Bayesian semiparametric copula estimation with application to psychiatric genetics

Ori Rosen[*,1] and **Wesley K. Thompson** [2]

[1] Department of Mathematical Sciences, University of Texas at El Paso, El Paso, Texas 79968, U.S.A.

[2] Department of Psychiatry, University of California at San Diego, La Jolla, California 92093, U.S.A.

This paper proposes a semiparametric methodology for modeling multivariate and conditional distributions. We first build a multivariate distribution whose dependence structure is induced by a Gaussian copula and whose marginal distributions are estimated nonparametrically via mixtures of B-spline densities. The conditional distribution of a given variable is obtained in closed form from this multivariate distribution. We take a Bayesian approach, using Markov chain Monte Carlo methods for inference. We study the frequentist properties of the proposed methodology via simulation and apply the method to estimation of conditional densities of summary statistics, used for computing conditional local false discovery rates, from genetic association studies of schizophrenia and cardiovascular disease risk factors.

Key words: B-spline densities; Cardiovascular Disease Risk Factors; Gaussian copula; Schizophrenia

1 Introduction

This paper proposes a semiparametric methodology for modeling multivariate and conditional densities. We first build a multivariate distribution whose dependence structure is induced by a Gaussian copula and whose marginal distributions are estimated nonparametrically via mixtures of B-spline densities. The conditional distribution of a given variable as a function of the remaining variables is then obtained from this multivariate distribution. We take a Bayesian approach, using Markov chain Monte Carlo (MCMC) methods for inference.

Our research is motivated by an application in psychiatric genetics. Individuals with schizophrenia have significantly higher mortality rates compared with the general population, corresponding to a 10–20 year reduction in life expectancy (Colton and Manderscheid, 2006; Laursen et al., 2012; Saha et al., 2007). Although the mortality rate from suicide is high, lifestyle and cardiovascular disease (CVD) risk factors contribute substantially to life expectancy reduction in schizophrenia (Marder et al., 2004; Mitchell et al., 2011). Epidemiological research has shown increased rates of dyslipidemia, type 2 diabetes, obesity, and a high prevalence of metabolic syndrome among people with schizophrenia (Mitchell et al., 2011; De Hert et al., 2006). This increase in CVD risk factors has been primarily

*Corresponding author: e-mail: ori@math.utep.edu, Phone: +1-915-747-6843, Fax: +1-915-747-6502

attributed to lifestyle factors such as unhealthy diets, sedentary habits, excessive smoking, and to antipsychotic medication side-effects (Laursen et al., 2012; De Hert et al., 2006; Kaddurah-Daouk et al., 2007). However, as suggested by studies (i) predating introduction of antipsychotics (Raphael and Parsons, 1921), (ii) on untreated first episode individuals as well as their healthy relatives (Ryan et al., 2003) and (iii) on overlapping candidate genes (Hansen et al., 2011), shared genetics between schizophrenia and CVD risk factors may also be of importance.

Both schizophrenia and CVD risk factors are heritable, and large genome-wide association studies (GWAS) have reported single nucelotide polymorphisms (SNPs) associated with schizophrenia (Ripke et al., 2011) as well as CVD risk factors , including systolic and diastolic blood pressure (Ehret et al., 2011), low- and high-density lipoprotein cholesterol and triglycerides (Teslovich et al., 2010), body mass index (Speliotes et al., 2010), and waist to hip ratio (Heid et al., 2010). Complex traits such as schizophrenia and CVD risk factors are also highly polygenic (Glazier et al., 2002; Hirschhorn and Daly, 2005; Hindorff et al., 2009; Manolio et al., 2009; Yang et al., 2010), and there is evidence that many of these genetic risk factors overlap across traits (Sivakumaran et al., 2011; Andreassen et al., 2013), i.e., are *pleiotropic*.

We investigate the shared genetic mechanisms of schizophrenia and CVD risk factors by analyzing summary statistics from large independent GWAS of schizophrenia (SCZ), triglycerides (TG), and systolic blood pressure (SBP). The goal of these analyses is to estimate the conditional probability that a given SNP is non-null for schizophrenia given its observed association with all three phenotypes simultaneously. This conditional version of the local false discovery rate (local fdr; Efron, 2007) is biologically informative about the shared genetic architecture of schizophrenia and CVD risk factors, as well as potentially increasing power to discover new genetic loci related to schizophrenia (Andreassen et al., 2013). Estimation of the conditional local fdr requires estimation of the conditional density of the schizophrenia summary statistics given values of the summary statistics of the pleiotropic phenotypes. As noted above, we accomplish this by modeling the marginal densities as a mixture of B-spline densities and linking the marginal densities via a Gaussian copula.

Copulas are defined as follows. Let $Y_1, \ldots, Y_p$ be random variables with cumulative distribution functions (cdf) $F_1(y_1)$, ..., $F_p(y_p)$, respectively. Sklar (1959) showed that there always exists a copula C, such that

$$F(y_1, \ldots, y_p) = C(F_1(y_1), \ldots, F_p(y_p)). \tag{1}$$

If $F_1, \ldots, F_p$ are continuous, then C is unique. Note that C is a cdf for p uniform random variables, and as such is a function from $[0,1]^p$ onto [0,1]. A copula models the dependence structure among the Y_j, $j = 1, \ldots, p$, but does not determine the marginal distributions. It can therefore be used as a tool for devising new multivariate distributions.

Genest and Favre (2007) discuss estimation methods for copula parameters, primarily in the bivariate case. Given pairs $(y_{11}, y_{12}), \ldots, (y_{n1}, y_{n2})$, they describe estimation of copula parameters using rank-based estimators. The justification for using rank-based methods is that the dependence structure captured by a copula is not related to the marginal distributions and is invariant under strictly monotonic transformations of the variables . This is a two-step

approach, where the marginal cdfs are first estimated by their empirical counterparts, and the copula parameters are then estimated by maximizing the rank-based log likelihood. Unlike the method of moments, maximum pseudolikelihood extends naturally to the multivariate case. As Genest and Favre (2007) acknowledge, there is no consensus in the literature on using only rank-based methods for estimating copula parameters. They also give references for kernel-based nonparametric estimation of the copula density. Chen et al. (2006) consider simultaneous estimation of both the marginal cdfs and the copula parameters. They argue that in general, empirical cdfs are inefficient estimators of the marginal cdfs. Furthermore, except for a few special cases, the two-step approach mentioned above is in general inefficient. Chen et al. (2006) propose a general sieve maximum likelihood estimation procedure for all the unknown parameters, where the marginal densities are approximated by linear combinations of finite-dimensional known basis functions with increasing complexity (sieves). Hoff (2007) proposes the extended-rank likelihood. For continuous data, this likelihood is equivalent to the distribution of the multivariate ranks. As a function of the parameters, the extended likelihood depends only on the copula parameters, not on the parameters of the marginal distributions. Pitt et al. (2006) use a Gaussian copula to handle multivariate dependence, assuming that the marginal distributions are specified. Their Bayesian formulation uses latent variables to transform each of the marginals to a standard normal distribution, and a multivariate normal distribution is assumed for these latent variables. The response variables can be either discrete or continuous or a combination of both. Pitt et al. (2006) also extend the idea of covariance selection to Gaussian copula models. Song et al. (2009) utilize Gaussian copulas to combine separate univariate generalized linear models into a joint regression model accommodating continuous, discrete or mixed outcomes. They specifically consider multidimensional logistic regression and a joint model for mixed normal and binary outcomes. Inference is based on maximum likelihood. Kolev and Paiva (2009) give a survey of copula-based regression models which includes the Gaussian copula regression model of Pitt et al. (2006), transition regression models and longitudinal models. Smith et al. (2010) model the dependence structure of continuous time series data using a sequence of bivariate copulas termed pair-copulas and take a Bayesian approach to inference.

Our proposed method is closest in spirit to that of Chen et al. (2006), described above, in that both the copula parameters and the parameters of the marginal distributions are estimated simultaneously. While Chen et al. (2006) use maximum likelihood estimation, our approach is Bayesian. Chen et al. (2006) first approximate the unknown marginal density functions by appropriate sieves, and then the sieve MLEs are obtained by maximization over a finite-dimensional parameter space. In their method, the smoothing parameters which control the smoothness of the estimated densities are selected by cross validation. Our method allows for joint estimation of all the parameters, including the smoothing parameters. Another emphasis of our proposed method is on conditional density estimation, i.e., the estimation of $f(Y_1|Y_2 = y_2, \ldots, Y_p = y_p)$. We derive an explicit formula for the conditional distribution corresponding to the Gaussian copula and use it to obtain the conditional density in closed form.

The remainder of the paper is organized as follows. Section 2 provides a brief background on Gaussian copulas. Section 3 presents our proposed model, the priors and an outline of the sampling scheme. Section 4 studies the

frequentist properties of the model via simulation, and Section 5 applies our model to data obtained from genome-wide asscociation studies of schizophrenia and cardiovascular risk factors. We conclude with some remarks and future directions in Section 6. Technical details are given in the appendices.

2 The Gaussian Copula

This paper focuses on the Gaussian copula (Song (2000)), which belongs to the family of elliptical copulas (Fang et al. (2002)). Let $\boldsymbol{u} = (u_1, \ldots, u_p)'$ such that $\boldsymbol{u} \in [0,1]^p$, then the cdf of the Gaussian copula is

$$C(u_1, \ldots, u_p) = \Phi_p(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_p); \Omega), \tag{2}$$

where $\Phi_p(\cdot; \Omega)$ is the cdf of a p-variate normal distribution with mean $\mathbf{0}$ and correlation matrix Ω, and Φ^{-1} is the inverse cdf of the standard normal distribution. Taking the derivative $\partial C(\boldsymbol{u})/\partial \boldsymbol{u}$ and evaluating it at $u_1 = F_1(y_1), \ldots, u_p = F_p(y_p)$ results in the Gaussian copula pdf

$$c(\boldsymbol{u}) = |\Omega|^{-1/2} \exp\Big\{-\frac{1}{2}\boldsymbol{q}'(\Omega^{-1} - I_p)\boldsymbol{q}\Big\}, \tag{3}$$

where $\boldsymbol{q} = (q_1, \ldots, q_p)'$, $q_j = \Phi^{-1}(F_j(y_j))$, $j = 1, \ldots, p$, and I_p is the $p \times p$ identity matrix. Partitioning Ω as

$$\Omega = \begin{pmatrix} 1 & \boldsymbol{\omega}' \\ \boldsymbol{\omega} & \Omega_{11} \end{pmatrix},$$

where Ω_{11} is the correlation matrix of $Y_2, \ldots, Y_p$ and $\boldsymbol{\omega} = (\omega_{12}, \ldots, \omega_{1p})'$, the conditional density of Y_1 given $Y_2, \ldots, Y_p$ is

$$f(y_1|y_2, \ldots, y_p) = \frac{1}{\sigma} f_1(y_1) \exp\Big\{-\frac{1}{2}\Big[\frac{(q_1 - \boldsymbol{\omega}'\Omega_{11}^{-1}\boldsymbol{q}_{-1})^2}{\sigma^2} - q_1^2\Big]\Big\}. \tag{4}$$

In (4), $f_1(y_1)$ is the marginal density of Y_1, $\boldsymbol{q}_{-1} = (q_2, \ldots, q_p)'$ and $\sigma^2 = 1 - \boldsymbol{\omega}'\Omega_{11}^{-1}\boldsymbol{\omega}$. Käärik and Käärik (2009) give without proof a less explicit version of the conditional density (4), while Crane and Van Der Hoek (2008) provide general conditional expectation formulas for continuous copulas. For a proof of (4), see Appendix A.

3 The model, priors and sampling scheme

3.1 The model

Similar to Chen et al. (2006), we propose simultaneous estimation of the marginal pdfs along with the copula parameters. We take a Bayesian approach and focus on the Gaussian copula. Given n observations on p continuous margins, $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_n')'$, the likelihood is

$$L(\boldsymbol{\Theta}, \Omega; \boldsymbol{y}) = |\Omega|^{-n/2} \exp\Big\{-\frac{1}{2}\sum_{i=1}^{n} \boldsymbol{q}_i'(\Omega^{-1} - I)\boldsymbol{q}_i\Big\} \prod_{i=1}^{n} f_1(y_{i1}) \ldots f_p(y_{ip}). \tag{5}$$

The marginal densities are expressed nonparametrically as mixtures of quadratic B-spline densities, which are regular B-splines that integrate to 1 over their support. This idea was first proposed by Ghidey et al. (2004) for modeling

the distribution of the random effects in linear mixed models. In fact, Ghidey et al. (2004) used Gaussian densities as the basis functions, citing a result on the convergence of a B-spline of degree q to a normal density, as $q \to \infty$. Staudenmayer et al. (2008) used the same idea with B-spline densities for density estimation in the presence of heteroscedastic measurement error. Pitt et al. (2006) use a Bayesian approach to copula modeling but assume known parametric margins.

More specifically, to facilitate nonparametric estimation of the marginal pdfs, we express $f_j(y_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, p$, as

$$f_j(y_{ij}) = \sum_{k=1}^{K_j} c_{jk} g_k(y_{ij}), \tag{6}$$

where $c_{jk} = \exp(\gamma_{jk}) / \sum_{l=1}^{K_j} \exp(\gamma_{jl})$ and $g_k(\cdot)$ are the B-spline densities. The number K_j of B-spline densities, specific to f_j, is fixed throughout the estimation process. In our experience, K_j of up to 30 is usually adequate for estimating the marginal densities. The corresponding knots are equally spaced along $y_{1j}, \ldots y_{nj}$, $j = 1, \ldots, p$. The γ_{jk}, $j = 1, \ldots, p$, $k = 1, \ldots, K_j$, are unknown parameters, and for identifiability, γ_{j1}, $j = 1, \ldots, p$, are set to zero. Although the γ_{jk}s are unconstrained parameters, the c_{jk}s satisfy $c_{jk} \in (0, 1)$ and $\sum_{k=1}^{K_j} c_{jk} = 1$, $j = 1, \ldots, p$. The marginal cdfs are

$$F_j(y_{ij}) = \sum_{k=1}^{K_j} c_{jk} G_k(y_{ij}), \tag{7}$$

where $G_k(\cdot)$, $k = 1, \ldots, K_j$, are the B-spline cdfs corresponding to the $g_k(\cdot)$s.

3.2 Priors

Priors on the $\boldsymbol{\gamma}_j$s

Let $\boldsymbol{\gamma}_j = (\gamma_{j2}, \ldots, \gamma_{jK_j})'$. We assume *a priori* independent $\boldsymbol{\gamma}_j$, $j = 1, \ldots, p$. To obtain smooth fits for the marginal pdfs, the c_{jk}s corresponding to neighboring B-splines must be close to one another (Eilers and Marx (1996)). This can be achieved by constraining the corresponding γ_{jk}s to be close to one another. By analogy to Eilers and Marx (1996)'s idea, Lang and Brezger (2004) require the γ_{jk}s to satisfy $\gamma_{j,\rho} = 2\gamma_{j,\rho-1} - \gamma_{j,\rho-2} + u_\rho$, where $u_\rho \sim N(0, \tau_j^2)$, and $p(\gamma_{j2}) \propto 1$, $p(\gamma_{j3}) \propto 1$. However, this results in an improper prior distribution on $\boldsymbol{\gamma}_j$ similar to the intrinsic Gaussian Markov random field prior used in spatial statistics. Chib and Jeliazkov (2006) place a joint normal prior on γ_{j2} and γ_{j3} which fixes the impropriety of the prior on $\boldsymbol{\gamma}_j$. For example, if $(\gamma_{j2}, \gamma_{j3})' \sim N_2(\mathbf{0}, c\,\tau_j^2 I_2)$, where c is a fixed constant, the prior on $\boldsymbol{\gamma}_j$ becomes

$$p(\boldsymbol{\gamma}_j | \tau_j^2) \propto (\tau_j^2)^{-\frac{1}{2}(K_j - 1)} \exp\Big\{ -\frac{1}{2\tau_j^2} \Big[\sum_{\rho=4}^{K_j} (\gamma_{j,\rho} - 2\gamma_{j,\rho-1} + \gamma_{j,\rho-2})^2 + c^{-1} \boldsymbol{\gamma}_{j,2:3}' \boldsymbol{\gamma}_{j,2:3} \Big] \Big\}, \tag{8}$$

where $\boldsymbol{\gamma}_{j,2:3}$ is a vector consisting of the first two entries of $\boldsymbol{\gamma}_j$. The summation in the exponent on the right-hand side of (8) can be expressed as $\boldsymbol{\gamma}_j' P_j \boldsymbol{\gamma}_j$, where $P_j = D_j' D_j$ and D_j is the $(K_j - 3) \times (K_j - 1)$ matrix

$$\begin{pmatrix} 1 & -2 & 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & 0 & \dots & 0 \\ \vdots & \dots & & & & & \vdots \\ 0 & 0 & \dots & & 1 & -2 & 1 \end{pmatrix}.$$

Equation (8) can now be re-expressed as

$$p(\boldsymbol{\gamma}_j|\tau_j^2) \propto (\tau_j^2)^{-\frac{1}{2}(K_j-1)} \exp\Big\{-\frac{1}{2\tau_j^2}\boldsymbol{\gamma}_j' P_j^* \boldsymbol{\gamma}_j\Big\},$$

where $P^*_{j,ll} = P_{j,ll} + c^{-1}$, for $l = 1, 2$.

Priors on the τ_j^2s

The τ_j^2s are assumed *a priori* independent with $\tau_j^2 \sim U(0, a_{\tau_j^2})$, where $a_{\tau_j^2}$ are large positive constants.

Prior on Ω

Following Danaher and Smith (2011), we express Ω as

$$\Omega = \text{diag}(\Sigma)^{-1/2}\, \Sigma\, \text{diag}(\Sigma)^{-1/2}, \tag{9}$$

where Σ is a non-unique positive definite matrix, and diag(Σ) is a diagonal matrix consisting of the leading diagonal of Σ. The matrix Σ^{-1} is decomposed as $\Sigma^{-1} = LL'$, where L is a lower triangular Cholesky factor with ones on its diagonal. The priors on the elements of L are independent $N(0, \sigma_L^2)$, where σ_L^2 is a fixed large number, reflecting vague knowledge on the elements of L. Alternative approaches include Chan and Jeliazkov (2009) who use the modified Cholesky decomposition and impose constraints on elements of the matrices making up this decomposition. Daniels and Pourahmadi (2009) parametrize correlation matrices in terms of partial autocorrelations as proposed by Joe (2006). These partial autocorrelations take values in the interval $[-1, 1]$ which makes them easy to work with. When the number of variables is large, it may be more appropriate to parameterize the correlation matrix parsimoniously via variable selection methods as in Pitt et al. (2006), for example.

3.3 The Sampling Scheme

This section gives an outline of the MCMC sampling scheme. More details are given in Appendix B. To simplify the sampling scheme, we introduce unobservable mixture indicators, z_{ijk}, such that $z_{ijk} = 1$ if y_{ij} came from the kth spline density. The sampling scheme consists of the following steps

1. Sample $\boldsymbol{\gamma}_j$ from $p(\boldsymbol{\gamma}_j|\boldsymbol{z}_j)$, $j = 1, \dots, p$, where $\boldsymbol{z}_j = \{z_{ijk}, i = 1, \dots, n, k = 1, \dots, K_j\}$, using a Metropolis-Hastings step.

2. Sample τ_j^2 from $p(\tau_j^2|\boldsymbol{\gamma}_j)$, $j = 1, \dots, p$.

3. Sample Ω from $p(\Omega|\{\boldsymbol{\gamma}_j\}_{j=1}^p, \text{data})$.

4. Sample the indicators, z_{ijk}, from $p(z_{ijk}|\{\boldsymbol{\gamma}_j\}_{j=1}^p, \text{data})$, $i = 1, \ldots, n$, $j = 1, \ldots, p$, $j = 1, \ldots, K_j$.

4 Simulation Study

In this section we study the frequentist properties of our model by Monte Carlo simulation studies. Two simulation settings are considered: (i) samples are generated from a trivariate normal distribution with mean vector $(3, 1, 2)'$ and correlation matrix

$$\Omega = \begin{pmatrix} 1 & 0 & .5774 \\ 0 & 1 & -.5774 \\ .5774 & -.5774 & 1 \end{pmatrix}, \tag{10}$$

and (ii) samples are generated from a trivariate Gaussian copula with marginal distributions

$$\begin{array}{rcl} F_1(y_1) & = & 0.3\,\Phi(y_1) + 0.7\,\Phi(y_1 - 5),\ -\infty < y_1 < \infty, \\ F_2(y_2) & = & 1 - \exp(-y_2/3),\ y_2 > 0, \\ F_3(y_3) & = & \Phi(y_3),\ -\infty < y_3 < \infty \end{array} \tag{11}$$

and the correlation matrix (10). Fifty samples, each of size 10,000, are generated from each setting. For each sample, a total of 10,000 MCMC iterations are run with a burn-in period of 2000. To evaluate the results, we utilize the Kullback-Liebler (KL) divergence

$$KL = -\int h(t) \log \frac{\hat{h}(t)}{h(x)} dt,$$

where $h(\cdot)$ is a probability density function and $\hat{h}(\cdot)$ is its estimate. In our case, $h(\cdot)$ is either a marginal density, $f(y_j)$, $j = 1, 2, 3$ or a conditional density, $f(y_1|y_2, y_3)$. In addition, the distance between the true correlation matrix Ω and its estimate is assessed using Stein's loss

$$\text{trace}(\hat{\Omega}\Omega^{-1}) - \log(\det(\hat{\Omega}\Omega^{-1})) - p,$$

see, for example, Daniels and Kass (2001).

Figure 1, panel (a), presents for each of the three marginal distributions boxplots of the KL divergence between the true and the estimated marginal densities based on the 50 samples from setting (i), while panel (b) displays similar boxplots for setting (ii). The KL values reflect good estimation of the marginal densities in both settings. Figure 2 displays in panels (a) and (b) boxplots of the KL divergence between the true and estimated conditional densities corresponding to settings (i) and (ii), respectively. Each panel includes nine boxplots corresponding to conditioning on quantiles of Y_2 and Y_3 according to

	1	2	3	4	5	6	7	8	9
Quantile of Y_2	25	50	75	25	50	75	25	50	75
Quantile of Y_3	25	25	25	50	50	50	75	75	75

For example, the fourth boxplot in each panel corresponds to conditioning on the 25th quantile of Y_2 and the 50th quantile of Y_3. It is seen in Figure 2 that the KL divergences for setting (ii) are slightly more variable than those for setting (i) as a function of the conditioning percentiles.

Figure 3 displays in panels (a) and (b) boxplots of the Stein loss values for setting (i) and (ii), respectively. From these plots, it is evident that the methodology provides good estimates of the underlying correlation matrix.

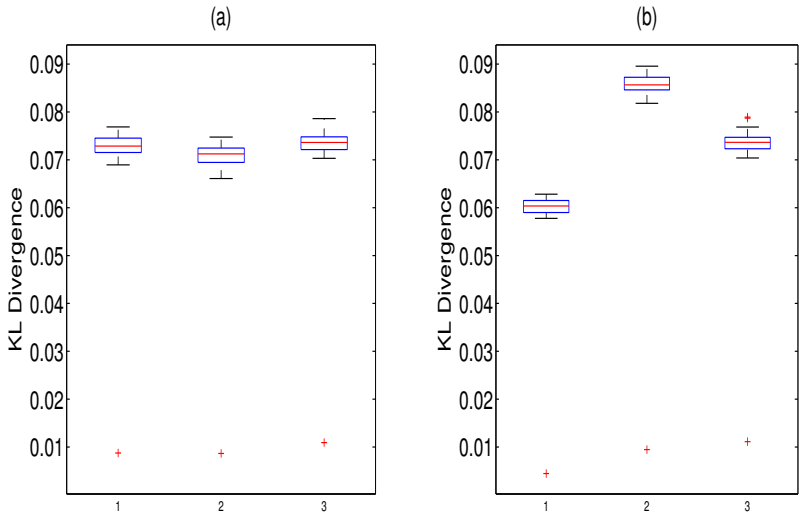

Figure 1 Panel (a): Boxplots of the Kullback-Liebler divergence between the true and the estimated marginal densities for setting (i). Panel (b): analogous plots for setting (ii).

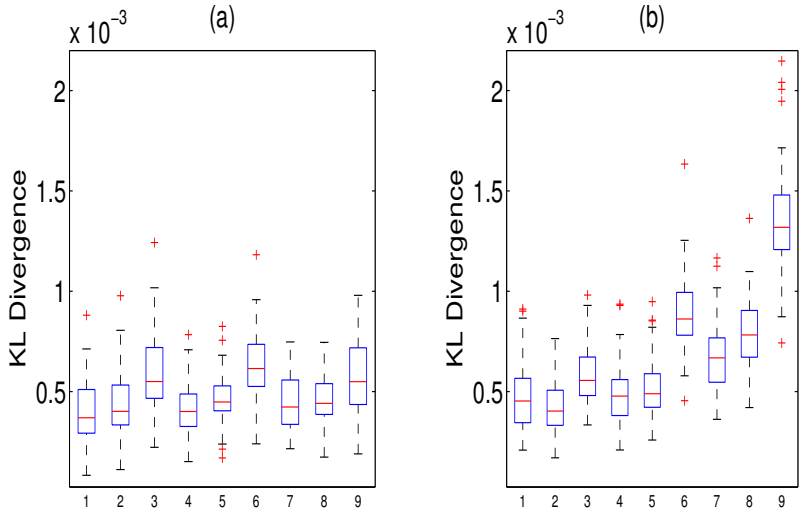

Figure 2 Panel (a): Boxplots of the Kullback-Liebler divergence between the true and the estimated conditional densities for setting (i). Panel (b): analogous plots for setting (ii).

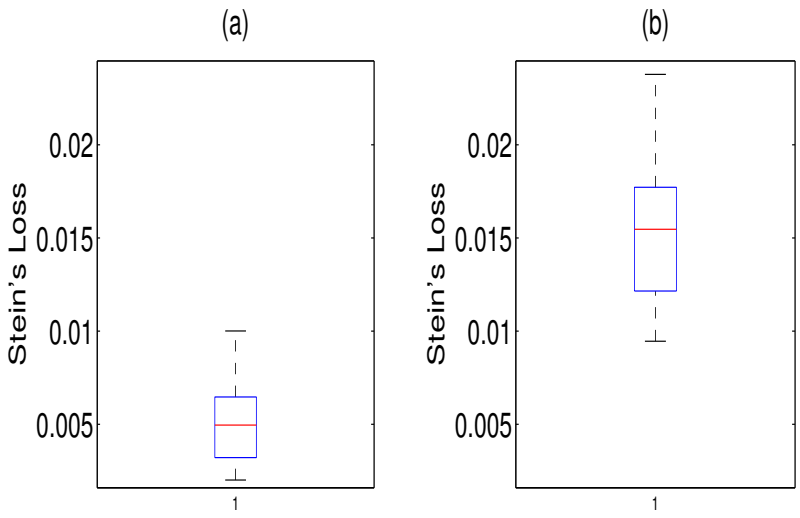

Figure 3 Panel (a): A Boxplot of Stein's loss based on the true and estimated correlation matrices for setting (i). Panel (b): An analogous boxplot for setting (ii).

5 Application

We investigate whether there are shared genetic mechanisms between schizophrenia (SCZ) and CVD risk factors by analyzing summary statistics from independent studies. We obtained two-tailed p-values from a schizophrenia GWAS (n=21,856 subjects; Ripke et al. 2011) and from independent GWAS on TG (n=96,568 subjects; Teslovich et al. 2010) and SBP (n= 203,056 subjects; Ehret et al. 2011). We selected all assayed SNPs in linkage disequilibrium with one or more $5'$ untranslated region loci, a category shown to be highly enriched for non-null effects across many phenotypes (Schork et al., 2013), resulting in 132,765 SNPs assessed in all three phenotypes simultaneously.

Since two-tailed p-values were made publicly available, we were able to recover the absolute z-scores of the test statistics but not their sign. For ease of notation, we denote absolute z-scores by $\boldsymbol{z} = (z_1, z_2, z_3)'$, where $z_j = \Phi^{-1}(1 - .5p_j)$, $j = 1, 2, 3$ and 1=SCZ, 2=TG and 3=SBP. Here, p_j is the two-tailed p-value for the jth phenotype and Φ is the standard normal cdf. Histograms of absolute z-scores for each phenotype are displayed in the top panel of Figure 4, along with the fitted marginal densities from the proposed model estimated using 20 equally spaced knots for each phenotype.

The estimates (posterior means) of the correlations making up the matrix Ω are $\hat{\omega}_{12} = .0467$, $\hat{\omega}_{13} = .0610$ and $\hat{\omega}_{23} = .0404$. The 95% credible intervals for ω_{12}, ω_{13} and ω_{23} are [.0466,.0473], [.0609,.0614] and [.0383,.0491], respectively. The lower and upper limits for these credible intervals are obtained as the 0.025 and 0.975 percentiles of the MCMC iterates of the correlations. Thus, quantiles of TG and SBP z-scores are postively correlated with quantiles of SCZ z-scores, demonstrating that SCZ has a partially overlapping genetic architecture with these two phenotypes.

Of primary interest is whether the probability that a SNP has a non-null association with SCZ depends on the strength of its relationship with TG and/or SBP. To investigate this question further, we utlize the local false discovery framework of Efron (2007). The local false discovery rate (*local fdr*) is defined as the posterior probability that a SNP is null for SCZ given its observed absolute z-score. By Bayes' rule this is given by

$$\text{fdr}(z_1) = \frac{\pi_0 \phi(z_1)}{f_1(z_1)} \tag{12}$$

where ϕ is the folded standard normal density (i.e., the distribution of the absolute value of a standard normal random variable), π_0 is the proportion of SNPs that are null for SCZ, and f_1 is the density of z_1. Use of the folded standard normal density is equivalent to the *theoretical null* of Efron (2007). This is justified in the current example since the summary statistics were made publicly available after performing usual genomic control procedures (Devlin and Roeder, 1999). A conservative estimate of the fdr is produced by assuming $\pi_0 = 1$ and utilizing the nonparametric estimate $\widehat{f_1}$ of f_1 produced from our proposed model. The assumption that $\pi_0 = 1$ in Equation (12) is reasonable, since for even highly polygenic traits such as SCZ, the proportion of null SNPs is considerably greater than 0.9 (Andreassen et al., 2013). Larger z-scores will have a higher probability of association with the phenotypes of interest. Here, a z-score cut-off of $z_1 \geq 3.3$ produces estimated $\widehat{\text{fdr}}(z_1) \leq .2$. The bottom panel of Figure 4 shows similar histograms to those in the top panel but for absolute z-scores ≥ 3.3.

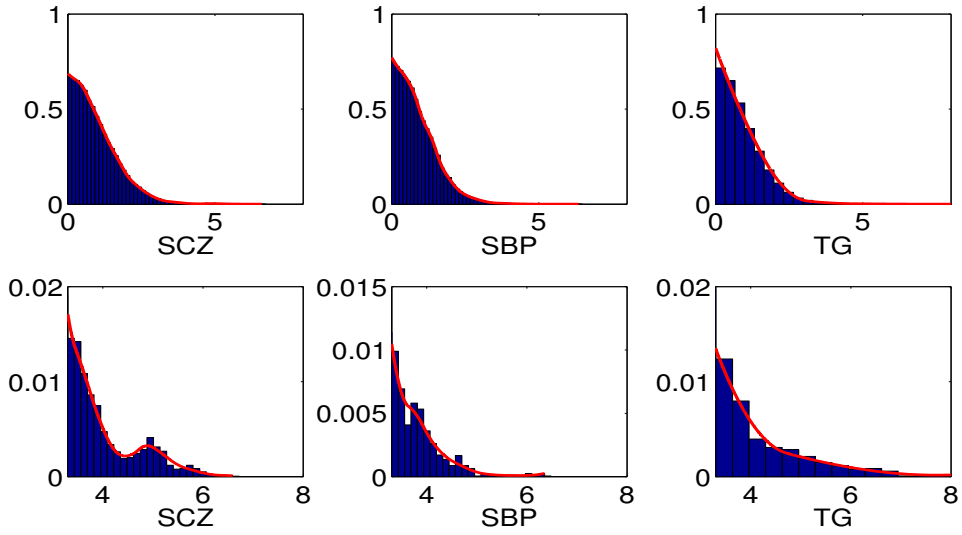

Figure 4 Top panel from left to right: marginal density histograms and density estimates of SCZ, SBP and TG. The solid lines are density estimates based on the method of Section 3. Bottom panel: similar plots for absolute z-scores ≥ 3.3.

The *conditional* local fdr is given by

$$\text{fdr}(z_1|z_2, z_3) = \frac{\pi_0(z_2, z_3)\phi(z_1)}{f(z_1|z_2, z_3)}, \tag{13}$$

where $\pi_0(z_2, z_3)$ is the conditional probability that z_1 is null given (z_2, z_3) and $f(z_1|z_2, z_3)$ is the conditional density of z_1 given (z_2, z_3). By setting $\pi_0(z_2, z_3) = 1$ in Equation (13) and using an estimate $\widehat{f}(z_1|z_2, z_3)$ based on Equation (4),

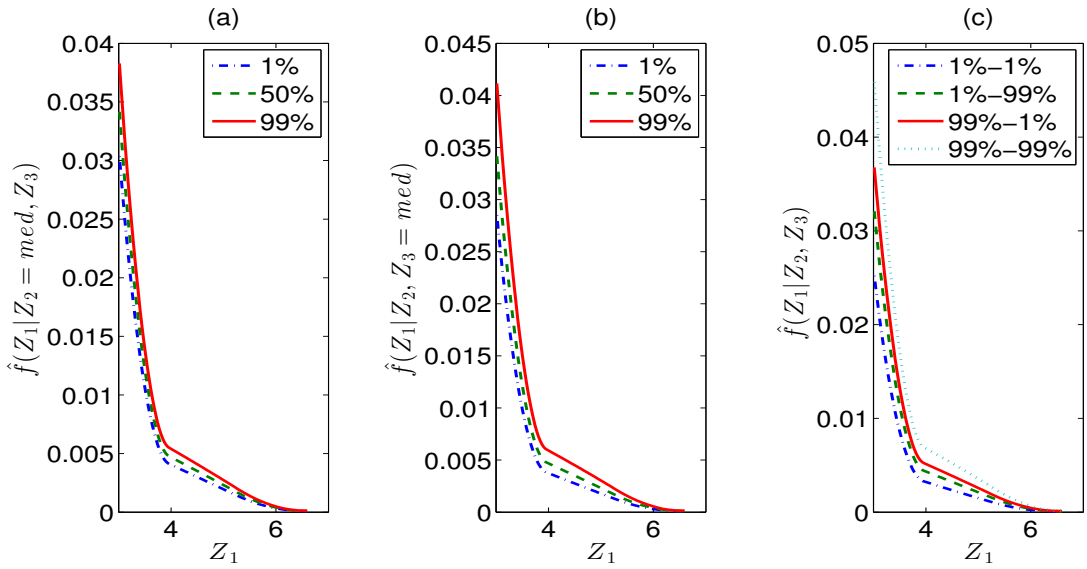

Figure 5 Conditional density estimates showing the tails only ($Z_1 \geq 3.3$).

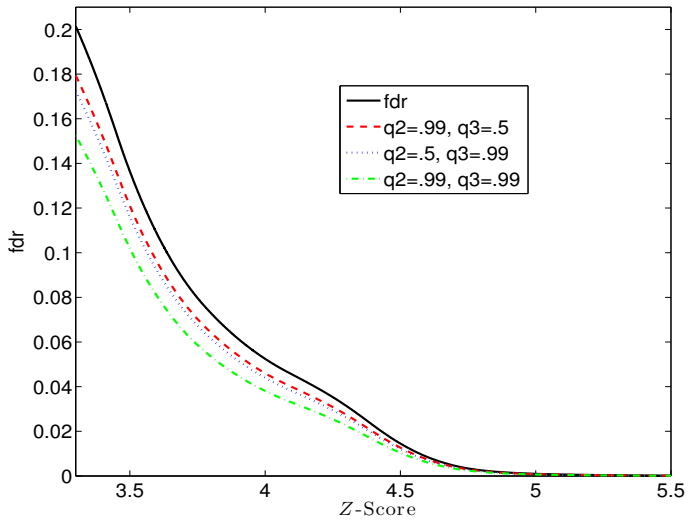

Figure 6 Conditional local fdr estimates $\widehat{f}_1(z_1|z_2, z_3)$ for different quantiles of z_2 and z_3.

we obtain a conservative estimate $\widehat{\text{fdr}}(z_1|z_2, z_3) = \phi(z_1)/\widehat{f}(z_1|z_2, z_3)$. Conditional density estimates for $z_1 \geq 3.3$ given quantiles of z_2 and z_3 are displayed in Figure 5, and conditional local fdr estimates are given in Figure 6. For a given z-score, the local fdr is lower when conditioning on larger quantiles of z_2 and z_3. For example, for $z_1 = 3.5$, the unconditional fdr estimate is $\widehat{\text{fdr}}(3.5) = .135$, whereas $\widehat{\text{fdr}}(3.5|z_2 = 3.03, z_3 = 3.63) = .099$, where $z_2 = 3.03$ and $z_3 = 3.63$ are the 99th quantiles of TG and SBP absolute z-scores, respectively. Using the commonly-used threshold of .2, there were 1,594 significant SNPs using local fdr, whereas there were 1,715 SNPs using conditional local fdr, for an increase of 7.6%.

6 Discussion

In this paper we propose a Bayesian methodology for estimating multivariate distributions wherein the marginal densities are estimated nonparametrically and linked via a Gaussian copula.

Various extensions and applications of our model are possible. For example, the use of Gaussian copulas assumes no tail dependence between pairs of variables (Embrechts et al. (2003)). Upper tail dependence between random variables Y_1 and Y_2, is defined as

$$\lim_{u\uparrow 1} \Pr\left(Y_2 > F_2^{-1}(u) | Y_1 > F_1^{-1}(u)\right). \tag{14}$$

Upper tail dependence is a measure of association in the upper-right quadrant, whereas lower tail dependence is defined analogously for the lower-left quadrant. Both upper and lower tail dependence assess the strength of relationships between extreme events of Y_1 and Y_2. While Gaussian copulas are upper- and lower-tail independent, other families of elliptical copulas (Fang et al. (2002)) are not. For example, t-copulas are a family of elliptical copulas allowing for tail dependence. Implementing our model using other families of elliptical copulas may provide better fits for random variables exhibiting dependence in extreme events.

Other extensions are possible. For example, the proposed model currently handles only continuous margins. Recent work has proposed Bayesian estimation of copula models with discrete margins by augmentation with uniform latent variables (Smith and Khaled (2012)). Additionally, Pitt et al. (2006) developed a graphical model selection prior on the Gaussian copula correlation matrix.

Finally, it is possible to adapt the copula model to estimate the multivariate local false discovery rate (Andreassen et al., 2013). This involves an extension of Efron's (2007) two-group mixture model to a 2^P-group mixture model, where P is the number of phenotypes of interest. Andreassen et al. (2013) have shown that bivariate local false discovery rate is potentially much more powerful than univariate local false discovery rate, and highly informative biologically.

Acknowledgements O. Rosen was supported in part by NSF grant DMS-0804140 and by the National Security Agency under Grant Number H98230-12-1-0246. The United States Government is authorized to reproduce and distribute reprints notwithstanding any copyright notation herein.

W. Thompson was supported by NIH grants R01HD061414, 5R01AG022381, and R01MH092793.

This project was also supported by NCRR grant no. 5G12RR008124 and NIMHD grant no. G12MD007592 from the National Institutes of Health.

Conflict of Interest

The authors have declared no conflict of interest.

Appendix A

Conditional Density Formula for Gaussian Copula

From Equation (1), it follows that the joint probability density function of $Y_1, \ldots, Y_p$ is

$$f(y_1, \ldots, y_p) = \prod_{i=1}^{p} f_i(y_i) c(F_1(y_1), \ldots, F_p(y_p)),$$

so that

$$f(y_1|y_2, \ldots, y_p) = f_1(y_1) \frac{c(F_1(y_1), \ldots, F_p(y_p))}{c_{-1}(F_2(y_2), \ldots, F_p(y_p))}, \tag{15}$$

where c_{-1} is the copula density with correlation matrix Ω_{11}, defined in Section 2. Differentiating (2), substituting $q_j = \Phi^{-1}(F_j(y_j))$ and plugging into (15) results in

$$f(y_1|y_2, \ldots, y_p) = f_1(y_1) \frac{\phi_p(q_1, \ldots, q_p)}{\phi(q_1)\phi_{p-1}(q_2, \ldots, q_p)}, \tag{16}$$

where ϕ_p and ϕ_{p-1} are the densities corresponding to $N(\mathbf{0}_p, \Omega)$ and $N(\mathbf{0}_{p-1}, \Omega_{11})$, respectively, and ϕ is the density of the univariate standard normal. The notation $\mathbf{0}_p$ means a vector of p zeros. Plugging the normal densities into (16) leads to

$$f(y_1|y_2, \ldots, y_p) = f_1(y_1) \frac{|\Omega|^{-1/2}}{|\Omega_{11}|^{-1/2}} \exp\Big\{-\frac{1}{2}\boldsymbol{q}'\Omega^{-1}\boldsymbol{q} + \frac{q_1^2}{2} + \frac{1}{2}\boldsymbol{q}'_{-1}\Omega_{11}^{-1}\boldsymbol{q}_{-1}\Big\}. \tag{17}$$

Applying Schur decomposition to Ω^{-1} gives

$$\Omega^{-1} = \begin{pmatrix} 1 & \mathbf{0}'_{p-1} \\ -\Omega_{11}^{-1}\boldsymbol{\omega} & I_{p-1} \end{pmatrix} \begin{pmatrix} (1 - \boldsymbol{\omega}'\Omega_{11}^{-1}\boldsymbol{\omega})^{-1} & \mathbf{0}'_{p-1} \\ \mathbf{0}_{p-1} & \Omega_{11}^{-1} \end{pmatrix} \begin{pmatrix} 1 & -\boldsymbol{\omega}'\Omega_{11}^{-1} \\ \mathbf{0}_{p-1} & I_{p-1} \end{pmatrix}, \tag{18}$$

where I_{p-1} is the $(p-1) \times (p-1)$ identity matrix. Using (18) implies

$$\boldsymbol{q}'\Omega^{-1}\boldsymbol{q} = (q_1 - \boldsymbol{\omega}'\Omega_{11}^{-1}\boldsymbol{q}_{-1})^2(1 - \boldsymbol{\omega}'\Omega_{11}^{-1}\boldsymbol{\omega})^{-1} + \boldsymbol{q}'_{-1}\Omega_{11}^{-1}\boldsymbol{\omega} \tag{19}$$

and

$$|\Omega|^{-1} = (1 - \boldsymbol{\omega}'\Omega_{11}^{-1}\boldsymbol{\omega})^{-1}|\Omega_{11}|^{-1}. \tag{20}$$

Plugging expressions (19) and (20) into (17) gives the conditional density (4).

Appendix B

Details of the Sampling Scheme

Section 3.3 describes the outline of the sampling scheme. This appendix provides some more details.

1. **Sampling the $\boldsymbol{\gamma}_j$s**

   The augmented conditional posterior of $\boldsymbol{\gamma}_j$ is

$$\begin{aligned} p(\boldsymbol{\gamma}_j|\{\boldsymbol{\gamma}_l\}_{l\neq j}, \Omega, \text{data}, \boldsymbol{z}_j) &\propto \exp\left\{-\tfrac{1}{2}\sum_{i=1}^n \boldsymbol{q}_i'(\Omega^{-1}-I)\boldsymbol{q}_i\right\}\prod_{i=1}^n\prod_k^{K_j} c_{jk}^{z_{ijk}} \\ &\times \exp\left\{-\tfrac{1}{2\tau_j^2}\boldsymbol{\gamma}_j' P_j^*\boldsymbol{\gamma}_j\right\}, \end{aligned}$$

   where $\boldsymbol{z}_j = \{z_{ijk}, i = 1,\ldots,n, k = 1,\ldots,K_j\}$. The marginal pdfs and cdfs $f_j(y_{ij})$ and $F_j(y_{ij})$ are modeled as in (6) and (7), respectively. With the latent indicators, $\boldsymbol{z}_j$, $q_{ij} = \Phi^{-1}(\prod_{k=1}^{K_j}(G_k(y_{ij}))^{z_{ijk}})$. Since this is not a function of $\boldsymbol{\gamma}_j$, the log conditional posterior of $\boldsymbol{\gamma}_j$ reduces to

$$\log p(\boldsymbol{\gamma}_j|\boldsymbol{z}_j) \stackrel{c}{=} \sum_k^{K_j} n_{jk}\log c_{jk} - \frac{1}{2\tau_j^2}\boldsymbol{\gamma}_j' P_j^*\boldsymbol{\gamma}_j, \tag{21}$$

   where $\stackrel{c}{=}$ denotes equality up to a constant, and $n_{jk} = \sum_{i=1}^n z_{ijk}$. To sample $\boldsymbol{\gamma}_j$, (21) is maximized at each iteration with respect to $\boldsymbol{\gamma}_j$. The maximizer and the negative inverse Hessian of (21), evaluated at the maximizer, are used in a Metropolis-Hastings step with a multivariate t distribution with small degrees of freedom (e.g., 4) as the proposal distribution.

2. **Sampling the τ_j^2s**

   The τ_j^2, $j = 1,\ldots,p$, are sampled from

$$p(\tau_j^2|\boldsymbol{\gamma}_j) \propto (\tau_j^2)^{-(K_j-1)/2}\exp\Big\{-\frac{1}{2\tau_j^2}\boldsymbol{\gamma}_j' P_j^*\boldsymbol{\gamma}_j\Big\}\mathbb{I}_{[0,a_{\tau_j^2}]}(\tau_j^2),$$

   where the indicator function $\mathbb{I}_{[a,b]}(x)$ is equal to 1 if $x \in [a,b]$, and is equal to zero, otherwise. Thus, τ_j^2 is drawn from an inverse gamma distribution, $IG(\frac{1}{2}(K_j-3), \frac{1}{2}\boldsymbol{\gamma}_j' P_j^*\boldsymbol{\gamma}_j)$, truncated at $a_{\tau_j^2}$.

3. **Sampling Ω**

$$\begin{aligned} p(\Omega|\{\boldsymbol{\gamma}_j\}_{j=1}^p, \text{data}) &\propto |\Omega|^{-n/2}\exp\Big\{-\tfrac{1}{2}\sum_{i=1}^n \boldsymbol{q}_i'(\Omega^{-1}-I)\boldsymbol{q}_i\Big\}p(\Omega) \\ &\propto |\Omega|^{-n/2}\exp\Big\{-\tfrac{1}{2}\text{tr}(S\Omega^{-1})\Big\}p(\Omega), \end{aligned}$$

   where $p(\Omega)$ is the prior on Ω, $S = \sum_{i=1}^n \boldsymbol{q}_i\boldsymbol{q}_i'$ and $q_{ij} = \Phi^{-1}(\sum_{k=1}^K c_{jk}G_k(y_{ij}))$. The matrix Ω is sampled in a random-walk Metropolis step via the decomposition described in Section 3.2. The elements of L are sampled individually using a normal proposal centered at the current value.

4. **Sampling the indicators**

   The indicators are sampled as Bernoulli random variables with probability

$$p(z_{ijk} = 1|\text{data}, \{\boldsymbol{\gamma}_j\}_{j=1}^p) \propto \exp\Big\{-\frac{1}{2}\boldsymbol{q}_i'(\Omega^{-1}-I)\boldsymbol{q}_i\Big\}c_{jk}g_k(y_{ij}),$$

   where $q_{ij} = \Phi^{-1}(\prod_{k=1}^K(G_k(y_{ij}))^{z_{ijk}})$.

References

Andreassen, O., Djurovic, S., Thompson, W., Schork, A., Kendler, K., ODonovan, M., Rujescu, D., Werge, T., van de Bunt, M., Morris, A., McCarthy, M., Roddey, J., McEvoy, L., Desikan, R., and Dale, A. (2013). Improved detection of common variants associated with schizophrenia by leveraging pleiotropywith cardiovascular disease risk factors. *American Journal of Human Genetics* **92,** 197–209.

Chan, J.-C. and Jeliazkov, I. (2009). MCMC estimation of restricted covariance matrices. *Journal of Computational and Graphical Statistics* **18,** 457–480.

Chen, X., Fan, Y., and Tsyrennikov, V. (2006). Efficient estimation of semiparametric multivariate copula models. *Journal of the American Statistical Association* **101,** 1228–1240.

Chib, S. and Jeliazkov, I. (2006). Inference in semiparametric dynamic models for binary longitudinal data. *Journal of the American Statistical Association* **101,** 685–700.

Colton, C. and Manderscheid, R. (2006). Congruencies in increased mortality rates, years of potential life lost, and causes of death among public mental health clients in eight states. *Preventing Chronic Diseases* **3,** A42.

Craiu, V. and Sabeti, A. (2012). In mixed company: Bayesian inference for bivariate conditional copula models with discrete and continuous outcomes. *Journal of Multivariate Analysis* **110,** 106–120.

Crane, G. and Van Der Hoek, J. (2008). Conditional expectation formulae for copulas. *Australian and New Zealand Journal of Statistics* **50,** 53–67.

Danaher, P. and Smith, M. (2011). Modeling multivariate distributions using copulas: applications in marketing. *Marketing Science* **30,** 4–21.

Daniels, M. and Kass, R. (2001). Shrinkage estimators for covariance matrices. *Biometrics* **57,** 1173–1184.

Daniels, M. and Pourahmadi, M. (2009). Modeling covariance matrices via partial autocorrelations. *Journal of Multivariate Analysis* **100,** 2352–2363.

De Hert, M., van Winkel, R., Van Eyck, D., Hanssens, L., Wampers, M., Scheen, A., and Peuskens, J. (2006). Prevalence of the metabolic syndrome in patients with schizophrenia treated with antipsychotic medication. *Schizophrenia Research* **83,** 87–93.

Devlin, D. and Roeder, K. (1999). Genomic control for association studies. *Biometrics* **55,** 997–1004.

Efron, B. (2007). Size, power and false discovery rates. *The Annals of Statistics* **35,** 1351–1377.

Ehret, G., Munroe, P., Rice, K., Bochud, M., Johnson, A., Chasman, D., Smith, A., Tobin, M., Verwoert, G., Hwang, S., and et al. (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **478,** 103–109.

Eilers, P. and Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11,** 89–121.

Embrechts, P., Lindskog, F., and McNeil, A. (2003). Modeling dependence with copulas and applications to risk management. In Rachev, S., editor, *Handbook of Heavy Tailed Distributions in Finance*, chapter 8, pages 329–384. Elsevier.

Fang, H.-B., Fang, K.-T., and Kotz, S. (2002). The meta-elliptical distributions with given marginals. *Journal of Multivariate Analysis* **82,** 1–16. (Corrigendum: Journal of Multivariate Analysis 94, 222-223, 2005).

Genest, C. and Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering* **12,** 347–368.

Ghidey, W., Lesaffre, E., and Eilers, P. (2004). Smooth random effects distribution in a linear mixed modeling. *Biometrics* **60,** 945–953.

Glazier, A., Nadeau, J., and Aitman, T. (2002). Finding genes that underlie complex traits. *Science* **298,** 2345–2349.

Hansen, T., Ingason, A., Djurovic, S., Melle, I., Fenger, M., Gustafsson, O., Jakobsen, K., Rasmussen, H., Tosato, S., Rietschel, M., and et al. (2011). At-risk variant in tcf7l2 for type ii diabetes increases risk of schizophrenia. *Biological Psychiatry* **70,** 59–63.

Heid, I., Jackson, A., Randall, J., Winkler, T., Qi, L., Steinthorsdottir, V., Thorleifsson, G., Zillikens, M., Speliotes, E., Magi, R., and et al. (2010). Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nature Genetics* **42,** 949–960.

Hindorff, L., Sethupathy, P., Junkins, H., Ramos, E., Mehta, J., Collins, F., and Manolio, T. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* **106,** 9362–9367.

Hirschhorn, J. and Daly, M. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Review Genetics* **6,** 95–108.

Hoff, P. (2007). Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics* **1,** 265–283.

Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis* **97,** 2177–2189.

Käärik, E. and Käärik, M. (2009). Modeling dropouts by conditional distribution, a copula-based approach. *Journal of Statistical Planning and Inference* **139,** 3830–3835.

Kaddurah-Daouk, R., McEvoy, J., Baillie, R., Lee, D., Yao, J., Doraiswamy, P., and Krishnan, K. (2007). Metabolomic mapping of atypical antipsychotic effects in schizophrenia. *Molecular Psychiatry* **12,** 934–945.

Kolev, N. and Paiva, D. (2009). Copula-based regression models: a survey. *Journal of Statistical Planning and Inference* **139,** 3847–3856.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics* **13,** 183–212.

Laursen, T., Munk-Olsen, T., and Vestergaard, M. (2012). Life expectancy and cardiovascular mortality in persons with schizophrenia. *Current Opinion Psychiatry* **25,** 83–88.

Manolio, T., Collins, F., Cox, N., Goldstein, D., Hindorff, L., Hunter, D., McCarthy, M., Ramos, E., Cardon, L., and Chakravarti, A., e. a. (2009). Finding the missing heritability of complex diseases. *Nature* **461,** 747–753.

Marder, S., Essock, S., Miller, A., Buchanan, R., Casey, D., Davis, J., Kane, J., Lieberman, J., Schooler, N., Covell, N., and et al. (2004). Physical health monitoring of patients with schizophrenia. *American Journal of Psychiatry* **161,** 1334–1349.

Mitchell, A., Vancampfort, D., Sweers, K., van Winkel, R., Yu, W., and De Hert, M. (2011). Prevalence of metabolic syndrome and metabolic abnormalities in schizophrenia and related disorders–a systematic review and meta-analysis. *Schizophrenia Bulletin* .

Pitt, M., Chan, D., and Kohn, R. (2006). Efficient Bayesian inference for Gaussian copula regression models. *Biometrika* **93,** 537–554.

Raphael, T. and Parsons, J. (1921). Blood sugar studies in dementia praecox and manic depressive insanity. *Archives of Neurology and Psychiatry* **5,** 687–709.

Ripke, S., Sanders, A., Kendler, K., Levinson, D., Sklar, P., Holmans, P., Lin, D., Duan, J., Ophoff, R., Andreassen, O., and et al. (2011). Genome-wide association study identifies five new schizophrenia loci. *Nature Genetics* **43,** 969–976.

Ryan, M., Collins, P., and Thakore, J. (2003). Impaired fasting glucose tolerance in first-episode, drug-naive patients with schizophrenia. *American Journal of Psychiatry* **160,** 284–289.

Saha, S., Chant, D., and McGrath, J. (2007). A systematic review of mortality in schizophrenia: is the differential mortality gap worsening over time? *Archives of General Psychiatry* **64,** 1123–1131.

Schork, A., Thompson, W., Pham, P., Torkamani, A., Roddey, J., Sullivan, P., Kelsoe, J., Purcell, S., ODonovan, M., Schork, N., Andreassen, O., and Dale, A. (2013). All snps are not created equal: Genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated snps. *PLoS Genetics* **9,** doi:10.1371/journal.pgen.1003449.

Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J., Zgaga, L., Manolio, T., Rudan, I., McKeigue, P., Wilson, J., and Campbell, H. (2011). Abundant pleiotropy in human complex diseases and traits. *American Journal of Human Genetics* **89,** 607–618.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leur marges. *Publ. Inst. Statist. Univ. Paris* **8,** 229–231.

Smith, M. and Khaled, M. (2012). Estimation of copula models with discrete margins via Bayesian data augmentation. *Journal of the American Statistical Association* **107,** 290–303.

Smith, M., Min, A., Almeida, C., and Czado, C. (2010). Modeling longitudinal data using a pair-copula ecomposition of serial dependence. *Journal of the American Statistical Association* **105,** 1467–1479.

Song, P. (2000). Multivariate dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics* **27,** 305–320.

Song, P. X.-K., Li, M., and Yuan, Y. (2009). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics* **65,** 60–68.

Speliotes, E., Willer, C., Berndt, S., Monda, K., Thorleifsson, G., Jackson, A., Allen, H., Lindgren, C., Luan, J., Magi, R., and et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics* **42,** 937–948.

Staudenmayer, J., Ruppert, D., and Buonaccorsi, J. (2008). Density estimation in the presence of heteroscedastic measurement error. *Journal of the American Statistical Association* **103,** 726–736.

Teslovich, T., Musunuru, K., Smith, A., and et al. (2010). Biological, clinical, and population relevance of 95 loci for blood lipids. *Nature* **466,** 707–713.

Yang, J., Benyamin, B., McEvoy, B., Gordon, S., Henders, A., Nyholt, D., Madden, P., Heath, A., Martin, N., and Montgomery, G.W., e. a. (2010). Common snps explain a large proportion of the heritability for human height. *Nature Genetics* **42,** 565–569.

