Statistica Sinica **23** (2013), 1505-1521
doi:http://dx.doi.org/10.5705/ss.2012.013s

A COPULA-MODEL BASED SEMIPARAMETRIC INTERACTION TEST UNDER THE CASE-CONTROL DESIGN

Hong Zhang[1], Jing Qin[2], Maria Landi[2], Neil Caporaso[2] and Kai Yu[2]

[1]*Fudan University and* [2]*National Institutes of Health*

Abstract: It is important to study the interaction between risk factors in molecular epidemiology studies. To improve the power for the detection of interaction, some statistical testing procedures have been proposed in the literature by incorporating certain assumptions on the underlying joint distribution of two risk factors. For example, the well known case-only test used in genetic epidemiology studies is derived under the assumption of independence between the two risk factors. However, such testing procedures could have detrimental effects on both false positive and false negative rates when assumptions are not met. We propose a parametric copula function to model the joint distribution while leaving the marginal distributions for the two risk factors unspecified. A unified approach is proposed to estimate/test the interaction effect. This approach is very flexible and can be applied to study the interaction between risk factors that are continuous or discrete. A simulation study finds that the proposed test is generally more powerful than the traditional robust test derived under the standard logistic regression, and without specifying the relationship between the two risk factors. The performance of the proposed approach is comparable with the case-only test when the two risk factors are indeed independent in the control population. Unlike the case-only test, the proposed test can still maintain the type I error rate when the independence assumption is not valid. The application of the proposed procedure is demonstrated through two cancer epidemiology studies.

Key words and phrases: Case-only design, gene-environment interaction, gene-gene interaction, pseudo likelihood.

1. Introduction

In epidemiology studies, it is usually of interest to evaluate whether there is any interaction between two risk factors for the disease of interest. For instance, in genetic epidemiology studies, it is important to study gene-gene and gene-environment interactions in order to better understand the etiology underlying the disease development. The case-control design is in wide use. Under such a retrospective design, information on risk factors and other covariates is collected at fixed numbers of cases and controls. It is well known that the prospective likelihood based on a logistic regression model can be used to estimate the log odds

ratio parameters due to the equivalence of the prospective and retrospective maximum likelihood estimates (Cornfield (1956); Prentice and Pyke (1979)). For the same reason, the prospective likelihood model can also be used to study the interaction by evaluating the coefficient of the product term of two risk factors. The standard logistic regression model is optimal for the assessment of the interaction when the joint distribution of two risk factors is fully nonparametric. However, this standard logistic regression method ignores the relationship between the risk factors and thus could loss power for detecting interaction if such information is available. Piegorsch, Weinberg, and Taylor (1994) proposed a case-only method for detecting interaction that is valid when the disease is rare enough and the two risk factors are independent in the general population, but this method does not allow for the adjustment of additional covariates. Umbach and Weinberg (1997) extended the case-only method to account for categorical covariates. Chatterjee and Carroll (2005) developed a general semiparametric method to detect gene-environment or gene-gene interaction by incorporating the independence assumption for the the risk factors. However, it has been shown that methods derived under the independence assumption can lead to serious inflation in type I error or loss of power if that assumption is not met in the application, and should be used with great caution when the independence assumption is in doubt. To relax the independence assumption, Mukherjee and Chatterjee (2008) proposed an empirical Bayes-type shrinkage estimator by combining the estimate from the standard logistic regression model with the one derived under the independence assumption.

In this paper, we develop a novel approach for detecting interaction by modeling the joint distribution of two risk factors in the control population (or equivalently, the general population if the disease is rare) through a copula model (Nelsen (1999)) while leaving the marginal distributions of the risk factors unspecified. Some authors have modeled the relationship between the risk factors in the general population (Chatterjee and Carroll (2005); Lin and Zeng (2009)), but the parameters might be nearly unidentifiable, as mentioned in their papers, while ours avoids the identifiability problem. The proposed model is very general and covers the scenario in which the two risk factors are independent in the control population. The theory underlying our approach is due to Sklar (1959), stating that there exists a unique copula function characterizing the joint distribution of any two continuous random variables. For discrete-continuous and discrete-discrete settings, we assume that the discrete risk factors are ordinal and can be derived through (unobservable) continuous random variables. Our approach can thus be used to study the interaction of two risk factors that are either continuous or discrete.

The rest of this paper is organized as follows. In Section 2, the logistic regression model and the copula model are described for continuous-continuous,

discrete-continuous, and discrete-discrete risk factors. In Section 3, a two-stage approach is developed for estimating unknown parameters under each setting, where the first stage is to estimate the marginal distributions of the risk factors and the second stage is to maximize the pseudo likelihood function with marginal distributions fixed as those estimated in the first stage. In Section 4, some large sample properties of the pseudo maximum likelihood estimator are established and a bootstrap-based procedure is proposed for estimating the associated variance-covariance matrix of the pseudo maximum likelihood estimators, then one can construct a confidence interval and a Wald test statistic for the interaction. In Sections 5 and 6, the proposed approach is illustrated with a simulation study and two applications. Some final conclusions and remarks are given in Section 7. Proofs are relegated to the appendices.

2. Model Assumption

We consider two risk factors for a disease of interest. The risk factors can be either continuous or discrete. Let X and Y denote the risk factors if continuous-continuous, X^* and Y if discrete-continuous, and X^* and Y^* if discrete-discrete. We first focus on the continuous-continuous case, then generalize the arguments to the other two cases.

A logistic regression model relating disease status D and the two risk factors X and Y is

$$Pr(D = 1|X = x, Y = y) = \frac{\exp(\alpha^* + \beta x + \gamma y + \xi xy)}{1 + \exp(\alpha^* + \beta x + \gamma y + \xi xy)}, \tag{2.1}$$

where α^* is the intercept, β and γ are the main effects, and ξ is the interaction effect. If $f_0(x, y)$ denotes the joint density function of (X, Y) in the control population, then the joint density function of (X, Y) in the case population can be written as (Qin and Zhang (1997))

$$f_1(x, y) = \exp(\alpha + \beta x + \gamma y + \xi xy) f_0(x, y), \tag{2.2}$$

where $\alpha = \alpha^* + \log\{Pr(D = 0)/Pr(D = 1)\}$. As $f_1(x, y)$ is a density function, we have $\exp(-\alpha) = \iint \exp(\beta x + \gamma y + \xi xy) f_0(x, y) dxdy$ so that

$$f_1(x, y) = \frac{\exp(\beta x + \gamma y + \xi xy) f_0(x, y)}{\iint \exp(\beta x + \gamma y + \xi xy) f_0(x, y) dxdy}. \tag{2.3}$$

Based on (2.3), the distribution of (X, Y) in cases can be treated as a re-weighted version of their distribution in controls, with weight $\exp(\alpha + \beta x + \gamma y + \xi xy)$.

If one assumes that X and Y are independent in the control population, then, under the null hypothesis of no interaction effect, X is also independent of Y in the case population. Therefore, to test interaction effect, one can apply the

Pearson chi-square test to the case data. This is the so-called case-only method.

If independence is in doubt, the application of a copula model is a natural choice to model the dependence. In the following, we specify the joint density function $f_0(\cdot,\cdot)$ through a copula function.

In the continuous-continuous case, we assume that X and Y are continuous random variables with marginal distribution functions in the control population of $F_X(x)$ and $F_Y(y)$, respectively. By Sklar's theorem (Sklar (1959)), we can take the joint distribution function of (X,Y) as

$$F_0(x,y) = C(F_X(x), F_Y(y); \theta), \tag{2.4}$$

where $C(u,v;\theta)$ is a copula function known up to a parameter vector θ of finite dimension. The joint density function of (X,Y) in the control population is then

$$f_0(x,y) = c(F_X(x), F_Y(y); \theta) f_X(x) f_Y(y), \tag{2.5}$$

where $c(x,y;\theta) = \partial^2 C(x,y;\theta)/\partial x \partial y$, and $f_X(x) = \partial F_X(x)/\partial x$ and $\partial F_Y(y)/\partial y$ are, respectively, the marginal density functions of X and Y in the control population.

In the discrete-continuous case, we can adopt the approach of de Leon and Wu (2011) and assume that there is a continuous random variable underlying the discrete risk factor. In detail, suppose that the discrete risk factor X^* is an ordinal random variable taking a finite number of values, say $1,\ldots,K$, and that Y is a continuous random variable with distribution function $F_Y(y)$. We assume that there exists an underlying continuous random variable X with distribution function $F_X(x)$ and that, for $-\infty = c_0 < c_1 < \cdots < c_K < c_{K+1} = \infty$,

$$X^* = k, \text{ if } c_k < X \le c_{k+1} \text{ for } k = 0, 1, \ldots, K. \tag{2.6}$$

If the joint distribution function of (X,Y) is $C(F_X(x), F_Y(y);\theta)$, then the joint density function of (X^*,Y) in the control population is

$$f_0^{(1)}(x^*,y) = \{C_2(F_X(c_{k+1}), F_Y(y);\theta) - C_2(F_X(c_k), F_Y(y);\theta)\} f_Y(y) \text{ if } x^* = k \tag{2.7}$$

for $k = 0,1,\ldots,K$ and $-\infty < y < \infty$, where $C_2(u,v;\theta) = \partial C(u,v;\theta)/\partial v$ and $f_Y(y) = \partial F_Y(y)/\partial y$. The joint density function of (X^*,Y) in the case population is

$$f_1^{(1)}(x^*,y) = \frac{f_0^{(1)}(x^*,y)\exp(\beta x^* + \gamma y + \xi x^* y)}{\sum_{k=0}^{K} \int f_0^{(1)}(k,y)\exp(\beta k + \gamma y + \xi k y) dy}. \tag{2.8}$$

In the discrete-discrete case, we further assume that Y^* is an ordinal random variable taking a finite number of values, say $1,\ldots,L$, that there exists an

underlying continuous random variable Y with distribution function $F_Y(y)$, and that for $-\infty = d_0 < d_1 < \cdots < d_L < d_{L+1} = \infty$,

$$Y^* = l, \text{ if } d_l < Y \le d_{l+1} \text{ for } l = 0, 1, \ldots, L. \tag{2.9}$$

If the joint distribution function of (X, Y) is $C(F_X(x), F_Y(y); \theta)$, then the joint density function of (X^*, Y^*) in the control population is

$$\begin{aligned} f_0^{(2)}(x^*, y^*) = {} & C(F_X(c_{k+1}), F_Y(d_{l+1}); \theta) - C(F_X(c_k), F_Y(d_{l+1}); \theta) \\ & - C(F_X(c_{k+1}), F_Y(d_l); \theta) + C(F_X(c_k), F_Y(d_l); \theta), \end{aligned} \tag{2.10}$$

if $(x^*, y^*) = (c_k, d_l)$ for $k = 0, 1, \ldots, K$ and $l = 0, 1, \ldots, L$. The joint density function of (X^*, Y^*) in the case population is

$$f_1^{(2)}(x^*, y^*) = \frac{f_0^{(2)}(x^*, y^*) \exp(\beta x^* + \gamma y^* + \xi x^* y^*)}{\sum_{k=0}^{K} \sum_{l=0}^{L} f_0^{(2)}(c_k, d_l) \exp(\beta c_k + \gamma d_l + \xi c_k d_l)}. \tag{2.11}$$

Remark 1. In the discrete-continuous case, the joint density functions $f_0^{(1)}(x^*, y)$ and $f_1^{(1)}(x^*, y)$ depend on the threshold values $\{c_1, \ldots, c_K\}$ only through $\{F_X(c_1), \ldots, F_X(c_K)\}$ which can be estimated, see the next section. Therefore, we do not need to estimate these threshold values. The same is true for the discrete-discrete case, because $f_0^{(2)}(x^*, y^*)$ and $f_1^{(2)}(x^*, y^*)$ depend on $\{c_1, \ldots, c_K; d_1, \ldots, d_L\}$ only through the estimable probabilities $\{F_X(c_1), \ldots, F_X(c_K); F_Y(d_1), \ldots, F_Y(d_K)\}$.

3. Parameter Estimation

The estimation of regression coefficients β, γ, and ξ is complicated by the presence of the high-dimensional nuisance parameters F_X and F_Y. We adopt a pseudo-likelihood based approach, with F_X and F_Y being estimated in the first stage. The pseudo likelihood method has been well developed for some widely used models, for instance, the parametric model used for a pseudo-likelihood estimation procedure (Gong and Samaniego (1981)), the copula model for multivariate data under a cross-sectional design (Genest, Ghoudi, and Rivest (1995)), and the bivariate survival model (Shih and Louis (1995)).

The detailed parameter estimation procedures for the various cases are presented in the next subsections.

3.1. Continuous-continuous case

Let the observed risk factors for cases and controls be, respectively, $\{(x_{1i}, y_{1i}), i = 1, \ldots, n_1\}$ and $\{(x_{0i}, y_{0i}), i = 1, \ldots, n_0\}$, and the pooled data be $\{(x_i, y_i),$

$i = 1, \ldots, n(= n_1 + n_0)\}$. The log likelihood is

$$l_n(\beta, \gamma, \xi, \theta, F_X, F_Y) = \sum_{i=1}^{n} \Big\{ \log f_X(x_i) + \log f_Y(y_i) + \log c(F_X(x_i), F_Y(y_i); \theta) \Big\} + \sum_{i=1}^{n_1} \Big\{ (\beta x_{1i} + \gamma y_{1i} + \xi x_{1i} y_{1i}) \log \iint c(F_X(x), F_Y(y); \theta) - \exp(\beta x + \gamma y + \xi xy) dF_X(x) dF_Y(y) \Big\}. \quad (3.1)$$

This is difficult to maximize with respect to all the unknown parameters. Instead, we adopt a two-stage algorithm. The first stage is to estimate F_X and F_Y without any constraint on their joint distribution. If the resulting estimators are $\hat{F}_X$ and $\hat{F}_Y$, the second stage maximizes $l_n(\beta, \gamma, \xi, \theta, \hat{F}_X, \hat{F}_Y)$ with respect to $(\beta, \gamma, \xi, \theta)$. Let the resulting pseudo-MLE be $(\hat{\beta}, \hat{\gamma}, \hat{\xi}, \hat{\theta})$.

We consider two types of estimates for F_X and F_Y. One method uses the empirical distribution functions for the control samples; these have some nice large sample properties but do not use the information in the case data. The other is a semiparametric method that utilizes both case and control data and is intuitively more efficient. The idea is to estimate the marginal distribution functions based on empirical likelihood estimate at each observation. The detailed algorithm for this is described as follows.

(i) Maximize

$$\sum_{i=1}^{n_1} (\alpha + \beta x_{1i} + \gamma y_{1i} + \xi x_{1i} y_{1i}) - \sum_{i=1}^{n} \log\{n_0 + n_1 \exp(\alpha + \beta x_i + \gamma y_i + \xi x_i y_i)\}$$

over $(\alpha, \beta, \gamma, \xi)$. Let the resulting estimator be $(\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}, \tilde{\xi})$.

(ii) Obtain the empirical maximum likelihood estimator of $p_i = Pr(x_i, y_i|\text{control})$: $\hat{p}_i = \{n_0 + n_1 \exp(\tilde{\alpha} + \tilde{\beta} x_i + \tilde{\gamma} y_i + \tilde{\xi} x_i y_i)\}^{-1}$.

(iii) Estimate F_X and F_Y by $\hat{F}_X(x) = \sum_{i=1}^{n} \hat{p}_i I(x_i \le x)$ and $\hat{F}_Y(y) = \sum_{i=1}^{n} \hat{p}_i I(y_i \le y)$, respectively.

Qin and Zhang (1997) showed that $\hat{F}(x, y) = \sum_{i=1}^{n} p_i I(x_i \le x, y_i \le y)$ is consistent for $F_0(x, y)$. Therefore, $\hat{F}_X(\cdot)$ and $\hat{F}_Y(\cdot)$ are consistent for $F_X(\cdot)$ and $F_Y(\cdot)$, respectively, and the resulting log pseudo likelihood $l_n(\beta, \gamma, \xi, \hat{F}_X, \hat{F}_Y)$ can be used for estimating/detecting the interaction effect.

The MLE $(\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}, \tilde{\xi})$ and $(\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\xi})$ can be easily obtained by the Newton-Raphson algorithm and nonlinear optimization algorithms.

3.2. Discrete-continuous case

Let the observed risk factors for cases and controls be $\{(x^*_{1i}, y_{1i}), i=1,\ldots,n_1\}$ and $\{(x^*_{1i}, y_{1i}),\ i=1,\ldots,n_0\}$, respectively, and the pooled data be $\{(x^*_i, y_i),\ i=1,\ldots,n(=n_1+n_0)\}$. The log likelihood is

$$l_n(\beta,\gamma,\xi,\theta,F_X,F_Y)=\sum_{i=1}^{n}\log f_0(x^*_i,y_i)+\sum_{i=1}^{n_1}\Bigg[(\beta x^*_{1i}+\gamma y_{1i}+\xi x^*_{1i}y_{1i}) - \log\sum_{k=0}^{K}\int_y f_0(k,y)\exp(\beta k+\gamma y+\xi k y)dy\Bigg], \tag{3.2}$$

where $f_0(\cdot,\cdot)$ is defined at (2.7). Now (3.2) depends on F_X only through $F_X(c_1)$, ..., $F_X(c_K)$, but it also depends on the high-dimensional parameter F_Y, making it difficult to maximize directly. One could adopt a two-stage approach as above. But our preliminary numerical study shows that this approach is numerically unstable due to the estimation of F_X (at $c_1,\ldots,c_K$) in the second stage. Instead, in the second stage we fix F_X at the estimator $\hat{F}_X$ (evaluated at $c_1,\ldots,c_K$) obtained in the first stage as in the continuous-continuous case, then maximize the log pseudo likelihood $l_n(\beta,\gamma,\xi,\theta,\hat{F}_X,\hat{F}_Y)$ with respect to $(\beta,\gamma,\xi,\theta)$ to get the pseudo-MLE $(\hat{\beta},\hat{\gamma},\hat{\xi},\hat{\theta})$. Here, the estimator $\hat{F}_X(c_k)$ is $\hat{F}_{X^*}(k)$, obtained as in the continuous-continuous case by treating F_{X^*} as a continuous distribution function that depends on k instead of the threshold value c_k.

3.3. Discrete-discrete case

Let the observed risk factors for cases and controls be, respectively, $\{(x^*_{1i}, y^*_{1i}),\ i=1,\ldots,n_1\}$ and $\{(x^*_{1i}, y^*_{1i}),\ i=1,\ldots,n_0\}$, the pooled data be $\{(x^*_i, y^*_i),\ i=1,\ldots,n(=n_1+n_0)\}$. The log likelihood is

$$l_n(\beta,\gamma,\xi,\theta,F_X,F_Y)=\sum_{i=1}^{n}\log f_0(x^*_i,y^*_i)+\sum_{i=1}^{n_1}\Bigg[(\beta x^*_{1i}+\gamma y^*_{1i}+\xi x^*_{1i}y^*_{1i}) - \log\sum_{k=0}^{K}\sum_{l=0}^{L} f_0(k,l)\exp(\beta k+\gamma l+\xi k l)\Bigg], \tag{3.3}$$

where $f_0(\cdot,\cdot)$ is defined at (2.10). The log likelihood (3.3) depends on F_X only through $F_X(c_1),\ldots,F_X(c_K)$, and F_Y only through $F_Y(d_1),\ldots,F_Y(d_L)$. Therefore, intuitively, one can maximize (3.3) with respect to all unknown parameters directly. Again, since this maximization can be difficult in practice, we adopt a two-stage approach. In the first stage we obtain the estimators $\hat{F}_X$ (at $c_1,\ldots,c_K$) and $\hat{F}_Y$ (at $d_1,\ldots,d_L$) based on a prospective likelihood as in the continuous-continuous case, and in the second stage we maximize the log pseudo likelihood

$l_n(\beta, \gamma, \xi, \theta, \hat{F}_X, \hat{F}_Y)$ with respect to $(\beta, \gamma, \xi, \theta)$. Let the resulting pseudo-MLE be $(\hat{\beta}, \hat{\gamma}, \hat{\xi}, \hat{\theta})$.

4. Large Sample Properties and Testing Procedure

We focus on the continuous-continuous case since the other cases can be addressed similarly.

In Supplementary Material S1, under some regularity conditions on F_{0X}, F_{0Y}, and $C(x, y; \theta)$, we show that there exists a local maximizer of the pseudo likelihood that is consistent for the true value of $(\beta, \gamma, \xi, \theta)$, and that this pseudo-MLE is asymptotically normally distributed with expectation 0 and a variance-covariance matrix given there. By virtue of the asymptotic normality, given the variance-covariance matrix of the pseudo-MLE, the $(1 - \alpha) \times 100\%$ confidence limit for ξ is $\hat{\xi} \pm z_{1-\alpha/2}\hat{\text{se}}(\hat{\xi})$ and the Wald test statistic for $H_0 : \xi = 0$ takes the form $\hat{\xi}\{\hat{\text{se}}(\hat{\xi})\}^{-1}$, where $z_{1-\alpha/2}$ is the upper α-quantile of the standard normal distribution and $\hat{\text{se}}(\hat{\xi})$ is an estimated standard error of $\hat{\xi}$.

In the Supplementary Material S1 one sees that the estimation of the variance-covariance matrix of the pseudo-MLE is quite complicated, even more complicated if F_X and F_Y are estimated by the algorithm in Subsection 3.1. Here, we consider two versions of the bootstrap. One is the nonparametric bootstrap that separately resamples case data and control data with replacement; the other is a semiparametric bootstrap based on the empirical distributions for the case group and the control group estimated from the retrospective likelihood (Qin and Zhang (1997)). In the semiparametric bootstrap method, since each subject in the pooled sample can be sampled for both cases and controls, the method fully uses all samples; it is more suitable than the nonparametric bootstrap method when n_0 or n_1 is relatively small. The detailed algorithm of the semiparametric bootstrap method is as follows.

(i) From $(x_i, y_i), i = 1, \ldots, n$, randomly generate n_0 risk factors $(x_i^{(0)}, y_i^{(0)})$ for controls, $i = 1, \ldots, n_0$, with weight $\hat{p}_i = \{n_0 + n_1 \exp(\tilde{\alpha} + \tilde{\beta}x_i + \tilde{\gamma}y_i + \tilde{\xi}x_iy_i)\}^{-1}$ for (x_i, y_i), and n_1 risk factors $(x_i^{(1)}, y_i^{(1)})$ for cases, $i = n_0 + 1, \ldots, n$, with weight $\hat{q}_i = \exp(\tilde{\alpha} + \tilde{\beta}x_i + \tilde{\gamma}y_i + \tilde{\xi}x_iy_i)\{n_0 + n_1 \exp(\tilde{\alpha} + \tilde{\beta}x_i + \tilde{\gamma}y_i + \tilde{\xi}x_iy_i)\}^{-1}$ for (x_i, y_i). Here $(\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}, \tilde{\xi})$ is the maximum likelihood estimator of $(\alpha, \beta, \gamma, \xi)$ without constraint, as defined in Subsection 3.1.

(ii) Obtain the pseudo maximum likelihood estimator $(\check{\beta}, \check{\gamma}, \check{\xi}, \check{\theta})$ using resampled data $\{(x_1^{(0)}, y_1^{(0)}), \ldots, (x_{n_0}^{(0)}, y_{n_0}^{(0)}), (x_1^{(1)}, y_1^{(1)}), \ldots, (x_{n_1}^{(1)}, y_{n_1}^{(1)})\}$.

(iii) Repeat (2.1) and (2.2) for B times to obtain some copies of $(\check{\beta}, \check{\gamma}, \check{\xi}, \check{\theta})$, say $(\check{\beta}_b, \check{\gamma}_b, \check{\xi}_b, \check{\theta}_b)$, $b = 1, \ldots, B$.

(iv) Estimate the variance-covariance matrix of the pseudo maximum likelihood estimator $(\hat{\beta}, \hat{\gamma}, \hat{\xi}, \hat{\theta})$ using the sample variance-covariance matrix of $(\check{\beta}_b, \check{\gamma}_b, \check{\xi}_b, \check{\theta}_b)$, $b = 1, \ldots, B$.

Because $(\hat{\beta}, \hat{\gamma}, \hat{\xi}, \hat{\theta})$ is asymptotically normal, a moderate number of resamplings can generate a good approximation of the variance of $(\hat{\beta}, \hat{\gamma}, \hat{\xi}, \hat{\theta})$, as shown in the subsequent simulation study with $B = 200$.

5. A Simulation Study

In the simulation study, we considered the Gaussian copula model (Li (2000))

$$C(x, y; \theta) = \Phi_\theta(\Phi^{-1}(x), \Phi^{-1}(y)), \tag{5.1}$$

where $\Phi^{-1}(x)$ is the inverse function of the standard normal distribution function and $\Phi_\theta(x, y)$ is the joint distribution function of the bivariate normal with means 0, variances 1, and correlation coefficient θ. We have

$$c(x, y; \theta) = \frac{\phi_\theta(\Phi^{-1}(x), \Phi^{-1}(y))}{\phi(\Phi^{-1}(x))\phi(\Phi^{-1}(y))} \text{ and } C_2(x, y; \theta) = \Phi\left(\frac{\Phi^{-1}(x) - \theta\Phi^{-1}(y)}{(1-\theta^2)^{1/2}}\right), \tag{5.2}$$

where $\phi(x)$ is the standard normal density function and $\phi_\theta(x, y) = \partial^2\Phi_\theta(x, y)/\partial x \partial y$. The first equation in (5.2) follows from the definition and the derivation of the second is given in Supplementary Material S2.

The risk factors (X, Y) in the control population were generated from the bivariate normal distribution with the copula function $C(x, y; \theta)$. For the discrete-continuous and discrete-discrete cases, the probability function of X^* was $Pr(X^* = k) = 1/4$, $k = 0, 1, \ldots, 3$, and for the discrete-discrete case, the probability function of Y^* was the same as that of X^*.

We considered three values of θ: 0, 0.2, and -0.2. The main effects were fixed at $\beta = \gamma = 0.5$. The interaction effect ξ either 0 for all the three cases or 0.25 for the continuous-continuous case and 0.5 for the discrete-continuous and discrete-discrete cases. For each combination of θ and ξ, we generated 10^6 risk factors for controls: $\{(X_i, Y_i), i = 1, \ldots, 10^6\}$, $\{X_i^*, i = 1, \ldots, 10^6\}$ according to (2.6), and $\{Y_i^*, i = 1, \ldots, 10^6\}$ according to (2.9). We used a biased sampling technique (Cochran (1977); Nair and Wang (1989)) to generate the case data. For instance, in the discrete-continuous case, we generated risk factors for cases from

$$Pr(X_i^*, Y_i | D_i = 1) = \frac{\exp(0.5X_i^* + 0.5Y_i + \xi X_i^* Y_i)}{\sum_{i=1}^{10^6} \exp(0.5X_i^* + 0.5Y_i + \xi X_i Y_i)}. \tag{5.3}$$

With the 10^6 observations, we randomly sampled 200 observations for controls and randomly sampled 200 observations for cases with weights (5.3). The simulation results were based on 2000 generated datasets. Then we applied the

Table 1. Interaction estimate/test results in the continuous-continuous case.

ξ	θ	Proposed					Logistic	Case-only
		Bias	SE	SEE	CP	Wald1	Wald2	Pearson
0	0.0	0.004	0.104	0.102	0.938	0.064	0.050	0.057
0	0.2	0.032	0.105	0.106	0.943	0.056	0.048	0.826
0	−0.2	−0.018	0.103	0.103	0.947	0.061	0.054	0.816
0.25	0.0	0.027	0.109	0.108	0.958	0.754	0.578	0.945
0.25	0.2	0.070	0.111	0.116	0.892	0.835	0.556	1.000
0.25	−0.2	−0.006	0.107	0.105	0.937	0.636	0.571	0.092

ξ, the interaction effect; θ, the correlation coefficient of marginal distributions; Bias, mean estimated ξ minus the ξ; SE, standard error of estimated ξ; SEE, mean estimated standard error of estimated ξ; CP, coverage probability of confidence interval; Wald1 and Wald2, Wald test; Pearson, Pearson chi-square test.

Table 2. Interaction estimate/test results in the discrete-continuous case.

ξ	θ	Proposed					Logistic	Case-only
		Bias	SE	SEE	CP	Wald1	Wald2	Pearson
0	0.0	0.005	0.096	0.096	0.949	0.051	0.054	0.048
0	0.2	0.010	0.100	0.100	0.948	0.055	0.049	0.645
0	−0.2	−0.006	0.097	0.097	0.962	0.052	0.052	0.698
0.5	0.0	0.049	0.164	0.156	0.937	0.936	0.802	0.997
0.5	0.2	0.063	0.162	0.164	0.938	0.925	0.722	1.000
0.5	−0.2	0.009	0.146	0.140	0.942	0.961	0.883	0.951

ξ, the interaction effect; θ, the correlation coefficient of marginal distribution; Bias, mean estimated ξ minus the ξ; SE, standard error of estimated ξ; SEE, mean estimated standard error of estimated ξ; CP, coverage probability of confidence interval; Wald1 and Wald2, Wald test; Pearson, Pearson chi-square test.

proposed approach, the standard logistic regression method, and the Pearson chi-square test for independence using the case data to each simulated dataset. In the proposed approach, we estimated the standard errors of pseudo-MLEs based on 200 semiparametric bootstrap samples.

For the proposed approach, the average value of the pseudo-MLE of $\hat{\xi}$ minus the true ξ (Bias), the empirical standard error (SE) of $\hat{\xi}$, the mean estimated standard error (SEE) of $\hat{\xi}$, and the coverage probability (CP) of the 95% confidence interval of ξ are reported in Tables 1, 2, and 3 for the continuous-continuous case, the discrete-continuous case, and the discrete-discrete case, respectively. The type I error rates and powers at 0.05 level for the proposed approach (Wald1), the standard logistic regression method (Wald2), and the case-only method (Pearson) are also reported in Tables 1−3.

In most situations, the biases are small, the estimated standard errors are close to the empirical standard errors, and the coverage probabilities are close

Table 3. Interaction estimate/test results in the discrete-discrete case.

ξ	θ	Proposed					Logistic	Case-only
		Bias	SE	SEE	CP	Wald1	Wald2	Pearson
0	0.0	−0.002	0.093	0.092	0.942	0.050	0.048	0.046
0	0.2	−0.005	0.093	0.094	0.965	0.048	0.056	0.479
0	−0.2	−0.005	0.089	0.091	0.953	0.05	0.050	0.550
0.5	0.0	−0.053	0.254	0.241	0.902	0.469	0.437	0.543
0.5	0.2	−0.045	0.276	0.256	0.884	0.441	0.414	0.995
0.5	−0.2	−0.038	0.217	0.223	0.921	0.549	0.511	0.920

ξ, the interaction effect; θ, the correlation coefficient of marginal distributions; Bias, mean estimated ξ minus the ξ; SE, standard error of estimated ξ; SEE, mean estimated standard error of estimated ξ; CP, coverage probability of confidence interval; Wald1 and Wald2, Wald test; Pearson, Pearson chi-square test.

to the nominal level 95%. In a few situations, absolute biases are greater than 0.05. Both the proposed approach and the logistic regression method have well-controlled type I error rates ($\xi = 0$). The type I error rates of the case-only method are also under control when the risk factors X and Y are indeed independent ($\theta = 0$), but the type I error rates are dramatically inflated when X and Y are correlated ($\theta \neq 0$). Under the alternative hypothesis ($\xi \neq 0$), the proposed test is uniformly more powerful than the logistic regression method, with relative power gains ranging from 11% to 50% in the continuous-continuous case, from 8% to 28% in the discrete-continuous case, and from 6.5% to 7.4% in the discrete-discrete case. When the independence assumption is met, the power of the proposed approach is slightly lower than the case-only method which is the most powerful test under the independence assumption. However, when the independence assumption is not met, the case-only method can lose substantial power. For instance, in the continuous-continuous case with $\xi = 0.25$ and $\theta = -0.2$, the power for the case-only method is 0.092, dramatically lower than 0.636 for the proposed approach.

We also considered different marginal distributions for the two risk factors, normal and uniform. The results are similar to those in Tables 1-3 and are not presented here.

Finally, we conducted a sensitivity analysis by misspecifying the copula function in the continuous-continuous case. We considered the Clayton, Frank, and t (with 10 degrees of freedom) copula functions. The parameters characterizing the copula functions were chosen such that the correlation coefficients were around 0.24, and the marginal distributions of the risk factors were again standard normal. The other settings were the same as those for Table 1. The corresponding results are presented in Table 4. When the true copula function was Clayton

Table 4. Interaction estimate/test results for non-Gaussian copula functions.

Copula	θ	ξ	Proposed					Logistic	Case-only
			Bias	SE	SEE	CP	Wald1	Wald2	Pearson
Clayton	0.23	0	−0.011	0.103	0.104	0.947	0.053	0.052	0.073
	0.23	0.25	0.002	0.111	0.106	0.933	0.670	0.583	0.976
Frank	0.24	0	−0.001	0.104	0.102	0.931	0.069	0.055	0.088
	0.24	0.25	0.012	0.103	0.106	0.950	0.711	0.566	0.982
t_{10}	0.24	0	0.084	0.115	0.113	0.889	0.112	0.048	0.344
	0.24	0.25	0.239	0.134	0.115	0.533	0.982	0.628	1.000

Copula, true copula function; Proposed, the proposed approach with copula function specified to be Gaussian; ξ, the interaction effect; θ, the correlation coefficient of marginal distributions; Bias, mean estimated ξ minus the true of ξ; SE, standard error of estimated ξ; SEE, mean estimated standard error of estimated ξ; CP, coverage probability of confidence interval; Wald1 and Wald2, Wald test; Pearson, Pearson chi-square test.

or Frank, the proposed approach produces minor bias in estimates, and well-controlled type I error rates and coverage probabilities. When the true copula function was t, the proposed approach produced relatively larger biases, inflated type I error rates, and poorer coverage probabilities.

6. Applications

6.1 Prostate cancer example

The dataset is from a nested case-control study (Ahn et al. (2008)) within the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO), where cases and controls were frequency matched by age at cohort entry, time since initial screening, and calendar year of cohort entry. In this study, the effect of some risk factors on the prostate cancer were examined. We considered two continuous risk factors: vitamin D level [25(OH)D concentrations] and body mass index (BMI). After removing individuals with extreme 25(OH)D concentrations, 749 case patients and 781 control subjects remained. The vitamin D measure determined by 25(OH)D concentrations (nmol/L) strongly depended on the season when the blood was drawn, so we removed this seasonal variation pattern using locally weighted scatterplot smoothing (Cleveland, Grosse, and Shyu (1991)). Let B and V denote the normalized values of BMI and the adjusted 25(OH)D concentrations with seasonal pattern removed, respectively. We modeled the relationship between the disease status D and (B, V) by the logistic regression model

$$Pr(D = 1|B, V) = \frac{\exp(\alpha + \beta B + \gamma V + \xi BV)}{1 + \exp(\alpha + \beta B + \gamma V + \xi BV)}.$$

Table 5. Analysis results for SNP rs12913946 in lung cancer data set.

Parameter	Estimate	SE	Z-value	P-value
θ	-0.036	0.024	-1.503	0.137
β	-0.239	0.112	-2.143	0.033
γ	1.183	0.094	12.557	3.65×10^{-36}
ξ	0.238	0.088	2.722	6.5×10^{-3}

To test the interaction effect $\xi = 0$, we applied the standard logistic regression, the case-only, and the proposed copula-model methods. In the proposed approach, we assumed that the copula characterizing the joint distribution was the Gaussian copula (5.1) with correlation coefficient parameter θ, and the standard errors of the pseudo-MLEs were obtained with 1,000 semiparametric bootstrap samples. The pseudo-MLE of θ was -0.207 (p-value $= 1.8 \times 10^{-10}$), showing a very significant negative correlation between BMI and vitamin D level. The resulting estimates of ξ from the proposed approach and the standard logistic regression method were 5.8×10^{-3} (p-value $= 0.916$) and 4.2×10^{-2} (p-value $=$ 0.443), respectively, both of which indicated the absence of the interaction. On the other hand, the case-only method gave a p-value of 4.2×10^{-8}, indicating a very statistically significant but most likely false positive finding of the interaction as the independence assumption was clearly violated. When the age at cohort entry, the time since initial screening, and the calendar year of cohort entry were further adjusted for in the standard logistic regression, we obtained very similar result, with the interaction effect being 4.4×10^{-2} (p-value $= 0.422$).

6.2. Lung cancer example

In recent genome-wide association studies (GWAS), a few chromosome regions (e.g., chromosomes 15q25, 5p15, and 6p21) have been identified to be associated with lung cancer (Hung et al. (2008)); Amos et al. (2008); Thorgeirsson et al. (2008); McKay et al. (2008); Wang et al. (2008); Rafnar et al. (2009); and Landi et al. (2009)). Among these chromosome regions, the chromosome 15q25 region was shown to be associated with both lung cancer and smoking behavior. It is of great interest to test whether there is interaction between the genetic variants in the 15q25 region and smoking on the risk of lung cancer. We used the data from Environment and Genetics in Lung Cancer Etiology Study (EAGLE; Landi et al. (2009)), and focused on the genotypes on 39 relatively common single-nucleotide polymorphisms (SNPs) within the 15q25 region and smoking intensity, measured by the average number of packs of cigarette per day (CPD). The numbers of individuals were 460 for CPD < 0.5, 965 for $0.5 \leq$ CPD < 1, 1393 for $1 \leq$ CPD < 2, and 256 for CPD ≥ 2. We evaluated the interaction between CPD and each of the 39 SNPs.

We modeled the relationship between the lung cancer status D and any of the genetic variants (coded by G, the number of minor alleles) and CPD measure C by the logistic regression model

$$Pr(D = 1|G, C) = \frac{\exp(\alpha + \beta G + \gamma C + \xi GC)}{1 + \exp(\alpha + \beta G + \gamma C + \xi GC)}.$$

In the proposed approach, we assumed a Gaussian copula (5.1) for the joint distribution of CPD and the continuous variable underlying the SNP genotype.

For our analysis, we only considered those subjects with a smoking history, and focused on the SNP rs12913946 which had the most significant interaction effect with CPD from the standard logistic regression analysis (p-value=0.042) and the case only method (p-value = 0.011). This left 1,738 lung cancer cases and 1,336 controls. We applied the proposed approach to study this interaction with standard errors of pseudo-MLEs being obtained with 1,000 semiparametric bootstrap samples. The pseudo-MLE of the interaction effect ξ was 0.238 (standard error = 0.088) with a two-sided p-value of 6.5×10^{-3} that was smaller than the one obtained by the standard logistic regression. More detailed results are summarized in Table 5. Clearly, further investigation is needed to validate this interaction.

7. Discussion

The majority of common diseases result from complex interplay of genetic and environmental risk factors. It is important to study gene-gene and gene-environment interactions in order to better understand the mechanism underlying the disease development. We develop a copula-model based semiparametric test for interaction detection. Our proposed approach strikes a balance between robustness and efficiency by modeling the correlation between the two risk factors while keeping their marginal distributions fully unspecified.

We have found the efficiency gain of the proposed approach is much higher when both factors are continuous than when they are both discrete, this being related to the numbers of degrees of freedom of associated tests.

Simulation results show that the copula approach provides valid results even if the underlying copula model is misspecified mildly. As a precaution, in applications one has to make sure that the copula model is not terribly misspecified.

Although we only consider two risk factors in the current manuscript, the proposed approach may be extended to more than two risk factors.

Supplementary Material

Supplementary material is provided for the consistency and the asymptotic normality of the pseudo-MLE, and for a proof of the second equation in (5.2).

Acknowledgement

We are grateful to two referees, an associate editor, and a joint editor for insightful comments. This research was supported by the State Key Development Program for Basic Research of China (Grant No. 2012CB316500) (HZ) and the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health (HZ, ML, NC, KY).

References

Ahn, J., Peters, U., Albanes, D., Purdue, M. P., Abnet, C. C., Chatterjee, N., Horst, R. L., Hollis, B. W., Huang, W. Y., Shikany, J. M., Hayes, R. B., Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial Project Team (2008). Serum vitamin D concentration and prostate cancer risk: a nested case-control study. *J. Nat. Cancer Inst.* **100**, 796-804.

Amos, C. I., Wu, X., Broderick, P., Gorlov, I. P., Gu, J., Eisen, T., Dong, Q., Zhang, Q., Gu, X. and Vijayakrishnan, J. (2008). Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat. Genet.* **40**, 616-622.

Cochran, W. G. (1977). *Sampling Techniques*, 3rd edition. John Wiley & Sons, New York.

Cornfield, J. (1956). A statistical problem arising from retrospective studies. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol 4* (Edited by J. Neyman), 135-148. University of California Press, Berkeley, CA.

Chatterjee, N. and Carroll, R. J. (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika* **92**, 399-418.

Cleveland, W. S., Grosse, E. and Shyu, W. M. (1991). Local regression models. Chapter 8 of *Statistical Models in S* (Edited by J. M. Chambers and T. J. Hastie), 309-376. Chapman & Hall, London.

de Leon, A. R. and Wu, B. (2011). Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Statist. Med.* **30**, 175-185.

Genest, C., Ghoudi, K. and Rivest, L. P. (1995). A semiparametrie estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika* **82**, 543-552.

Gong, G. and Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: theory and applications. *Ann. Statist.* **9**, 861-869.

Hung, R. J., McKay, J. D., Gaborieau, V., Boffetta, P., Hashibe, M., Zaridze, D., Mukeria, A., Szeszenia-Dabrowska, N., Lissowska, J. and Rudnai, P. (2008). A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* **452**, 633-637.

Landi, M. T., Chatterjee, N., Yu, K., Goldin, L. R., Goldstein, A. M., Rotunno, M., Mirabello, L., Jacobs, K., Wheeler, W., Yeager, M., Bergen, A. W., Li, Q., Consonni, D., Pesatori, A. C., Wacholder, S., Thun, M., Diver, R., Oken, M., Virtamo, J., Albanes, D., Wang, Z., Burdette, L., Doheny, K. F., Pugh, E. W., Laurie, C., Brennan, P., Hung, R., Gaborieau, V., McKay, J. D., Lathrop, M., McLaughlin, J., Wang, Y., Tsao, M. S., Spitz, M. R., Krokan, H., Vatten, L., Skorpen, F., Arnesen, E., Benhamou, S., Bouchard, C., Metsapalu, A., Vooder, T., Nelis, M., Valk, K., Field, J. K., Chen, C., Goodman, G., Sulem, P., Thorleifsson, G., Rafnar, T., Eisen, T., Sauter, W., Rosenberger, A., Bickeboller, H., Risch, A., Chang-Claude, J., Wichmann, H. E., Stefansson, K., Houlston, R., Amos, C. I., Fraumeni, J. F., Savage, S. A., Bertazzi, P. A., Tucker, M. A., Chanock, S. and Caporaso, N. E. (2009). A genome-wide association study of lung cancer identifies a

region of chromosome 5p15 associated with risk for adenocarcinoma. *Am. J. Hum. Genet.* **85**, 679-691.

Li, D. X. (2000). On default correlation: a copula function approach. *J. Financial Intermediation* **9**, 43-54.

Lin, D. and Zeng, D. (2009). Proper analysis of secondary phenotype data in case-control association studies. *Genet. Epi.* **33**, 256-265.

McKay, J. D., Hung, R. J., Gaborieau, V., Boffetta, P., Chabrier, A., Byrnes, G., Zaridze, D., Mukeria, A., Szeszenia-Dabrowska, N. and Lissowska, J. (2008). Lung cancer susceptibility locus at 5p15.33. *Nat. Genet.* **40**, 1404-1406.

Mukherjee, B. and Chatterjee, N. (2008). Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics* **64**, 685-694.

Nair, V. N. and Wang, P. C. C. (1989). Maximum likelihood estimation under a successive sampling discovery model. *Technometrics* **31**, 423-436.

Nelsen, R. B. (1999). *An Introduction to Copulas*. Springer, New York.

Piegorsch, W. W., Weinberg, C. R. and Taylor, J. A. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population based case-control studies. *Statist. Med.* **13**, 153-162.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403-411.

Qin, J. and Zhang, B. (1997). A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika* **84**, 609-618.

Rafnar, T., Sulem, P., Stacey, S. N., Geller, F., Gudmundsson, J., Sigurdsson, A., Jakobsdottir, M., Helgadottir, H., Thorlacius, S. and Aben, K. K. (2009). Sequence variants at the TERTCLPTM1L locus associate with many cancer types. *Nat. Genet.* **41**, 221-227.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Statistical Institute of the University of Paris* **8**, 229-231.

Shih, J. H. and Louis, T. A. (1995). Association parameter in copula models for bivariate survival data. *Biometrics* **51**, 1384-1399.

Thorgeirsson, T. E., Geller, F., Sulem, P., Rafnar, T., Wiste, A., Magnusson, K. P., Manolescu, A., Thorleifsson, G., Stefansson, H. and Ingason, A. (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* **452**, 638-642.

Umbach, D. M. and Weinberg, C. M. (1997). Designing and analyzing case-control studies to exploit independence of genotype and exposure. *Statist. Med.* **16**, 1731-1743.

Wang, Y., Broderick, P., Webb, E., Wu, X., Vijayakrishnan, J., Matakidou, A., Qureshi, M., Dong, Q., Gu, X. and Chen, W. V. (2008). Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat. Genet.* **40**, 1407-1409.

Institute of Biostatistics, School of Life Science, Fudan University, 220 HanDan Road, Shanghai 200433, China.

E-mail: zhanghfd@fudan.edu.cn

Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, NIH MSC 7630, 6700A Rockledge Dr, Bethesda, MD 20892, USA.

E-mail: jingqin@mail.nih.gov

Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH MSC 9769, 9609 Medical Center Dr, Rockville, MD 20850, USA.

E-mail: landim@mail.nih.gov

Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH MSC 9769, 9609 Medical Center Dr, Rockville, MD 20850, USA.

E-mail: caporasn@mail.nih.gov

Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH MSC 9780, 9609 Medical Center Dr, Rockville, MD 20850, USA.

E-mail: yuka@mail.nih.gov

(Received January 2012; accepted September 2012)