Statistics and Its Interface Volume 9 (2016) 47–56

Modeling multiple responses via bootstrapping margins with an application to genetic association testing

Jiwei Zhao and Heping Zhang*,†

The need for analysis of multiple responses arises from many applications. In behavioral science, for example, comorbidity is a common phenomenon where multiple disorders occur in the same person. The advantage of jointly analyzing multiple correlated responses has been examined and documented. Due to the difficulties of modeling multiple responses, nonparametric tests such as generalized Kendall's Tau have been developed to assess the association between multiple responses and risk factors. These procedures have been applied to genomewide association studies of multiple complex traits. Unfortunately, those nonparametric tests only provide the significance of the association but not the magnitude. We propose a Gaussian copula model with discrete margins for modeling multivariate binary responses. This model separates marginal effects from between-trait correlations. We use a bootstrapping margins approach to constructing Wald's statistic for the association test. Although our derivation is based on the fully parametric Gaussian copula framework for simplicity, the underlying assumptions to apply our method can be weakened. The bootstrapping margins approach only requires the correct specification of the model margins. Our simulation and real data analysis demonstrate that our proposed method not only increases power over some existing association tests, but also provides further insight into genetic association studies of multivariate traits.

AMS 2000 subject classifications: Primary 62P10; secondary 62G09.
Keywords and phrases: Multiple traits, Marginal approach, Bootstrap, Gaussian copula.

1. INTRODUCTION

The advent of high-throughput genotyping technology has led to discoveries of numerous disease genes, commonly through genomewide association studies (GWAS). Most of the GWAS data are analyzed using a single disease trait at a time. However, comorbidity often occurs in genetic studies of complex diseases, particularly mental illness and substance use [33]. Here, comorbidity refers to the occurrence of multiple disorders in the same subject. For example, more than half of the persons with one substance use disorder suffer from another form of mental illness [3]. Thus, it is scientifically important to consider comorbidity in genetic studies. From the statistical perspective, [37] conducted comprehensive simulation studies and demonstrated that analyzing multiple traits together generally improves the statistical power over single-trait based tests.

While it is important and well motivated to analyze multivariate traits jointly, it also raises theoretical and computational challenges. [34] proposed a generalized Kendall's Tau to test the association of any hybrid of dichotomous, ordinal or quantitative traits with a genetic marker. Later, [17] and [36] extended this method to consider the adjustments of covariates. However, the nonparametric tests can only report the significance of the association but not the magnitude. Therefore, it is desirable, and the goal of this article, to establish a parametric framework for genetic association studies of multiple traits.

We propose Gaussian copula model [23, 28] with discrete margins to analyze multivariate binary traits. Our model has several advantages. First, this is a rich class of parametric models that includes some commonly-used models such as the multivariate probit model [2, 1]. Second, our model makes much weaker assumptions than the multivariate probit model does, and hence is more broadly applicable. Third, under the Gaussian copula framework, the model components that characterize the marginal effects and the correlations among the traits are readily separated. Before we can deliver these useful features, we need to resolve the computational challenge. To this end, we propose to fit this model using a two-step semi-parametric approach. In the first step, we compute the maximum marginal likelihood estimator (MMLE) of association coefficients, say, $\tilde{\beta}$. In the second step, we estimate the variance of $\tilde{\beta}$ using the bootstrapping technique [11]. This bootstrapping margins approach does not assume independence of the traits. Since it only requires the correct specification of model margins, this approach is robust to the misspecification of the correlation information among the traits, and it can be extended to any case as long as the model margins are correctly specified.

*Corresponding author.
†The work is supported by the grant R01 DA016750-09 from the National Institute on Drug Abuse.

Although there is a rich literature on modeling multivariate discrete data through copula structure, including [18], [23], [30], and the references therein, our work has distinctive features in model feasibility and computational cost, especially in genetic association testings. As discussed above, our modeling strategy only requires the correct specification of the margins, and is more flexible to use. Also, as shown in Sections 4 and 5, from the computational perspective, our approach is much faster than multivariate probit model, especially when the dimension of multivariate discrete data is large. This feature is particularly appealing in analyzing high-throughput genotyping data.

Copula has become a useful tool in genetic studies. For example, [19] considered a Gaussian copula variance-components method for linkage analysis with nonnormal quantitative traits. [15] introduced a Gaussian copula based approach to modeling the dependence between disease status and secondary phenotypes in case-control association studies. We exploit copula-based methods for further use in genetic studies. In GWAS, the need of analyzing millions of single nucleotide polymorphisms (SNPs) requires the algorithm for each single SNP would be extremely fast, which is the main motivation of our work. Our proposal of Gaussian copula framework guarantees the MMLE is consistent and asymptotically normal [28]. To account for the ignorance of potential correlations among multiple traits, we further propose Bootstrap method correcting for the variance estimation. Although our method of using MMLE seems simple for Gaussian copula model itself, it works very fast and shows substantial power gain over some nonparametric tests in our numerical studies. More importantly, through the analysis of a real data set on comorbidity, our proposed method identifies some significant SNP biomarkers reported in previous related studies, illustrating the usefulness of our proposed method.

This paper is organized as follows. We establish our model in Section 2. In Section 3, we describe our two-step semiparametric estimation method for association testing. We present our simulation studies in Section 4 and the SAGE data analysis in Section 5. We compare our analysis results with those based on multivariate probit model and another existing nonparametric method [36]. We also provide the estimates and their standard errors for genetic associations, which reveal further scientific details for GWAS. The article ends with a discussion in Section 6.

2. THE GAUSSIAN COPULA MODEL

Copula, a multivariate distribution function with uniformly distributed margins, is a useful tool for modeling correlated variables. For the general introduction and application of copula, we refer the readers to [23]. A common choice is Gaussian copula, which is constructed from a multivariate normal distribution using Sklar's theorem. Specifically, the d-variate Gaussian copula is

$$C_\Phi(u_1,\ldots,u_d|\Gamma)=\Phi_d\{\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_d)|\Gamma\},$$

where $u_i\in[0,1]$, Φ is the cumulative distribution function (c.d.f.) of a standard normal distribution, and Φ_d represents the c.d.f. of d-dimensional normal vector with mean zero and covariance matrix Γ. For instance, [28] used the Gaussian copula to construct a class of multivariate dispersion models for d-dimensional multivariate data $(y_1,\ldots,y_d)$ with marginal distributions $F_1,\ldots,F_d$. That is,

$$\begin{aligned}&C_\Phi(F_1(y_1),\ldots,F_d(y_d)|\Gamma)\\&=\Phi_d\{\Phi^{-1}(F_1(y_1)),\ldots,\Phi^{-1}(F_d(y_d))|\Gamma\}.\end{aligned}\tag{1}$$

Motivated by genetic case-control studies of complex diseases, here we concentrate on the modeling of multiple binary traits $W=(W^{(1)},\ldots,W^{(L)})^T$. By taking Radon-Nikodym derivative for $C_\Phi(F_1(y_1),\ldots,F_L(y_L)|\Gamma)$ in (1) with respect to the counting measure, we can show that

$$\begin{aligned}&P(W^{(1)}=w_1,\ldots,W^{(L)}=w_L)\\&=\sum_{j_1=0}^{1}\cdots\sum_{j_L=0}^{1}(-1)^{j_1+\cdots+j_L}C_\Phi(u_{1j_1},\ldots,u_{Lj_L}|\Gamma),\end{aligned}\tag{2}$$

where $w_l=0$ or 1, $u_{l0}=F_l(w_l)$, $u_{l1}=F_l(w_l-1)$, and F_l is the c.d.f. for $W^{(l)}$, i.e.,

$$F_l(s)=\begin{cases}0 & s<0\\ 1-p_l & 0\le s<1\\ 1 & s\ge 1,\end{cases}$$

where $p_l=P(W^{(l)}=1)$.

This model setting includes many commonly-used models. For example, the bivariate probit model has the following probability mass function:

$$\begin{aligned}&P(W^{(1)}=w_1,W^{(2)}=w_2)\\&=\begin{cases}\Phi_2(\Phi^{-1}(1-p_1),\Phi^{-1}(1-p_2)|\Gamma)\\ \qquad\qquad\text{if } w_1=0,\ w_2=0,\\ 1-p_1-\Phi_2(\Phi^{-1}(1-p_1),\Phi^{-1}(1-p_2)|\Gamma)\\ \qquad\qquad\text{if } w_1=0,\ w_2=1,\\ 1-p_2-\Phi_2(\Phi^{-1}(1-p_1),\Phi^{-1}(1-p_2)|\Gamma)\\ \qquad\qquad\text{if } w_1=1,\ w_2=0,\\ p_1+p_2+\Phi_2(\Phi^{-1}(1-p_1),\Phi^{-1}(1-p_2)|\Gamma)-1\\ \qquad\qquad\text{if } w_1=1,\ w_2=1.\end{cases}\end{aligned}$$

Let G denote a variable of interest (e.g., a genetic marker) and X be a p-vector of covariates. To model the marginal effects of G and X on $W^{(l)}$, we consider a generalized linear model [GLM, 22], i.e., for each l,

$$g(p_l)=\eta_l=\alpha_l+\beta_l G+\gamma_l^T X,\tag{3}$$

where g is the link function. The choices include $\log\{t/(1-t)\}$ (logit), $\Phi^{-1}(t)$ (probit), and $\log\{-\log(1-t)\}$ (complementary log-log). Note that different choices of link functions, i.e., models for margins, will not affect the Gaussian copula correlation structure. Our model is flexible in that

every single $W^{(l)}$ can have its own distinct choice of link function. In addition, $\alpha = (\alpha_1, \ldots, \alpha_L)^T$, $\beta = (\beta_1, \ldots, \beta_L)^T$, and γ is $L \times p$ matrix with γ_l^T as its l-th row. For convenience, we also introduce $\theta = (\alpha, \beta, \gamma)$ as the $L \times (p+2)$ matrix, with θ_l^T as its l-th row.

The β-coefficients reflect the association between W and G. The hypothesis of great interest is

$$H_0 : \beta = 0 \text{ versus } H_1 : \beta \neq 0. \tag{4}$$

3. BOOTSTRAPPING THE MARGINS

The maximum likelihood based approaches, based on the maximum joint likelihood estimator $\hat{\beta}$, are generally used to test the hypothesis in (4). However, these approaches are difficult to implement because the likelihood function under the Gaussian copula model can be so complicated that there do not exist effective methods to compute $\hat{\beta}$ and $\mathrm{Var}(\hat{\beta})$. To overcome this difficulty, we adopt the following two-step procedure:

1. Compute $\tilde{\beta}$, the maximum marginal likelihood estimator (MMLE);
2. Bootstrap samples, and execute Step 1 repeatedly to obtain an estimator of $\mathrm{Var}(\tilde{\beta})$, $\widehat{\mathrm{Var}(\tilde{\beta})}$. Then use the following Wald statistic to test (4):

$$\tilde{\beta}^T \widehat{\mathrm{Var}(\tilde{\beta})}^{-1} \tilde{\beta}.$$

3.1 Step 1

From the general discussion of Gaussian copula model, the MMLE $\tilde{\beta}$ is consistent and asymptotically normal [28]. The assumed regularity conditions are quite standard and easily satisfied. Specifically,

$$\tilde{\beta} \xrightarrow{p} \beta, \ \sqrt{n}(\tilde{\beta} - \beta) \xrightarrow{d} N(0, \Omega_m), \tag{5}$$

where Ω_m is asymptotic covariance matrix of $\tilde{\beta}$.

Let $\hat{\beta}$ be the maximum likelihood estimate of β using the joint likelihood and Ω_j be its asymptotic covariance matrix. It is of interest to find out under what conditions that $\tilde{\beta} = \hat{\beta}$ and $\Omega_m = \Omega_j$. In general, we expect a tradeoff in computation and efficiency. In theory, $\hat{\beta}$ is more efficient than $\tilde{\beta}$, but we prefer $\tilde{\beta}$ for the computational sake provided that the efficiency loss is relatively small. Interestingly, our numerical studies reveal a very small level of the efficiency loss by $\tilde{\beta}$.

An alternative of $\tilde{\beta}$ in the first step could be the estimator from maximizing pairwise composite likelihood using the method of inference functions for margins (IFM), as discussed in [18] and [35]. If this alternative is adopted, correct specification of pairwise composite likelihood is required, which is stronger than the correct specification of univariate margin likelihood. In addition, our method is different from IFM on two aspects: first, IFM computes the estimators from inference functions, while we concentrate on the marginal likelihood; second, IFM uses Jackknife for variance estimation, whose computational cost is getting greater as the sample size increases, while we propose the Bootstrap, as discussed in the following.

3.2 Step 2

Under H_0, we have

$$\tilde{\beta}^T \mathrm{Var}(\tilde{\beta})^{-1} \tilde{\beta} \xrightarrow{d} \chi^2_L.$$

Here we propose to use the bootstrap procedure [11] to estimate $\mathrm{Var}(\tilde{\beta})$. The simplicity of bootstrap makes it very straightforward to use in various applications for deriving standard errors and confidence intervals. Asymptotically, bootstrap is more accurate than the standard intervals obtained using sample variance and assumptions of normality [9]. Bootstrap can be easily extended to more complex scenarios. For example, it can be applied if our null hypothesis H_0 in (4) has a more complicated structure of unknown parameters. It can also be applied for the variance estimation of maximum pairwise composite likelihood estimator.

We compute $\tilde{\beta}_{*b}$ in the b-th bootstrap sample (W_{*b}, G_{*b}, X_{*b}), $b = 1, \ldots, B$. Then,

$$\widehat{\mathrm{Var}(\tilde{\beta})} = \mathrm{Var}(\tilde{\beta}_{*1}, \ldots, \tilde{\beta}_{*B}). \tag{6}$$

Under our Gaussian copula model and standard regularity conditions [26, 27], $\tilde{\beta}$ is consistent and asymptotically normal as presented in (5), and the Bootstrap variance estimator $\widehat{\mathrm{Var}(\tilde{\beta})}$ is also consistent. As suggested by [12], we chose $B = 200$ and also tried $B = 500$ and $B = 1{,}000$ to validate that B is sufficiently large.

4. SIMULATION STUDIES

In this section, we conduct simulation studies to compare the finite sample performance of our proposed method: Bootstrapping the Margins with Probit link (BMP), with multivariate probit model (mvProbit) and nonparametric test based on generalized Kendall's Tau (gKT, 36). Note that, from the construction of our Gaussian copula model in Section 2, the proposed method can be applied with any link function at each margin. In the simulation studies, we concentrate on probit link function only for comparison reasons. Different link functions will be explored in SAGE data analysis. The objective is two-fold. First, we compare the performance of the three methods in terms of the type I error rate and power. Three nominal levels of significance, $\alpha = 0.05$, 0.01 and 0.001, are used. Second, we examine and compare the parameter estimates obtained from the two parametric methods.

4.1 Settings

We generate the multivariate binary trait W following the Gaussian copula model. We use two sample sizes: 500 and

1,000. We first generate the environmental covariate X from normal distribution $N(1,1)$ and the test-locus genotype G from the distribution with the following probability mass function

$$P(G=0)=0.5,\ P(G=1)=0.4,\ P(G=2)=0.1,$$

which mimics the distribution of SNP biomarker `rs1573178` in Gene STXBP1 of Chromosome 9, one of the most interesting findings in previous SAGE data studies. For simplicity, we only consider $L=2$ in this section, and we set

$$\theta=\begin{pmatrix}-0.5 & 0.25 & 0.5\\ -0.5 & 0.25 & 0.5\end{pmatrix}.$$

The probability mass function of bivariate trait W can be written as

$$\begin{aligned}&P(W^{(1)}=w_1,W^{(2)}=w_2)\\ &=\begin{cases}C_\Phi(1-p_1,1-p_2|\Gamma) & \text{if } w_1=0,\ w_2=0\\ 1-p_1-C_\Phi(1-p_1,1-p_2|\Gamma) & \text{if } w_1=0,\ w_2=1\\ 1-p_2-C_\Phi(1-p_1,1-p_2|\Gamma) & \text{if } w_1=1,\ w_2=0\\ p_1+p_2+C_\Phi(1-p_1,1-p_2|\Gamma)-1 & \text{if } w_1=1,\ w_2=1,\end{cases}\end{aligned}$$

where $\Gamma=\left(\begin{smallmatrix}1&\rho\\ \rho&1\end{smallmatrix}\right)$. We set $\rho=0.5$ in this section.

For the marginal models, we consider two scenarios with probit and logit link functions respectively, as

$$p_l=\Phi(\eta_l),\ p_l=\frac{\exp(\eta_l)}{1+\exp(\eta_l)},$$

where $\eta_l=\alpha_l+\beta_l G+\gamma_l X$, $l=1,2$. We should note that, in the scenario with the probit link, both mvProbit and BMP happen to use the correct link function, which is not the case with the logit link.

4.2 Results

To evaluate the performance of the tests, we first report the empirical type I error rate for both scenarios based on 10,000 simulation replications in Table 1. Although gKT is slightly more conservative, whose type I error is almost universally smaller than the corresponding α value, all three methods approximately control the type I error at each nominal level, for both correct and misspecified link scenarios. For the misspecified link scenario, the probit link (used in the model fitting) and logit link (used in the data generation) can be numerically similar [6]. As a result, mvProbit and BMP perform similarly, and can control their respective type I error rates.

In Table 2, we summarize the power comparison results based on 1,000 replications. With the correct link, the power of mvProbit is slightly greater than BMP; however, the difference fades as the sample size increases. With a misspecified link, it reduces the power for both mvProbit and BMP. The two methods perform similarly, and are superior to the nonparametric gKT method.

Tables 3–4 report the results for the parameter estimates for the two scenarios, respectively, based on 1,000 replications. In each replication, we calculate the estimate, the bias, the standard error (SE), and coverage probability (CP) of approximately 95% confidence interval of the parameter, using estimate ± 1.96SE. The reported Bias, SE and CP are averaged across 1,000 simulation runs. We also report Monte Carlo approximation of the standard deviation (SD) of the parameter estimate across 1,000 runs. In Table 3, with the

Table 1. Type I Error Rate in Two Scenarios Based on 10,000 Replications. gKT: nonparametric test based on generalized Kendall's Tau. mvProbit: multivariate Probit modeling approach. BMP: the proposed Bootstrapping Margins method with Probit link

Scenario	Sample Size	Method	α		
			0.050	0.010	0.001
Correct Link	$n=500$	gKT	0.040	0.007	0.001
		mvProbit	0.049	0.009	0.001
		BMP	0.049	0.011	0.001
	$n=1000$	gKT	0.042	0.008	0.001
		mvProbit	0.049	0.009	0.001
		BMP	0.052	0.011	0.002
Misspecified Link	$n=500$	gKT	0.045	0.009	0.001
		mvProbit	0.051	0.010	0.001
		BMP	0.051	0.0112	0.001
	$n=1000$	gKT	0.046	0.009	0.001
		mvProbit	0.051	0.012	0.001
		BMP	0.054	0.012	0.001

Table 2. Power Comparison in Two Scenarios Based on 1,000 Replications. gKT: nonparametric test based on generalized Kendall's Tau. mvProbit: multivariate Probit modeling approach. BMP: the proposed Bootstrapping Margins method with Probit link

Scenario	Sample Size	Method	α		
			0.050	0.010	0.001
Correct Link	$n=500$	gKT	0.808	0.601	0.309
		mvProbit	0.873	0.722	0.460
		BMP	0.875	0.708	0.450
	$n=1000$	gKT	0.994	0.959	0.820
		mvProbit	0.998	0.979	0.906
		BMP	0.998	0.976	0.906
Misspecified Link	$n=500$	gKT	0.451	0.229	0.062
		mvProbit	0.488	0.256	0.088
		BMP	0.484	0.259	0.087
	$n=1000$	gKT	0.781	0.552	0.245
		mvProbit	0.795	0.584	0.314
		BMP	0.788	0.578	0.301

Table 3. Parameter Estimates Comparison for Correct Link Scenario. mvProbit: multivariate Probit modeling approach. BMP: the proposed Bootstrapping Margins method with Probit link. Bias: the average of biases across 1,000 simulation runs. SE: the average of standard errors across 1,000 simulation runs. SD: Monte Carlo approximation of the standard deviation. CP: coverage probability of approximately 95% confidence intervals

			α_1	β_1	γ_1	α_2	β_2	γ_2	ρ
$n = 500$	mvProbit	Bias	-0.0034	0.0057	0.0047	-0.0051	0.0073	0.0030	0.0022
		SE	0.1031	0.0913	0.0654	0.1028	0.0910	0.0651	0.0623
		SD	0.1038	0.0926	0.0662	0.1031	0.0911	0.0654	0.0601
		CP(%)	94.7	95.6	94.3	95.7	95.1	95.1	95.5
	BMP	Bias	-0.0029	0.0056	0.0042	-0.0039	0.0068	0.0018	
		SE	0.1024	0.0907	0.0647	0.1024	0.0906	0.0646	
		SD	0.1040	0.0928	0.0663	0.1034	0.0913	0.0654	
		CP(%)	94.1	95.5	94.0	95.2	94.9	94.9	
$n = 1000$	mvProbit	Bias	-0.0012	0.0024	0.0015	-0.0026	0.0020	0.0029	0.0036
		SE	0.0724	0.0640	0.0458	0.0722	0.0639	0.0457	0.0439
		SD	0.0729	0.0634	0.0457	0.0719	0.0614	0.0457	0.0460
		CP(%)	95.1	95.5	95.2	95.6	96.3	95.2	92.7
	BMP	Bias	-0.0012	0.0024	0.0015	-0.0018	0.0016	0.0021	
		SE	0.0722	0.0638	0.0456	0.0722	0.0638	0.0456	
		SD	0.0729	0.0637	0.0458	0.0720	0.0616	0.0459	
		CP(%)	95.2	95.4	94.8	95.5	96.4	95.2	

Table 4. Parameter Estimates Comparison for Misspecified Link Scenario. mvProbit: multivariate Probit modeling approach. BMP: the proposed Bootstrapping Margins method with Probit link. Bias: the average of biases across 1,000 simulation runs. SE: the average of standard errors across 1,000 simulation runs. SD: Monte Carlo approximation of the standard deviation. CP: coverage probability of approximately 95% confidence intervals

			α_1	β_1	γ_1	α_2	β_2	γ_2	ρ
$n = 500$	mvProbit	Bias	0.1848	-0.0928	-0.1871	0.1914	-0.0970	-0.1913	0.0014
		SE	0.0979	0.0875	0.0600	0.0976	0.0874	0.0597	0.0595
		SD	0.1025	0.0886	0.0619	0.1000	0.0893	0.0603	0.0600
		CP(%)	51.1	81.3	14.6	50.4	79.8	12.1	94.7
	BMP	Bias	0.1852	-0.0929	-0.1874	0.1923	-0.0973	-0.1920	
		SE	0.0975	0.0871	0.0596	0.0974	0.0871	0.0595	
		SD	0.1027	0.0887	0.0620	0.1002	0.0896	0.0605	
		CP(%)	50.2	80.9	13.8	49.7	79.7	11.7	
$n = 1000$	mvProbit	Bias	0.1931	-0.0958	-0.1907	0.1893	-0.0943	-0.1882	0.0043
		SE	0.0689	0.0616	0.0421	0.0688	0.0614	0.0420	0.0419
		SD	0.0707	0.0605	0.0421	0.0701	0.0627	0.0418	0.0432
		CP(%)	20.6	67.1	0.5	20.6	65.7	0.5	93.2
	BMP	Bias	0.1932	-0.0958	-0.1908	0.1898	-0.0947	-0.1888	
		SE	0.0688	0.0615	0.0420	0.0688	0.0615	0.0421	
		SD	0.0706	0.0605	0.0421	0.0702	0.0628	0.0418	
		CP(%)	20.6	66.4	0.5	20.5	65.9	0.6	

correct link, mvProbit and BMP perform well and similarly, and the CP is always around 95%. However, in Table 4, with a misspecified link, the two methods perform similarly again although not surprisingly, they yield large biases and unreliable CPs.

In terms of computation time, the benefit from BMP in the bivariate trait case is not obvious, and it becomes greater as the dimension of the trait increases. For example, when we consider a 6-dimension binary trait as presented in Section 5, mvProbit uses about 12 times of computing time as BMP does. We did not investigate the comparison between BMP and mvProbit for a higher-dimensional-trait case in this section, mainly due to the computational complexity of mvProbit, when the dimension is high.

5. APPLICATION TO THE SAGE DATA

5.1 Background

The Study of Addiction: Genetics and Environment (SAGE) aims to identify susceptible genetic factors that contribute to substance dependence through a large scale

Table 5. Descriptive Statistics of Substance Dependence for Each Subpopulation in SAGE. alc: alcohol dependence; coc: cocaine dependence; mj: marijuana dependence; nic: nicotine dependence; op: opiates dependence; oth: dependence on other drugs

Subpopulation	Total	Substance Dependence					
		alc(%)	coc(%)	mj(%)	nic(%)	op(%)	oth(%)
Black Men	535	332(62.1)	248(46.4)	136(25.4)	254(47.5)	44(8.2)	61(11.4)
Black Women	568	224(39.4)	206(36.3)	78(13.7)	271(47.7)	35(6.2)	37(06.5)
White Men	1131	704(62.3)	309(27.3)	285(25.2)	528(46.7)	112(9.9)	203(18.0)
White Women	1393	433(31.1)	174(12.5)	121(08.7)	572(41.1)	67(4.8)	131(09.4)
Overall	3627	1693(46.7)	937(25.8)	620(17.1)	1625(44.8)	258(7.1)	432(11.9)

genomewide association study. The SAGE data include 4,121 European and African Americans for whom the addiction of alcohol, nicotine, marijuana, cocaine, opiates, and other drugs and genomewide SNP data (ILLUMINA Human 1M platform) are available. The SAGE data set is composed of three separate studies: the Collaborative Study on the Genetics of Alcoholism (COGA), the Family Study of Cocaine Dependence (FSCD), and the Collaborative Genetic Study of Nicotine Dependence (COGEND). The dependence of each subject on these six categories of substances was diagnosed in accordance with the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV). The main results of SAGE data can be accessed from [24], [14], [21], [4].

We hypothesize that there exist common genetic factors (SNPs) for the comorbidity including the addiction to all six categories of substances. We thus use multivariate binary trait, each representing for whether or not the subject is addicted to a single substance.

In our study, we exclude 60 duplicate genotype samples and remove nine subjects with ethnic backgrounds other than African-origin (black) or European-origin (white). After excluding the samples with call rate below 90%, we have 3,627 unrelated subjects for whom we have both genotype and phenotype data. Motivated by [7], we separate the whole data set for both race (black or white) and gender (female or male), to allow for racial and gender specific genetic effects. The whole data set is composed of: 1,393 white women, 1,131 white men, 568 black women, and 535 black men [7]. We filter SNPs by setting thresholds for call rate (>90%), minor allele frequency (MAF) (>1%), and Hardy-Weinberg equilibrium (p-value > 0.001).

5.2 Data analysis

We first provide the substance dependence distribution by gender and race in Table 5. It can be seen that the dependence to some substances, for example, nicotine dependence, is homogeneous across the four subpopulations, while other substance dependencies differ by gender (e.g., alcohol dependence, marijuana dependence) or race (e.g., cocaine dependence). Therefore, we concur with the strategy of [7] by examining the association within each of the four racial and gender groups, removing racial and gender heterogeneity.

Following [32], we concentrate on chromosome 9. To incorporate the effect of age and population stratification, we include the first principal component score and age as two covariates, which constitute the X variable in our model. We analyze the data using our proposed method with probit link (BMP), logit link (BML), multivariate probit model (mvProbit), and the nonparametric method based on generalized Kendall's Tau (gKT). The top SNPs with their p-values are summarized in Table 6, and the genetic association coefficients estimates are summarized in Table 7.

Previous animal studies suggested that STXBP1 (an ortholog of human STXBP1) may be linked to an alcohol preference drinking locus on a mouse chromosome [13]. Moreover, more recent studies [25, 8] suggested that mutations in gene STXBP1 are associated with early infantile epileptic encephalopathy with suppression-burst (EIEE), also known as Ohtahara syndrome, which is one of the most severe and earliest forms of epilepsy. [16] identified that SNP rs1573178 in gene STXBP1 is significant for substance use dependence in black men, although the corresponding p-value, 6.44×10^{-7}, is slightly larger than the commonly accepted genomewide significance level 5×10^{-7} [5]. Thus, we specifically assess the association of this SNP with substance dependence. With 200 bootstrap samples, the p-values from our methods are 7.81×10^{-7} and 7.24×10^{-7} when the probit and logit links are used. The p-values are relatively stable when we choose more bootstrap samples. We should note that, the p-values calculated from our method only need an adjustment for multiple genetic markers across the whole genome, thus, in this section, we simply compare each single p-value with the genomewide significance level, demonstrated by [5].

PTPRD is another gene that received a great deal of attention in the literature. The protein encoded by this gene is a member of the protein tyrosine phosphatase (PTP) family. In neuroblastoma, PTPRD was identified as a candidate tumor suppressor gene [29]. [20] reported that PTPRD was associated with nicotine dependence through a genome wide linkage scan. More recently, through GWAS approaches, PTPRD was identified to be associated with

Table 6. SNPs with p-values on Chromosome 9 for Multivariate Substance Dependencies in SAGE. gKT: nonparametric test based on generalized Kendall's Tau. mvProbit: multivariate Probit modeling approach. BMP2, BMP5, BMP10: Bootstrapping Margins method with Probit link and 200, 500, 1,000 bootstrap samples respectively. BML2, BML5, BML10: Bootstrapping Margins method with Logit link and 200, 500, 1,000 bootstrap samples respectively

SNP	MAF	Gene	p-values							
			gKT	mvProbit	BMP2	BMP5	BMP10	BML2	BML5	BML10
					Black Men					
rs1573178	0.295	STXBP1	6.44e-07	1.27e-05	7.81e-07	1.79e-06	4.69e-06	7.24e-07	1.52e-06	4.96e-06
					Black Women					
rs10977327	0.042	PTPRD	3.66e-02	2.78e-02	3.93e-08	3.63e-07	5.04e-07	1.08e-07	1.14e-06	1.60e-06
rs716573	0.068		1.78e-02	4.28e-02	6.62e-07	2.57e-07	4.32e-07	8.90e-07	6.01e-07	8.46e-07
rs2596412	0.019		4.02e-02	1.33e-02	9.42e-07	1.13e-07	6.15e-08	1.35e-06	2.17e-07	1.25e-07
					White Women					
rs7856948	0.018	PTPRD	6.04e-02	2.20e-02	6.12e-11	1.47e-10	5.05e-12	1.22e-09	2.01e-09	1.08e-10
rs12000151	0.013		7.54e-02	7.01e-01	8.85e-09	1.86e-08	8.19e-09	1.65e-08	4.68e-08	1.52e-08
rs7019602	0.047	C9orf3	3.69e-02	6.48e-02	8.60e-08	1.78e-07	1.36e-07	3.34e-07	6.53e-07	4.96e-07
rs12004497	0.013	PALM2-AKAP2	5.94e-02	4.89e-01	1.09e-07	1.44e-06	2.84e-07	7.05e-08	6.88e-07	1.73e-07
rs1543185	0.020	FBP2	3.19e-02	3.91e-01	1.44e-07	2.36e-07	9.10e-08	2.13e-07	7.85e-07	3.03e-07

smoking cessation success by [31]. Comorbid depressive syndrome and alcohol dependence were reported to be associated with PTPRD in [10]. Based on such mounting evidence, the PTPRD gene is believed to be strongly associated with addiction-related traits [10]. Again, we re-evaluate this association using our methods.

It turns out that our findings are consistent with the existing literature. Specifically, based on our bootstrapping margins approach with the probit link and 1,000 bootstrap samples, SNP rs10977327 in the PTPRD gene is significant in African-origin women (p-value $= 5.04 \times 10^{-7}$) and SNP rs7856948 in the PTPRD gene is highly significant in European-origin women (p-value $= 5.05 \times 10^{-12}$). Note that no SNP markers in the PTPRD gene are identified to be significant in the men cohort, indicating a gender specific effect.

In summary, we successfully identified gene PTPRD as strongly associated with substance dependence in African-origin and European-origin women, while other methods failed to uncover this association. It is worth mentioning that this finding is consistent with previous studies, where PTPRD was found to be associated with alcohol dependence, nicotine dependence, and other addiction-related traits.

6. DISCUSSION

Modeling multivariate dichotomous outcomes is important and challenging. Although a fully parametric approach is possible in principle, the dependence among the outcomes makes it complicated to compute the joint likelihood and obtain the maximum likelihood estimates. We propose a Gaussian copula model with discrete margins that enables us to model the binary outcomes jointly, and more distinctly, separate the estimation of the marginal parameters pertinent to each trait from the complicated and unknown dependence structure. This semi-parametric approach utilizes commonly used link functions including probit and logit for binary responses in the marginal model, for which there exist well-tested computational methods and algorithms, avoiding the complications with the joint likelihood. Not only do we obtain consistent estimates for the marginal parameters, but also we adopt the bootstrap method to obtain consistent variance estimation based on the marginal parameter estimates. While we present our method for multivariate binary outcomes, we expect it can be extended to analyze a hybrid of continuous and discrete outcomes. The specific development warrants a separate effort in the future.

Our simulation studies demonstrate that our proposed method and mvProbit perform similarly and the difference is negligible for both power and parameter estimation comparisons, regardless of whether the correct link function is used. This is reassuring because the advantage of our proposed method is to dramatically reduce the computation without compromising the performance. Our simulation results suggest that we accomplished this objective.

As expected, our proposed method is computationally more efficient when the dimension of the trait increases, as evident from the analysis of the SAGE data. Compared to parametric mvProbit, our method is 12 times faster; and compared to nonparametric gKT, our method provides additional and essential results that are not available from the nonparametric test. Specifically, our method estimates the strength and direction of the association as well as its precision. Besides, our method is convenient for real data anal-

Table 7. Genetic Association Coefficients Estimates Comparison for SNPs on Chromosome 9

SNP	Gene			alc	coc	mj	nic	op	oth
				Black Men					
rs1573178	STXBP1	mvProbit	Est	-0.134	0.146	0.149	0.270	0.062	0.370
			SE	0.089	0.089	0.094	0.090	0.143	0.120
		BMP	Est	-0.111	0.145	0.148	0.265	0.057	0.375
			SE	0.087	0.086	0.092	0.086	0.123	0.110
		BML	Est	-0.179	0.232	0.255	0.423	0.122	0.711
			SE	0.139	0.138	0.156	0.139	0.245	0.205
				Black Women					
rs10977327	PTPRD	mvProbit	Est	-0.566	-0.190	-1.266	-0.418	0.001	0.020
			SE	0.242	0.221	0.532	0.197	0.493	0.359
		BMP	Est	-0.547	-0.200	-0.978	-0.404	0.002	0.006
			SE	0.212	0.202	0.415	0.195	0.295	0.297
		BML	Est	-0.906	-0.329	-2.102	-0.655	0.005	0.021
			SE	0.366	0.338	1.019	0.319	0.617	0.618
rs716573		mvProbit	Est	-0.444	-0.342	-0.003	-0.337	-1.015	-0.831
			SE	0.175	0.162	0.220	0.165	0.497	0.375
		BMP	Est	-0.443	-0.326	0.065	-0.316	-0.743	-0.780
			SE	0.166	0.165	0.191	0.157	0.398	0.400
		BML	Est	-0.726	-0.543	0.114	-0.504	-1.714	-1.779
			SE	0.280	0.277	0.351	0.254	1.020	1.021
rs2596412		mvProbit	Est	-0.337	-0.655	0.277	-0.526	-1.693	-0.043
			SE	0.903	0.987	1.555	0.500	0.794	2.858
		BMP	Est	-0.156	-0.125	0.593	-0.417	-0.167	-0.147
			SE	0.292	0.300	0.306	0.292	0.483	0.483
		BML	Est	-0.253	-0.211	1.035	-0.672	-0.343	-0.319
			SE	0.478	0.499	0.516	0.477	1.052	1.054
				White Women					
rs7856948	PTPRD	mvProbit	Est	-0.881	0.186	-0.143	-0.403	-0.495	-0.579
			SE	0.314	0.292	0.316	0.203	0.495	0.386
		BMP	Est	-0.682	0.207	-0.021	-0.389	-0.383	-0.438
			SE	0.231	0.219	0.266	0.193	0.420	0.326
		BML	Est	-1.230	0.351	-0.044	-0.633	-0.870	-0.925
			SE	0.440	0.400	0.534	0.322	1.021	0.729
rs12000151		mvProbit	Est	-0.182	-0.284	-0.497	-0.314	-0.371	0.309
			SE	0.251	0.406	0.421	0.244	0.445	0.332
		BMP	Est	-0.141	-0.160	-0.601	-0.275	-0.346	0.290
			SE	0.221	0.277	0.424	0.215	0.443	0.246
		BML	Est	-0.229	-0.287	-1.345	-0.445	-0.712	0.528
			SE	0.369	0.526	1.011	0.352	1.008	0.443
rs7019602	C9orf3	mvProbit	Est	-0.230	-0.088	0.221	-0.044	-0.806	0.013
			SE	0.130	0.149	0.161	0.120	0.391	0.171
		BMP	Est	-0.223	-0.080	0.234	-0.048	-0.832	0.019
			SE	0.123	0.148	0.144	0.114	0.377	0.153
		BML	Est	-0.367	-0.154	0.425	-0.078	-1.972	0.028
			SE	0.208	0.281	0.270	0.184	1.008	0.298
rs12004497	PALM2-AKAP2	mvProbit	Est	0.023	-0.311	-0.235	0.149	0.686	-0.237
			SE	0.242	0.341	0.385	0.240	0.411	0.332
		BMP	Est	0.079	0.079	0.026	0.147	0.760	0.026
			SE	0.218	0.257	0.289	0.213	0.251	0.283
		BML	Est	0.139	0.167	0.094	0.235	1.494	0.056
			SE	0.354	0.462	0.546	0.341	0.450	0.542
rs1543185	FBP2	mvProbit	Est	0.107	0.146	-0.648	-0.092	0.100	-0.744
			SE	0.184	0.209	0.391	0.182	0.269	0.466
		BMP	Est	0.119	0.130	-0.719	-0.074	0.135	-0.787
			SE	0.180	0.217	0.406	0.177	0.283	0.409
		BML	Est	0.201	0.242	-1.630	-0.121	0.217	-1.736
			SE	0.293	0.397	1.016	0.287	0.612	1.014

ysis. We allow different link functions for different margins, if necessary and reasonable for the data.

Lastly, we analyze one single SNP each time in our proposed method. The generalization to incorporating multiple SNPs warrants further effort in the future. Additionally, an application of our method to the SAGE data reveals some significant SNP markers in the genes C9orf3, PALM2-AKAP2 and FBP2, that have not been reported before, demonstrating the usefulness of the proposed parametric framework in genetic association studies.

ACKNOWLEDGEMENTS

We would like to thank a referee and an associate editor for providing helpful comments and suggestions. The work is supported by the grant R01 DA016750-09 from the National Institute on Drug Abuse.

Received 8 April 2014

REFERENCES

[1] Amemiya, T. (1974). Bivariate probit analysis: minimum chi-square methods. *Journal of the American Statistical Association* **69** 940–944.

[2] Ashford, J. and Sowden, R. (1970). Multi-variate probit analysis. *Biometrics* 535–546.

[3] Bierut, L. J., Agrawal, A., Bucholz, K. K., Doheny, K. F., Laurie, C., Pugh, E., Fisher, S., Fox, L., Howells, W., Bertelsen, S. et al. (2010). A genome-wide association study of alcohol dependence. *Proceedings of the National Academy of Sciences* **107** 5082–5087.

[4] Bierut, L. J., Strickland, J. R., Thompson, J. R., Afful, S. E. and Cottler, L. B. (2008). Drug use and dependence in cocaine dependent subjects, community-based individuals, and their siblings. *Drug and Alcohol Dependence* **95** 14.

[5] Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W. H., Samani, N. J. et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447** 661–678.

[6] Chambers, E. A. and Cox, D. R. (1967). Discrimination between alternative binary response models. *Biometrika* **54** 573–578. MR0223058

[7] Chen, X., Cho, K., Singer, B. H. and Zhang, H. (2011). The nuclear transcription factor PKNOX2 is a candidate gene for substance dependence in European-origin women. *PLoS One* **6** e16002.

[8] Deprez, L., Weckhuysen, S., Holmgren, P., Suls, A., Van Dyck, T., Goossens, D., Del-Favero, J., Jansen, A., Verhaert, K., Lagae, L. et al. (2010). Clinical spectrum of early-onset epileptic encephalopathies associated with STXBP1 mutations. *Neurology* **75** 1159–1165.

[9] DiCiccio, T. J. and Efron, B. (1996). Bootstrap confidence intervals (with Discussion). *Statistical Science* 189–228. MR1436647

[10] Edwards, A. C., Aliev, F., Bierut, L. J., Bucholz, K. K., Edenberg, H., Hesselbrock, V., Kramer, J., Kuperman, S., Nurnberger Jr, J. I., Schuckit, M. A. et al. (2012). Genome-wide association study of comorbid depressive syndrome and alcohol dependence. *Psychiatric Genetics* **22** 31–41.

[11] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics* 1–26. MR0515681

[12] Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap* **57**. Chapman & Hall/CRC. MR1270903

[13] Fehr, C., Shirley, R. L., Crabbe, J. C., Belknap, J. K., Buck, K. J. and Phillips, T. J. (2005). The syntaxin binding protein 1 gene (Stxbp1) is a candidate for an ethanol preference drinking locus on mouse chromosome 2. *Alcoholism: Clinical and Experimental Research* **29** 708–720.

[14] Hartel, D. M., Schoenbaum, E. E., Lo, Y. and Klein, R. S. (2006). Gender differences in illicit substance use among middle-aged drug users with or at risk for HIV infection. *Clinical Infectious Diseases* **43** 525–531.

[15] He, J., Li, H., Edmondson, A. C., Rader, D. J. and Li, M. (2012). A Gaussian copula approach for the analysis of secondary phenotypes in case–control genetic association studies. *Biostatistics* **13** 497–508.

[16] Jiang, Y., Li, N. and Zhang, H. (2014). Identifying genetic variants for addiction via propensity score adjusted generalized Kendall's tau. *Journal of the American Statistical Association*. MR3265665

[17] Jiang, Y. and Zhang, H. (2011). Propensity score-based nonparametric test revealing genetic variants underlying bipolar disorder. *Genetic Epidemiology* **35** 125–132.

[18] Joe, H. (1997). *Multivariate models and dependence concepts* **73**. CRC Press. MR1462613

[19] Li, M., Boehnke, M., Abecasis, G. R. and Song, P. X.-K. (2006). Quantitative trait linkage analysis using Gaussian copulas. *Genetics* **173** 2317–2327.

[20] Li, M., Sun, D., Lou, X., Beuten, J., Payne, T. and Ma, J. (2006). Linkage and association studies in African-and Caucasian-American populations demonstrate that SHC3 is a novel susceptibility locus for nicotine dependence. *Molecular Psychiatry* **12** 462–473.

[21] Luo, Z., Alvarado, G. F., Hatsukami, D. K., Johnson, E. O., Bierut, L. J. and Breslau, N. (2008). Race differences in nicotine dependence in the Collaborative Genetic study of Nicotine Dependence (COGEND). *Nicotine & Tobacco Research* **10** 1223–1230.

[22] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2 ed. Chapman & Hall/CRC. MR3223057

[23] Nelsen, R. B. (2006). *An introduction to copulas*. Springer Verlag. MR2197664

[24] Reich, T., Edenberg, H. J., Goate, A., Williams, J. T., Rice, J. P., Van Eerdewegh, P., Foroud, T., Hesselbrock, V., Schuckit, M. A., Bucholz, K. et al. (1998). Genome-wide search for genes affecting the risk for alcohol dependence. *American Journal of Medical Genetics* **81** 207–215.

[25] Saitsu, H., Kato, M., Mizuguchi, T., Hamada, K., Osaka, H., Tohyama, J., Uruno, K., Kumada, S., Nishiyama, K., Nishimura, A. et al. (2008). De novo mutations in the gene encoding STXBP1 (MUNC18-1) cause early infantile epileptic encephalopathy. *Nature Genetics* **40** 782–788.

[26] Shao, J. (1990). Bootstrap estimation of the asymptotic variances of statistical functionals. *Annals of the Institute of Statistical Mathematics* **42** 737–752. MR1089473

[27] Shao, J. and Tu, D. (1995). *The jackknife and bootstrap*. Springer New York. MR1351010

[28] Song, P. X.-K. (2000). Multivariate dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics* **27** 305–320. MR1777506

[29] Stallings, R. L., Nair, P., Maris, J. M., Catchpoole, D., McDermott, M., O'Meara, A. and Breatnach, F. (2006). High-resolution analysis of chromosomal breakpoints and genomic instability identifies PTPRD as a candidate tumor suppressor gene in neuroblastoma. *Cancer Research* **66** 3673–3680.

[30] Trivedi, P. K. and Zimmer, D. M. (2007). *Copula modeling: an introduction for practitioners*. Now Publishers Inc.

[31] Uhl, G. R., Drgon, T., Johnson, C., Li, C.-Y., Contoreggi, C., Hess, J., Naiman, D. and Liu, Q.-R. (2008). Molecular genetics of addiction and related heritable phenotypes. *Annals of the New York Academy of Sciences* **1141** 318–381.

[32] Xu, K., Anderson, T., Neyer, K., Lamparella, N., Jenkins, G., Zhou, Z., Yuan, Q., Virkkunen, M. and Lipsky, R.

(2007). Nucleotide sequence variation within the human tyrosine kinase B neurotrophin receptor gene: association with antisocial alcohol dependence. *The Pharmacogenomics Journal* **7** 368–379.
[33] Zhang, H. (2011). Statistical analysis in genetic studies of mental illnesses. *Statistical Science* **26** 116–129. MR2866281
[34] Zhang, H., Liu, C.-T. and Wang, X. (2010). An association test for multiple traits based on the generalized Kendall's tau. *Journal of the American Statistical Association* **105** 473–481. MR2724840
[35] Zhao, Y. and Joe, H. (2005). Composite likelihood estimation in multivariate data analysis. *Canadian Journal of Statistics* **33** 335–356. MR2193979
[36] Zhu, W., Jiang, Y. and Zhang, H. (2012). Nonparametric covariate-adjusted association tests based on the generalized Kendall's tau. *Journal of the American Statistical Association* **107** 1–11. MR2949337
[37] Zhu, W. and Zhang, H. (2009). Why do we test multiple traits in genetic association studies? (with discussion). *Journal of the Korean Statistical Society* **38** 1–10. MR2656857

Jiwei Zhao
Department of Biostatistics
University at Buffalo
The State University of New York
Buffalo, NY 14214
USA
E-mail address: zhaoj@buffalo.edu

Heping Zhang
Department of Biostatistics
Yale School of Public Health
Yale University
New Haven, CT 06511
USA
E-mail address: heping.zhang@yale.edu