract grant sponsor: NIH; Contract grant number: HG02899.
*Correspondence to: Charalampos Papachristou, Department of Mathematics, Physics, and Statistics, University of the Sciences, 600 South 43rd Street, Mailbox 64, Philadelphia, PA 19104. E-mail: c.papach@usp.edu
Received 15 September 2010; Revised 21 December 2010; Accepted 31 January 2011
Published online 4 April 2011 in Wiley Online Library (wileyonlinelibrary.com/journal/gepi).
DOI: 10.1002/gepi.20577

## INTRODUCTION

Identifying the genetic and environmental factors that influence susceptibility to common complex diseases, such as T2D, is a major goal of current biomedical research. Although environmental risk factors can often be estimated by studying unrelated individuals, estimation of genetic risks requires related individuals. In fact, the presence of familial aggregation of a disease is often taken as indicative of a genetic factor in susceptibility. A common measure of the degree of familial aggregation is the sibling relative risk $\lambda_s$, the ratio of the risk of the disease in the sibling of a case to the risk in the general population. The relative risk to a sibling (or other relationship type) holds significant interest because the pattern of risk across different relatives may reveal information about the mode of transmission of the disease and because it is a critical parameter in determining the power to map disease loci [Risch, 1990a,b]. A typical study design for estimating $\lambda_s$ would be to ascertain affected individuals from the population and measure the rate of disease among the probands' siblings, making adjustments for ascertainment bias, as needed [Olson and Cordell, 2000; Zou and Zhao, 2004]. This approach, however, does not easily account for individual specific risk factors, for example, age and gender, or common environmental effects. Apparent familial aggregation, then, may in fact be the result of shared nongenetic factors among families, but estimates of risk are uninformed by knowledge of these factors.

One approach to the analysis of binary traits that incorporate both environmental covariates and genetic correlation due to familial relatedness makes use of the generalized linear mixed model (GLMM) statistical framework. Although using GLMMs provides the necessary flexibility in modeling, it comes at the cost of large computational demands. In particular, the likelihood under such a model typically involves an integral of dimension equal to the number of individuals in the study. This difficulty, though, is ameliorated when the subjects can be grouped into nuclear families as the integral reduces to a product of integrals each of which has dimension equal to the size of the nuclear family. Recent work has addressed the case of binary trait data in nuclear families using either a Gibbs sampling approach [Burton et al., 1999] or an h-likelihood [Noh et al., 2006].

Within the context of using a GLMM to model a binary trait, however, a sample of many related individuals presents significant computational challenges. Unlike the

case of using nuclear families, where the likelihood separates into many independent likelihoods, with each being an integral of relatively small dimension, the likelihood for a large family will have a high-dimensional integral. Finding maximum likelihood estimates (MLE), then, becomes a daunting task. The particular application we undertake here involves the analysis of binary traits in hundreds of individuals, all of whom are joined together in a single, complex genealogy. Our approach is to use a Monte Carlo expectation-maximization (MCEM) algorithm to find MLEs of the GLMM [McCulloch, 1994, 1997; Booth and Hobert, 1999]. The Monte Carlo (MC) approach to the EM algorithm is computationally demanding because each calculation of the expectation calls for an MC sample. Here, we speed up the MCEM through the use of importance sampling [Levine and Casella, 2001], which allows relatively few MC samples while maintaining the efficacy of the MCEM.

In this article, we first describe the GLMM framework and the likelihood associated with it, as well as the development of the MCEM and importance sampling implementations. We then demonstrate the validity of the method by applying it to a simple case where exact MLEs can also be computed. Simulations are used to show that the method gives reliable estimates in the more computationally challenging case of many related individuals. Finally, we demonstrate this method in a Hutterite pedigree in which 491 related individuals have been assessed for T2D, 36 of whom were affected.

# METHODS

## HIERARCHICAL MODEL

Consider a family of arbitrary structure consisting of $n$ members, and let $\mathbf{y}^t = (y_1,\ldots,y_n)$ be a vector of zeros and ones indicating whether an individual is affected (1), or not (0) by a binary trait of interest. We assume that for each person $i$ the susceptibility to the trait is influenced by an underlying (unobserved) random polygenic effect $u_i$ and (potentially) by some known covariates $\mathbf{x}_i$ (e.g., age, sex, etc.). Furthermore, suppose that given the random effects $u_i$, each pedigree member $i$ independently has the trait with probability $p_i$. In other words, given $\mathbf{u} = (u_1,\ldots,u_n)$, the $y_i$'s are assumed to be independent Bernoulli trials, each with parameter $p_i$. Finally, the probabilities $p_i$ are modeled as functions of the covariates and the random effects as follows

$$p_i = h(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + u_i),$$

where $\beta$ is a vector of fixed unknown regression coefficients and $h$ is an appropriate function that takes values between zero and one. Although there are a variety of functions that satisfy this condition, throughout we will assume that $h$ is the inverse logit function, that is $h(t) = e^t/(1+e^t)$. To completely specify our model we assume that the random effects $\mathbf{u}$ follow a multivariate normal distribution with mean zero and some covariance matrix $\boldsymbol{\Omega}$, i.e., $\mathbf{u}\sim\mathrm{MVN}(0,\boldsymbol{\Omega})$. Under additivity across the effects of the trait contributing loci, it can be shown [Jaquard, 1974; Abney et al., 2000] that $\Omega$ has the following form

$$\boldsymbol{\Omega} = 2\Phi\sigma_a^2 + \mathbf{V}_d\sigma_d^2 + \mathbf{V}_h\sigma_h^2 + \mathbf{V}_{ad}\mathrm{Cov}_h(a,d) + \mathbf{V}_{\mu_h}SS_{\mu_h}, \qquad (1)$$

where $\sigma_a^2$ and $\sigma_d^2$ are the additive and dominance genetic variance in the population, $\sigma_h^2$ and $\mathrm{Cov}_h(a,d)$ are the dominance variance and additive-dominance covariance in homozygous populations, and $SS_{\mu_h}$ is the inbreeding depression. Finally, $\boldsymbol{\Phi}$ is a known matrix whose $(i,j)$th element is the kinship coefficient between individuals $i$ and $j$, while the elements of the $\mathbf{V}$ matrices are also known and are functions of the condensed coefficients of identity [Jaquard, 1974]. Here we make the simplifying assumption that all genetic variances other than additive are zero so that Equation (1) only retains the first term and in that case the total genetic variance $\sigma_g$ is equal to the additive component $\sigma_a$. Also, note that unlike in the case of a quantitative trait, this formulation does not include residual environmental variance. Here, this random residual effect is captured by the Bernoulli trials for the affection status.

## THE LIKELIHOOD

In order to estimate the genetic parameters of interest $\boldsymbol{\sigma}^2 = \sigma_g^2$, as well as the covariate parameters $\boldsymbol{\beta}$ that correspond to the fixed effects, we need to formulate the likelihood function given the observed data under our model. If the random genetic effects were observed, then this likelihood function would simply be

$$\begin{aligned} L_c(\boldsymbol{\beta},\boldsymbol{\sigma}^2;\mathbf{y},\mathbf{u}) &= \frac{1}{(2\pi)^{n/2}}\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{(1-y_i)}\exp\left(-\frac{1}{2}\mathbf{u}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{u}\right)|\boldsymbol{\Omega}|^{-\frac{1}{2}} \\ &= \prod_{i=1}^{n} B(p_i;\boldsymbol{\beta},y_i,u_i)\phi_{\mathbf{u}}(0,\boldsymbol{\Omega};\boldsymbol{\sigma}^2), \end{aligned} \qquad (2)$$

where $B(\cdot)$and $\phi_{\mathbf{u}}(0,\boldsymbol{\Omega})$ denote the density functions for the Bernoulli and the multivariate normal with mean zero and variance-covariance matrix $\Omega$, respectively. However, in practice we only observe the affection statuses $\mathbf{y}$. Thus, we need to integrate out the $u_i$'s to obtain the marginal likelihood

$$L_m(\boldsymbol{\beta},\boldsymbol{\sigma}^2;\mathbf{y}) = \int\prod_{i=1}^{n} B(p_i;\boldsymbol{\beta},y_i,u_i)\phi_{\mathbf{u}}(0,\boldsymbol{\Omega};\boldsymbol{\sigma}^2)d\mathbf{u}. \qquad (3)$$

Evaluation, as well as maximization of this marginal likelihood requires the computation of an $n$-dimensional integral, where $n$ is the number of individuals. In general, unless $n$ is small or $\boldsymbol{\Omega}$ has a fairly simple form, exact computation of the above integral is infeasible [Booth and Hobert, 1999]. Nevertheless, there are several approaches available that can be useful in overcoming this obstacle, such as the Monte Carlo Newton-Raphson (MCNR) method [Lange, 1995], the Simulated Maximum Likelihood (SML) approach [Geyer and Thompson, 1992], the Stochastic Approximation Expectation-Maximization (SAEM) [Delyon et al., 1999], or the Monte Carlo EM (MCEM) [Wei and Tanner, 1990], to name a few. We opt to use the last one since simulation studies suggest that it performs at least as well as the other approaches for a variety of models [McCulloch, 1997; Booth and Hobert, 1999].

An EM algorithm is a natural method to use for this type of problem because the loglikelihood of the complete data (2) is considerably simpler than the observed data likelihood (3). In this case the E-step entails calculating the conditional expectation over $\mathbf{u}$ of the complete data loglikelihood given the observed $\mathbf{y}$. It is this step that presents the primary computational challenge of this problem. Our approach combines MCMC and importance sampling techniques and is detailed in the Appendix.

## CONVERGENCE OF THE ALGORITHM

In order to study the convergence properties of our algorithm we perform the following study. We created an artificial super-family of 208 individuals by concatenating 52 nuclear families. Note that the fact that not all the individuals are related to each other should not affect the behavior of the MCEM algorithm, except that the matrix $\Phi$ will simply have a block diagonal form. However, using the independence across families allows us to compute the exact likelihood curve for this super-family and obtain the true value of the MLE's for the model parameters. We analyzed this super family assuming a model with no covariates and only with additive genetic variance. The left graph in Figure 1 displays the likelihood curve around the vicinity of the MLE of the $\sigma_g$, 1.42, marked by the solid gray line. We can see that the likelihood seems to be well behaved in this region.

To gauge the effect of the choice of the starting point on the convergence of the algorithm, we maximized the likelihood 250 times each time randomly selecting a staring point in the interval from 0.5 to 6.5. We avoided starting points closer to 0, since from preliminary analyses such starting points seemed to result in an MCEM that takes too long to move away from that neighborhood. For all runs we used a burn-in period of 1,000 iterations before we switched to the importance sampling version. For the stopping rule we required that the criterion be met three consecutive times to avoid premature stopping of the algorithm. The resulting MCEM estimates from those runs are summarized in the middle and right graphs of Figure 1. The solid gray line corresponds to the true value of the MLE, while the black dashed line on the third graph marks the 1,000 iteration where the algorithm switches to the importance sampling. As we can see from these graphs, the estimates of the MLE from these 250 runs were very close to the true MLE demonstrating that, as long as the starting point is not too far away from the neighborhood of the true MLE, the algorithm will converge to the right value with high accuracy with a mean relative error from the true MLE of 0.04. Furthermore, we can see that typically, regardless of the starting point, we need about 500 iterations for the MCEM to bring us in the neighborhood of the MLE. Finally, note that our model had only one parameter, the genetic variance component. As a result, the likelihood function was very sharp in the vicinity of the MLE, allowing the algorithm to converge quickly and with high accuracy. Due to the limited information inherent in binary data, adding more parameters (covariates) in the model would likely result in a flatter likelihood function [McCulloch, 1997; Sung and Geyer, 2007]. In such a case, the MCEM algorithm would require larger number of MC realizations to converge to the true MLE with the same accuracy, thereby resulting in longer runs of the algorithm.

## A SIMULATION STUDY

To explore the properties of our method we performed a simulation study. We considered three different types of multi-generational pedigrees, all modeled after specific sub-branches of the Hutterite pedigree [Abney et al., 2000; Ober et al., 2001]. The first two pedigrees resembled a five and seven generation family with 220 and 382 members, respectively. The third pedigree, of 491 individuals, had exactly the same covariance structure as the Hutterite subpedigree we analyze below and consisted of the four most recent generations of a 1,623 member pedigree founded by 64 individuals of European descent (see description below). The genetic model we considered included two fixed effects, intercept ($\beta_0$) and gender ($\beta_1$), and a genetic additive random effect ($\sigma_g$). The values of the parameters $\beta_0$, $\beta_1$, and $\sigma_g$, of the simulation model were set to $-1$, $-0.25$, and 4, respectively, and they were chosen in such a way as to yield an average disease prevalence in the family close to 12%, while making sure that the model would not generate pedigrees with too few affected individuals that could potentially cause problems in the algorithm (Table I).

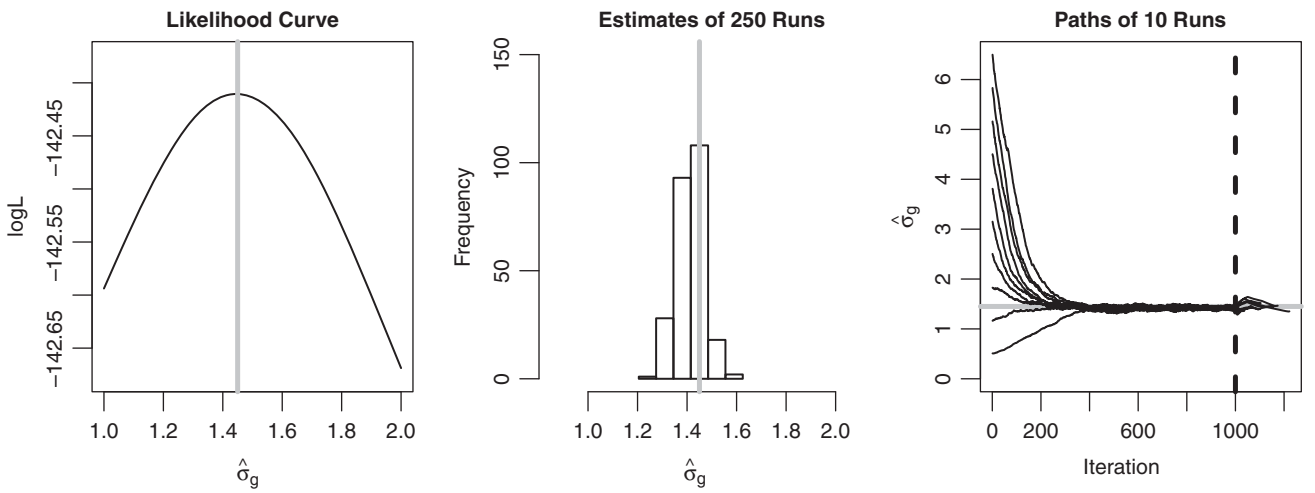


**Fig. 1. Convergence of the MCEM algorithm: The graph to the left displays the likelihood curve in the vicinity of the MLE, vertical solid grey line, of the genetic variance component ($\sigma_g$). The graph in the middle displays the observed distribution of the resulting MLE values of 250 runs of the MCEM. The vertical solid grey line marks the true value of the MLE. The graph to the right displays the paths of a sample of 10 runs for varying starting points in the interval [0.5, 6.5]. The dashed vertical line marks the end of the burn-in period, while the solid horizontal grey line corresponds to the true value of the MLE.**

**TABLE I. Average number of affected individuals in the pedigrees under the simulation model (based on 10,000 simulated families)**

| Pedigree size | Number of affected members (%) | |
|---|---|---|
| | Mean | SD |
| 220 | 25.7 (11.7) | 11.0 (5.0) |
| 382 | 46.5 (12.2) | 16.8 (4.4) |
| 491 | 60.4 (12.3) | 27.9 (5.7) |

The values of the parameters $\beta_0$, $\beta_1$, and $\sigma_g$ of the simulation model were set to $-1$, $-0.25$, and 4, respectively.

For each pedigree type we generated phenotypes for 200 replicates and analyzed them according to the following scheme. First, values for the underlying quantitative risk for all family members were randomly generated from a multivariate normal with mean zero and covariance matrix given in (1). Then, for each member of the family we computed the person-specific probability of being affected by the trait using the logit function and the information on the covariates. These probabilities were then used to determine the affection statuses of the family members through conditionally independent Bernoulli trials. For the MCEM computations, we set the number of initial burn-in iterations $B$ to 500. The stopping criterion used was the one described in the Appendix, where we required that the criterion be met for three consecutive times before convergence was declared. Finally, in order to moderate the effect of potential local maxima, we decided to run the MCEM algorithm five times, each time using a different starting point, and take as MLE's for each model parameter the median of the resulting values from these five runs.

The results from the simulations are summarized in Table II. We can see that when the family size was 220 individuals, the estimates of the parameters showed some bias. In particular, the estimates for the random polygenic effect seemed to be slightly positively biased. However, as the family size increased, the bias was significantly reduced. This is not entirely surprising. Biased estimates, especially for the random effects, may be expected in the GLMM context [Moreno et al., 1997; Burton et al., 1999; Noh et al., 2006; Yun and Lee, 2004]. Bias can be particularly pronounced in situations in which the trait prevalence is less than 15%, as it is in this case [Yi and Xu, 1999; Stock et al., 2007]. Furthermore, restricting the genetic variance to be positive was also expected to positively bias the MLEs [Burton et al., 1999]. Including more individuals in the analysis should, in principal, help moderate the bias on the estimates, whether this bias comes from low prevalence or from the constraint that the parameter has to be positive, however, at the cost of greater computational intensity.

The last column of Table II provides the observed coverage of the 95% confidence intervals based on the asymptotic normality of the MLEs. We can see that the coverage for the genetic variance component was lower than the nominal one. This is again not surprising. Low coverage probabilities of the confidence intervals for the variance components have also been observed before in the MCEM context [Burton et al., 1999]. The very low coverage probability may indicate issues with the starting points leading to local maxima. Such problems are not unusual in GLMMs, where non-unimodal likelihood functions are observed quite often [McCulloch, 1997; Sung and Geyer, 2007]. Furthermore, the coverage probably might have also been affected by the fact that the estimates of the parameters, and especially those for the genetic component, are biased. An additional reason for this low coverage probability may also be that the distribution of the MLE's seemed to not have reached asymptotic normality. Indeed, as we can see from Figure 2, the observed distribution of the MLEs, and especially of that for the genetic variance, seem to be rather skewed even when the family comprised 491 individuals, implying that this size may not be sufficiently large for the central limit theorem to take effect. This could be the result of estimating the target function to be maximized using an MCMC sampling scheme, which can affect the validity of the asymptotic normality of the resulting MLEs [Sung and Geyer, 2007]. Moreover, we observed flatness in the likelihood function in the neighborhood of the MLE, suggesting asymptotic normality has not yet been achieved in the data set. To investigate the effect of the choice of the starting point on the bias of the estimates and the coverage of the asymptotic confidence intervals, we decided to increase the MCEM runs to 21, and again take the median estimates as MLEs. As we can see from Table III increasing the number of runs of the MCEM significantly reduced the bias. Even though the coverage probability was also significantly increased, it remained below the nominal value, indicating that the problem with the non-normality of the asymptotic distribution of the MLEs is probably still an issue.

**TABLE II. Simulation results for the MCEM with five runs per replicate based on 200 replicates**

| Size[a] | Parameter | True value[b] | Mean[c] | Median[c] | SD[c] | 95% CP[d] |
|---|---|---|---|---|---|---|
| 220 | $\beta_0$ | $-1.00$ | $-0.69$ | $-0.88$ | 3.08 | 0.975 |
| | $\beta_1$ | $-0.25$ | $-0.35$ | $-0.28$ | 0.29 | 0.935 |
| | $\sigma_g$ | 4.00 | 5.33 | 4.30 | 3.54 | 0.660 |
| 382 | $\beta_0$ | $-1.00$ | $-0.82$ | $-0.62$ | 2.65 | 0.955 |
| | $\beta_1$ | $-0.25$ | $-0.30$ | $-0.27$ | 0.22 | 0.840 |
| | $\sigma_g$ | 4.00 | 4.56 | 4.01 | 2.31 | 0.525 |
| 491 | $\beta_0$ | $-1.00$ | $-1.18$ | $-0.90$ | 1.80 | 0.920 |
| | $\beta_1$ | $-0.25$ | $-0.27$ | $-0.26$ | 0.11 | 0.880 |
| | $\sigma_g$ | 4.00 | 4.32 | 4.14 | 1.69 | 0.555 |

For each replicate the MLE values were taken to be the median values of the resulting MLEs from the five runs.
[a]Number of family members in the pedigree.
[b]Target value of parameter under simulation model.
[c]Observed mean, median, and SD of the resulting MLEs.
[d]Observed coverage probability defined as the proportion of the CIs that captured the target value.

## TYPE 2 DIABETES IN THE HUTTERITES

The Hutterites are a religious isolate of more than 40,000 members living on approximately 400 communal farms in the northern United States and western Canada. Due to the small number of founders and their communal lifestyle, the Hutterites have been the focus of genetic studies for over 50 years [Ober et al., 2001, 2008, 2009; Steinberg et al., 1967; Hostetler, 1974]. Here we consider 491 Hutterites of age 15 years or older who were evaluated for T2D between 1996 and 1997. Information on pertinent covariates, such

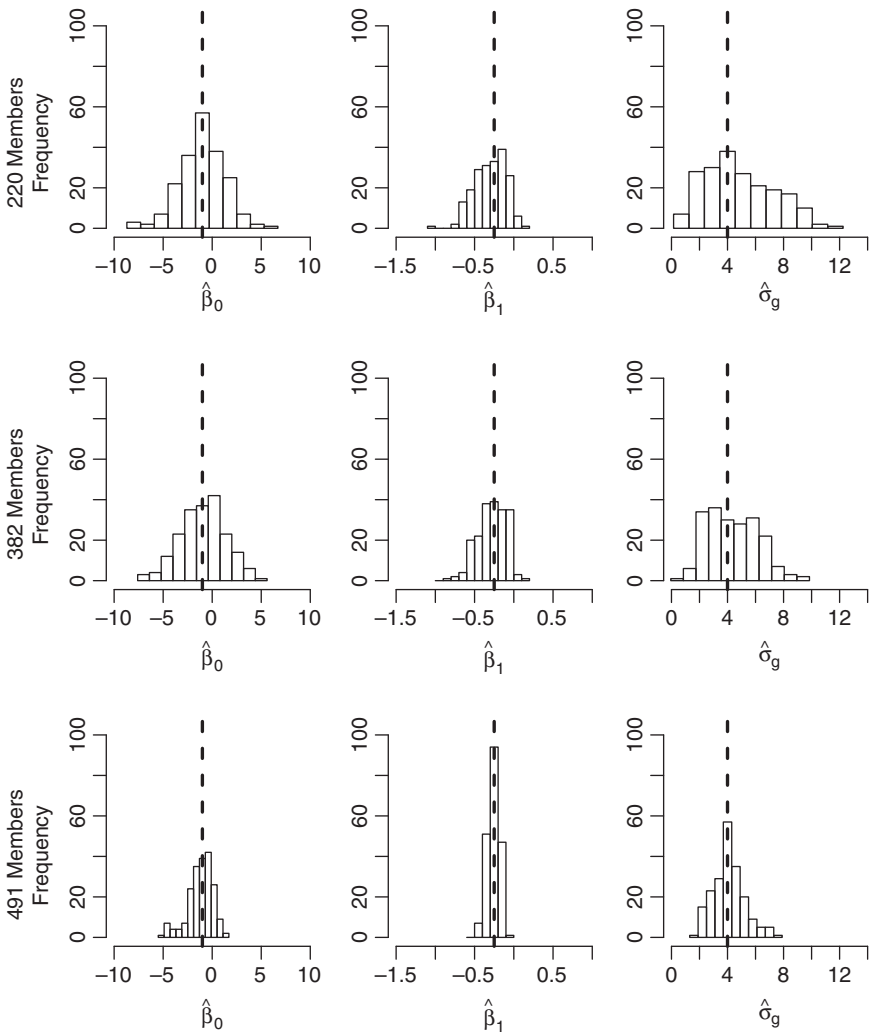


**Fig. 2. Distributions of the MLEs from 200 replicates with five runs per replicate. The vertical dashed lines represent the true value of the parameter under the simulation model.**

**TABLE III. Simulation results for the MCEM with 21 runs per replicate based on 200 replicates**

| Size[a] | Parameter | True value[b] | Mean[c] | Median[c] | SD[c] | 95% CP[d] |
|---|---|---|---|---|---|---|
| 220 | $\beta_0$ | −1.00 | −1.03 | −1.10 | 2.33 | 0.980 |
| | $\beta_1$ | −0.25 | −0.30 | −0.28 | 0.20 | 0.955 |
| | $\sigma_g$ | 4.00 | 4.90 | 4.48 | 2.41 | 0.795 |
| 382 | $\beta_0$ | −1.00 | −0.97 | −0.72 | 2.31 | 0.975 |
| | $\beta_1$ | −0.25 | −0.28 | −0.27 | 0.18 | 0.915 |
| | $\sigma_g$ | 4.00 | 4.39 | 4.14 | 1.79 | 0.650 |
| 491 | $\beta_0$ | −1.00 | −1.21 | −1.04 | 1.26 | 0.980 |
| | $\beta_1$ | −0.25 | −0.26 | −0.26 | 0.07 | 0.970 |
| | $\sigma_g$ | 4.00 | 4.07 | 4.01 | 1.11 | 0.745 |

For each replicate the MLE values were taken to be the median values of the resulting MLEs from the 21 runs.
[a]Number of family members in the pedigree.
[b]Target value of parameter under simulation model.
[c]Observed mean, median, and SD of the resulting MLEs.
[d]Observed coverage probability defined as the proportion of the CIs that captured the target value.

as gender, age, and BMI, were available for all subjects. All individuals in our study are related to each other through a complex pedigree of 1,623 individuals that traces back to 64 ancestors who were born between the early 1700s and the early 1800s in Europe [Abney et al., 2000]. For more information on the history and the relationship of the pedigree members, the reader is referred to Ober et al. [2001] or Steinberg et al. [1967].

A total of 36 individuals were clinically diagnosed with T2D based on their fasting glucose levels, according to the American Diabetes Association criteria [2010], yielding an overall prevalence of 7.5% (Table IV). There were slightly more women with T2D than men, 20 (8%) versus 16 (6%). Hutterites with T2D were significantly older than those without T2D (average age of 60.1 versus 33.6 years, respectively), and more overweight (mean BMI = 31.1 and 25.8, respectively). Only seven individuals with T2D were younger than 50 years at the time of study. Because of the highly polarized distributions of age and BMI, and possible differences in gender between the Hutterites with and without T2D, we fit several models that included

different combinations of these covariates: (a) Model I included gender and age, (b) Model II included gender, age, and BMI, (c) Model III included only age, (d) Model IV included age and BMI, and (e) Model V included age, BMI, and the interaction between age and BMI (Age × BMI). For the MCEM computations, we used the same values for all tuning parameters as in the simulations. That is, we set $B$ to 500 and we required that the stopping criterion be met for three consecutive times before convergence was declared. Based on the simulation results we decided to run the MCEM algorithm 21 times using different starting points and took as MLE's for the model parameters the medians of the resulting values from these runs.

Table V displays the MLE estimates for the parameters for all five models we considered as well as their estimated standard errors. The results suggest that gender is not an important risk factor for T2D in the Hutterites. Models II and IV suggest that the other two covariates are significant risk factors for T2D, as expected [Narayan et al., 2007; Bays et al., 2007], while Model V further indicates that age and BMI are independent because the interaction term was not statistically significant. Finally, all models clearly indicate a significant genetic component to the trait in the Hutterites, as for all models the estimate of the genetic variance component was statistically significant.

Interpretation of the coefficients of the fixed effects in our model is similar to that of the coefficients of the fixed effects in a logistic model and they are subject-specific and not population specific [Burton et al., 1999]. The genetic variance component can be interpreted in a similar fashion as in the case of variance components of a continuous trait but on the log scale [Burton et al., 1999]. Traditionally, in the context of continuous traits one gauges the contribution of genetics using a measure such as broad or narrow heritability, which are ratios of functions of the genetic variance components, over the total variance at the trait (sum of the genetic and the residual – environmental – variance). In our case estimation of heritabilities is hindered because our model does not include an explicit environmental variance component. One might consider using the binomial residual variance, appropriately transformed on the log scale, as an estimate of the environmental effect. However, this is not appropriate because the binomial residual variance is not analogous to the environmental variance in the continuous traits and as such it can lead to misleading conclusions [Burton et al., 1999].

Instead, in Figure 3 we plot the population prevalence of the trait and the recurrence probability (risk) for siblings as a function of the age of a person. We only considered models III and IV (Table V) because neither sex nor the Age × BMI interaction term was a significant predictor of risk. For the sibling recurrence risk we computed the probability of a person having T2D at a certain age, given that a sibling had T2D at the same age. To compute the risks under each model we worked as follows. First, using the MLEs of the parameters and for a given age, we computed the marginal probability of a single person being affected (i.e. the prevalence) using equation (3). This was an easy task because the data consist of a single individual resulting in a one-dimensional integral. In a similar fashion, we used the same equation to compute the joint probability of two siblings being affected by setting their phenotypes to $y_1 = y_2 = 1$ in Equation (3) and using the same age value for both siblings. This was also easily computed as it involved a two-dimensional integral and computations were feasible using standard software packages. Finally, we divided the joint probability by the marginal probability to obtain the sibling recurrence risk. For the models that included BMI as a covariate, we computed the risks assuming the same BMI for both siblings. We selected two different BMIs that corresponded to the average BMI observed among Hutterites with T2D (BMI = 31, which is considered obese) and among those without T2D (BMI = 26, which is considered normal) at the time of examination.

From Figure 3A we can see that all models suggest a significant increase in the risk for T2D with increasing age, as expected [Bays et al., 2007]. Consistent with our observations, the risk of T2D is low for Hutterites younger than age 50 years and increasing sharply thereafter. Obesity is also a significant risk factor, as previously reported [Narayan et al., 2007; Bays et al., 2007]. A BMI of 31 increases one's risk for T2D between 30 and 100% relative to a person of the same age and with a BMI of 26. For instance, based on the model that included BMI (Model IV), a 50-year-old person with a BMI of 26 has a 5% chance of having T2D, while a 50-year-old person with a BMI of 31 has twice this risk (10%) of having T2D. Similarly, two individuals who are 60 years old have 24 and 32% chances of having T2D, depending on whether

**TABLE IV. Numerical summaries of the features of the Hutterite Pedigree**

| | *N* | No. with diabetes | Age | BMI |
|---|---|---|---|---|
| Total | 491 | 36 (7.3%) | 35.6 (16.0) | 26.2 (5.6) |
| Male | 267 | 16 (6.0%) | 35.3 (16.0) | 25.6 (5.6) |
| Female | 224 | 20 (8.0%) | 35.9 (16.2) | 26.2 (5.5) |
| Diabetics | 36 | – | 60.1 (10.8) | 31.1 (4.7) |
| Nondiabetics | 455 | – | 33.6 (14.7) | 25.8 (5.5) |
| Age <50 years | 394 | 7 (1.8%) | 29.2 (9.8) | 25.3 (5.4) |
| Age ≥0 years | 97 | 29 (30.0%) | 61.4 (8.9) | 29.9 (5.1) |

**TABLE V. MLE estimates of the model parameters and their estimated standard error for the four models considered**

| | Fixed effects | | | | | Genetic effect |
|---|---|---|---|---|---|---|
| Model | Constant | Gender | Age | BMI | Age × BMI | $\sigma_g$ |
| I | −11.2 (2.6) | −0.98 (0.67) | 0.16 (0.04) | – | – | 2.42 (0.74) |
| II | −32.7 (8.8) | −2.33 (1.37) | 0.31 (0.08) | 0.34 (0.16) | – | 6.06 (1.98) |
| III | −12.2 (1.7) | – | 0.17 (0.03) | – | – | 2.58 (0.35) |
| IV | −19.9 (4.1) | – | 0.18 (0.04) | 0.21 (0.09) | – | 3.24 (0.77) |
| V | −14.9 (6.9) | – | 0.13 (0.12) | 0.11 (0.22) | 0.001 (0.004) | 2.45 (0.71) |

The estimates represent the median values from 21 runs of the MCEM algorithm.

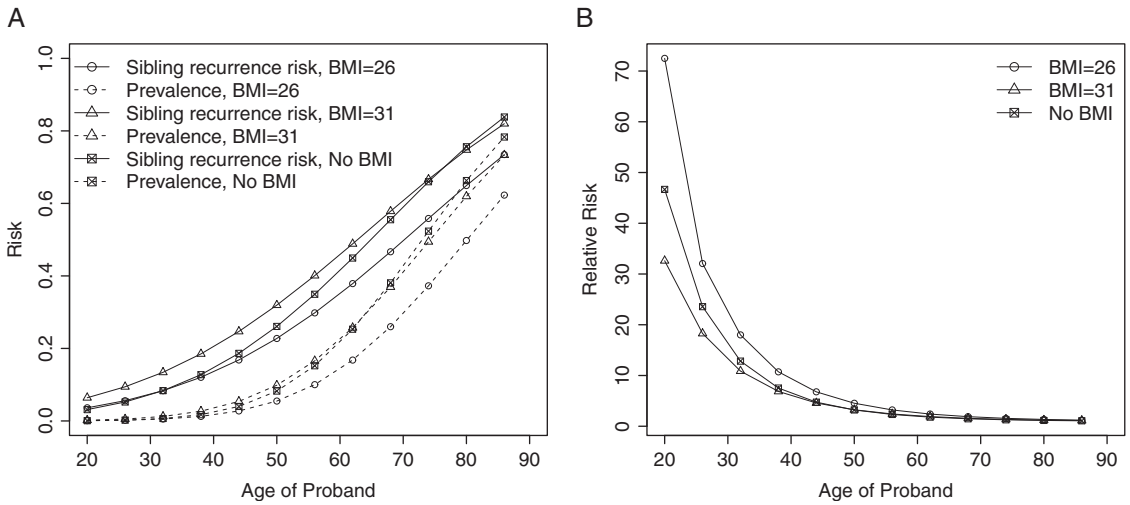


**Fig. 3. Estimated population and sibling risks for T2D (Graph A) and Sibling Relative Risk (Graph B) as functions of the age of the individual, based on the MLEs of Models III and IV in Table IV. The lines labeled "No BMI" correspond to model III in which BMI was not included as a covariate.**

their BMI is 26 or 31, respectively, marking a 33% increase in the risk. Thus, the absolute risk for T2D in the Hutterites increases with increasing age and increasing BMI.

We next specifically examined the contribution of family history (i.e., genetics) to T2D risk. Having a sibling with T2D increases the risk of T2D at all ages, regardless of BMI, in the Hutterites (Fig. 3B). However, the increased risk for T2D is highest when the affected sibling is young and with lower BMI. For example, having a sibling with T2D by the age of 40 years is associated with a five- to eight-fold increased risk for having T2D at the same age, with absolute risk increasing from 1.5–3.5% to 15–18.5%, depending on BMI. The fold increase in risk is less pronounced at older ages, dropping to around 1.5 at 70 years or older. This indicates a potentially strong genetic component for early onset T2D, with a lesser contribution of genetics to T2D risk at later ages. Similar to the age effect, there is a stronger genetic component to T2D risk at lower BMI. Thus, the highest relative risk for T2D in the Hutterites are for siblings of T2D cases who are young and of normal weight.

## DISCUSSION

We have described a likelihood method, developed in the context of the GLMMs [McCulloch, 1997], for modeling dichotomous traits using data from a large complex pedigree. Our formulation provides a flexible tool that can capitalize on information on pertinent covariates by treating them as fixed effects, while at the same time it takes into account shared genetic background among related individuals through random effects. In particular, under our formulation, the common genetic background is captured by the covariance matrix of the joint unobserved genetic components. Usually, the structure of the covariance matrix is easily computed based on the known relationships of the individuals in the sample. When the relationship between individuals are not completely known, as in most association studies, the structure of the matrix can be estimated from marker data, thus uncovering any cryptic relatedness in the sample [Astle and Balding, 2009]. Maximization of the likelihood function requires the computation of a multidimensional integral that is not feasible in analytical form. To overcome this hurdle we have devised an efficient automated MCEM algorithm.

Simulation results suggest that our method performs well, especially for sample sizes of more than 400 individuals. The estimates of the model parameters appear biased with respect to the target values, especially for the genetic variance parameters. This is not uncommon in the context of the GLMMs where biased estimates are expected due to the sparse nature of the data that often result in very flat likelihood functions or likelihoods that are non-unimodal [Sung and Geyer, 2007; Burton et al., 1999; Noh et al., 2006; McCulloch, 1997]. In order for the MCEM algorithm to distinguish between values of the parameters with similar likelihood values, one would need prohibitively large Monte Carlo sample sizes that would render computations cumbersome. This problem can be partially mitigated by running the MCEM multiple times with different starting points and setting as MLEs the means or medians of the MLE values from these runs. In addition, use of penalized likelihood functions may further reduce the magnitude of the bias [Noh et al., 2006; Yun and Lee, 2004]. Alternatively, one could consider using liability threshold models, such as the one implemented in SOLAR [Blangero and Almasy, 1996]. Even though such models may also potentially result in biased estimates in some cases [Moreno et al., 1997], depending on the quality of the information on the fixed effects, they usually tend to provide essentially unbiased estimates of the genetic parameters [Williams et al., 1999; Duggirala et al., 1997]. Notice though that the interpretation of the genetic variance parameters of the liability model is different than that of the variance parameters in our modeling, since the liability-threshold model also includes an environmental variance component.

In order to ease the computational burden associated with the Monte Carlo part of the MCEM, we implemented an importance sampling scheme since it can significantly reduce the workload. Although importance sampling techniques work well when the MCEM is very close to the neighborhood of the MLEs, they may drift when this is

not the case. To minimize this risk, we start our algorithm with a number of regular MCEM burn-in iterations to move in the vicinity of the MLE before switching to the IS version. A burn-in iteration period of around 500 seemed to have worked well in all the scenarios we examined. Another factor that seemed to have some effect on the convergence of the algorithm was the number of MCMC realizations used to estimate the target function for the EM algorithm in the burn-in period. We found that a sample size of 500 MCMC realizations were sufficient for the algorithm to behave appropriately, while smaller sample sizes tended to move the algorithm close to the boundary where the MCEM was eventually trapped.

As in Burton et al. [1999], in our simulations we observed a low coverage probability of the asymptotic confidence intervals for the true value of the genetic variance component. The low coverage could be partially attributed to the bias introduced in the parameter estimates because the random component parameter is constrained to be non-negative. An additional explanation may rest on the asymptotic distribution of the MLEs that might not have reached normality either due to flatness of the likelihood in the neighborhood of the MLE, or due to the use of MCMC techniques for generating observations used to estimate the target function in the MCEM [Sung and Geyer, 2007]. Indeed, our simulations seem to corroborate the latter scenario, as the empirical distribution of the MLEs of the random effects displayed a clear non-normal pattern even when the family had almost 500 members. Nevertheless, the confidence intervals for the fixed effects seemed to have maintained the correct nominal coverage probability even for family sizes as small as 220 individuals.

We implemented our method using data for 491 Hutterites who were evaluated for T2D. We fit five models that included different combinations of relevant covariates such as BMI, age, and gender of the participants. Our analyses indicated that gender is not a significant factor for T2D in this population. On the other hand, similar to other studies [Narayan et al., 2007; Bays et al., 2007], BMI and age are significant and independent risk factors for T2D. Overall, irrespective of the number and type of covariates used in the models we considered, our results suggest a strong genetic component contributing to risk, particularly among younger and leaner cases. In contrast, genetics appears to play a less important role in the development of diabetes among older and more obese individuals, in which case the disease is more likely due to environmental factors (i.e., such as the obesity itself).

Our current approach assumes an additive genetic model. This does not pose significant limitations to the method, as in most situations the ability to uncover non-additive effects is limited [Pawitan et al., 2004]. Nevertheless, our model can be easily modified to include additional genetic components such as dominance. From a theoretical point of view this is easily achieved. However, including more genetic components can potentially lead to problems with the convergence of the MCEM. Due to the sparseness of the binary data that often result in flat likelihoods, there is a great risk that the MCEM will drive the estimates of the fixed and the random effects to extreme opposites, increasing the bias in the estimates and resulting in singular fisher information [Sung and Geyer, 2007]. Instead, a penalized likelihood approach might be more preferable in that case.

Finally, a great advantage of our method is that it can serve as a framework for developing mapping techniques for loci contributing to complex traits. This can be easily accomplished in our method by including the genotype as a covariate in the model. One can then maximize the likelihood, obtain the MLE of the parameter corresponding to the locus of interest, and use this MLE to test the effect of the marker on the trait. Even though this is a simple approach, its practical utility is limited to candidate gene studies where the number of markers to be tested is small, thereby allowing for the analysis to be performed in a reasonable time frame. However, the computational intensity of the algorithm (typically around 4 hr for a pedigree of size 491 members on a single core E5530, 2.4 GHz, Dell Precision T7500 personal computer) prohibits genome-wide scans since they involve a very large number of markers that need to be tested. Instead, one can use an efficient scoring function as in Abney et al. [2000] that can significantly alleviate the computational burden of the method, thus making GWAS more amenable. The properties of this approach will be explored in future work.

## SOFTWARE

The software package BVC implementing the methods described here is freely available at http://code.google.com/p/papachristou-free-genetics-software/downloads/list.

## REFERENCES

Abney M, McPeek MS, Ober C. 2000. Estimation of variance components of quantitative traits in inbred populations. Am J Hum Genet 66:629–650.

American Diabetes Association. 2010. Risk factors for the development of diabetes mellitus. Diabetes Care 33:1S62–1S69.

Astle W, Balding D. 2009. Population structure and cryptic relatedness in genetic association studies. Stat Sci 24:451–471.

Bays HE, Bazata DD, Clark NG, Gavin JR, Green AJ, Lewis SJ, Reed ML, Stewart W, Chapman RH, Fox KM, Grandy S. 2007. Prevalence of self-reported diagnosis of diabetes mellitus and associated risk factors in a national survey in the us population: Shield (study to help improve early evaluation and management of risk factors leading to diabetes). BMC Public Health 7:277.

Blangero J, Almasy LA. 1996. SOLAR: Sequential Oligogenic Linkage Analysis Routines. Technical Notes No. 6 Population Genetics Laboratory, Southwest Foundation for Biomedical Research, San Antonio, TX.

Booth J, Hobert J. 1999. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. J R Stat Soc Ser B Stat Methodol 61:265–285.

Burton PR, Tiller KJ, Gurrin LC, Cookson WO, Musk AW, Palmer LJ. 1999. Genetic variance components analysis for binary phenotypes using generalized linear mixed models (glmms) and Gibbs sampling. Genet Epidemiol 17:118–140.

Delyon B, Lavielle M, Moulines E. 1999. Convergence of a stochastic approximation version of the EM algorithm. Ann Stat 27:94–128.

Duggirala R, Williams JT, Williams-Blangero S, Blangero J. 1997. A variance component approach to dichotomous trait linkage analysis using a threshold model. Genet Epidemiol 14:987–992.

Geyer CJ, Thompson EA. 1992. Constrained Monte Carlo maximum likelihood for dependent data. J Roy Stat Soc Ser B 54:657–699.

Hostetler JA. 1974. Hutterite Society. Baltimore: Johns Hopkins University Press.

Jaquard A. 1974. The Genetic Structure of Populations. New York: Springer.

Lange K. 1995. A gradient algorithm locally equivalent to the em algorithm. J Roy Stat Soc Ser B 57:425–437.

Levine R, Casella G. 2001. Implementations of the Monte Carlo EM algorithm. J Comput Graph Stat 10:422–439.
Levine R, Fan J. 2004. An automated (Markov Chain) Monte Carlo EM algorithm. J Stat Comput Simul 74:349–359.
McCullagh P, Nelder J. 1989. Generalized Linear Models. London: Chapman and Hall.
McCulloch CE. 1994. Maximum-likelihood variance-components estimation for binary data. J Am Stat Assoc 89:330–335.
McCulloch CE. 1997. Maximum likelihood algorithms for generalized linear mixed models. J Am Stat Assoc 92:162–170.
Moreno C, Sorensen D, Garcia-Cortes L, Varona L, Altarriba J. 1997. On biased inferences about variance components in the binary threshold model. Genet Select Evol 29:145–160.
Narayan K, Boyle J, Thompson T, Gregg E, Williamson D. 2007. Effect of bmi on lifetime risk for diabetes in the U.S. Diabetes Care 30: 1562–1566.
Noh M, Yip B, Lee Y, Pawitan Y. 2006. Multicomponent variance estimation for binary traits in family-based studies. Genet Epidemiol 30:37–47.
Ober C, Abney M, McPeek MS. 2001. The genetic dissection of complex traits in a founder population. Am J Hum Genet 69:1068–1079.
Ober C, Tan Z, Sun Y, Possick J, Pan L, Nicolae R, Radford S, Parry R, Heinzmann A, Deichmann K, Lester L, Gern J, Lemanske R, Nicolae D, Elias J, Chupp G. 2008. Variation in the chi3l1 gene influences serum ykl-40 levels, asthma risk, and lung function. NEJM 358:1682–1691.
Ober C, Nord A, Thompson E, Pan L, Tan Z, Cusanovich D, Sun Y, Nicolae R, Edelstein C, Schneider D, Billstrand C, Pfånger D, Phillips N, Anderson R, Philips B, Rajagopalan R, Hatsukami T, Rieder M, Heagerty P, Nickerson D, Abney M, Marcovina S, Jarvik G, Scanu A, Nicolae D. 2009. Genome-wide association study of plasma lp(a)levels identifies multiple genes on chromosome 6q. J Lipid Res 30:798–806.
Olson JM, Cordell HJ. 2000. Ascertainment bias in the estimation of sibling genetic risk parameters. Genet Epidemiol 18:217–235.
Pawitan Y, Reilly M, Nilsson E, Cnattingius S, Lichtenstein P. 2004. Estimation of genetic and environmental factors for binary traits using family data. Stat Med 23:449–466.
Risch N. 1990a. Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222–228.
Risch N. 1990b. Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 46:229–241.
Robert CP, Ryden T, Titterington DM. 1999. Convergence controls for MCMC algorithms, with applications to hidden Markov chains. J Stat Comput Simul 64:327–355.
Steinberg AG, Bleibtreu HK, Kurczynski TW, Martin AO, Kurczynski EM. 1967. Genetic studies in an inbred human isolate. Proceedings of the Third International Congress of Human Genetics. Baltimore: Johns Hopkins University Press. p 267–290.
Stock K, Distl O, Hoeschele I. 2007. Bayesian estimation of genetic parameters for multivariate threshold and continuous phenotypes and molecular genetic data in simulated horse populations using Gibbs sampling. BMC Genet 8:19.
Sung YJ, Geyer CJ. 2007. Monte Carlo likelihood inference for missing data models. Ann Stat 35:990–1011.
Wei GCG, Tanner MA. 1990. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. J Am Stat Assoc 85:699–704.
Williams JT, Van Eerdewegh P, Almasy L, Blangero J. 1999. Joint multipoint linkage analysis of multivariate qualitative and quantitative traits. I. Likelihood formulation and simulation results. Am J Hum Genet 65:1134–1147.
Yi N, Xu S. 1999. Mapping quantitative trait loci for complex binary traits in outbred populations. Heredity 82:668–676.
Yun S, Lee Y. 2004. Comparison of hierarchical and marginal likelihood estimators for binary outcomes. Comput Stat Data Anal 45:639–650.
Zou G, Zhao H. 2004. The estimation of sibling genetic risk parameters revisited. Genet Epidemiol 26:286–293.

# APPENDIX: MONTE CARLO EXPECTATION-MAXIMIZATION

The EM algorithm is a routine commonly employed for obtaining MLEs of parameters of interest in the presence of missing data. It iteratively alternates between an E-step and an M-step until convergence of the parameters is reached. The E-step entails the computation of the expected value of the complete data log-likelihood function with respect to the distribution of the missing data, conditional on the current estimates of the parameters and the observed data. On the M-step, the conditional expectation computed on the E-step is maximized to provide the new estimates of the parameters.

We can easily adapt the EM algorithm to our specific needs by noticing that our modeling resembles a missing data problem where the complete likelihood function is given by (2) and the missing data are simply the unobserved random effects. Under this scenario, the expectation in the E-step on the $(t+1)$th iteration has the following form

$$\begin{aligned} Q(\boldsymbol{\beta}, \boldsymbol{\sigma}^2 | \hat{\boldsymbol{\beta}}^t, \hat{\boldsymbol{\sigma}}^{2^t}) &= E_{\mathbf{u}}[\log L_c(\boldsymbol{\beta}, \boldsymbol{\sigma}^2; \mathbf{y}, \mathbf{u}) | \mathbf{y}, \hat{\boldsymbol{\beta}}^t, \hat{\boldsymbol{\sigma}}^{2^t}] \\ &= \int \left[ \sum_{i=1}^{n} \log B(p_i; \boldsymbol{\beta}, y_i, u_i) + \log(f(\mathbf{u}; \boldsymbol{\sigma}^2)) \right] \\ &\quad \times g(\mathbf{u} | \mathbf{y}, \hat{\boldsymbol{\beta}}^t, \hat{\boldsymbol{\sigma}}^{2^t}) d\mathbf{u}, \end{aligned} \tag{A1}$$

where $f(\mathbf{u}; \boldsymbol{\sigma}^2)$ is the multivariate normal density, $g(\mathbf{u} | \mathbf{y}, \hat{\boldsymbol{\beta}}^t, \hat{\boldsymbol{\sigma}}^{2^t})$ is the conditional distribution of $\mathbf{u}$ given the observed affection statuses $\mathbf{y}$ and the estimates $\hat{\boldsymbol{\beta}}^t$ and $\hat{\boldsymbol{\sigma}}^{2^t}$ of the parameters from the previous iteration. Obviously, the above expectation still involves the computation of a high dimension integral, and although its integrand has a much simpler form than that in (3), its computation is also intractable.

Wei and Tanner [1990] suggested substituting the $Q$ function in (4) with its estimate obtained from a random sample of $M$ realizations $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(M)}$ from the distribution $g(\mathbf{u} | \mathbf{y}, \hat{\boldsymbol{\beta}}^t, \hat{\boldsymbol{\sigma}}^{2^t})$, i.e.,

$$\begin{aligned} Q_M(\boldsymbol{\beta}, \boldsymbol{\sigma}^2 | \hat{\boldsymbol{\beta}}^t, \hat{\boldsymbol{\sigma}}^{2^t}) &= \frac{1}{M} \sum_{m=1}^{M} \log L_c(\boldsymbol{\beta}, \boldsymbol{\sigma}^2; \mathbf{y}, \mathbf{u}^{(m)}) \\ &= \frac{1}{M} \sum_{m=1}^{M} \left[ \sum_{i=1}^{n} \log B(p_i; \boldsymbol{\beta}, y_i, u_i^{(m)}) + \log f(\mathbf{u}^{(m)}; \boldsymbol{\sigma}^2) \right] \\ &= \frac{1}{M} \sum_{m=1}^{M} \left[ \sum_{i=1}^{n} \log B(p_i; \boldsymbol{\beta}, y_i, u_i^{(m)}) \right] \\ &\quad + \frac{1}{M} \sum_{m=1}^{M} \log f(\mathbf{u}^{(m)}; \boldsymbol{\sigma}^2) \\ &= Q_{\boldsymbol{\beta}, M}(\boldsymbol{\beta} | \hat{\boldsymbol{\beta}}^t) + Q_{\boldsymbol{\sigma}^2, M}(\boldsymbol{\sigma}^2 | \hat{\boldsymbol{\sigma}}^{2^t}). \end{aligned} \tag{A2}$$

Due to the use of the random sample, it is not guaranteed that $Q_M$ will increase from iteration to iteration, but, under some mild regularity conditions,

the MCEM will converge to the true MLEs [Booth and Hobert, 1999].

Finally, it is worth mentioning that in (5), $Q_{\boldsymbol{\beta},M}$ involves only the fixed effect parameters $\boldsymbol{\beta}$, while $Q_{\boldsymbol{\sigma}^2,M}$ is a function of only the parameters $\boldsymbol{\sigma}^2$ of the random genetic effects. Thus, the task of maximizing $Q_M(\boldsymbol{\beta}, \boldsymbol{\sigma}^2 | \hat{\boldsymbol{\beta}}^t, \hat{\boldsymbol{\sigma}}^{2^t})$ reduces to that of separately maximizing $Q_{\boldsymbol{\beta},M}$ and $Q_{\boldsymbol{\sigma}^2,M}$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$, respectively. But, $Q_{\boldsymbol{\beta},M}$ is simply the log-likelihood function of a generalized linear model (GLM) whose maximization can be easily done using iterative weighted least squares [McCullagh and Nelder, 1989]. In addition, $Q_{\boldsymbol{\sigma}^2,M}$ is the log-likelihood function of a multivariate normal and its maximization may be achieved quickly through a standard maximization routine (e.g. Newton-Raphson), as long as the number of variance components is small.

# A METROPOLIS-HASTINGS (MH) ALGORITHM

The ability of the MCEM to quickly converge depends heavily on our ability to efficiently sample realizations from $g(\mathbf{u}|\mathbf{y}, \hat{\boldsymbol{\beta}}^t, \hat{\boldsymbol{\sigma}}^{2^t})$, that is the conditional distribution of the random effects $u$ given the current estimates of the parameters, $\hat{\boldsymbol{\beta}}^t$ and $\hat{\boldsymbol{\sigma}}^{2^t}$, and the affection statuses of the individuals in the pedigree, $\mathbf{y}$. The density of this distribution has the form

$$g(\mathbf{u}|\mathbf{y}, \hat{\boldsymbol{\beta}}^t, \hat{\boldsymbol{\sigma}}^{2^t}) = \frac{\prod_{i=1}^{n} B(p_i; \boldsymbol{\beta}, y_i, u_i) f(\mathbf{u}; \boldsymbol{\sigma}^2)}{\int \prod_{i=1}^{n} B(p_i; \boldsymbol{\beta}, y_i, u_i) f(\mathbf{u}; \boldsymbol{\sigma}^2) d\mathbf{u}}, \qquad \text{(A3)}$$

and obviously involves the likelihood function of the observed data, that we are trying to avoid computing. Thus, sampling directly from this distribution is not feasible. However, it is possible to obtain a sample either through a multivariate rejection sampling [Booth and Hobert, 1999], or Gibbs sampling [McCulloch, 1994], or a Metropolis-Hastings Markov Chain Monte Carlo (MCMC) sampler [McCulloch, 1997].

Due to the very low acceptance rate of the rejection sampler observed in our applications, we employ an efficient MH sampler. Generally, the MH algorithm requires the specification of a candidate distribution $e(\mathbf{u})$ from which a potential new observation $\mathbf{u}^*$ is drawn, and a function $a(\mathbf{u}, \mathbf{u}^*)$ that will provide the probability of accepting the proposed value, $\mathbf{u}^*$, over the current one, $\mathbf{u}$, as a possible realization from the target distribution.

The MH version that we implement is a component-wise sampler, where sequentially one by one we update all random effects to obtain a new realization from the target distribution [McCulloch, 1997]. Assuming that $\mathbf{u}$ was the previous draw from $g(\mathbf{u}|\mathbf{y}, \hat{\boldsymbol{\beta}}^t, \hat{\boldsymbol{\sigma}}^{2^t})$, then, we obtain the next sample value $\mathbf{u}^*$ by performing an $n$-step scan. At the $i$th step of the scan, we consider updating only the component $u_i$ by proposing a new value $u_i^*$, while the rest of the $u_i$'s, denoted by $\mathbf{u}_{-i}$, remain unchanged. For each step $i$, we choose $e(\mathbf{u})$ to be $f_{u_i|(\mathbf{u}_{-i}, \hat{\boldsymbol{\sigma}}^{2^t})}$, the conditional distribution of the $u_i$ given the $\mathbf{u}_{-i}$ and the current estimates of the variance components parameters $\hat{\boldsymbol{\sigma}}^{2^t}$.

There are two major advantages to this choice of $e(\mathbf{u})$. First, sampling from this univariate distribution is extremely easy because it is normal with mean $-\mathbf{u}_{-i}^{\mathrm{T}} W_{-i,i} W_{i,i}^{-1}$ and variance $W_{i,i}^{-1}$, where $W_{ij}$ is the $ij$th block of the inverse of the current estimate of the matrix $\boldsymbol{\Omega}$. Second, the acceptance probability takes the simple form

$$a(\mathbf{u}, \mathbf{u}^*) = \min\left\{1, \frac{g(\mathbf{u}^*|\mathbf{y}, \hat{\boldsymbol{\beta}}^t, \hat{\boldsymbol{\sigma}^2}^t) e(\mathbf{u})}{g(\mathbf{u}|\mathbf{y}, \hat{\boldsymbol{\beta}}^t, \sigma^{2^t}) e(\mathbf{u}^*)}\right\}$$
$$= \min\left\{1, \frac{B(p_i^*; \hat{\boldsymbol{\beta}}^t, y_i, u_i^*)}{B(p_i; \hat{\boldsymbol{\beta}}^t, y_i, u_i)}\right\}, \qquad \text{(A4)}$$

that is, the ratio of two univariate Bernoulli distributions. Hence, this sampling scheme allows us to quickly and efficiently generate large numbers of realizations for the target distributions $g(\mathbf{u}|\mathbf{y}, \hat{\boldsymbol{\beta}}^t, \hat{\boldsymbol{\sigma}}^{2^t})$, thereby significantly reducing the time needed to obtain the estimates of the MLEs of the parameters of interest.

## AN AUTOMATED MCEM

The use of the random sample to approximate the expectation on the E-step of the EM algorithm introduces a random error in the estimation of the model parameters. Due to this error, holding the MC sample size $M$ fixed through all iterations would not lead to convergence, unless it is significantly large causing the random error to become negligible [Booth and Hobert, 1999]. However, there is a trade-off between accurately estimating the parameters of interest and the required time for generating realizations from the target distribution. In general, it is wasteful and unnecessary to use a large $M$ from the beginning when the MCEM steps are relatively large and, thus, larger random errors can be tolerated [Booth and Hobert, 1999]. On the other hand, it is inefficient to use small sample sizes when we are in the vicinity of the MLEs and greater precision is needed. Typically, one addresses this issue by gradually increasing the number of MC sample as the MCEM routine progresses. For example, one can choose to let the sample linearly increase with the number of iterations [McCulloch, 1994], or even choose a fixed set of iteration $r_1, r_2, \ldots$, where the MC sample size will be increased to predetermined fixed values $M_1, M_2, \ldots$, [McCulloch, 1994]. Nevertheless, these strategies are not very efficient since they can potentially lead to many wasted MCEM iterations [Booth and Hobert, 1999].

Booth and Hobert [1999] proposed a more systematic approach that allows the algorithm to assess the adequacy of the current MC sample size at each iteration and to dynamically adjust it, should there is a need for it. The advantage of this approach is that early iterations use small values of $M$, thus saving computational cost, while late iterations use larger values of $M$, resulting in higher accuracy. For their approach, they assume the availability of an independent random sample from the distribution of interest. Based on this sample, they derive an estimate of the MC error at each iteration $t$ and, assuming asymptotic normality, they use it to construct a confidence ellipsoid around the current estimates of the parameters $(\hat{\boldsymbol{\beta}}^t, \hat{\boldsymbol{\sigma}}^{2^t})$ for the true value of the parameters that we would obtain if we were able to perform the deterministic EM algorithm. Then, they perform one more iteration to get the new estimates of the parameters of interest $(\hat{\boldsymbol{\beta}}^{t+1}, \hat{\boldsymbol{\sigma}}^{2^{t+1}})$. If these new estimates fall within the limits of the confidence region constructed from the previous iteration, then the

MC error has swamped the EM step and an increase in $M$ is needed.

Recently, Levine and Fan [2004] extended the approach of [6] to allow for non-independent samples, such as those obtained by MCMC samplers. Their approach assumes that we have available a sample of $N$ dependent realizations $\mathbf{u}^1,\ldots,\mathbf{u}^N$ from the distribution of interest. Using a Poisson sampling scheme [Robert et al., 1999] we draw a sub-sample $\mathbf{u}^{s_1},\ldots,\mathbf{u}^{s_M}$, where $s_i = x_1 + \cdots + x_i$, $x_i - 1 \sim$ Poisson($\nu_i$), and $\nu_i = \nu i^b$, for some $\nu \geq 1$ and $b > 0$, $i = 1,\ldots, M$. Under this sub-sampling scheme, the authors showed that $(\hat{\boldsymbol{\beta}}^t, \hat{\boldsymbol{\sigma}}^{2^t})$, the MCEM estimates of the parameters based on the sub-sample, follow (approximately) normal distribution, that is,

$$\sqrt{M}((\hat{\boldsymbol{\beta}}^t, \hat{\boldsymbol{\sigma}}^{2^t}) - (\dot{\boldsymbol{\beta}}^t, \dot{\boldsymbol{\sigma}}^{2^t})) \xrightarrow{\text{approx.}} \text{MVN}(\mathbf{0}, \Sigma), \tag{A5}$$

where $(\dot{\boldsymbol{\beta}}^t, \dot{\boldsymbol{\sigma}}^{2^t})$ are the values of the parameters that would have resulted on the $t$th step of the deterministic EM algorithm, and $\boldsymbol{\Sigma}$ is an appropriate covariance matrix. Levine and Fan [2004] argued that a good approximation to this covariance matrix can be easily obtained from the sample itself as follows

$$\begin{aligned}\hat{\Sigma} \approx & \{Q_M^{(2)}(\dot{\boldsymbol{\beta}}^{t+1}, \dot{\boldsymbol{\sigma}}^{2^{t+1}} | \dot{\boldsymbol{\beta}}^t, \dot{\boldsymbol{\sigma}}^{2^t})\}^{-1} \\ & \times \left[\frac{1}{N}\sum_{m=1}^{N}\left\{\frac{\partial \log\, L_c(\boldsymbol{\beta}, \boldsymbol{\sigma}^2; \mathbf{y}, \mathbf{u}^m)}{\partial(\boldsymbol{\beta}, \boldsymbol{\sigma}^2)} - \hat{\mu}_N\right\}\right. \\ & \left.\times \left\{\frac{\partial \log\, L_c(\boldsymbol{\beta}, \boldsymbol{\sigma}^2; \mathbf{y}, \mathbf{u}^m)}{\partial(\boldsymbol{\beta}, \boldsymbol{\sigma}^2)} - \hat{\mu}_N\right\}^{\mathrm{T}}\right] \\ & \times \{Q_M^{(2)}(\dot{\boldsymbol{\beta}}^{t+1}, \dot{\boldsymbol{\sigma}}^{2^{t+}} | \dot{\boldsymbol{\beta}}^t, \dot{\boldsymbol{\sigma}}^{2^t})^{-1}\}|_{(\boldsymbol{\beta},\boldsymbol{\sigma}^2)=(\hat{\boldsymbol{\beta}}^{t+1}, \hat{\boldsymbol{\sigma}}^{2^{t+1}})}\end{aligned} \tag{A6}$$

where $(\dot{\boldsymbol{\beta}}^t, \dot{\boldsymbol{\sigma}}^{2^t})$ are the MCEM estimates based on all $N$ dependent draws,

$$Q_M^{(2)}(\boldsymbol{\beta}, \boldsymbol{\sigma}^2 | \boldsymbol{\beta}', \boldsymbol{\sigma}^{2'}) = \frac{1}{M}\sum_{m=1}^{M}\frac{\partial^2\, \log\, L_c(\boldsymbol{\beta}, \boldsymbol{\sigma}^2; Y, u^{s_m})}{\partial(\boldsymbol{\beta}, \boldsymbol{\sigma}^2)\partial(\boldsymbol{\beta}, \boldsymbol{\sigma}^2)^{\mathrm{T}}}, \tag{A7}$$

and

$$\hat{\mu}_N = \frac{1}{N}\sum_{m=1}^{N}\frac{\partial\, \log\, L_c(\boldsymbol{\beta}, \boldsymbol{\sigma}^2; Y, u^m)}{\partial(\boldsymbol{\beta}, \boldsymbol{\sigma}^2)}. \tag{A8}$$

Based on this result, the authors proposed an automated MCEM algorithm that allows not only for dynamically deciding whether an increase in the MC sample size is needed, but also how much this increase should be. More specifically, based on the current estimates of the parameters at the $t+1$ iteration, they construct an (1–α) confidence ellipsoid for the true value of the parameters on the $(t+1)$th iteration, had the deterministic EM algorithm been used. This is done by aggregating all the values $(\boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ that satisfy

$$M[(\hat{\boldsymbol{\beta}}^{t+1}, \hat{\boldsymbol{\sigma}}^{2^{t+1}}) - (\boldsymbol{\beta}, \boldsymbol{\sigma}^2)]^{\mathrm{T}}\Sigma^{-1}[(\hat{\boldsymbol{\beta}}^{t+1}, \hat{\boldsymbol{\sigma}}^{2^{t+1}}) - (\boldsymbol{\beta}, \boldsymbol{\sigma}^2)] \leq \chi^2_{d;1-\alpha} \tag{A9}$$

where $d$ is the number of model parameters and $\chi^2_{d;1-\alpha}$ is the (1–α) percentile of the $\chi^2$ distribution with $d$ degrees of freedom. An increase in the MCMC sample size is necessary every time the above ellipsoid includes the point defined by the values of the parameters on the previous iteration. The recommended new value for $N$, the overall number of depended realizations, is set to

$$N = \left(\frac{M}{a}\right)^{1+b} \tag{A10}$$

where $a = \{(1+b)/v\}^{1/(1+d)}$ and

$$M = \left\lceil \frac{\chi^2_{d;1-\alpha}}{[(\hat{\boldsymbol{\beta}}^{t+1}, \hat{\boldsymbol{\sigma}}^{2^{t+1}}) - (\boldsymbol{\beta}^t, \boldsymbol{\sigma}^{2^t})]^{\mathrm{T}}\Sigma^{-1}[(\hat{\boldsymbol{\beta}}^{t+1}, \hat{\boldsymbol{\sigma}}^{2^{t+1}}) - (\boldsymbol{\beta}^t, \boldsymbol{\sigma}^{2^t})]} \right\rceil, \tag{A11}$$

where $\lceil x \rceil$ is the smallest integer greater or equal to $x$. Following the suggestions of Levine and Fan [2004] and Booth and Hobert [1999] we also set $\alpha = 0.25$, $\nu = 1$ and $b = 0.5$ on all of our applications. Our simulation results indicate that these values seem to be working quite satisfactory in the context of our method.

## IMPORTANCE SAMPLING (IS)

Even though the component-wise MH sampler described earlier is very fast, it is still very likely that it will require a significant amount of time to generate the large samples necessary at the late stages of the MCEM algorithm, especially when the number of individuals $n$ is large. Moreover, it is inefficient to spend so much time in generating a large amount of data only to be used on a single iteration of the MCEM algorithm and then be discarded. Instead, we use importance sampling.

Booth and Hobert [1999] and Levine and Casella [2001] are among those who have explored the utility of IS in the context of the GLMMs. Implementation of their paradigm in our situation amounts to the following. First, we choose some fixed values for the model parameters, say $(\hat{\boldsymbol{\beta}}^0, \hat{\boldsymbol{\sigma}}^{2^0})$, and based on these values we generate a random sample $\mathbf{u}^{(1)},\ldots,\mathbf{u}^{(M)}$ from $g(\mathbf{u}|\mathbf{y}, \hat{\boldsymbol{\beta}}^0, \hat{\boldsymbol{\sigma}}^{2^0})$. This is done only once at the beginning of the MCEM algorithm. For each subsequent MCEM iteration, instead of generating a new sample from the $g(\mathbf{u}|\mathbf{y}, \hat{\boldsymbol{\beta}}^t, \hat{\boldsymbol{\sigma}}^{2^t})$, we continue to use the one obtained at the beginning. To make up for the fact that the sample we use does not come from the correct distribution, we substitute $Q_M(\boldsymbol{\beta}, \boldsymbol{\sigma}^2 | \hat{\boldsymbol{\beta}}^t, \hat{\boldsymbol{\sigma}}^{2^t})$ with its importance sampling estimate

$$\begin{aligned}& Q_M^{(IS)}(\boldsymbol{\beta}, \boldsymbol{\sigma}^2 | \hat{\boldsymbol{\beta}}^t, \hat{\boldsymbol{\sigma}}^{2^t}, \hat{\boldsymbol{\beta}}^0, \hat{\boldsymbol{\sigma}}^{2^0}) \\ & = \frac{1}{W}\sum_{m=1}^{M}\sum_{m=1}^{M} w_m^t\left[\sum_{i=1}^{n} B(p_i; \boldsymbol{\beta}, y_i, u_i^{(m)}) + \log(f(u^{(m)}; \boldsymbol{\sigma}^2))\right] \\ & = \frac{1}{W}\sum_{m=1}^{M} w_m^t\left[\sum_{i=1}^{n} B(p_i; \boldsymbol{\beta}, y_i, u_i^{(m)})\right] + \frac{1}{W}\sum_{m=1}^{M} w_m^t[\log(f(u^{(m)}; \boldsymbol{\sigma}^2))] \\ & = Q_{\boldsymbol{\beta},M}^{(IS)}(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}^t) + Q_{\boldsymbol{\sigma}^2,M}^{(IS)}(\boldsymbol{\sigma}^2|\hat{\boldsymbol{\sigma}}^{2^t}),\end{aligned} \tag{A12}$$

where $W = \sum_{m=1}^{M} w_m^t$, and the weights $w_m^t$ are given by

$$w_m^t = \frac{g(\mathbf{u}|\mathbf{y}, \hat{\boldsymbol{\beta}}^t, \hat{\boldsymbol{\sigma}}^{2^t})}{g(\mathbf{u}|\mathbf{y}, \hat{\boldsymbol{\beta}}^0, \hat{\boldsymbol{\sigma}}^{2^0})}. \tag{A13}$$

A key factor for the successful implementation of the IS is the choice of $(\hat{\boldsymbol{\beta}}^0, \hat{\boldsymbol{\sigma}}^{2^0})$. A good choice can result in a great deal of time savings. On the other hand, our experience showed that bad values can mean longer run-time than the simple MCEM, or even a poor convergence performance. To minimize these

risks, one needs to choose values for $(\hat{\boldsymbol{\beta}}^0, \hat{\boldsymbol{\sigma}}^{2^0})$ that are close to the true MLEs. Levine and Casella [2001] recommended the use of a burn-in period, that is, they suggested that we first run a few, say $K$, regular MCEM iterations before switching to the importance sampling version, with $(\hat{\boldsymbol{\beta}}^0, \hat{\boldsymbol{\sigma}}^{2^0})$ chosen to be the estimates at last iteration.

There is no universally optimal choice of the number of burn-in iterations $K$. Its value depends on how close to the true value of the MLE's we need to be and thus it is problem specific. Usually, a relatively small value of $K$, say 20, is sufficient since in most cases the regular MCEM converges to the neighborhood of the true MLE's relatively fast. However, our preliminary results suggested that for our method longer burn-in periods may be needed. Therefore, we chose to implement a slightly more flexible strategy. As in Levine and Casella [2001] we also start with an initial burn-in period with a relatively small $K$ and then we switch to the IS version of the MCEM. However, for each of the subsequent iteration we check if the current estimates $(\hat{\boldsymbol{\beta}}^t, \hat{\boldsymbol{\sigma}}^{2^t})$ are still in the "vicinity" of $(\hat{\boldsymbol{\beta}}^0, \hat{\boldsymbol{\sigma}}^{2^0})$ that were used for generating the random sample. If the current estimates are far from the initial ones, then we reset the IS by first discarding the old MCMC sample, and then generating a new one by setting $(\hat{\boldsymbol{\beta}}^0, \hat{\boldsymbol{\sigma}}^{2^0}) = (\hat{\boldsymbol{\beta}}^t, \hat{\boldsymbol{\sigma}}^{2^t})$. For deciding whether we are still in the neighborhood of the null parameters, we decided to implement the following criterion

$$\max\left\{\max_i\left\{\frac{|\hat{\boldsymbol{\beta}}_i^t - \hat{\boldsymbol{\beta}}_i^0|}{|\hat{\boldsymbol{\beta}}_i^0|}\right\}, \max_j\left\{\frac{|\hat{\boldsymbol{\sigma}}_j^{2^t} - \hat{\boldsymbol{\sigma}}_j^{2^0}|}{|\hat{\boldsymbol{\sigma}}_j^{2^0}|}\right\}\right\} < \lambda, \quad \text{(A14)}$$

where $\lambda$ is some positive number. From our experience, a value of $\lambda = 0.2$ seems to give satisfactory results, and thus we opt to use this value for all of our simulation analyses and the analyses of the T2D data.

## CONVERGENCE CRITERIA

The classic EM algorithm usually monitors convergence by checking if the relative change in the parameter estimates in two consecutive iterations is less than a predetermined small value [Booth and Hobert, 1999]. That is, at each iteration $t+1$ it tests if

$$\max\left\{\max_i\left\{\frac{|\hat{\boldsymbol{\beta}}_i^{t+1} - \hat{\boldsymbol{\beta}}_i^t|}{|\hat{\boldsymbol{\beta}}_i^t| + \delta_1}\right\}, \max_j\left\{\frac{|\hat{\boldsymbol{\sigma}}_j^{2^{t+1}} - \hat{\boldsymbol{\sigma}}_j^{2^t}|}{|\hat{\boldsymbol{\sigma}}_j^{2^t}| + \delta_1}\right\}\right\} < \delta_2 \quad \text{(A15)}$$

where $\delta_1$ and $\delta_2$ are predetermined constants with their most commonly employed values being 0.001 and 0.002, respectively [Booth and Hobert, 1999]. However, this criterion cannot be implemented in the context of the MCEM without taking into consideration the random error introduced by the MCMC sample, since it could lead to premature stop of the algorithm before convergence has been reached. This is why Booth and Hobert [1999] suggested to stop the MCEM iterations either when the above criterion has been met three consecutive times, or when

$$\max\left\{\max_i\left\{\frac{|\hat{\boldsymbol{\beta}}_i^{t+1} - \hat{\boldsymbol{\beta}}_i^t|}{\sqrt{\mathrm{Var}\,(\hat{\boldsymbol{\beta}}_i)} + \delta_1^*}\right\}, \max_j\left\{\frac{|\hat{\boldsymbol{\sigma}}_j^{2^{t+1}} - \hat{\boldsymbol{\sigma}}_j^{2^t}|}{\sqrt{\mathrm{Var}(\hat{\boldsymbol{\sigma}}_j^2)} + \delta_1^*}\right\}\right\} < \delta_2^* \quad \text{(A16)}$$

where $\mathrm{Var}(\hat{\boldsymbol{\beta}}_i)$ and $\mathrm{Var}(\hat{\boldsymbol{\sigma}}_j^2)$ are the estimates of the variance of the parameters obtained through the use of the MCMC sample on the tth iteration, and $\delta_1^*$ and $\delta_2^*$ are some user-defined constants, which need not be the same ones as in (18). However, for our implementations we chose to use the same values for both criteria, namely we set $\delta_1 = \delta_1^* = 0.001$ and $\delta_2 = \delta_2^* = 0.002$.