Efficient estimation in the semiparametric normal regression-copula model with a focus on QTL mapping

Basrak, B.; Klaassen, C.

DOI
10.1214/12-IMSCOLL903

Publication date
2013

Document Version
Final published version

Published in
Institute of Mathematical Statistics Collections

IMS Collections
From Probability to Statistics and Back: High-Dimensional Models and Processes
Vol. 9 (2013) 20–32

DOI: 10.1214/12-IMSCOLL903

Efficient estimation in the semiparametric normal regression-copula model with a focus on QTL mapping

Bojan Basrak[1] **and Chris A. J. Klaassen**[2]

University of Zagreb and University of Amsterdam and EURANDOM

Abstract: The semiparametric normal copula model is studied with a correlation matrix that depends on a covariate. The bivariate version of this regression-copula model has been proposed for statistical analysis of Quantitative Trait Loci (QTL) via twin data. Appropriate linear combinations of Van der Waerden's normal scores rank correlation coefficients yield $\sqrt{n}$-consistent estimators of the coefficients in the correlation function, i.e. of the regression parameters. They are used to construct semiparametrically efficient estimators of the regression parameters.

1. Normal regression-copula model

Identification of the location of genes contributing to so-called *quantitative traits* is an important problem in genetics. Quantitative traits are phenotypes that can be measured numerically and show continuous variation in living organisms, such as height or cholesterol level in humans. One of the first steps in this identification process is *linkage analysis*. In human studies this is typically performed using the normal variance components approach; see e.g. Cherny, Sham and Cardon [3]. Basrak et al. [1] have proposed to incorporate not necessarily normally distributed traits using the semiparametric normal regression-copula model discussed in the present paper.

We consider a so-called sib-pair study, where the siblings form twins, but non-identical twins. Data are collected about n independent sib-pairs and we measure a particular quantitative trait for each person in the pair, resulting in (Y_1, Y_2), say. Additionally, the IBD (Identical By Descent) status of the two siblings is determined at a given marker (or a 'gene' loosely speaking). It is the number Z' of alleles that the siblings have in common by descent. So, if they inherited the same alleles both from their mother and their father, then Z' equals 2. This happens with probability 1/4. With the same probability they inherited different alleles from both parents; so, $P(Z' = 0) = 1/4$. For any two individuals this number is an element of the set $\{0, 1, 2\}$ and hence $P(Z' = 1) = 1/2$. The number Z' represents the degree of

[1]Department of Mathematics, University of Zagreb, Bijenička 30, Zagreb, Croatia, e-mail: bbasrak@math.hr

[2]Korteweg-de Vries Institute for Mathematics, University of Amsterdam, P.O. Box 94248, 1090 GE Amsterdam, The Netherlands, EURANDOM, Eindhoven, The Netherlands, e-mail: c.a.j.klaassen@uva.nl

AMS 2000 subject classifications: Primary 62G05, 62G20; secondary 62P10

Keywords and phrases: Semiparametric inference, Van der Waerden, $\sqrt{n}$-consistency, linkage analysis, QTL mapping

relatedness between the two siblings at the given marker. Note, however, that the siblings could have both alleles the same, but the IBD status Z' still equal to 0.

In practice Z' has to be estimated, so geneticists sometimes work with an estimate of it that can take any value in the interval $[0, 2]$. For convenience, we will assume the IBD status can be determined exactly and we will work with the numbers $Z = Z' - 1 \in \{-1, 0, 1\}$ instead.

Let $G(\cdot)$ be the distribution function of Y_1 and Y_2, and let us assume this distribution function is continuous. Consequently, the random variables

$$T_1 = \psi(Y_1) = \Phi^{-1}(G(Y_1)) \quad \text{and} \quad T_2 = \psi(Y_2) = \Phi^{-1}(G(Y_2)) \tag{1.1}$$

have a standard normal distribution. The fundamental assumption is that, conditionally on Z, the 2-vector $T = (T_1, T_2)^T$ has a bivariate normal distribution with correlation coefficient $\rho + \gamma Z$, more precisely,

$$\mathcal{L}\left(\begin{pmatrix}\psi(Y_1)\\ \psi(Y_2)\end{pmatrix} \,\middle|\, Z\right) = \mathcal{L}\left(\begin{pmatrix}T_1\\ T_2\end{pmatrix} \,\middle|\, Z\right) = \mathcal{N}\left(0, \begin{pmatrix}1 & \varrho+\gamma Z\\ \varrho+\gamma Z & 1\end{pmatrix}\right). \tag{1.2}$$

Since Z takes its values in $\{-1, 0, 1\}$, we assume that the unknown values of the parameters ϱ and γ satisfy $|\varrho| + |\gamma| < 1$.

The main assumption of our statistical model is that there is a monotone, differentiable, and invertible transformation $\psi : \mathbb{R} \to \mathbb{R}$ such that we observe n i.i.d. copies of the random vector $X = (Y^T, Z)^T = (Y_1, Y_2, Z)^T$ with

$$\mathcal{L}\left((\psi(Y_1), \psi(Y_2), Z)\right) = \mathcal{L}\left((T_1, T_2, Z)\right). \tag{1.3}$$

The main problem of linkage analysis in this setting, is inference about the regression parameter γ, which tells us if higher IBD status translates into closer phenotypic values. More specifically, one would like to determine if for a certain marker, the parameter γ differs from 0. A test for the null hypothesis $\gamma = 0$ has been proposed, and the appropriate approximate significance levels have been found; see Basrak et al. [1] and cf. Dupuis and Siegmund [6]. Here we will focus on semiparametrically efficient estimation of the regression parameters γ and ϱ.

In Section 2 we will construct an asymptotic bound on the performance of estimators of these parameters. Based on the Van der Waerden normal scores rank correlation coefficient we will construct $\sqrt{n}$-consistent estimators in Section 3. We also need an estimator of the transformation $\psi(\cdot)$ that has a sufficiently high convergence rate in an appropriate squared distance. We present such an estimator in Section 4. Finally, by a sample splitting technique and based on the preliminary $\sqrt{n}$-consistent estimators of the regression parameters and the estimator of the transformation, we are able to construct estimators that attain the bound from Section 2. A technical result about empirical distributions is proved in the Appendix.

2. Asymptotic bound

Let us consider model (1.2) and (1.3) with $\gamma = 0$, which means that, in fact, we observe n i.i.d. copies of the random vector $X = (Y_1, Y_2)^T$ with

$$\mathcal{L}\left(\begin{pmatrix}\psi(Y_1)\\ \psi(Y_2)\end{pmatrix}\right) = \mathcal{L}(T) = \mathcal{L}\left(\begin{pmatrix}T_1\\ T_2\end{pmatrix}\right) = \mathcal{N}\left(\mathbf{0}, \begin{pmatrix}1 & \varrho\\ \varrho & 1\end{pmatrix}\right) \tag{2.1}$$

for $|\varrho| < 1$. This classic semiparametric normal copula model has been studied in Klaassen and Wellner [9]. They show that at $(\varrho_0, \psi_0(\cdot))$ with $|\varrho_0| < 1$ the least favorable parametric submodel of the semiparametric model from (2.1) for estimating the correlation coefficient ϱ is the correlation-scale model that we get by restricting the nonparametric class of transformations $\psi(\cdot)$ to the one-dimensional parametric class of transformations $\psi_0(\cdot)/\sigma$ with $\sigma > 0$. So, $(Y_1, Y_2)^T$ is ruled by this parametric model, if

$$\mathcal{L}\left(\begin{pmatrix}\psi_0(Y_1)/\sigma\\ \psi_0(Y_2)/\sigma\end{pmatrix}\right) = \mathcal{L}\left(\begin{pmatrix}T_1\\ T_2\end{pmatrix}\right) = \mathcal{N}\left(\mathbf{0}, \begin{pmatrix}1 & \varrho\\ \varrho & 1\end{pmatrix}\right) \tag{2.2}$$

or equivalently

$$\mathcal{L}\left(\begin{pmatrix}\psi_0(Y_1)\\ \psi_0(Y_2)\end{pmatrix}\right) = \mathcal{N}\left(\mathbf{0}, \sigma^2\begin{pmatrix}1 & \varrho\\ \varrho & 1\end{pmatrix}\right), \qquad |\varrho| < 1,\ \sigma > 0, \tag{2.3}$$

where (ϱ_0, σ_0) with $\sigma_0 = 1$ corresponds to $(\varrho_0, \psi_0(\cdot))$.

In the terminology of Drost, Klaassen and Werker [4, 5] one may say that in the presence of the nuisance parameter σ the parameter of interest ϱ can be estimated adaptively. The Van der Waerden normal scores rank correlation coefficient is such an adaptive estimator of ϱ in the presence of σ, as has been shown by Klaassen and Wellner [9].

One wonders if the semiparametric normal regression-copula model has an analogous structure, i.e., if at $(\gamma_0, \varrho_0, \psi_0(\cdot))$ with $|\varrho_0| + |\gamma_0| < 1$ the least favorable parametric submodel of the semiparametric model from (1.2) for estimating the correlation coefficients γ and ϱ is the regression-correlation-scale model

$$\begin{aligned}&\mathcal{L}\left(\begin{pmatrix}\psi_0(Y_1)\\ \psi_0(Y_2)\end{pmatrix}\,\middle|\, Z\right)\\ &\quad = \mathcal{N}\left(\mathbf{0}\,,\, \sigma^2\begin{pmatrix}1 & \varrho+\gamma Z\\ \varrho+\gamma Z & 1\end{pmatrix}\right), \qquad |\varrho| + |\gamma| < 1,\ \sigma > 0,\end{aligned} \tag{2.4}$$

at $(\gamma_0, \varrho_0, \sigma_0)$ with $\sigma_0 = 1$.

Least favorable or not, this parametric submodel (2.4) yields an asymptotic lowerbound to the performance of estimators by e.g. the Hájek-LeCam convolution theorem. This bound is determined by the Fisher information matrix, which we will compute first.

For notational convenience we write $\rho(Z) = \varrho + \gamma Z$ and we take $\psi_0(\cdot)$ to be the identity function. We study the density of $(Y_1, Y_2, Z)^T$ with respect to Lebesgue measure on $\mathbb{R}^2$ times counting measure on $\{-1, 0, 1\}$. The score functions or logarithmic derivatives of this density with respect to γ, ϱ, and σ^2 equal

$$\begin{aligned}\dot{\ell}_\gamma &= Z\dot{\ell}_\varrho\,,\\ \dot{\ell}_\varrho &= \frac{1}{\sigma^2(1-\rho^2(Z))^2}\left\{-\rho(Z)\left[Y_1^2 + Y_2^2\right] + \left(1+\rho^2(Z)\right)Y_1Y_2\right.\\ &\qquad\qquad \left. + \sigma^2\rho(Z)\left(1-\rho^2(Z)\right)\right\},\\ \dot{\ell}_{\sigma^2} &= \frac{1}{2\sigma^4(1-\rho^2(Z))}\left\{\left[Y_1^2 + Y_2^2\right] - 2\rho(Z)Y_1Y_2\right.\\ &\qquad\qquad \left. - 2\sigma^2\left(1-\rho^2(Z)\right)\right\}.\end{aligned} \tag{2.5}$$

The Fisher information matrix is the covariance matrix of these score functions and equals

$$I = \begin{pmatrix} E\left(\frac{1+\rho^2(Z)}{(1-\rho^2(Z))^2}\ Z^2\right) & E\left(\frac{1+\rho^2(Z)}{(1-\rho^2(Z))^2}\ Z\right) & -\frac{1}{\sigma^2}E\left(\frac{\rho(Z)}{1-\rho^2(Z)}\ Z\right) \\ E\left(\frac{1+\rho^2(Z)}{(1-\rho^2(Z))^2}\ Z\right) & E\left(\frac{1+\rho^2(Z)}{(1-\rho^2(Z))^2}\right) & -\frac{1}{\sigma^2}E\left(\frac{\rho(Z)}{1-\rho^2(Z)}\right) \\ -\frac{1}{\sigma^2}E\left(\frac{\rho(Z)}{1-\rho^2(Z)}\ Z\right) & -\frac{1}{\sigma^2}E\left(\frac{\rho(Z)}{1-\rho^2(Z)}\right) & \frac{1}{\sigma^4} \end{pmatrix}. \tag{2.6}$$

Taking $\sigma = \sigma_0 = 1$ we denote this information matrix by

$$I = \begin{pmatrix} \mu_2 & \mu_1 & -\nu_1 \\ \mu_1 & \mu_0 & -\nu_0 \\ -\nu_1 & -\nu_0 & 1 \end{pmatrix}. \tag{2.7}$$

We note that the score functions from (2.5) are linear combinations of $Y_1^2 + Y_2^2$, Y_1Y_2, and 1 with coefficients that are rational functions of Z, γ, and ϱ. In view of formulae (2.4.3) and (2.4.4) of Bickel, Klaassen, Ritov and Wellner [2] we may conclude that the efficient influence functions for estimation of the regression parameters γ and ϱ are also linear combinations of $Y_1^2 + Y_2^2$, Y_1Y_2, and 1 with coefficients that are functions of Z, γ, and ϱ. Therefore, we may write the efficient influence function for γ as

$$\tilde{\ell}_\gamma = a_\gamma(Z;\gamma,\varrho)[Y_1^2 + Y_2^2] + b_\gamma(Z;\gamma,\varrho)Y_1Y_2 + c_\gamma(Z;\gamma,\varrho) \tag{2.8}$$

and the efficient influence function for ϱ as

$$\tilde{\ell}_\varrho = a_\varrho(Z;\gamma,\varrho)[Y_1^2 + Y_2^2] + b_\varrho(Z;\gamma,\varrho)Y_1Y_2 + c_\varrho(Z;\gamma,\varrho). \tag{2.9}$$

Some computation shows that with

$$D = \mu_0\mu_2 - \mu_0\nu_1^2 + \mu_1\nu_0\nu_1 - \mu_2\nu_0^2 \tag{2.10}$$

the coefficients satisfy

$$\begin{aligned} D\, a_\gamma(Z;\gamma,\varrho) &= -(\mu_0 - \nu_0^2)\frac{Z\rho(Z)}{(1-\rho^2(Z))^2} + (\mu_1 - \nu_0\nu_1)\frac{\rho(Z)}{(1-\rho^2(Z))^2} \\ &\quad + (\mu_0\nu_1 - \mu_1\nu_0)\frac{1}{2(1-\rho^2(Z))}, \\ D\, b_\gamma(Z;\gamma,\varrho) &= (\mu_0 - \nu_0^2)\frac{Z(1+\rho^2(Z))}{(1-\rho^2(Z))^2} - (\mu_1 - \nu_0\nu_1)\frac{1+\rho^2(Z)}{(1-\rho^2(Z))^2} \\ &\quad - (\mu_0\nu_1 - \mu_1\nu_0)\frac{\rho(Z)}{1-\rho^2(Z)}, \\ D\, c_\gamma(Z;\gamma,\varrho) &= (\mu_0 - \nu_0^2)\frac{Z\rho(Z)}{1-\rho^2(Z)} - (\mu_1 - \nu_0\nu_1)\frac{\rho(Z)}{1-\rho^2(Z)} \\ &\quad - \mu_0\nu_1 + \mu_1\nu_0, \\ D\, a_\varrho(Z;\gamma,\varrho) &= (\mu_1 - \nu_0\nu_1)\frac{Z\rho(Z)}{(1-\rho^2(Z))^2} - (\mu_2 - \nu_1^2)\frac{\rho(Z)}{(1-\rho^2(Z))^2} \\ &\quad - (\mu_1\nu_1 - \mu_2\nu_0)\frac{1}{2(1-\rho^2(Z))}, \end{aligned} \tag{2.11}$$

$$D\,b_\varrho(Z;\gamma,\varrho) = -(\mu_1-\nu_0\nu_1)\frac{Z(1+\rho^2(Z))}{(1-\rho^2(Z))^2} + (\mu_2-\nu_1^2)\frac{1+\rho^2(Z)}{(1-\rho^2(Z))^2} + (\mu_1\nu_1-\mu_2\nu_0)\frac{\rho(Z)}{1-\rho^2(Z)},$$

$$D\,c_\varrho(Z;\gamma,\varrho) = -(\mu_1-\nu_0\nu_1)\frac{Z\rho(Z)}{1-\rho^2(Z)} + (\mu_2-\nu_1^2)\frac{\rho(Z)}{1-\rho^2(Z)} + \mu_1\nu_1-\mu_2\nu_0.$$

We will not compute these coefficients explicitly as functions of γ and ϱ, but we just mention that in the special case of $\varrho = 0$

$$E(\tilde{\ell}_\gamma^2) = \frac{4(1-\gamma^2)^2}{2+\gamma^2} \tag{2.12}$$

holds.

Notice that (2.6), (2.7), and (2.11) imply

$$Ec_\gamma(Z;\gamma,\varrho) = Ec_\varrho(Z;\gamma,\varrho) = 0.$$

Since an influence function has mean zero, this yields in view of (2.8) and (2.9) the equalities

$$E\left(2a_\gamma(Z;\gamma,\varrho) + \rho(Z)b_\gamma(Z;\gamma,\varrho)\right) = 0 \tag{2.13}$$

and

$$E\left(2a_\varrho(Z;\gamma,\varrho) + \rho(Z)b_\varrho(Z;\gamma,\varrho)\right) = 0, \tag{2.14}$$

respectively. These equalities may also be derived straightforwardly.

The functions in (2.8) and (2.9) are called efficient influence functions, since any estimators of γ and ϱ that are efficient within the parametric submodel in the sense of e.g. the convolution theorem, are asymptotically linear in these influence functions.

The regression-correlation-scale model is least favorable for the normal regression-copula model if there exist estimators in this semiparametric model that are asymptotically linear in the influence functions from (2.8) and (2.9) with Y_j replaced by $\psi_0(Y_j)$.

3. $\sqrt{n}$-consistent estimators

In the normal regression-copula model of (1.2) the Van der Waerden normal scores rank correlation coefficient $\hat{\rho}_n(z)$ that is based on all observations with $Z_i = z$, is semiparametrically efficient in estimating $\varrho + \gamma z$, $z = -1, 0, 1$, as follows from Klaassen and Wellner [9]. One would guess then that

$$\tilde{\gamma}_n = \frac{1}{2}\left\{\hat{\rho}_n(1) - \hat{\rho}_n(-1)\right\} \tag{3.1}$$

estimates γ efficiently, because this is the linear combination of the $\hat{\rho}_n(z)$s with minimal asymptotic variance in estimating γ. If $\tilde{\gamma}_n$ were efficient at $(\gamma, \varrho, \psi(\cdot))$, it would be asymptotically linear in the efficient influence function (2.8) with Y_j replaced by $\psi(Y_j)$, and its asymptotic variance would equal (2.12) for $\rho = 0$. However, at $\rho = 0$ the asymptotic variance of $\tilde{\gamma}_n$ equals

$$2\left(1-\gamma^2\right)^2 \geq \frac{4(1-\gamma^2)^2}{2+\gamma^2} = E(\tilde{\ell}_\gamma^2), \tag{3.2}$$

which is an equality only at $\gamma = 0$. We conclude that either $\tilde{\gamma}_n$ is not semiparametrically asymptotically efficient or the bound in terms of (2.8) is not sharp, i.e., the regression-correlation-scale model is not least favorable.

However, we will show that there exist estimators of the regression parameters γ and ϱ that are asymptotically linear in (2.8) and (2.9) with Y_j replaced by $\psi(Y_j)$, and hence that the regression-correlation-scale model (2.4) is least favorable.

To this end we need $\sqrt{n}$-consistent estimators of both γ and ϱ. Since $\sqrt{n}(\tilde{\gamma}_n - \gamma)$ is asymptotically normal, $\tilde{\gamma}_n$ is $\sqrt{n}$-consistent. Similarly,

$$\tilde{\varrho}_n = \frac{1}{4}\{\hat{\rho}_n(-1) + 2\hat{\rho}_n(0) + \hat{\rho}_n(1)\} \tag{3.3}$$

is a preliminary $\sqrt{n}$-consistent estimator of ϱ.

4. Estimator of the transformation

In the next Section 5 we will present semiparametrically efficient estimators of the regression parameters. For their construction we need estimators of the efficient influence functions (2.8) and (2.9) with Y_j replaced by $\psi(Y_j)$, based on n i.i.d. copies of the random vector $X = (Y_1, Y_2, Z)^T$ for the artificial situation that we know the true values of γ and ϱ. Let us denote by

$$\tilde{\ell}_\gamma(X; \gamma, \varrho, \psi) \quad \text{and} \quad \tilde{\ell}_\varrho(X; \gamma, \varrho, \psi) \tag{4.1}$$

the efficient influence functions (2.8) and (2.9) with Y_j replaced by $\psi(Y_j)$.

Let $X_1, \ldots, X_n$ with $X_i = (Y_{1i}, Y_{2i}, Z_i)^T$ be the n i.i.d. copies of $X = (Y_1, Y_2, Z)^T$ that represent the observations. Fix γ_0 and ϱ_0 with $|\varrho_0| + |\gamma_0| < 1$, and let (γ_n) and (ϱ_n) be local sequences, in the sense that $\sqrt{n}(\gamma_n - \gamma_0) = \mathcal{O}(1)$ and $\sqrt{n}(\varrho_n - \varrho_0) = \mathcal{O}(1)$ hold. We need an estimator $\hat{\ell}_{\gamma,n}(x; \gamma, \varrho; X_1, \ldots, X_n)$ for $\tilde{\ell}_\gamma(x; \gamma, \varrho, \psi)$ satisfying the consistency condition

$$\int \left[\hat{\ell}_{\gamma,n}(x; \gamma_n, \varrho_n; X_1, \ldots, X_n) - \tilde{\ell}_\gamma(x; \gamma_n, \varrho_n, \psi)\right]^2 dP_{\gamma_n,\varrho_n,\psi}(x) \xrightarrow{P_{\gamma_n,\varrho_n,\psi}} 0 \tag{4.2}$$

and the $\sqrt{n}$-unbiasedness condition

$$\sqrt{n} \int \hat{\ell}_{\gamma,n}(x; \gamma_n, \varrho_n; X_1, \ldots, X_n)\, dP_{\gamma_n,\varrho_n,\psi}(x) \xrightarrow{P_{\gamma_n,\varrho_n,\psi}} 0, \tag{4.3}$$

as $n \to \infty$. We also need an estimator $\hat{\ell}_{\varrho,n}(x; \gamma, \varrho; X_1, \ldots, X_n)$ for $\tilde{\ell}_\varrho(x; \gamma, \varrho, \psi)$ satisfying the analogous consistency and $\sqrt{n}$-unbiasedness conditions.

Studying (2.5), (2.6), (2.8), and (2.9) we conclude that the a, b, and c coefficients in (4.1) are known rational, continuous functions of Z, γ, and ϱ, which do not depend on the unknown transformation $\psi(\cdot)$. Consequently it suffices to construct an estimator $\hat{\psi}_n(\cdot; X_1, \ldots, X_n) = \hat{\psi}_n(\cdot)$ of $\psi(\cdot)$ such that

$$\begin{aligned}
&\hat{\ell}_{\gamma,n}((y_1, y_2, z); \gamma, \varrho; X_1, \ldots, X_n) \\
&\quad = a_\gamma(Z; \gamma, \varrho)[\hat{\psi}_n^2(Y_1) + \hat{\psi}_n^2(Y_2)] \\
&\qquad + b_\gamma(Z; \gamma, \varrho)\hat{\psi}_n(Y_1)\hat{\psi}_n(Y_2) + c_\gamma(Z; \gamma, \varrho)
\end{aligned} \tag{4.4}$$

and

$$\begin{aligned}
&\hat{\ell}_{\varrho,n}((y_1, y_2, z); \gamma, \varrho; X_1, \ldots, X_n) \\
&\quad = a_\varrho(Z; \gamma, \varrho)[\hat{\psi}_n^2(Y_1) + \hat{\psi}_n^2(Y_2)] \\
&\qquad + b_\varrho(Z; \gamma, \varrho)\hat{\psi}_n(Y_1)\hat{\psi}_n(Y_2) + c_\varrho(Z; \gamma, \varrho)
\end{aligned} \tag{4.5}$$

satisfy (4.2) and (4.3). In view of (1.1) we need an appropriate estimator of the marginal distribution $G(\cdot)$ of Y_1 and Y_2. We choose a modification of the empirical distribution function, namely

$$\hat{G}_n(y) = \frac{1}{n+2}\left\{1 + \sum_{i=1}^{n} \frac{1}{2}\left(\mathbf{1}_{[Y_{1i}\le y]} + \mathbf{1}_{[Y_{2i}\le y]}\right)\right\}, \qquad y \in \mathbb{R}, \tag{4.6}$$

which yields the following result.

Theorem 4.1. *The estimator*

$$\hat{\psi}_n(y) = \hat{\psi}_n(y; X_1, \ldots, X_n) = \Phi^{-1}(\hat{G}_n(y)), \qquad y \in \mathbb{R}, \tag{4.7}$$

based on (4.6) of the transformation $\psi(\cdot)$ in the semiparametric normal regression-copula model (1.3) yields estimators (4.4) and (4.5) of the semiparametrically efficient influence functions for the regression parameters γ and ϱ that satisfy the consistency and $\sqrt{n}$-unbiasedness conditions (4.2) and (4.3).

Proof. In order to prove consistency (4.2) for the efficient influence function for both γ and ϱ we show that the expectation of the left hand side of (4.2) converges to 0, and for this it suffices to prove

$$\lim_{n\to\infty} E_{\gamma_n,\varrho_n,\psi}\left(\hat{\psi}_n^2(Y_1) - \psi^2(Y_1)\right)^2 = 0, \tag{4.8}$$

$$\lim_{n\to\infty} E_{\gamma_n,\varrho_n,\psi}\left(|\, b_\gamma(Z;\gamma_n,\varrho_n)\,|^2 \left|\, \hat{\psi}_n(Y_1)\hat{\psi}_n(Y_2) - \psi(Y_1)\psi(Y_2)\,\right|^2\right) = 0, \tag{4.9}$$

and similarly for $b_\varrho(Z;\gamma_n,\varrho_n)$, where we have taken $X = (Y_1, Y_2, Z)^T$ independent of $X_1, \ldots, X_n$. By Cauchy-Schwarz and the triangle inequality the square of the expectation in (4.8) may be bounded by

$$\begin{aligned} &E_{\gamma_n,\varrho_n,\psi}\big(\hat{\psi}_n(Y_1) - \psi(Y_1)\big)^4 E_{\gamma_n,\varrho_n,\psi}\big(\hat{\psi}_n(Y_1) + \psi(Y_1)\big)^4 \\ &\quad\le E_{\gamma_n,\varrho_n,\psi}\big(\hat{\psi}_n(Y_1) - \psi(Y_1)\big)^4 \\ &\qquad\times \left[E^{1/4}_{\gamma_n,\varrho_n,\psi}\big(\hat{\psi}_n(Y_1) - \psi(Y_1)\big)^4 + 2E^{1/4}_{G}\big(\psi(Y_1)\big)^4\right]^4. \end{aligned} \tag{4.10}$$

In view of $E_G(\psi(Y_1))^4 = 3$, this shows that for (4.8) to hold it is sufficient to prove

$$\lim_{n\to\infty} E_{\gamma_n,\varrho_n,\psi}\big(\hat{\psi}_n(Y_1) - \psi(Y_1)\big)^4 = 0. \tag{4.11}$$

Writing $\hat{\psi}_n(Y_1)\hat{\psi}_n(Y_2) - \psi(Y_1)\psi(Y_2) = [\hat{\psi}_n(Y_1) - \psi(Y_1)]\hat{\psi}_n(Y_2) + \psi(Y_1)[\hat{\psi}_n(Y_2) - \psi(Y_2)]$ we see that independence of Y_1 and Z and of Y_2 and Z shows that (4.11) also suffices to prove (4.9).

For $j = 1, 2$ let

$$\hat{G}_{jn}(y) = \frac{1}{n+2}\left\{1 + \sum_{i=1}^{n} \mathbf{1}_{[Y_{ji}\le y]}\right\}, \qquad y \in \mathbb{R}, \tag{4.12}$$

be a modified empirical distribution function based on $Y_{j1}, \ldots, Y_{jn}$. The Glivenko-Cantelli theorem implies

$$\lim_{n\to\infty} \sup_{y\in\mathbb{R}} \left|\, \hat{G}_{jn}(y) - G(y)\,\right| = 0 \quad \text{a.s.}, \tag{4.13}$$

and, since $\hat{G}_n(\cdot)$ is the average of $\hat{G}_{1n}(\cdot)$ and $\hat{G}_{2n}(\cdot)$,

$$\lim_{n\to\infty} \sup_{y\in\mathbb{R}} \big| \hat{G}_n(y) - G(y) \big| = 0 \quad \text{a.s.} \tag{4.14}$$

This yields

$$\lim_{n\to\infty} \big| \hat{G}_n(Y_1) - G(Y_1) \big| = 0 \quad \text{a.s.} \tag{4.15}$$

and hence

$$\lim_{n\to\infty} \big| \hat{\psi}_n(Y_1) - \psi(Y_1) \big| = 0 \quad \text{a.s.} \tag{4.16}$$

Since $u \mapsto [\Phi^{-1}(u)]^4$ is convex, we obtain

$$\begin{aligned} E_{\gamma_n,\varrho_n,\psi}[\hat{\psi}_n(Y_1)]^4 &\le \frac{1}{2}\big(E[\hat{\psi}_{1n}(Y_1)]^4 + [E\hat{\psi}_{2n}(Y_1)]^4\big) \\ &= \int_0^1 \sum_{i=0}^n \left[\Phi^{-1}\left(\frac{i+1}{n+2}\right)\right]^4 \binom{n}{i} u^i (1-u)^{n-i}\, du \\ &= \frac{1}{n+1}\sum_{i=0}^n \left[\Phi^{-1}\left(\frac{i+1}{n+2}\right)\right]^4 \le \int_0^1 [\Phi^{-1}(u)]^4\, du, \end{aligned} \tag{4.17}$$

where the first equality follows by taking the conditional expectation given $G(Y_1) = u$. Combining (4.16) and (4.17) we obtain (4.11) and hence (4.2) by Vitali's theorem; see Lemma A.7.5 of Bickel et al. [2].

Since $\Phi^{-1}(\cdot)$ is monotone, we have

$$\begin{aligned} &\big\{\Phi^{-1}(\hat{G}_n(y)) - \Phi^{-1}(G(y))\big\}^2 \\ &\quad= \left\{\Phi^{-1}\left(\frac{1}{2}\big[\hat{G}_{1n}(y) + \hat{G}_{2n}(y)\big]\right) - \Phi^{-1}(G(y))\right\}^2 \\ &\quad\le \big\{\big|\Phi^{-1}(\hat{G}_{1n}(y)) - \Phi^{-1}(G(y))\big| \vee \big|\Phi^{-1}(\hat{G}_{2n}(y)) - \Phi^{-1}(G(y))\big|\big\}^2 \\ &\quad\le \big\{\Phi^{-1}(\hat{G}_{1n}(y)) - \Phi^{-1}(G(y))\big\}^2 + \big\{\Phi^{-1}(\hat{G}_{2n}(y)) - \Phi^{-1}(G(y))\big\}^2. \end{aligned} \tag{4.18}$$

By Proposition 6.1 and the integral transform this yields

$$\lim_{n\to\infty} E\left(\sqrt{n}\int \big[\hat{\psi}_n(y) - \psi(y)\big]^2\, dG(y)\right) = 0. \tag{4.19}$$

By the independence of Y_1 and Z and of Y_2 and Z and by symmetry this implies

$$\begin{aligned} &E_{\gamma_n,\varrho_n,\psi}\left(\int \hat{\ell}_{\gamma,n}(x;\gamma_n,\varrho_n;X_1,\dots,X_n)\, dP_{\gamma_n,\varrho_n,\psi}(x)\right) \\ &\quad= E_{\gamma_n,\varrho_n,\psi}\big(\hat{\ell}_{\gamma,n}(X;\gamma_n,\varrho_n;X_1,\dots,X_n) - \tilde{\ell}_\gamma(X;\gamma_n,\varrho_n,\psi)\big) \\ &\quad= E_{\gamma_n,\varrho_n,\psi}\big(a_\gamma(Z;\gamma_n,\varrho_n)\big[\hat{\psi}_n^2(Y_1) - \psi^2(Y_1) + \hat{\psi}_n^2(Y_2) - \psi^2(Y_2)\big] \\ &\qquad\qquad + b_\gamma(Z;\gamma_n,\varrho_n)\big[\hat{\psi}_n(Y_1)\hat{\psi}_n(Y_2) - \psi(Y_1)\psi(Y_2)\big]\big) \\ &\qquad - 2E_{\gamma_n,\varrho_n,\psi}\big(a_\gamma(Z;\gamma_n,\varrho_n)\big[\hat{\psi}_n(Y_1) - \psi(Y_1)\big]^2\big) \\ &\qquad - E_{\gamma_n,\varrho_n,\psi}\big(b_\gamma(Z;\gamma_n,\varrho_n)\big[\hat{\psi}_n(Y_1) - \psi(Y_1)\big]\big[\hat{\psi}_n(Y_2) - \psi(Y_2)\big]\big) \\ &\qquad + o_P\left(\frac{1}{\sqrt{n}}\right) \end{aligned} \tag{4.20}$$

$$
\begin{aligned}
&= 4E_{\gamma_n,\varrho_n,\psi}\big(a_\gamma(Z;\gamma_n,\varrho_n)\big[\hat\psi_n(Y_1)-\psi(Y_1)\big]\psi(Y_1)\big)\\
&\quad + 2E_{\gamma_n,\varrho_n,\psi}\big(b_\gamma(Z;\gamma_n,\varrho_n)\big[\hat\psi_n(Y_1)-\psi(Y_1)\big]\psi(Y_2)\big) + o_P\left(\frac{1}{\sqrt{n}}\right)\\
&= o_P\left(\frac{1}{\sqrt{n}}\right),
\end{aligned}
$$

where the last equality holds in view of (2.13). The analogous result holds for the efficient influence function for ϱ. This completes the proof of the $\sqrt{n}$-unbiasedness condition (4.3) and of the theorem. □

Unlike the consistency condition, the $\sqrt{n}$-unbiasedness condition can be satisfied only if the structure of the normal regression-copula model is used. This structure is captured by (2.13) and (2.14), apparently.

5. Efficient estimators

By the technique of sample splitting efficient estimators of the regression parameters can be constructed now.

Theorem 5.1. *In the semiparametric normal regression-copula model of (1.2) and (1.3) there exist efficient estimators of the regression parameters γ and ϱ, i.e. there exist $\hat\gamma_n$ and $\hat\varrho_n$ satisfying*

$$
\sqrt{n}\left(\hat\gamma_n - \left[\gamma_n + \frac{1}{n}\sum_{i=1}^n \tilde\ell_\gamma(X_i;\gamma_n,\varrho_n,\psi)\right]\right) = o_{\gamma_n,\varrho_n,\psi}(1), \tag{5.1}
$$

and

$$
\sqrt{n}\left(\hat\varrho_n - \left[\varrho_n + \frac{1}{n}\sum_{i=1}^n \tilde\ell_\varrho(X_i;\gamma_n,\varrho_n,\psi)\right]\right) = o_{\gamma_n,\varrho_n,\psi}(1), \tag{5.2}
$$

for every γ_0 and ϱ_0 with $|\varrho_0| + |\gamma_0| < 1$ and for all sequences (γ_n) and (ϱ_n) satisfying $\sqrt{n}(\gamma_n-\gamma_0) = \mathcal{O}(1)$ and $\sqrt{n}(\varrho_n-\varrho_0) = \mathcal{O}(1)$.

Proof. Note that Theorem 7.8.1 of Bickel et al. [2] can be applied and the paragraph below it; cf. Klaassen [7]. Contiguity and smoothness follow from the regularity of the least favorable parametric submodel (2.4), preliminary estimators have been constructed in Section 3, and appropriate estimators of the efficient influence functions have been presented in Section 4. □

The asymptotic linearity in the efficient influence function from (5.1) suggests that

$$
\hat\gamma_n = \tilde\gamma_n + \frac{1}{n}\sum_{i=1}^n \hat\ell_\gamma(X_i;\tilde\gamma_n,\tilde\varrho_n;X_1,\dots,X_n) \tag{5.3}
$$

might work as an estimator of γ satisfying (5.1). Typically, this works in practice indeed. However, because of the dependence between $\tilde\gamma_n$ and the estimator of the influence function, this is difficult to prove; cf. Schick [10, 11]. Although sample splitting might look artificial at first sight, it will hardly influence the asymptotic performance, as suggested by Edgeworth expansions in Klaassen [8].

6. Appendix

The following result is crucial to our proof of the main result of the paper.

Proposition 6.1. *Let $U_1, U_2, \ldots$ be i.i.d. random variables with the uniform distribution on the unit interval. With the modified empirical distribution function defined by*

$$G_n^*(u) = \frac{1}{n+2}\left\{1 + \sum_{i=1}^{n} \mathbf{1}_{[U_i \le u]}\right\} \tag{6.1}$$

we have

$$\lim_{n\to\infty} E\left(\sqrt{n}\int_0^1 \left\{\Phi^{-1}(G_n^*(u)) - \Phi^{-1}(u)\right\}^2 du\right) = 0. \tag{6.2}$$

Proof. To prove this convergence we need a bound on the distance between quantiles of the standard normal distribution. We prove that the standard normal distribution is less spread out than a logistic distribution with variance $\pi^3/6$, and we also present another bound on the distance between standard normal quantiles in the following Lemma.

Lemma 6.1. *For all $u, v \in [0, 1$ we have*

$$\left|\Phi^{-1}(v) - \Phi^{-1}(u)\right| \le \sqrt{\frac{\pi}{2}}\left|\log\left(\frac{v}{1-v}\right) - \log\left(\frac{u}{1-u}\right)\right| \tag{6.3}$$

and

$$\left|\Phi^{-1}(v) - \Phi^{-1}(u)\right|^2 \le \frac{\pi}{2}(v-u)^2\left(\frac{1}{uv} + \frac{1}{(1-u)(1-v)}\right). \tag{6.4}$$

We notice that $u \mapsto \log(u/(1-u))$ is the quantile function of the logistic distribution with mean 0 and variance $\pi^2/3$.

Proof of Lemma 6.1. First note

$$1 - \Phi(x) \le \frac{1}{2}e^{-x^2/2}, \qquad x \ge 0. \tag{6.5}$$

Indeed, by differentiation we see that

$$x \mapsto \psi(x) = \frac{1}{2}e^{-x^2/2} - 1 + \Phi(x)$$

is increasing-decreasing on $[0, \infty)$ with $\psi(0) = \psi(\infty) = 0$.

By symmetry of the standard normal density (6.5) implies

$$\varphi(\Phi^{-1}(u)) \ge \sqrt{\frac{2}{\pi}}\big(u \wedge (1-u)\big)$$

and hence

$$\begin{aligned}
\left|\Phi^{-1}(v) - \Phi^{-1}(u)\right| &= \left|\int_u^v \frac{1}{\varphi\left(\Phi^{-1}(w)\right)}\, dw\right| \\
&\le \sqrt{\frac{\pi}{2}}\left|\int_u^v \frac{1}{w \wedge (1-w)}\, dw\right| \\
&\le \sqrt{\frac{\pi}{2}}\left|\int_u^v \left(\frac{1}{w} + \frac{1}{(1-w)}\right) dw\right| \\
&= \sqrt{\frac{\pi}{2}}\left|\log\left(\frac{v}{1-v}\right) - \log\left(\frac{u}{1-u}\right)\right|.
\end{aligned} \tag{6.6}$$

Similarly the Cauchy-Schwarz inequality implies

$$\begin{aligned}\left|\Phi^{-1}(v)-\Phi^{-1}(u)\right|^2 &\le \frac{\pi}{2}\left|\int_u^v \frac{1}{w\wedge(1-w)}\,dw\right|^2\\ &\le \frac{\pi}{2}(v-u)\int_u^v\left(\frac{1}{w^2}+\frac{1}{(1-w)^2}\right)dw\\ &= \frac{\pi}{2}(v-u)^2\left(\frac{1}{uv}+\frac{1}{(1-u)(1-v)}\right).\end{aligned}\tag{6.7}$$

□

Let $0=U_{(0)}\le U_{(1)}\le\cdots\le U_{(n)}\le U_{(n+1)}=1$ denote uniform order statistics. We note that the expectation in (6.2) equals

$$\begin{aligned}&\sqrt{n}\sum_{i=0}^{n}\int_0^1\left\{\Phi^{-1}\left(\frac{i+1}{n+2}\right)-\Phi^{-1}(u)\right\}^2 P\left(U_{(i)}\le u<U_{(i+1)}\right)\,du\\ &\quad=\sqrt{n}\sum_{i=0}^{n}\int_0^1\left\{\Phi^{-1}\left(\frac{i+1}{n+2}\right)-\Phi^{-1}(u)\right\}^2\binom{n}{i}u^i(1-u)^{n-i}\,du.\end{aligned}\tag{6.8}$$

Note that this expression is symmetric under the map $u\mapsto 1-u,\ i\mapsto n-i$. Consequently, in view of (6.4) and (6.3) expression (6.8) may be bounded by

$$\begin{aligned}&\frac{\pi}{2}\sqrt{n}\sum_{i=1}^{n-1}\int_0^1\left(\frac{i+1}{n+2}-u\right)^2\left(\frac{n+2}{u(i+1)}+\frac{n+2}{(1-u)(n-i+1)}\right)\\ &\qquad\times\binom{n}{i}u^i(1-u)^{n-i}\,du\\ &\quad+\pi\sqrt{n}\int_0^1\log^2\left(\frac{1-u}{u(n+1)}\right)(1-u)^n\,du.\end{aligned}\tag{6.9}$$

Tedious computation shows that the first term from (6.9) equals

$$\begin{aligned}&\frac{\pi}{2}\sqrt{n}\sum_{i=1}^{n-1}\frac{1}{(n+1)(n+2)}\left(\frac{n-i+1}{i}+\frac{i+1}{n-i}\right)\\ &\quad=\frac{\pi\sqrt{n}}{(n+1)(n+2)}\sum_{i=1}^{n-1}\frac{n-i+1}{i}\le\frac{\pi\sqrt{n}\,n}{(n+1)(n+2)}\sum_{i=1}^{n-1}\frac{1}{i}\\ &\quad\le\frac{\pi\sqrt{n}}{n+2}\left(1+\int_1^n\frac{1}{x}\,dx\right)=\frac{\pi\sqrt{n}}{n+2}\left(1+\log n\right),\end{aligned}\tag{6.10}$$

which converges to 0 as $n\to\infty$.

The second term from (6.9) may be bounded by

$$3\pi\sqrt{n}\int_0^1\left(\log^2(1-u)+\log^2 u+\log^2(n+1)\right)(1-u)^n\,du.\tag{6.11}$$

Note

$$\begin{aligned}\int_0^1\log^2(1-u)\,(1-u)^n\,du&=\int_0^1\left\{\int_0^u\frac{1}{1-y}\,dy\right\}^2(1-u)^n\,du\\ &\le\int_0^1 u\int_0^u\frac{1}{(1-y)^2}\,dy\,(1-u)^n\,du\end{aligned}\tag{6.12}$$

$$\leq \int_0^1 \int_y^1 \frac{(1-u)^n}{(1-y)^2}\, du\, dy$$
$$= \frac{1}{n+1} \int_0^1 (1-y)^{n-1}\, dy = \frac{1}{n(n+1)}.$$

For positive α we have

$$\int_0^1 \log^2 u\ (1-u)^n\, du \leq \int_0^{n^{-\alpha}} \log^2 u\, du + \alpha^2 \log^2 n \int_{n^{-\alpha}}^1 (1-u)^n\, du$$
$$\leq \left[u \log^2 u - 2u \log u + 2u\right]_0^{n^{-\alpha}} + \frac{\alpha^2 \log^2 n}{n+1} \tag{6.13}$$
$$= \frac{\alpha^2 \log^2 n + 2\alpha \log n + 2}{n^\alpha} + \frac{\alpha^2 \log^2 n}{n+1}.$$

Choosing α larger than $1/2$ and combining (6.11), (6.12), and (6.13), we see that also the second term from (6.9) converges to 0 as $n \to \infty$. □

Acknowledgments

Chris Klaassen would like to thank Jon Wellner for his friendship and pleasant collaboration during more than the last quarter century. The paper Klaassen and Wellner [9] is fundamental to the present paper, is based on Jon's ideas, and was written on his initiative. Bojan Basrak gratefully acknowledges friendship and hospitality of Eurandom staff and colleagues during his postdoctoral studies at the institute. His work is supported in part by the research grants MZOŠ nr. 037-0372790-2800 and 037-0982913-2762 of the Croatian government.

References

[1] Basrak, B., Klaassen, C. A. J., Beekman, M., Martin, N. and Boomsma, D. (2004). Copulas in QTL mapping. *Behavior Genetics* **34** 161–171.

[2] Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* The Johns Hopkins University Press, Baltimore. (1998) Springer, New York, revised paperbound edition.

[3] Cherny, S. S., Sham, P. C. and Cardon, L. R. (2004). Introduction to the special issue on variance components methods for mapping quantitative trait loci. *Behavior Genetics* **34** 125–126.

[4] Drost, F. C., Klaassen, C. A. J. and Werker, B. J. M. (1994). Adaptiveness in time series models. *Asymptotic Statistics*, P. Mandl, M. Hušová (eds.) 203–211. Physica-Verlag, Heidelberg.

[5] Drost, F. C., Klaassen, C. A. J. and Werker, B. J. M. (1997). Adaptive estimation in time-series models. *The Annals of Statistics* **25** 786–817.

[6] Dupuis, J. and Siegmund, D. (1999). Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics* **151** 373–386.

[7] Klaassen, C. A. J. (1987). Consistent estimation of the influence function of locally asymptotically linear estimators. *The Annals of Statistics* **15** 1548–1562.

[8] Klaassen, C. A. J. (2001). Discussion to "Inference for Semiparametric Models: Some Current Frontiers" by P. J. Bickel and J. Kwon. *Statistica Sinica* **11** 906–909.

[9] Klaassen, C. A. J. and Wellner, J. A. (1997). Efficient estimation in the bivariate normal copula model: Normal margins are least favorable. *Bernoulli* **3** 55–77.

[10] Schick, A. (1986). On asymptotically efficient estimation in semiparametric models. *Annals of Statistics* **14** 1139–1151.

[11] Schick, A. (1987). A note on the construction of asymptotically linear estimators. *Journal of Statistical Planning and Inference* **16** 89–105. Correction (1989) **22** 269–270.