DOI: 10.1534/genetics.105.054650

Quantitative Trait Linkage Analysis Using Gaussian Copulas

Mingyao Li,*,†,[1] Michael Boehnke,† Gonçalo R. Abecasis† and Peter X.-K. Song‡

**Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104,* †*Department of Biostatistics and Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, Michigan 48109 and* ‡*Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada*

Manuscript received December 12, 2005
Accepted for publication May 23, 2006

ABSTRACT

Mapping and identifying variants that influence quantitative traits is an important problem for genetic studies. Traditional QTL mapping relies on a variance-components (VC) approach with the key assumption that the trait values in a family follow a multivariate normal distribution. Violation of this assumption can lead to inflated type I error, reduced power, and biased parameter estimates. To accommodate nonnormally distributed data, we developed and implemented a modified VC method, which we call the "copula VC method," that directly models the nonnormal distribution using Gaussian copulas. The copula VC method allows the analysis of continuous, discrete, and censored trait data, and the standard VC method is a special case when the data are distributed as multivariate normal. Through the use of link functions, the copula VC method can easily incorporate covariates. We use computer simulations to show that the proposed method yields unbiased parameter estimates, correct type I error rates, and improved power for testing linkage with a variety of nonnormal traits as compared with the standard VC and the regression-based methods.

VARIANCE-COMPONENTS (VC) linkage analysis (Amos 1994; Almasy and Blangero 1998) plays an important role in mapping quantitative trait loci (QTL) that influence quantitative traits in humans. Unlike the original Haseman–Elston regression (Haseman and Elston 1972), which lacks flexibility in modeling variance–covariance structures and covariates, the VC method can analyze pedigrees essentially of any configuration and provides increased linkage power (Amos *et al.* 1996; Williams and Blangero 1999). In the simplest implementation of the VC method, trait values are assumed to follow a multivariate normal distribution with the variances and covariances depending on identical-by-descent (IBD) sharing between relative pairs and major gene, shared polygenes, and environmental variance components.

A key assumption in the VC method is that the quantitative traits follow a multivariate normal distribution within a family. Violation of this assumption can lead to inflated type I error, reduced linkage power, and biased parameter estimates (Allison *et al.* 1999; Epstein *et al.* 2003). Several solutions have been proposed when the trait distribution is not normal. Most simply, one can transform the data to univariate normal and apply the standard VC method. For example, for continuous traits, this can be achieved by the inverse-normal transformation using the empirical distribution of the trait. This transformation is quite accurate when the sample size is large and ensures that the trait is approximately distributed as univariate normal. Alternatively, a semiparametric model that jointly estimates an empirical transformation and genetic model parameters can be used (Diao and Lin 2005). A weakness of these transformation-based approaches is that they are not appropriate for the analysis of discrete or censored traits.

To analyze the data without transformation, one could use an approach based on generalized estimating equations (GEEs) (Liang and Zeger 1986). Chen *et al.* (2004) described a GEE framework for linkage analysis that includes the Haseman–Elston regression and the standard VC methods as special cases, where the different methods result from different choices of a working covariance matrix. This approach allows for various robust score tests to be defined and can be extended to take higher moments of the trait distribution into account (Chen *et al.* 2005). If sufficient computing resources are available, another option is to use the standard VC method and assess significance through gene-dropping simulations. For settings where computing resources are limited, Blangero *et al.* (2000) proposed a robust estimator of the covariance matrix that controls type I error but may lead to loss of efficiency.

As far as efficiency is concerned, the maximum-likelihood procedure with the proper distribution is the method of choice. For example, Lange *et al.* (1989), Wan *et al.* (1998), and Epstein *et al.* (2003) developed VC models for data from *t*-, log-normal, and censored normal distributions, respectively. In each case,

[1] *Corresponding author:* Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, 624 Blockley Hall, 423 Guardian Dr., Philadelphia, PA 19104.
E-mail: mli@cceb.med.upenn.edu

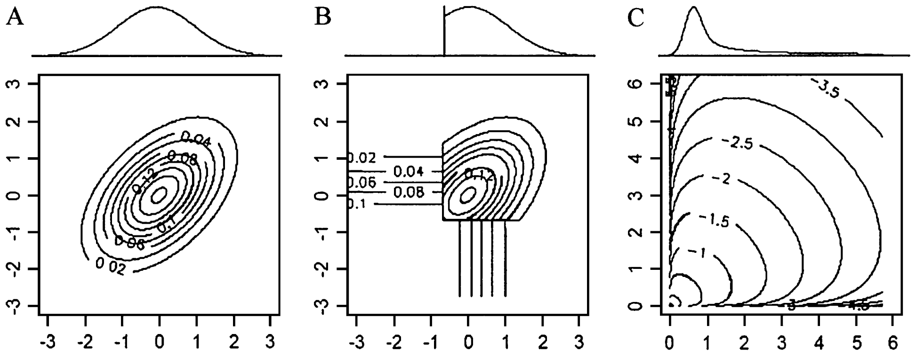

FIGURE 1.—Contour plot of densities for (A) bivariate normal, (B) bivariate censored-normal, and (C) bivariate gamma distributions (plotted in the $\log_{10}$ scale).

parameters are estimated by maximum likelihood and the methods outperform alternatives for the analysis of data with *t*-, log-normal, and censored distributions, respectively. Given that quantitative traits of interest in genetic studies follow many different distributions, it is desirable to develop a unified likelihood framework that allows the analysis of a broad range of distributions.

In the statistical literature, the most general way to describe dependence between correlated random variables is to use copulas (SKLAR 1959). Copulas are multivariate distribution functions whose one-dimensional margins are uniform on the [0, 1] interval (NELSON 1999). Copulas are useful for constructing joint distributions, especially when working with nonnormal random variables (JOE 1997). Here, we use Gaussian copulas, which share many similarities with the multivariate normal distribution. We show three examples in Figure 1, where the bivariate normal, the bivariate censored normal, and the bivariate gamma distributions are modeled by Gaussian copulas. In principle, copulas can be used to model the joint distributions of any continuous or discrete traits and even mixed continuous and discrete traits.

Copulas have been employed in previous genetic mapping studies. TREGOUET *et al.* (1999) developed a parametric copula model for the analysis of familial binary data and conducted a combined segregation–linkage analysis of levels of plasma angiotensin. WANG and HUANG (2002) proposed a score test for QTL mapping with sibships of arbitrary size using transformed data based on Gaussian copulas. BASRAK *et al.* (2004) described a bivariate Gaussian copula approach to relax the normality assumption in QTL mapping. Both WANG and HUANG (2002) and BASRAK *et al.* (2004) implemented their copulas using a two-stage approach in which the inverse-normal transformation was first used to standardize the data, and the transformed traits were later tested for linkage assuming multivariate normality. However, the main strength of the copula is not transformation but its ability to describe the joint distribution of multivariate random variables and to characterize their dependence structure. Compared to the two-stage approach, we expect that methods based on the joint distributions of the original traits will be more efficient.

In this article, we describe a unified method for mapping genes that influence quantitative traits by use of the Gaussian copulas in the VC framework. We call this the "copula VC method." Our method allows the analysis of continuous, discrete, and censored traits, and the standard VC method is a special case of our method when the data are distributed as multivariate normal. The copula VC method shares several features of the standard VC method; it is (i) likelihood based, (ii) applicable to pedigrees of any configuration, and (iii) readily incorporates covariates through the use of link functions. We evaluated the performance of the copula VC method by simulating data from multivariate Poisson and censored normal distributions. We compared our method with the standard VC method and a regression-based method (SHAM *et al.* 2002), which is equivalent to a robust score test derived under the GEE framework (CHEN *et al.* 2004). Our simulation results indicate that the copula VC method yields unbiased parameter estimates, correct type I error, and modest improvement of power for testing linkage.

MATERIALS AND METHODS

We consider the problem of identifying genetic variants that influence quantitative traits, which may be continuous, discrete, or censored, and have distributions that may not be normal. Here, we develop a unified likelihood framework for the analysis of quantitative traits with a broad range of distributions. We seek to (i) identify major genetic loci that influence the quantitative traits and (ii) estimate the major gene heritability, overall genetic heritability, and regression coefficients of measured environmental factors. In the following sections, we review likelihood calculation in the standard VC method, briefly describe Gaussian copulas, and provide details of our approach.

Likelihood of the standard VC method: A critical assumption in the standard VC method is that the trait values within a family are distributed as multivariate normal. For a family with m related individuals, denote their trait values by $\mathbf{y} = (y_1, \ldots, y_m)$. Let $\mathbf{x}_j$ denote a vector of observed covariates for individual j, and let the mean of the trait value be $E(y_j) = \mu_j = \mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}$. In the standard VC method, the trait value is modeled as the sum of independent effects due to measured covariates, such as age and gender, and unmeasured factors that can be modeled as random effects, such as the effect of the major-gene (mg), polygenes (pg), and individual-specific environmental (e) factors. The covariance matrix $\sum = (\Sigma_{jk})$ for the m individuals has elements

$$\Sigma_{jk} = \begin{cases} \sigma^2_{\mathrm{mg}} + \sigma^2_{\mathrm{pg}} + \sigma^2_{\mathrm{e}}, & \text{if } j = k \\ \pi_{jk}\sigma^2_{\mathrm{mg}} + 2\phi_{jk}\sigma^2_{\mathrm{pg}}, & \text{if } j \neq k, \end{cases}$$

where π_{jk} denotes the proportion of alleles at the major gene shared IBD between individuals j and k, ϕ_{jk} is the kinship coefficient, σ^2_{mg} is the additive genetic variance of the major gene, σ^2_{pg} is the additive genetic variance due to polygenes, and σ^2_{e} is the residual environmental variance. The IBD sharing probabilities are typically unobservable, but can be estimated from genetic marker data by use of the Lander–Green algorithm (Lander and Green 1987), as implemented in software packages such as GENEHUNTER (Kruglyak *et al.* 1996), ALLEGRO (Gudbjartsson *et al.* 2000), and MERLIN (Abecasis *et al.* 2002). Under the multivariate normality assumption, the likelihood of the family is

$$L = (2\pi)^{-m/2}\left|\sum\right|^{-1/2}\exp\left\{-\tfrac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\}. \quad (1)$$

Gaussian copulas: The standard VC method uses the Pearson correlation, a measure of linear dependence, to model phenotypic similarity between a pair of individuals. However, when the traits are nonnormally distributed, linear dependence may not be suitable due to the presence of higher-order correlation, especially when the traits are highly skewed or discrete. A more flexible way to describe dependence is to use copulas. Consider m possibly dependent uniform random variables $U_1, \ldots, U_m$ on the [0, 1] interval. The dependence relationship can be modeled through copula $C(u_1, \ldots, u_m) = P(U_1 \leq u_1, \ldots, U_m \leq u_m)$, where C is the joint distribution function of $U_j, j = 1, \ldots, m$.

A copula of particular interest is the Gaussian copula, defined as $C(u_1, \ldots, u_m) = \Phi_m(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_m) \mid \Gamma)$, where Φ and Φ_m are the standard univariate and multivariate normal cumulative distribution functions (CDFs), and Γ is an $m \times m$ correlation matrix. By using the Gaussian copulas, handling of a multivariate distribution can be separated into a marginal model for the inverse normal score $\Phi^{-1}(G_j(y_j))$ and a model for the joint distribution of the inverse normal scores with $\Phi_m(\Phi^{-1}(G_1(y_1)), \ldots, \Phi^{-1}(G_m(y_m)))$, where $G_j(y_j)$ is the CDF of y_j. If the trait is continuous, then the CDF is uniformly distributed, and the corresponding inverse-normal score is distributed as standard univariate normal. In genetic linkage studies of univariate traits, we are typically interested in trait values that follow the same marginal distribution. For now, we assume that all marginal distributions are the same and denote $G_j = G$.

Gaussian copula VC models: We assume that the marginal distribution of each trait value comes from an exponential family with distribution function $g(y; \eta, \varphi) = \exp\{(y\eta - b(\eta))/a(\varphi) + c(y, \varphi)\}$ (McCullagh and Nelder 1989), where a, b, and c are known functions, φ is the dispersion parameter, and η is the canonical parameter. The mean and variance of the trait value are given by $E(y) = \mu = b'(\eta)$ and $\mathrm{var}(y) = b''(\eta)a(\varphi)$, respectively. Given a set of covariates $\mathbf{x}$, the mean is related to $\mathbf{x}$ through a known link function $h(\mu) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$. The specification of $h(\cdot)$ depends on the trait distribution. For example, for a normally distributed trait, $h(\mu) = \mu$; for a count-related trait, $h(\mu) = \log(\mu)$; for a binary trait, $h(\mu) = \log[\mu/(1-\mu)]$; for a gamma-distributed trait, one can use either a reciprocal link function $h(\mu) = 1/\mu$ or a log link function $h(\mu) = \log(\mu)$. For a gamma-distributed trait, the reciprocal link is the canonical link, but it prohibits negative mean values and can lead to unstable estimation of parameters (McCullagh and Nelder 1989), so that the log link is typically preferred.

Given the marginal trait distributions, the Gaussian copula gives rise to the following joint distribution of $\mathbf{y} = (y_1, \ldots, y_m)$,

$$\begin{aligned} &F(\mathbf{y}; \boldsymbol{\eta}, \varphi, \Gamma) \\ &\quad = \Phi_m(\Phi^{-1}(G(y_1; \eta_1, \varphi)), \ldots, \Phi^{-1}(G(y_m; \eta_m, \varphi)) \mid \Gamma), \quad (2) \end{aligned}$$

where the correlation matrix Γ has elements 1 on the diagonal and $(\pi_{jk}\sigma^2_{\mathrm{mg}} + 2\phi_{jk}\sigma^2_{\mathrm{pg}})/(\sigma^2_{\mathrm{mg}} + \sigma^2_{\mathrm{pg}} + \sigma^2_{\mathrm{e}}), j \neq k$, on the off-diagonal. The correlation matrix Γ characterizes the pairwise nonlinear dependence of the trait values, $\mathrm{corr}(\Phi^{-1}(G(y_j; \eta_j, \varphi)), \Phi^{-1}(G(y_k; \eta_k, \varphi)))$, among the components of $\mathbf{y}$.

Joint probability/density functions: Given the joint distribution function of $\mathbf{y}$, the corresponding joint probability/density function can be obtained by taking derivatives with respect to (2) (Song 2000). When the trait is continuous, the joint density function of $\mathbf{y}$ is

$$\begin{aligned} &f(\mathbf{y}; \boldsymbol{\eta}, \varphi, \Gamma) \\ &\quad = c(G(y_1; \eta_1, \varphi), \ldots, G(y_m; \eta_m, \varphi) \mid \Gamma)\prod_{j=1}^{m} g(y_j; \eta_j, \varphi), \quad (3) \end{aligned}$$

where $c(u_1, \ldots, u_m|\Gamma) = |\Gamma|^{-1/2}\exp\{\frac{1}{2}\mathbf{q}^{\mathrm{T}}(I_m - \Gamma^{-1})\mathbf{q}\}$ and $u_j = G(y_j; \eta_j, \varphi)$, $\mathbf{q} = (q_1, \ldots, q_m)^{\mathrm{T}}$ is a vector of inverse-normal scores $q_j = \Phi^{-1}(u_j)$, and I_m is an m-dimensional identity matrix.

When the trait is discrete, the joint probability function of $\mathbf{y}$ is obtained by taking the Radon–Nikodym derivative for $F(\mathbf{y}; \boldsymbol{\eta}, \varphi, \Gamma)$ in (2) with respect to counting measure (Song 2000),

$$\begin{aligned} &f(\mathrm{y}; \eta, \varphi, \Gamma) \\ &\quad = \sum_{j_1=1}^{2} \cdots \sum_{j_m=1}^{2} (-1)^{j_1+\cdots+j_m}\Phi_m(\Phi^{-1}(u_{1,j_1}), \ldots, \Phi^{-1}(u_{m,j_m}) \mid \Gamma), \quad (4) \end{aligned}$$

where $u_{j,1} = G(y_j-; \eta_j, \varphi)$ and $u_{j,2} = G(y_j; \eta_j, \varphi)$. Here, $G(y_j-; \eta_j, \varphi)$ is the left-hand limit of G at y_j, which is equal to $G(y_j - 1; \eta_j, \varphi)$ when y_j takes integer values as for the Poisson and Binomial distributions.

Finally, when the m margins include m_1 continuous and $m_2 = m - m_1$ discrete outcomes, the joint density function can be obtained as follows. Let $\mathbf{u}_1 = (u_1, \ldots, u_{m_1})^{\mathrm{T}}$ and $\mathbf{u}_2 = (u_{m_1+1}, \ldots, u_m)^{\mathrm{T}}$. The same partition and notation are applied for vectors $\mathbf{q}$ and $\mathbf{y}$. Let

$$\begin{aligned} C^*(\mathbf{u}_1, \mathbf{u}_2 \mid \Gamma) = (2\pi)^{-(m_2/2)}|\Gamma|^{-(1/2)}\int_{-\infty}^{\Phi^{-1}(u_{m_1+1})} \cdots \int_{-\infty}^{\Phi^{-1}(u_m)} \\ \times \exp\left\{-\tfrac{1}{2}(\mathbf{q}_1^{\mathrm{T}}, \mathbf{y}_2^{\mathrm{T}})\Gamma^{-1}(\mathbf{q}_1^{\mathrm{T}}, \mathbf{y}_2^{\mathrm{T}}) + \tfrac{1}{2}\mathbf{q}_1^{\mathrm{T}}\mathbf{q}_1\right\} d\mathbf{y}_2, \end{aligned}$$

and then the joint density of $\mathbf{y}$ is given by

$$\begin{aligned} f(\mathbf{y}; \boldsymbol{\eta}, \varphi, \Gamma) = \prod_{j=1}^{m_1} g(y_j; \eta_j, \varphi_1) \times \sum_{j_{m_1+1}=1}^{2} \cdots \sum_{j_m=1}^{2} (-1)^{j_{m_1+1}+\cdots+j_m} \\ \times C^*(G(y_1; \eta_1, \varphi_1), \ldots, G(y_{m_1}; \eta_{m_1}, \varphi_1), \\ u_{m_1+1, j_{m_1+1}}, \ldots, u_{m,j_m} \mid \Gamma), \quad (5) \end{aligned}$$

where φ_1 is the dispersion parameter for the first m_1 continuous outcomes and u_{t,j_t}'s are defined as above.

It is worth noting that when the marginal distribution G is continuous, the transformed trait vector $(\Phi^{-1}(G(y_1;\eta_1,\varphi)),\ldots,\Phi^{-1}(G(y_m;\eta_m,\varphi)))^{\mathrm{T}}$ is distributed as multivariate normal with mean vector $\mathbf{0}$ and correlation matrix Γ. However, this is not true for discrete traits. By taking Radon–Nikodym derivatives with respect to counting measure, Equation 4 allows us to calculate explicitly the joint probability mass function of the correlated discrete traits. This makes our method different from those of Wang and Huang (2002) and Basrak *et al.* (2004) who assumed that the inverse-normal transformed traits are distributed as multivariate normal regardless of the trait types. Without deriving the joint probability mass function for the original traits, their methods should be used for the analysis of continuous traits only.

Given the joint density/probability function for one family in Equations 3, 4, or 5, the construction of the full likelihood for the trait data is simply the product of the joint density/probability functions over all families:

$$L=\prod f(\mathbf{y};\boldsymbol{\eta},\varphi,\Gamma). \qquad (6)$$

The joint probability/density functions allow us to analyze traits with a variety of distributions. For example, for traits that have skewed distributions, one might assume that the trait values within a family follow a multivariate gamma distribution and model their joint density by Equation 3; for count data, one might assume that the trait values within a family follow a multivariate Poisson or negative binomial distribution and model their joint probability by Equation 4.

It is interesting to note that our likelihood calculation also allows the analysis of censored data that may arise in genetic studies, for example, due to assay limitations or when some subjects are taking medication (Epstein *et al.* 2003). Here, we assume that the latent distribution of the censored data is continuous. To illustrate how one could obtain joint densities of censored data, consider a sib pair with trait values (y_1, y_2) that have a bivariate censored normal distribution. For convenience, assume that censoring results in all trait values less than a threshold value y^* are equal to y^*. A censored trait value can be regarded as being generated from a Bernoulli distribution with the probability parameter being the proportion of censoring. If both sibs are censored, their joint probability function is given by the discrete-type Equation 4; if neither observation is censored, their joint density function is given by the continuous-type Equation 3; and if one sib is censored and the other sib is not, their joint density function is given by the mixed-type Equation 5. The joint density for bivariate censored normal data is given in appendix a. Similar derivations apply to censored data in higher dimensions and with other latent distributions.

Test of linkage: Testing linkage is the central task in QTL mapping studies. Invoking the above copula joint models, we can establish a test for linkage at the major gene $\mathrm{H_0}$: $\sigma^2_{\mathrm{mg}}=0$ *vs.* $\mathrm{H_A}$: $\sigma^2_{\mathrm{mg}}>0$ within the framework of likelihood-ratio tests, where the likelihood-ratio statistic is $2\ln(\hat{L}_{\mathrm{A}}/\hat{L}_0)$, with $\hat{L}_{\mathrm{A}}$ and $\hat{L}_0$ being the likelihood (6) maximized under the alternative and null hypotheses, respectively. Since the value of σ^2_{mg} is on the boundary of the parameter space under the null hypothesis, the asymptotic null distribution of the likelihood-ratio statistic will be approximated by a 50:50 mixture of χ^2_1 and a point mass at 0 (Self and Liang 1987). The LOD score at the locus being tested is $\log_{10}(\hat{L}_{\mathrm{A}}/\hat{L}_0)$, which is equivalent to $2\ln 10\approx 4.6$ units of the likelihood-ratio statistic. We maximized these likelihoods via a Gauss–Newton type algorithm (Ruppert 2005) (appendix b), which requires the first derivatives of the log-likelihood.

Simulations: We conducted several simulations to examine and compare the performance of the copula VC, the standard VC, and the regression-based method as implemented in MERLIN-REGRESS for the analysis of nonnormal data. We specified the true population mean, variance, and heritability of the traits in MERLIN-REGRESS, therefore allowing this method to achieve optimal performance. The method implemented in MERLIN-REGRESS has been shown to be equivalent to a robust score test using the GEE framework (Chen *et al.* 2005). For illustration purposes, we examined Poisson- and censored normal-distributed traits. We first simulated data sets of 400 sib trios according to the copula model (4) to generate trait values with a Poisson distribution. We simulated a map of 10 markers each with four equally frequent alleles evenly spaced at 11.16-cM intervals, corresponding to recombination fraction 0.10 under Haldane's (1919) no interference map function. A QTL with two equally frequent alleles was placed in the middle of the map. We generated data with different values of major-gene heritability $h^2_{\mathrm{mg}}=\sigma^2_{\mathrm{mg}}/(\sigma^2_{\mathrm{mg}}+\sigma^2_{\mathrm{pg}}+\sigma^2_{\mathrm{e}})$ and overall genetic heritability $h^2=(\sigma^2_{\mathrm{mg}}+\sigma^2_{\mathrm{pg}})/(\sigma^2_{\mathrm{mg}}+\sigma^2_{\mathrm{pg}}+\sigma^2_{\mathrm{e}})$. We removed the QTL genotypes prior to data analysis.

To determine whether the tested methods have correct type I error under the null hypothesis of no linkage, we simulated 10,000 replicate data sets. We also conducted simulations to compare the power of the three methods using 5000 replicate data sets. We consider trait models with combinative parameter values of $\{\lambda=(0.5,1.0)\}\times\{h^2_{\mathrm{mg}}=(0,0.25,0.4)\}\times\{h^2=(0.6,0.8)\}$, where λ is the mean parameter of Poisson distribution. To make fair power comparisons between the methods, we used empirical significance thresholds obtained from the null distribution simulations with the same total heritability, h^2, but assuming that the major gene effect was $h^2_{\mathrm{mg}}=0$. To determine the impact of discreteness on the estimation of covariate effects, we conducted additional simulations including a covariate that was generated from the standard normal distribution. We set the regression coefficient $\beta=0.5$ or 1 and simulated 2000 additional replicate data sets in each setting. The trait values were connected with the covariates using a log link function. In this setting, we analyzed the simulated data using the copula VC method only since the standard VC method assumes an identity link and thus it is not appropriate to compare covariate estimates for these two methods.

We repeated the simulation procedure for censored normal traits. For ease of computation, we considered sib pairs only. We simulated 800 sib pairs in each data set. To obtain censored trait values, we first simulated latent bivariate normal traits in accordance with the copula model (3) and then censored those values below a threshold of the latent trait distribution. We determined the threshold by the proportion of censoring, denoted by c. We considered trait models with combinative parameter values of $\{c=(10\%,25\%)\}\times\{h^2_{\mathrm{mg}}=(0,0.25,0.4)\}\times\{h^2=(0.6,0.8)\}$. Without loss of generality, for all the trait models we considered, the total variance of the latent trait values was set to be 1.0. To determine the impact of censoring on the estimation of covariate effects, we conducted additional simulations including a covariate that was generated from the standard normal distribution. We set the regression coefficient $\beta=0.5$ or 1 and simulated 2000 additional replicate data sets in each setting. The trait values were connected with the covariates by an identity link function.

RESULTS

Poisson-distributed traits: *Empirical type I error and power for detecting linkage:* Figure 2 shows the empirical type I error for Poisson (count)-distributed traits when

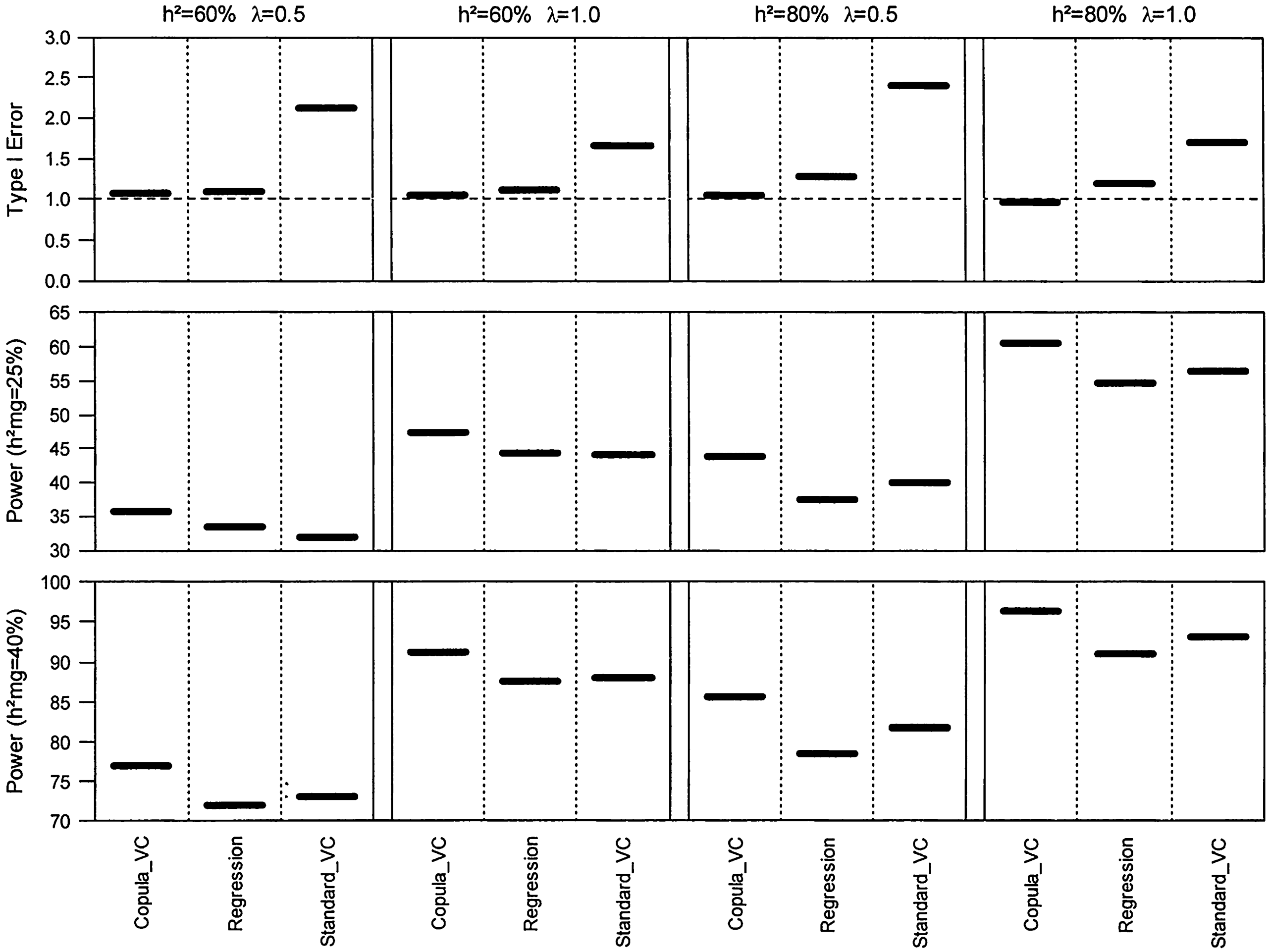

Figure 2.—Empirical type I error and power of the standard and copula VC methods and the regression-based method for Poisson-distributed traits. Empirical type I error is based on 10,000 replicates of data sets of 400 sib trios. Empirical power is determined by adjusting for the size under the null model that assumed no major gene effect on the basis of 5000 replicates of data sets. λ is the mean parameter of Poisson distribution. Significance was assessed at the 1% level.

significance was evaluated using the asymptotic null distribution. The standard VC method gives inflated type I error for testing linkage, especially when the mean parameter of the Poisson distribution, λ, is small, corresponding to greater departure from normality. For example, when $\lambda = 0.5$ and $h^2 = 0.8$, the type I error is 2.4% at the 1% significance level. We also found that the type I error for the standard VC method increases as the overall genetic heritability h^2 increases. In contrast, the copula VC method gives type I error close to the nominal levels for all four trait models that we considered. The type I error for the regression-based method is slightly higher than that for the copula VC method, but is nearly under control.

The empirical power of the copula and standard VC methods and the regression-based method for testing linkage is shown in Figure 2. Since the standard VC method has inflated type I error when using the asymptotic null distribution, we determined the critical values for testing linkage using the empirical null distributions generated with the same parameters except that $h^2_{\mathrm{mg}} = 0$ was assumed for all three methods. As expected, the power to detect linkage of all three methods increases as the major gene heritability and overall genetic heritability increase. Our results also show that, for all the eight trait models considered, the copula VC method has modest improvement of power to detect linkage compared to the standard VC method and the regression-based method.

Trait model parameter and regression coefficient estimates: The mean parameter estimates and the square root of mean square errors (MSEs) of the major gene heritability and the overall genetic heritability for both VC methods are shown in Figure 3. Our results indicate that the standard VC method underestimates heritability, especially when the mean count of the Poisson distribution is small. Compared to the overall genetic heritability, the major gene heritability appears to be less influenced by discreteness of the Poisson distribution. Mean estimates averaged 79–91% of the true values for the major gene heritability and 75–87% of the true values for the overall genetic heritability. In contrast, the

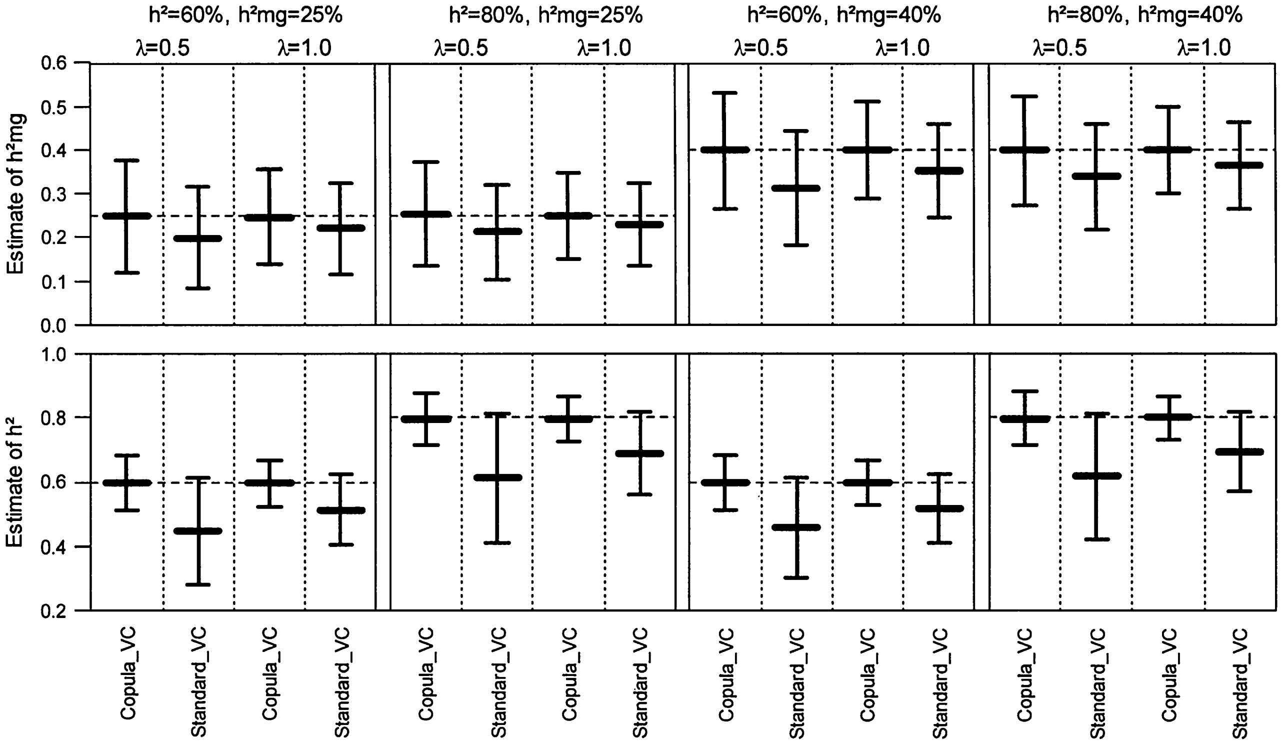

FIGURE 3.—Parameter estimates of the standard and copula VC methods for Poisson-distributed traits (without covariates). λ is the mean of the Poisson distribution. The error bar represents the square root of the mean squared error. Results are based on 2000 replicate data sets of 400 sib trios.

copula VC method yields unbiased parameter estimates for all trait models that we considered. Further, the accuracy of the parameter estimates, as measured by the square root of the MSEs, improves as the major gene heritability and the overall genetic heritability increase.

We also investigated the effects of the discreteness of the Poisson distribution on regression coefficient estimates for the copula VC method (Figure 4). We are not able to evaluate the regression coefficient estimates using the standard VC method, since the covariate is linked with the Poisson-distributed trait values through a log link function, which is different from the identity link assumed by the standard VC method. Hence, the two VC methods are not directly comparable in terms of regression coefficient estimates. Figure 4 indicates that the copula VC method yields unbiased estimates of the regression coefficient for the eight trait models that we considered. As expected, the accuracy of the parameter estimates improves as the major gene and overall trait heritabilities increase.

Censored normal distributed traits: *Empirical type I error and power for detecting linkage:* In Figure 5, we show the empirical type I error for the test of linkage with censored normal data. As noted by EPSTEIN *et al.* (2003), the standard VC method yields inflated type I error, especially when the proportion of censored observations is large. Further, the type I error for the standard VC method increases as the overall genetic heritability increases. For example, the type I error ranges between 1.2% (when $h^2 = 0.6$ and $c = 10\%$) and 1.7% (when $h^2 = 0.8$ and $c = 25\%$) when significance was assessed at the 1% level. In contrast, the copula VC method yields type I error that is close to the nominal level.

The empirical power of the standard and the copula VC methods for testing linkage is shown in Figure 5. Again, we determined the empirical power by simulating data under the null model to estimate critical values. As expected, the power to detect linkage of all three methods increases as the major gene and overall genetic heritabilities increase. The power to detect linkage of all three methods diminishes as the percentage of censored observations increases, corresponding to less information about the underlying trait distribution. Further, the copula VC method provides a modest increase in power to detect linkage over the standard VC method, consistent with the results of EPSTEIN *et al.* (2003), who developed the Tobit VC model to handle censored normal data. The copula VC method is also more powerful than the regression-based method.

Trait model parameter and regression coefficient estimates: In Figure 6, we show the mean parameter estimates and the square root of the MSEs of the major gene heritability and the overall genetic heritability for both VC methods. As previously noted by EPSTEIN *et al.* (2003), we found that the standard VC method underestimates the true values of the heritability parameters. On average, the estimates of the major gene heritability

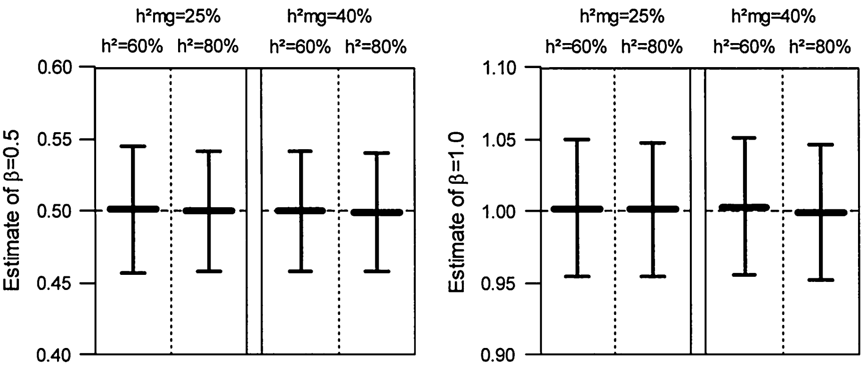

FIGURE 4.—Parameter estimates of the copula VC method for Poisson-distributed traits with one covariate generated from the standard normal distribution. The error bar represents the square root of the mean squared error. Results are based on 2000 replicate data sets of 400 sib trios.

are 94–99% of the true values and the estimates of the overall genetic heritability are 92–98% of the true values for the eight trait models that we considered. Compared to the overall genetic heritability, the major gene heritability appears to be less influenced by censoring. As expected, the copula VC method yields unbiased estimates for both heritability parameters and the accuracy of parameter estimation improves as the percentage of censored observations decreases.

We also examined the effects of censoring on estimation of regression coefficients for both VC methods (Figure 7). We found that the regression coefficient estimates are notably attenuated toward zero using the standard VC method. The estimated values are ~75% of

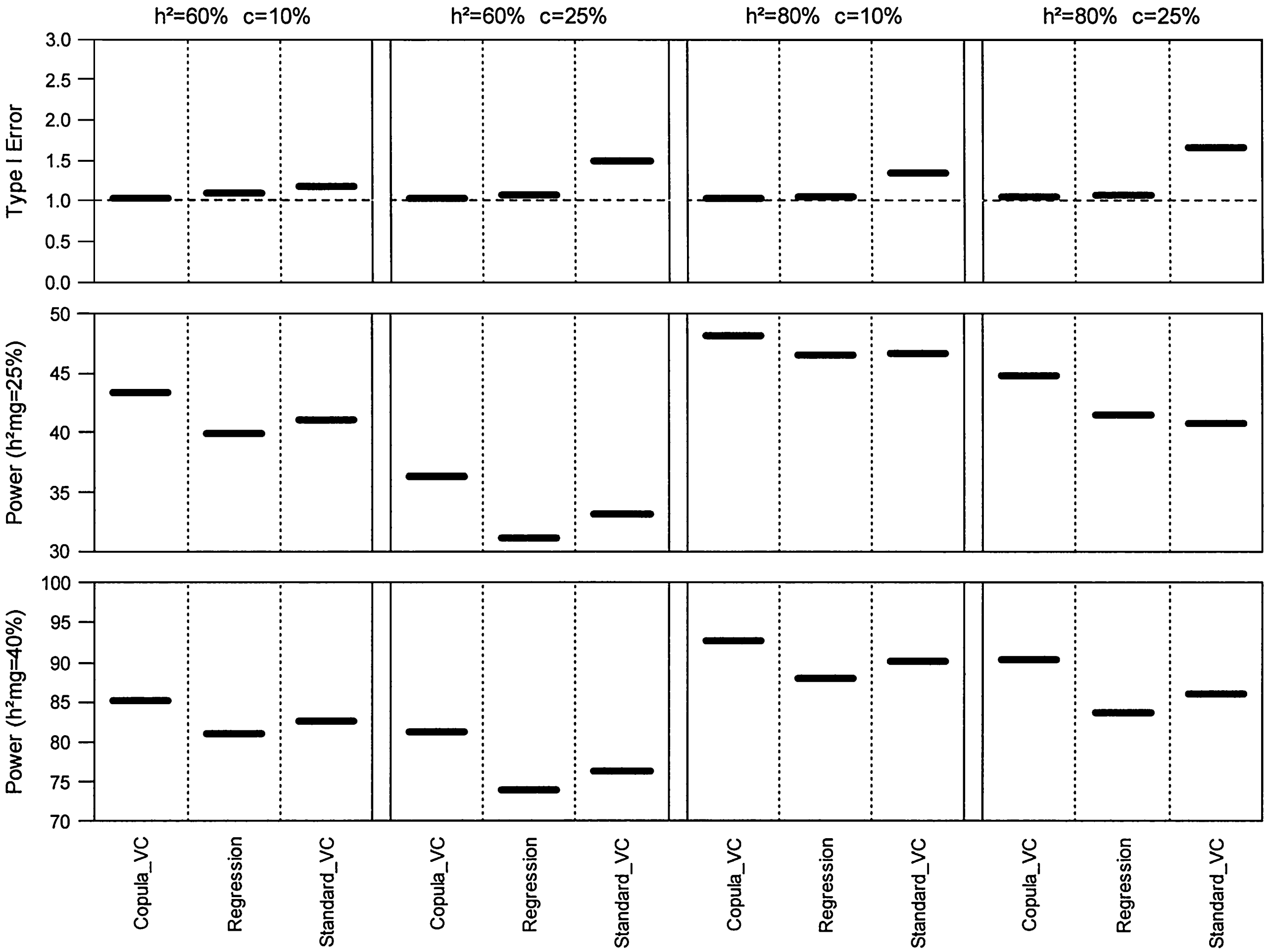

FIGURE 5.—Empirical type I error and power of the standard and copula VC methods and the regression-based method for censored normal traits. Empirical type I error is based on 10,000 replicates of data sets of 800 sib pairs. Empirical power is determined by adjusting for the size under the null model that assumed no major gene effect on the basis of 5000 replicates of data sets. *c* is the percentage of censoring. Significance was assessed at the 1% level.

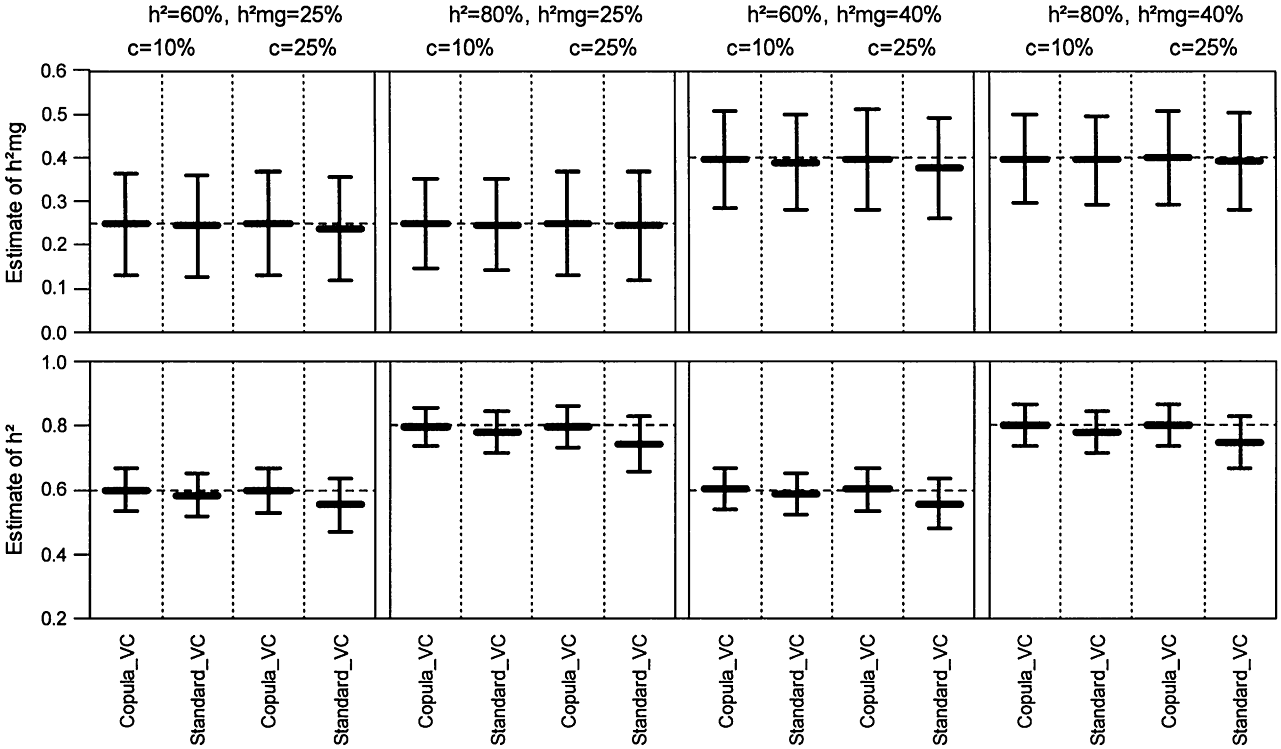

FIGURE 6.—Parameter estimates of the standard and copula VC methods for censored normal traits (without covariates). *c* is the percentage of censoring. The error bar represents the square root of the mean squared error. Results are based on 2000 replicate data sets of 400 sib trios.

the true values for all the eight trait models that we considered. In contrast, the copula VC method not only gives unbiased parameter estimates but also shows less variability.

DISCUSSION

Many traits of scientific interest have nonnormal distributions. Using Gaussian copulas, we developed a unified likelihood framework that allows the analysis of traits with a wide range of distributions. Unlike the standard VC method, our copula VC method does not require the traits to be normally distributed. In particular, the standard VC method is a special case of the copula VC method when the traits are distributed as multivariate normal. Our method allows the analysis of continuous, discrete, and censored traits, including binary, polychotomous, count, and continuous skewed data. Through the use of link functions, the method can easily incorporate covariates to study the influence of

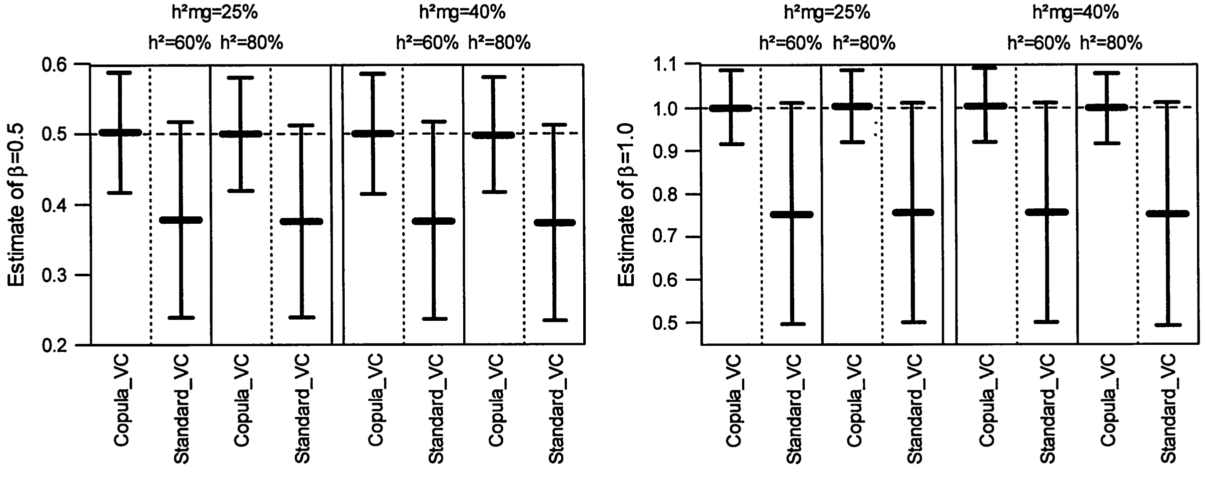

FIGURE 7.—Parameter estimates of the standard and copula VC methods for censored normal traits with one covariate generated from the standard normal distribution. Results are based on 2000 replicate data sets of 800 sib pairs. The error bar represents the square root of the mean squared error. The total variance of the trait is fixed at 1. The percentage of censoring is 25%.

environmental effects. For simplicity, we simulated sibship data in this article, but our method can be employed for the analysis of large pedigrees, although this becomes computationally more challenging.

The copula VC method yields unbiased parameter estimates and the correct type I error for testing linkage with a variety of nonnormal traits. In contrast, the standard VC method gives biased parameter estimates and inflated type I error and therefore requires intensive simulations to evaluate significance levels appropriately. The type I error of the regression-based method (Sham *et al.* 2002) is generally under control; however, that method is less powerful than the copula VC method. We found that for Poisson and censored normal traits, the major gene and overall genetic heritabilities are underestimated using the standard VC method. Further, for censored normal traits, the estimated regression coefficients for covariate effects are attenuated toward zero.

Although our article is focused on linkage analysis, our method can be employed in association studies by including the genetic markers as covariates. It is unclear that Wang and Huang's (2002) and Basrak *et al.*'s (2004) methods can be extended to test for association since the covariates have been regressed out in the data standardization procedure. Our results suggest that the copula VC method could be used to improve power of the association tests in the quantitative trait transmission/disequilibrium test (Abecasis *et al.* 2000), which incorporates the tested markers as covariates in a standard VC model that assumes multivariate normality.

For continuous traits, the copula VC method implicitly assumes that the inverse-normal transformed traits are multivariate normal, an assumption that is also made by the two-stage approach (Wang and Huang 2002; Basrak *et al.* 2004) and the semiparametric approach (Diao and Lin 2005). In contrast to these transformation-based methods, by taking Radon–Nikodym derivatives, we derived the joint probability/density function of the original trait values, including discrete and mixed outcomes. Moreover, by using generalized linear models, the copula VC method can easily accommodate any traits with marginal distributions belonging to the exponential family (McCullagh and Nelder 1989). Compared to other VC-based methods for nonnormal traits (Lange *et al.* 1989; Wan *et al.* 1998; Epstein *et al.* 2003), the method proposed here is more general and flexible.

In this article, we compared the copula VC method with the regression-based method for Poisson and censored normal traits and found that although both methods controlled type I error rates, our method appeared to be slightly more powerful. Chen *et al.* (2004, 2005) used GEEs to develop two robust score tests based on higher moments of the trait distributions and showed that the regression-based method (Sham *et al.* 2002) was equivalent to a robust score test derived from the GEE framework. Using the regression-based method as a benchmark, we expect that the copula VC method might be more powerful than the GEE-based method if the joint distributions are nearly correctly specified. However, if the joint distributions of the traits are hard to specify, then the GEE-based method may be preferred. We also expect that methods that incorporate higher moments of the trait distribution into a robust score test framework may improve the performance of the basic GEE approach so that it performs nearly as well as our maximum-likelihood approach.

The copula VC method assumes that the marginal distributions of the traits are known and, for practical data analysis, this assumption should be confirmed through model diagnostic procedures. One approach is to use the Q–Q plot to validate the parametric distribution assumption. If the parametric assumption is in question, to make our method practical, one could replace $G(y)$ with the empirical CDF $\hat{G}(y)$. The strong law of large numbers ensures that $\hat{G}(y)$ converges to $G(y)$ almost surely, which means that the empirical CDF can be used to estimate the marginal distribution of any trait of interest. This approach is not applicable to discrete traits.

In this article, we employed Gaussian copulas to model the joint distributions of the traits. Gaussian copulas are a powerful tool for modeling multiple correlated variables and enjoy the flexibility of allowing arbitrary correlation structures and modeling high-dimensional data, whereas many other copulas are restricted to two dimensions. Moreover, Gaussian copulas are conceptually simple and have intuitive connections with the familiar multivariate normal distribution. We recognize that the Gaussian copulas are just one way to model nonnormally distributed traits. Similar approaches might be employed for other families of copulas, such as the *t*- or Archimedian copulas (Nelson 1999). It might be worth comparing the performance of the Gaussian copulas with these copulas, and such a study might offer even more flexibility in analyzing quantitative traits with different distributions.

In summary, we have developed a unified copula VC approach that allows the analysis of traits with a variety of distributions. Our method relaxes the multivariate normality assumption as employed by the standard VC method. We illustrated the utility of the copula VC method by simulating Poisson and censored normal traits. We simulated sibship data for simplicity, but our method can be employed for the analysis of larger pedigrees. We believe that the copula VC method provides a useful tool for the mapping of genes that influence nonnormally distributed quantitative traits.

We thank Wei-Min Chen for helpful discussions. We also thank Karl Broman and two anonymous reviewers for their valuable comments. This research was supported by National Institutes of Health grants HG00376 (to M.B.), HG02651, and EY12562 (to G.R.A.) and by the Natural Sciences and Engineering Research Council of Canada (to P.X-K.S.). M.L. was previously supported by a University of Michigan Rackham predoctoral fellowship.

LITERATURE CITED

Abecasis, G. R., L. R. Cardon and W. O. Cookson, 2000 A general test of association for quantitative traits in nuclear families. Am. J. Hum. Genet. **66:** 279–292.

Abecasis, G. R., S. S. Cherny, W. O. Cookson and L. R. Cardon, 2002 Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nat. Genet. **30:** 97–101.

Allison, D. B., M. C. Neale, R. Zannolli, N. J. Schork, C. I. Amos *et al.*, 1999 Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. Am. J. Hum. Genet. **65:** 531–544.

Almasy, L., and J. Blangero, 1998 Multipoint quantitative-trait linkage analysis in general pedigrees. Am. J. Hum. Genet. **62:** 1198–1211.

Amos, C. I., 1994 Robust variance-components approach for assessing genetic linkage in pedigrees. Am. J. Hum. Genet. **54:** 535–543.

Amos, C. I., D. K. Zhu and E. Boerwinkle, 1996 Assessing genetic linkage and association with robust components of variance approaches. Ann. Hum. Genet. **60:** 143–160.

Basrak, B., C. A. J. Klaassen, M. Beekman, N. G. Martin and D. Boomsma, 2004 Copulas in QTL mapping. Behav. Genet. **34:** 161–171.

Blangero, J., J. T. Williams and L. Almasy, 2000 Robust LOD scores for variance component-based linkage analysis. Genet. Epidemiol. **19**(S1): 8–14.

Chen, W. M., K. W. Broman and K. Y. Liang, 2004 Quantitative trait linkage analysis by generalized estimating equations: unification of variance components and Haseman-Elston regression. Genet. Epidemiol. **26:** 265–272.

Chen, W. M., K. W. Broman and K. Y. Liang, 2005 Power and robustness of linkage tests for quantitative traits in general pedigrees. Genet. Epidemiol. **28:** 11–23.

Diao, G., and D. Y. Lin, 2005 A powerful and robust method for mapping quantitative trait loci in general pedigrees. Am. J. Hum. Genet. **77:** 97–111.

Epstein, M. P., X. Lin and M. Boehnke, 2003 A Tobit variance-component method for linkage analysis of censored trait data. Am. J. Hum. Genet. **72:** 611–620.

Gudbjartsson, D., K. Jonasson, M. Frigge and A. Kong, 2000 Alegro, a new computer program for multipoint linkage analysis. Nat. Genet. **25:** 12–13.

Haldane, J. B. S., 1919 The combination of linkage values and the calculation of distances between the loci of linked factors. J. Genet. **8:** 299–309.

Haseman, J. K., and R. C. Elston, 1972 The investigation of linkage between a quantitative trait and a marker locus. Behav. Genet. **2:** 3–19.

Joe, H., 1997 *Multivariate Models and Dependence Concepts.* Chapman & Hall, New York.

Kruglyak, L., M. J. Daly, M. P. Reeve-Daly and E. S. Lander, 1996 Parametric and nonparametric linkage analysis: a unified multipoint approach. Am. J. Hum. Genet. **58:** 1347–1363.

Lander, E. S., and P. Green, 1987 Construction of multilocus genetic linkage maps in humans. Proc. Natl. Acad. Sci. USA **84:** 2363–2367.

Lange, K., R. J. A. Little and J. M. G. Taylor, 1989 Robust statistical modeling using the t distribution. J. Am. Stat. Assoc. **84:** 881–896.

Liang, K. Y., and S. L. Zeger, 1986 Longitudinal data analysis using generalized linear models. Biometrika **73:** 13–22.

McCullagh, P., and J. A. Nelder, 1989 *Generalized Linear Models.* Chapman & Hall, London.

Nelson, R. B., 1999 *An Introduction to Copulas.* Springer-Verlag, New York.

Ruppert, D., 2005 Discussion of "Maximization by parts in likelihood inference." J. Am. Stat. Assoc. **100:** 1161–1163.

Self, S. G., and K. Y. Liang, 1987 Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J. Am. Stat. Assoc. **82:** 605–610.

Sham, P. C., S. Purcell, S. S. Cherny and G. R. Abecasis, 2002 Power regression-based quantitative-trait linkage analysis of general pedigrees. Am. J. Hum. Genet. **71:** 238–253.

Sklar, A., 1959 Fonctions de repartition a n dimensions et leurs marges. Publ. Inst. Stat. Univ. Paris **8:** 229–231.

Song, X.-K. P., 2000 Multivariate dispersion models generated from Gaussian copulas. Scand. J. Stat. **27:** 305–320.

Tregouet, D. A., P. Ducimetiere, V. Bocquet, S. Visvikis, F. Soubrier *et al.*, 1999 A parametric copula model for analysis of familial binary data. Am. J. Hum. Genet. **64:** 886–893.

Wan, Y., M. Deandrade, L. Yu, J. Cohen and C. I. Amos, 1998 Genetic linkage analysis using lognormal variance components. Ann. Hum. Genet. **62:** 521–530.

Wang, K., and J. Huang, 2002 A score-statistic approach for the mapping of quantitative-trait loci with sibships of arbitrary size. Am. J. Hum. Genet. **70:** 412–424.

Williams, J. T., and J. Blangero, 1999 Comparison of variance components and sibpair-based approaches to quantitative trait linkage analysis in unselected samples. Genet. Epidemiol. **16:** 113–134.

Communicating editor: K. W. Broman

APPENDIX A

We briefly describe the derivation of the joint density function for bivariate censored normal variables. Let $(z_1, z_2)^{\mathrm{T}}$ follow a bivariate normal distribution with mean $(\mu_1, \mu_2)^{\mathrm{T}}$ and variance–covariance matrix

$$\begin{pmatrix} \sigma_1^2 & \gamma\sigma_1\sigma_2 \\ \gamma\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

Assume that $(z_1, z_2)^{\mathrm{T}}$ are censored below a threshold y^* such that $y_1 = y^*$ if $z_1 \le y^*$ and $y_1 = z_1$ otherwise, and y_2 is similarly defined. Then $(y_1, y_2)^{\mathrm{T}}$ are censored bivariate normal variables. The density of $(y_1, y_2)^{\mathrm{T}}$ takes the following form:

1. If both y_1 and y_2 are censored at y^*, then both of them can be regarded as coming from a Bernoulli distribution with the probability parameter being the proportion of censoring $\Phi((y^* - \mu_j)/\sigma_j)$, $j = 1, 2$. By Equation 4, the joint density function of $(y_1, y_2)^{\mathrm{T}}$ is

$$f(y_1, y_2) = \Phi_{2,\gamma}\left(\frac{y^* - \mu_1}{\sigma_1}, \frac{y^* - \mu_2}{\sigma_2}\right),$$

where $\Phi_{2,\gamma}$ is the standard bivariate normal cumulative distribution function with correlation parameter γ.

2. If y_1 is censored at y^* and y_2 is not, then $(y_1, y_2)^{\mathrm{T}}$ can be regarded as mixed type outcomes with y_1 being Bernoulli and y_2 being normal. By Equation 5, the joint density function of $(y_1, y_2)^{\mathrm{T}}$ is

$$f(y_1, y_2) = \frac{1}{\sigma_2}\phi\left(\frac{y_2 - \mu_2}{\sigma_2}\right)\Phi\left(\frac{(y^* - \mu_1)/\sigma_1 - \gamma(y_2 - \mu_2)/\sigma_2}{\sqrt{1-\gamma^2}}\right).$$

3. If y_2 is censored at y^* and y_1 is not, then $(y_1, y_2)^{\mathrm{T}}$ can be regarded as mixed type outcomes with y_2 being Bernoulli and y_1 being normal. By Equation 5, the joint density function of $(y_1, y_2)^{\mathrm{T}}$ is

$$f(y_1, y_2) = \frac{1}{\sigma_1}\phi\left(\frac{y_1 - \mu_1}{\sigma_1}\right)\Phi\left(\frac{(y^* - \mu_2)/\sigma_2 - \gamma(y_1 - \mu_1)/\sigma_1}{\sqrt{1-\gamma^2}}\right).$$

4. If both y_1 and y_2 are not censored, then the joint density of them is

$$f(y_1, y_2) = \frac{1}{\sigma_1\sigma_2}\phi_{2,\gamma}\left(\frac{y_1 - \mu_1}{\sigma_1}, \frac{y_2 - \mu_2}{\sigma_2}\right),$$

   where $\phi_{2,\gamma}$ is the density function for standard bivariate normal with correlation parameter γ.

APPENDIX B

We describe a Gauss–Newton type algorithm (Ruppert 2005) to maximize the likelihood L in (6). In the $(l+1)$th iteration, the parameters $\boldsymbol{\theta}$ are updated by

$$\boldsymbol{\theta}^{(l+1)} = \boldsymbol{\theta}^{(l)} + \delta\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \log L_i(\boldsymbol{\theta}^{(l)})}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial \log L_i(\boldsymbol{\theta}^{(l)})}{\partial\boldsymbol{\theta}}\right)^{\mathrm{T}}\right]^{-1}\frac{\partial \log L_i(\boldsymbol{\theta}^{(l)})}{\partial\boldsymbol{\theta}},$$

where δ is the step-halving term that starts at 1 and halves until $\log L(\boldsymbol{\theta}^{(l+1)}) > \log L(\boldsymbol{\theta}^{(l)})$ at iteration l. This algorithm guarantees that the likelihood increases progressively over iterations. The algorithm stops when the increase in the likelihood is no longer possible or the difference between two consecutive updates is smaller than a prespecified precision level.