IMS Collections
Vol. 0 (2009) 65–91

arXiv: math.PR/0000003

Semiparametric Models and Likelihood - The Power of Ranks

Kjell Doksum[1,*] **and Akichika Ozeki**[2,*]

University of Wisconsin, Madison

Abstract: We consider classes of models related to those introduced by Lehmann in 1953 and Sklar in 1959. Recently developed algorithms for finding profile NP likelihood procedures are discussed, extended and implemented for such models by combining them with the MM algorithm. In particular we consider statistical procedures for a regression model with proportional expected hazard rates, and for transformation models including the normal copula. A variety of likelihoods introduced to deal with semiparametric models are considered. They all generate rank results, not only tests, but also estimates, confidence regions, and optimality theory, thereby, to paraphrase Lehmann (1953), demonstrating "the power of ranks".

Contents

[1]Department of Statistics, University of Wisconsin, Madison, WI 53706, email: doksum@stat.wisc.edu
*Supported in part by NSF grant DMS-0505651.
[2]Department of Statistics, University of Wisconsin, Madison, WI 53706
AMS 2000 subject classifications: Primary 62G05, 62G20; secondary 62N02
Keywords and phrases: Lehmann model, proportional hazard model, profile NP likelihood, nonparametric maximum likelihood, MM algorithm, copula models, Box-Cox models

1. Introduction

We will focus on statistical inference for models where the distribution of the data can be expressed as a parametric function of unknown distribution functions.

1.1. Lehmann Type Models. Cox Regression

Suppose T is a random variable with a continuous distribution function F. For testing the null hypothesis $H_0 : F = F_0$, Lehmann (1953) considered alternatives of the form

$$F_\theta(\cdot) = C_\theta(F_0(\cdot)), \tag{1.1}$$

for some continuous distribution $C_\theta(\cdot)$ on $[0, 1]$, which is known except for the parameter θ. We consider the problem of estimating θ when $F_0(\cdot)$ is an unknown baseline distribution. In this case, if $T_1, \cdots, T_n$ are independent with $T_i \sim C_{\theta_i}(F_0(\cdot))$ and we set $U_i = F_0(T_i)$, then U_i has distribution $C_{\theta_i}(\cdot)$ Moreover $R_i \equiv Rank(T_i) = Rank(U_i)$, which implies that the distribution of any statistical method based on $R_1, \cdots, R_n$ will not depend on F_0.

For regression experiments with observations $(T_i, \boldsymbol{x}_i), i = 1, \cdots, n$, where T_i is a response and $\boldsymbol{x}_i$ is a vector of nonrandom covariates, Cox (1972) considered the parametrization $\theta_i = g(\boldsymbol{\beta}^T \boldsymbol{x}_i)$ with $g(\cdot)$ a known function and $\boldsymbol{\beta}$ a vector of regression coefficients. He considered statistical inference procedures based on the Cox (1972, 1975) partial likelihood in very general frameworks. These procedures are based on generalized ranks and show how powerful ranks are in generating statistical inference procedures.

In this paper we consider a special case of (1.1) obtained from the Lehmann models $[F_0(t)]^N$ and $1 - [1 - F_0(t)]^N$ by letting N be a zero truncated Poisson variable whose parameter depends on covariates and regression coefficients. We call this model "SINAMI" after SIbuya (1968) and NAbeya and MIura (1972). For a subset of the parameter space, the model has proportional expected hazard rate (PEHR). We show that semiparametric likelihood methods for the SINAMI model give more weight to intermediate survival times than the Cox proportional hazard model which heavily weights long survival times. Recently developed algorithms for finding profile nonparametric maximum likelihood estimates (profile NPMLE's) are combined with the MM algorithm to produce estimates. In the two sample case, we carry out a Monte Carlo comparison of the NPMLE with a parametric MLE and a method of moment (MOM) estimate of the two sample parameter. The NPMLE is nearly unbiased but only about 70% as efficient in terms of root MSE as the parametric estimate if the parametric model is true. The MOM estimate is slightly less efficient than the NPMLE.

1.2. Sklar Type Models. Copula Regression

Suppose X and Y are random variables with continuous joint distribution $H(\cdot, \cdot)$ and marginals $F_1(\cdot)$ and $F_2(\cdot)$. Sklar (1959) considered models that include models

of the form

$$H_\theta(\cdot,\cdot) = C_\theta(F_1(\cdot), F_2(\cdot)), \tag{1.2}$$

for some continuous distribution $C_\theta(\cdot,\cdot)$ on $[0,1]\times[0,1]$, which is known except for the parameter θ. We consider the problem of estimating θ when $F_1(\cdot)$ and $F_2(\cdot)$ are unknown baseline distributions. If we set $U = F_1(X)$, $V = F_2(Y)$, then (U,V) has distribution $C_\theta(\cdot,\cdot)$, and $C_\theta(\cdot,\cdot)$ is called a *copula*. Note that if $(X_1,Y_1),\cdots,(X_n,Y_n)$ are independent with $(X_i,Y_i) \sim H_{\theta_i}(\cdot,\cdot)$, then $R_i \equiv Rank(X_i) = Rank(F_1(X_i))$ and $S_i \equiv Rank(Y_i) = Rank(F_2(Y_i))$, which shows that the distribution of any statistical method based on these ranks will not depend on (F_1,F_2). This model extends in the natural way to the d-dimensional case.

In this paper we consider the bivariate normal copula model where $C_\theta(u,v) = \Phi_\theta(\Phi^{-1}(u),\Phi^{-1}(v))$ with C_θ the bivariate $N(0,0,1,1,\theta)$ distribution. We also consider the multivariate normal copula model and show that in regression experiments it can be used to construct a "transform both sides regression" transformation model (copula regression model.) Klaassen and Wellner (1997) have shown that the normal scores correlation coefficient is semiparametrically efficient for the bivariate normal copula. We use simulations to compare this estimate with the profile MLE for the transform both sides Box-Cox regression model and a nonparametric estimate based on splines thereby augmenting the comparisons made by Zou and Hall (2002). The normal scores estimate is nearly as efficient as the parametric MLE for estimating median regression when the transform both sides Box-Cox model is correct. We also consider the performance of the estimates for models outside the copula regression model and find that the normal scores based estimate of median regression is remarkably robust with respect to both bias and variance. On the other hand, the profile MLE of median regression derived from the transform both sides Box-Cox model is very sensitive to deviations from the model. The nonparametric spline estimate is the best for extreme deviations from the copula regression model.

2. Proportional Hazard and Proportional Expected Hazard Rate Models

Interesting special cases of (1.1) are obtained by considering the distributions of $T_1 = \min(T_{01},\cdots,T_{0k})$ and $T_2 = \max(T_{01},\cdots,T_{0k})$ where $T_{01},T_{02},\cdots$ are i.i.d. as $T_0 \sim F_0$. Then, for $k \geq 1$,

$$T_1 \sim F_1(t) = 1 - [1 - F_0(t)]^k, \tag{2.1}$$

and

$$T_2 \sim F_2(t) = [F_0(t)]^k, \tag{2.2}$$

with $t > 0, k = 1,2,\cdots$. More general forms (Lehmann (1953); Savage (1956)) are

$$T_1 \sim F_1(t) = 1 - [1 - F_0(t)]^\Delta, \tag{2.3}$$

and

$$T_2 \sim F_2(t) = [F_0(t)]^\Delta, \tag{2.4}$$

with $t > 0, \Delta > 0$. Here (2.3) can be derived by considering two-sample models where the two samples follow distributions of the form (2.1) with different k's (Bickel and Doksum (2007), Problem 1.1.12.)

For T_1, the hazard rate is

$$\lambda(t) \equiv \frac{f(t)}{1-F(t)} = \triangle \frac{f_0(t)}{1-F_0(t)} \equiv \triangle \lambda_0(t). \tag{2.5}$$

In regression experiments, we set $\triangle_i = g(\boldsymbol{\beta}^T \boldsymbol{x}_i)$, and note that (2.5) is the Cox proportional hazard (PH) model (Cox (1972)).

Nabeya and Miura (1972) proposed replacing k in (2.1) and (2.2) by a random variable. In particular, they considered $T_1 = \min(T_{01}, \cdots, T_{0N})$, where N is independent of $T_{01}, T_{02}, \cdots$, and has a zero truncated Poisson(θ) distribution with $\theta > 0$. They also considered $T_2 = \max(T_{01}, \cdots, T_{0M})$, $T_{0i} \sim F_0$ where M is independent of $T_{01}, T_{02}, \cdots$, and has a zero truncated Poisson($-\theta$) distribution with $\theta < 0$.

Using Sibuya (1968), they found

$$T_1 \sim F_1(t) = \frac{1-e^{-\theta F_0(t)}}{1-e^{-\theta}}, \quad \theta > 0, \tag{2.6}$$

$$T_2 \sim F_2(t) = \frac{1-e^{-\theta F_0(t)}}{1-e^{-\theta}}, \quad \theta < 0. \tag{2.7}$$

Combining (2.6) and (2.7), we get

$$\begin{aligned} T \sim F(t) = & \frac{1-e^{-\theta F_0(t)}}{1-e^{-\theta}}, \quad \theta \neq 0, \\ = & \ F_0(t), \qquad\qquad \theta = 0. \end{aligned} \tag{2.8}$$

Note that model (2.6) is a mixture of proportional hazard models for individuals with the same baseline hazard rate $\lambda_0(\cdot)$ but different hazard factors $\triangle$ in the factorization (2.5) of the hazard rate. Let $\lambda(t;k)$ denote the hazard rate of T_1 given $N = k$; then by (2.5)

$$E\lambda(t;N) = \sum_{k=1}^{\infty} k\lambda_0(t) p_\theta(k) = \tau(\theta)\lambda_0(t), \ \theta > 0, \tag{2.9}$$

where $p_\theta(x)$ is the zero truncated Poisson(θ) probability and

$$\tau(\theta) = E(N) = \frac{\theta}{1-\exp(-\theta)}. \tag{2.10}$$

Thus (2.6) is a model with proportional expected hazard rate. Note that (2.8) does not have this property for $\theta < 0$. We will refer to (2.6) and (2.8) as the PEHR and SINAMI models, respectively.

***Remark* 2.1 :** In regression experiments, the traditional frailty models are also constructed by introducing a random element in the PH model. However, these models are different from the PEHR and SINAMI models. To see this recall that in the frailty model the conditional hazard rate given the covariate vector $\boldsymbol{x}$ (see Oakes (1992)) for the history and interpretation of frailty models) is of the form

$$\lambda_W(t|\boldsymbol{x}) = \lambda_0(t) W \exp[\boldsymbol{\beta}^T \boldsymbol{x}], \tag{2.11}$$

where W is a random effect that incorporates potential unobservable covariates that represent frailties. Semiparametric optimality theory for model (2.11) has been developed by Kosorok, Lee, and Fine (2004).

Consider model (2.5) with $\Delta = N$ and N a zero truncated Poisson(θ) random variable with $\theta = g(\boldsymbol{\beta}^T \boldsymbol{x})$, i.e., the conditional hazard rate given $\boldsymbol{x}$ is

$$\lambda^{(N)}(t|\boldsymbol{x}) = N\lambda_0(t). \tag{2.12}$$

Here N plays the role of $W \exp[\boldsymbol{\beta}^T \boldsymbol{x}]$ in (2.11). However, (2.11) and (2.12) are different because N is an integer and $W \exp[\boldsymbol{\beta}^T \boldsymbol{x}]$ is not when $\boldsymbol{\beta} \neq 0$. In model (2.12), N represents the effect of both observed covariates and frailties. In deriving the distribution function (2.6), the unobservable covariates are averaged out, that is, we compute $P(T \leq t) = E[P(T \leq t|N)]$.

***Remark* 2.2 :** Model (2.8) was considered by Bell and Doksum (1966), Example 5.2 and Table 8.1) and Ferguson (1967, p.257, Problem 5.7.7) without any of the above interpretations. Nabeya and Miura (1972) did not use any proportional hazard or frailty interpretation. These concepts had not been invented yet.

Fig.1 gives a plot of the relative hazard rate $\lambda(t|x = 1)/\lambda(t|x = 0)$ with $\theta = -5, -3, 3, 5$ for model (2.8) with $\theta = \beta x$, $F_0(t) = 1 - \exp(-t)$, $t > 0$, and

$$\lambda(t|x) = \frac{\theta f_0(t)}{1 - e^{-\theta(1-F_0(t))}}. \tag{2.13}$$

In the PEHR and SINAMI models, the hazard ratio between two covariate values converge to unity as time increases. This explains why the likelihoods for these models give less weight to long survival times than the likelihood for the Cox model (see Section 3). The hazard rate is decreasing for the PEHR model for any continuous F_0.

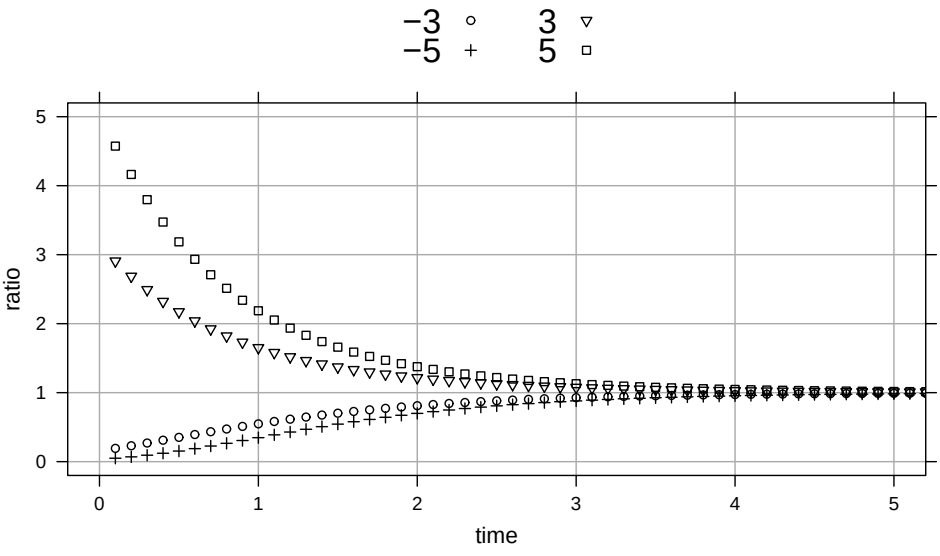


FIG 1. *SINAMI and PEHR hazard ratios for* $\theta = -5, -3, 3, 5$.

3. Rank, Partial and Marginal Likelihood

In regression experiments, we observe $(T_i, \boldsymbol{x}_i)$, $i = 1, \cdots, n$, where $T_1, \cdots, T_n$ are independent responses and $\boldsymbol{x}_i$ is a nonrandom covariate vector. In the proportional hazard model, it is customary to use model (2.5) for T_i with $\triangle_i = \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)$ because $\triangle_i$ needs to be positive. In the PEHR model, $\theta_i = \boldsymbol{\beta}^T \boldsymbol{x}_i$ is a possible parametrization, but $\theta_i = \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)$ could also be used. Let $\boldsymbol{R} = (R_1, \cdots, R_n)$ where $R_i = \text{Rank}(T_i)$, then $l_{\boldsymbol{r}}(\boldsymbol{\beta}) = P(\boldsymbol{R} = \boldsymbol{r})$ is the *rank likelihood* (Hoeffding (1951)).

We first consider the one covariate case. Using the rank likelihood, the locally most powerful (LMP) rank test statistic for $H_0 : \beta = 0$ versus $H_1 : \beta > 0$ is (approximately) for the Cox model (Savage (1956), 1957), Cox (1964)), Oakes and Jeong (1998)):

$$\sum_{i=1}^{n} [-\log(1 - \frac{R_i}{n+1})](x_i - \bar{x}) \quad \text{(Savage or log rank)}, \tag{3.1}$$

and for the PEHR and SINAMI models, the LMP rank test statistics is (Bell and Doksum (1966), Ferguson (1967), Nabeya and Miura (1972)):

$$\sum_{i=1}^{n} \frac{R_i}{n+1}(x_i - \bar{x}) \quad \text{(Wilcoxon type)}, \tag{3.2}$$

where R_i =Rank(T_i). The log rank statistic gives more weights to large observations, that is, in survival analysis, to those that live longer, while the Wilcoxon statistics is even handed.

In order to compare how much relative weight is given to the small, in between, and large observed survival times for the PH and PEHR models, we next consider the rank likelihood for d covariates. Note that if $h(\cdot)$ is decreasing, then $\text{Rank}(h(T_i)) = n + 1 - R_i$. For the proportional hazard model, transform T_i by $U_i = 1 - F_0(T_i)$, then by (2.3) we have $f_{U_i}(u) = \triangle_i u^{\triangle_i - 1}, 0 < u < 1$. Hoeffding (1951) formula shows,

$$\text{Rank lkhd} = \Pi_{i=1}^n \triangle_i \int \cdots \cdots \int_{0<u_1<u_2<\cdots<u_n<1} \Pi_{i=1}^n u_i^{\delta_i - 1} du_1 \cdots du_n, \tag{3.3}$$

where $\delta_i = \triangle_{b_i}$ and b_i= index on the T with rank $n + 1 - i$= reverse anti-rank. It follows that

$$\text{Rank lkhd} \propto \Pi_{i=1}^n \frac{\triangle_i}{\sum_{k:T_k \geq T_{(i)}} \triangle_k}, \tag{3.4}$$

that is, the familiar Cox (1972, 1975) *partial likelihood* formula. Here $\{k : T_k \geq T_{(i)}\}$= patients at risk at time $T_{(i)}$ where $T_{(i)}$ is the ith ordered survival time. Kalbfleich and Prentice (1973, 2002) called the rank likelihood the *marginal likelihood* and extended it to censored data.

For the PEHR model, transform T_i by the decreasing function $U_i = a^{-1}\{\exp\{-F_0(T_i)\} - b\}$, then, we have $f_i(u) = a\tau_i(au + b)^{\theta_i - 1}$, $0 < u < 1$, where $b = e^{-1}$, $a = 1 - b$, and $\tau_i \equiv \tau(\theta_i)$. Let $\gamma_i = \theta_{b_i}$, b_i=index on the T with rank $n + 1 - i$. Then,

$$\text{Rank lkhd} \propto (\Pi_{i=1}^n \tau_i) \int \cdots \cdots \int_{0<u_1<u_2<\cdots<u_n<1} \Pi_{i=1}^n (au_i + b)^{\gamma_i - 1} du_1 \cdots du_n. \tag{3.5}$$

If we perform the integration, we find that the likelihood for the PEHR model is similar to the likelihood for the Cox model except that in addition to terms involving $\{k : T_k \geq T_{(i)}\}$, $i = 1, \cdots, n$, it includes terms involving $\{k : T_{(i)} \leq T_k \leq T_{(j)}\}$, $i = 1, \cdots, n$, $j = 1, \cdots, n$, $i \neq j$. That is, the PEHR likelihood gives more weight to the intermediate survival times than the Cox likelihood.

Computationally, the Cox rank likelihood is easier than the PEHR rank likelihood. However, we can handle the PEHR rank likelihood with available algorithms and software (e.g. MATLAB.) More generally, $F(x) = C_\theta(F_0(x))$ type models, originally considered by Lehmann (1953), can be handled effectively by considering the profile NP likelihood of the next section (e.g. Tsodikov and Gabribotti (2007)), Zeng and Lin (2007)).

4. Profile NP Likelihood

Andersen, Borgan, Gill, and Keiding (1996), Bickel, Klaassen, Ritov, and Wellner (1993, 1996), van der Vaart (1998), Murphy and van der Vaart (2000), Tsodikov and Garibotti (2007), Zeng and Lin (2007) and many others considered the problem of finding the MLE of all the parameters in a semiparametric model. It is useful to divide the procedure into two steps by grouping parameters into two groups. Suppose the distribution function of T is of the form $P(T \leq t) = F(\theta, \eta(t))$, where $\theta \in R^d$ and $\eta(\cdot)$ is a nondecreasing function. If we assume temporarily that $\eta(\cdot)$ has a positive derivative $\eta'(t)$ for $t \in \{t_1, \cdots, t_n\}$, then the likelihood is

$$\prod_{i=1}^{n} \eta'(t_i) f(\theta, \eta(t_i)),$$

where $f(\theta, \eta) = \partial F(\theta, \eta)/\partial \eta$. The NP likelihood we consider is of the form

$$L_{NP}(\theta, \eta) = \prod_{i=1}^{n} \eta\{t_i\} f(\theta, \eta[t_i]),$$

where $\eta[t_i] = \sum_{j \leq i} \eta\{t_i\}$ is a step function with positive jumps $\eta\{t_i\}$ at the data points t_i, $i = 1, \cdots, n$.

We assume that

$$af(\theta, a) \to 0 \quad as\ a \to \infty$$

Next we fix θ, and define $\hat{\eta}_\theta\{t_i\}$ as

$$\hat{\eta}_\theta\{.\} = ARG\ MAX_{\eta\{.\}} L_{NP}(\theta, \eta). \tag{4.1}$$

Set

$$PROF\ NPLIK = l(\theta) = MAX_{\eta\{.\}} L_{NP}(\theta, \eta), \tag{4.2}$$

and solve

$$\hat{\theta} = ARG\ MAX\ l(\theta). \tag{4.3}$$

Next estimate $\eta\{\cdot\}$ as $\hat{\eta}_{\hat{\theta}}\{\cdot\}$. In the Lehmann model (1.1), the NP likelihood is

$$\prod_{i=1}^{n} F_0\{x_i\} C'_\theta\Big(\sum_{j \leq i} F_0\{x_j\}\Big),$$

The method is similar to finding the empirical MLE, Owen (1988, 2001), and profile (partial) MLE's as in Andersen, et al. (1996), and Murphy and van der Vaart (2000).

***Remark* 4.1 :** Note that when $P(T \leq t) = F(\theta, \eta(t))$, (4.1), (4.2), and (4.3) do not depend on the values of $t_1, \cdots, t_n$. In regression experiments, they will depend on the ranks of $t_1, \cdots, t_n$. For example, see (4.4) and (5.1). This is in contrast to the Hodges and Lehmann (1963) approach that uses estimating equations based on rank test statistics to obtain estimates of parameters. In this Hodges-Lehmann "rank inversion" approach, estimates are functions of the "raw" data rather than the ranks.

As an example that will guide the algorithm for the PEHR model, consider the Cox model. Set $\Lambda(t) = -\log(1 - F_0(t))$, then,

$$L_{NP}(\boldsymbol{\beta}, \Lambda) = \prod_{i=1}^{n} e^{\beta^T x_i} \Lambda\{t_i\} e^{-(e^{\beta^T x_i})\Lambda[t_i]},$$

where

$$\Lambda[t_i] = \sum_{j:t_j \leq t_i} \Lambda\{t_j\}.$$

Using calculus, we find

$$\hat{\Lambda}_\beta\{t_i\} = ARG\ MAX_{\Lambda\{t_i\}}\ L_{NP}(\boldsymbol{\beta}, \Lambda) = (\sum_{j:t_j \geq t_i} e^{\beta^T x_j})^{-1},$$

and

$$l(\boldsymbol{\beta}) = PROF\ NPLIK = \prod_{i=1}^{n} \frac{e^{\beta^T x_i}}{\sum_{j:t_j \geq t_i} e^{\beta^T x_j}}. \tag{4.4}$$

This is exactly the same as the rank, the partial, and the marginal likelihood.

5. Profile NP Likelihood for the PEHR Model

Consider model (2.6) with $\theta > 0$. Set $\tau(\theta) = \theta[1 - e^{-\theta}]^{-1}$, then

$$L_{NP}(\boldsymbol{\theta}, F_0) = [\prod_{i=1}^{n} \tau(\theta_i)] \prod_{i=1}^{n} F_0\{t_i\} e^{-\theta_i F_0[t_i]}, \tag{5.1}$$

where $F_0[t_i] = \sum_{j:t_j \leq t_i} F_0\{t_j\}$. Now set $p_i \equiv F_0\{t_i\}$ and maximize with respect to $p_1, \cdots, p_n$ with θ fixed. The maximization problem looks very similar to Cox model maximization except for the constraint $\sum p_i = 1$. We handle this constraint by writing $F_0(t) = 1 - \exp[-\Lambda(t)]$ with $\Lambda(t)$ unconstrained except for $\Lambda(t) \geq 0$ and by using a new approach based on the MM algorithm.

5.1. The MM Algorithm

Lang, Hunter and Yang (2000) introduced a concept called the MM algorithm. Its idea is that instead of maximizing a complicated original objective function, use a

simpler surrogate function so that each iteration is faster and guarantees that the original objective function increases. Given the original objective function $l(\boldsymbol{h})$ for a maximization problem, a surrogate function $g(\boldsymbol{h}|\boldsymbol{h}_{old})$ must satisfy two properties:

$$l(\boldsymbol{h}_{old}) = g(\boldsymbol{h}_{old}|\boldsymbol{h}_{old}), \tag{5.2}$$

$$l(\boldsymbol{h}) \geq g(\boldsymbol{h}|\boldsymbol{h}_{old}). \tag{5.3}$$

The EM algorithm is a special case of the MM algorithm. A practical implementation issue of the MM algorithm is that we have to find a nice surrogate function case by case.

Now we construct a surrogate function based on Tsodikov (2003). Suppose we can write $l(\cdot)$ in the form $l(\boldsymbol{h}) = B(\boldsymbol{h}) - A(\boldsymbol{h})$ for some parameter vector $\boldsymbol{h} > \boldsymbol{0}$, where A and B are differentiable concave functions. Then by the concavity property,

$$g(\boldsymbol{h}|\boldsymbol{h}_{old}) = B(\boldsymbol{h}) - A(\boldsymbol{h}_{old}) - \nabla^T A(\boldsymbol{h}_{old})(\boldsymbol{h} - \boldsymbol{h}_{old}), \tag{5.4}$$

where $\nabla^T A(\boldsymbol{h}) = \partial A/\partial \boldsymbol{h}$ is the gradient of A, satisfies (5.2) and (5.3), and $g(\boldsymbol{h}|\boldsymbol{h}_{old})$ is a surrogate function for $l(\boldsymbol{h})$. Differentiating (5.4) gives

$$\nabla^T B(\boldsymbol{h}_{new}) = \nabla^T A(\boldsymbol{h}_{old}). \tag{5.5}$$

Solve (5.5) for $\boldsymbol{h}_{new}$. Iterate the procedure until there is a minimal change in $\boldsymbol{h}$.

5.2. The MM Algorithm for the PEHR Model with $\boldsymbol{\theta} \geq \boldsymbol{0}$ (SINAMI with $\boldsymbol{\theta} \geq \boldsymbol{0}$)

Let $\theta_i = g(\boldsymbol{\beta}^T \boldsymbol{x}_i)$, for some known function $g(\cdot) \geq 0$. When $\theta_i = 0$, the distribution function of T_i is $F_0(t)$. Let $F_0(t) = 1 - \exp[-\Lambda(t)]$, $h_k = \Lambda\{t_k\}$, and $\Lambda[t_i] = \sum_{k=1}^{i} h_k$ with $\Lambda(t) \geq 0$ and $h_k \geq 0$. Then for a temporarily fixed numerical vector $\boldsymbol{\beta}$,

$$\begin{aligned} l(\boldsymbol{h}) = log[L_{NP}(\boldsymbol{\beta}, \boldsymbol{h})] &= \sum_{i=1}^{n} \log \tau(\theta_i) + \sum_{i=1}^{n} \log h_i \\ &\quad - \sum_{i=1}^{n} [\sum_{k=1}^{i} h_k + \theta_i (1 - \exp(-\sum_{k=1}^{i} h_k))]. \end{aligned} \tag{5.6}$$

Now we can write $l(\boldsymbol{h}) = B(\boldsymbol{h}) - A(\boldsymbol{h})$ with

$$B(\boldsymbol{h}) = \sum_{i=1}^{n} \log h_i, \tag{5.7}$$

$$A(\boldsymbol{h}) = \sum_{i=1}^{n} [\sum_{k=1}^{i} h_k + \theta_i (1 - \exp(-\sum_{k=1}^{i} h_k))]. \tag{5.8}$$

Here we may ignore $\Sigma \log \tau(\theta_i)$ because we maximize (5.6) w.r.t. $\boldsymbol{h}$.

$B(\boldsymbol{h})$ and $A(\boldsymbol{h})$ are concave, because for $0 \leq t \leq 1$, $B(t\boldsymbol{h}_a + (1-t)\boldsymbol{h}_b) \geq tB(\boldsymbol{h}_a) + (1-t)B(\boldsymbol{h}_b)$ and by mathematical induction, $A(t\boldsymbol{h}_a + (1-t)\boldsymbol{h}_b) \geq tA(\boldsymbol{h}_a) + (1-t)A(\boldsymbol{h}_b)$ hold. Note that

$$\partial B/\partial h_j = 1/h_j, \ j = 1, ..., n, \tag{5.9}$$

$$\partial A/\partial h_j = \sum_{i=1}^{n}(1+\theta_i(1-\exp(-\sum_{k=1}^{i} h_k)))1(j \le i). \tag{5.10}$$

Using (5.5), (5.9), and (5.10), update $h_j,\ j = 1, \cdots, n$, at the same time,

$$h_{j,new} = [\sum_{i=1}^{n}(1+\theta_i(1-\exp(-\sum_{k=1}^{i} h_{k,old})))1(j \le i)]^{-1}. \tag{5.11}$$

Iterate (5.11) until there is a minimal change in $l(\hat{\boldsymbol{h}}_{new})$; call the result $\hat{\boldsymbol{h}}_A$ ($\boldsymbol{A}$pproximated profile NPMLE). Note that we call $\hat{\boldsymbol{h}}_A$ approximated profile NPMLE because $\hat{\boldsymbol{h}}_A$ is obtained by fixing $\boldsymbol{\beta}$. This approximation is necessary because there is no closed form $\hat{\boldsymbol{h}}$ w.r.t. $\boldsymbol{\beta}$. Next set $l(\boldsymbol{\beta}) = log[L_{NP}(\boldsymbol{\beta}, \hat{\boldsymbol{h}}_A)]$ and maximize w.r.t. $\boldsymbol{\beta}$.

5.3. The MM Algorithm for the SINAMI Model with $\theta \le 0$

Consider model (2.7) with $\theta_i = g(\boldsymbol{\beta}^T \boldsymbol{x}_i)$, for some known function $g(\cdot) \le 0$. In this case we can use the algorithm of Section 5.2. To see this, suppose T satisfies model (2.7) with parameter $\theta_2 < 0$. Set $V = 1 - F_0(T)$, then V satisfies model (2.6) with parameter $\theta_1 = -\theta_2$. Moreover, the rank of $1 - F_0(T_i)$ is $n + 1 - R_i$.

5.4. The MM Algorithm for the SINAMI Model with $\theta \in R$

Consider model (2.7) with $\theta_i = g(\boldsymbol{\beta}^T \boldsymbol{x}_i)$, for some known function $g(\cdot) \in R$. In this case we can not use the transformation in Section 5.3 because it changes the likelihood and the monotonicity of the likelihood as a function of θ does not necessarily hold. Instead, we modify the algorithm as follows: If the value $\hat{\theta}_j$ in the jth iteration is positive, use the MM algorithm in Section 5.2. to find $\hat{\boldsymbol{h}}_j$. If $\hat{\theta}_j < 0$, then (5.6) implies that finding the maximizer $\boldsymbol{h}$ is a convex optimization problem which produces $\hat{\boldsymbol{h}}_j$.

5.5. Profile NPMLE Implementation

Successful convergence of the MM algorithm depends on a good starting point $(\hat{\theta}, \hat{\boldsymbol{h}})$. We consider the two sample problem:

$$T_{0,i} \sim F_0(t),\ i = 1, ..., n_0, \tag{5.12}$$

$$T_{1,i} \sim F(t) = \begin{cases} \frac{1-e^{-\theta F_0(t)}}{1-e^{-\theta}}, & (\theta > 0),\ i = 1, ..., n_1, \\ F_0(t), & (\theta = 0),\ i = 1, ..., n_1, \end{cases} \tag{5.13}$$

where $F_0(\cdot)$ is an unknown distribution with density f_0. Note that the density of $F(t)$ is

$$f(t;\theta) = \begin{cases} \tau(\theta) f_0(t) e^{-\theta F_0(t)}, & (\theta > 0), \\ f_0(t), & (\theta = 0). \end{cases} \tag{5.14}$$

We use an algorithm to find $(\hat{\theta}, \hat{\boldsymbol{h}})$ where $\theta = \beta$ in this case. For fixed F_0, (5.14) gives an MOM estimating equation for θ. We plug in an estimate $\hat{F}_0$ for F_0, and use the following algorithm:

$$step\ (1) : \hat{F}_0 \rightarrow\ \ step\ (2) : \hat{\theta} \rightarrow\ \ step\ (3) : (\hat{\theta}_A \leftrightarrow \hat{\boldsymbol{h}}_A)_{until\ \hat{\theta}_A\ converges.}$$

Here $\hat{\theta}_A$ and $\hat{h}_A$ are approximated profile NPMLE's from Section 5.2. The details are as follows:

***Step* (1) :** Compute the empirical distribution $\hat{F}_0(t)$ based on $T_{0,i}$ only:

$$\hat{F}_{0,[0]}(t) \equiv \hat{F}_0(t) = \frac{1}{1+n_0}\sum_{i=1}^{n_0} 1(T_{0,i} \le t). \tag{5.15}$$

Here $\hat{F}_0(t) \rightarrow_{a.s.} F_0(t)$ uniformly in t as $n_0 \rightarrow \infty$. The subscript [0] indicate iteration zero (starting point) for *step* (3). Note that the one-to-one relation between $\hat{F}_0$ and $\hat{\Lambda}_0$ is used to obtain $\hat{\boldsymbol{h}}$ by solving for $\boldsymbol{h}$ in the equations:

$$\begin{aligned} \hat{F}_0(t_i) \equiv &\quad \hat{F}_{0,i} = \exp(-\hat{\Lambda}_i),\ \ where \\ \hat{\Lambda}[t_i] \equiv &\quad \hat{\Lambda}_i = \sum_{k=1}^{i} \hat{h}_k,\ \ i = 1, ..., n_0. \end{aligned} \tag{5.16}$$

***Step* (2) :** Solve for $\hat{\theta}$ based on $\hat{F}_{0,[0]}(t)$:

$$\begin{aligned} \bar{T}_1 = &\quad \tau(\theta)\int y e^{-\theta \hat{F}_{0,[0]}(y)} d\hat{F}_{0,[0]}(y) \\ = &\quad \tau(\theta)\sum_{i=1}^{n_0} T_{0,(i)} \exp(-\theta\frac{i}{n_0})\frac{1}{n_0}, \end{aligned} \tag{5.17}$$

where $T_{0,(i)}$ is an order statistics of $T_{0,1}, \cdots, T_{0,n_0}$ and $\bar{T}_1 = \sum_{i=1}^{n_1} T_{1,i}/n_1$. The solution is uniquely determined because the distribution function (5.13) is monotone increasing in θ and hence its mean is monotone decreasing in θ. If a model is $\theta \ge 0$ and $\hat{\theta} < 0$, set $\hat{\theta} = 0$. If a model is $\theta < 0$ and $\hat{\theta} > 0$, set $\hat{\theta} = 0$.

***Step* (3) :** Compute $\hat{\theta}_A$ and $\hat{\boldsymbol{h}}_A$ as follows:

$$(\hat{\theta}_A \leftrightarrow \hat{\boldsymbol{h}}_A)_{until\ \hat{\theta}_A\ converges.} \tag{5.18}$$

The first iteration $\hat{\theta}_{A,[1]}$, is obtained by maximizing (5.6) with $\hat{\theta}$ as a starting point, i.e., $\hat{\theta}_{A,[0]} = \hat{\theta}$ and with fixed $\hat{\boldsymbol{h}}_{[0]}$ obtained from (5.16) and $\hat{F}_{0,[0]}$, i.e.,

$$\hat{\theta}_{A,[1]} = \arg\max_{\theta} \{l(\theta, \hat{\boldsymbol{h}}_{[0]}) : \theta \ge 0\}. \tag{5.19}$$

Then by the MM algorithm in Section 5.2 with $\boldsymbol{\beta}_0 = \theta_0 = \hat{\theta}_{A,[1]}$ (see (5.11)), obtain $\hat{\boldsymbol{h}}_{[1]}$ using the starting point $\hat{\boldsymbol{h}}_{[0]}$. Next obtain

$$\hat{\theta}_{A,[2]} = \arg\max_{\theta} \{l(\theta, \hat{\boldsymbol{h}}_{A,[1]}) : \theta \ge 0\}, \tag{5.20}$$

with starting point $\hat{\theta}_{A,[1]}$. Then by the MM algorithm in Section 5.2 with $\boldsymbol{\beta}_0 = \theta_0 = \hat{\theta}_{A,[2]}$, obtain $\hat{\boldsymbol{h}}_{A,[2]}$ with starting point $\hat{\boldsymbol{h}}_{A,[1]}$. Repeat the procedure to get $\hat{\theta}_{A,[j]}$ and $\hat{\boldsymbol{h}}_{A,[j]}$ until convergence, i.e., $|\hat{\theta}_{A,[j]} - \hat{\theta}_{A,[j-1]}| < \epsilon$ for some small ϵ.

Numerical optimizations for $\hat{\theta}$ and $\hat{\theta}_A$ are carried out by the MATLAB fmincon() function.

***Remark* 5.1 :** For fixed $\hat{\boldsymbol{h}}$, $l(\theta, \hat{\boldsymbol{h}})$ is strictly concave and have a unique maximum.

***Remark* 5.2 :** The estimate of $\boldsymbol{\beta}$ in the Cox model that we have discussed is asymptotically optimal in the semiparametric sense (Begun, Hall, Huang, and Wellner (1983), Bickel et al. (1993, 1998), van der Vaart (1998)), Murphy and van der Vaart (2000)). These references and others give results that can be used to check the semiparametric asymptotic optimality of the profile NPMLE in the PEHR model.

***Remark* 5.3 :** *Transformation models.* We can show that the PEHR is a special case of transformation models as follows: Let F_λ be the exponential (λ) distribution function and define

$$G_0(y|\boldsymbol{x}) = \frac{1 - e^{-\theta F_\lambda(y)}}{1 - e^{-\theta}}, \quad y > 0, \tag{5.21}$$

where $\theta = g(\boldsymbol{x}, \boldsymbol{\beta})$. Let ψ be an increasing function from $[0, \infty)$ to $[0, \infty)$ and define the transformation model

$$G(y|\boldsymbol{x}) = G_0(\psi(y)|\boldsymbol{x}). \tag{5.22}$$

This model is of the form (2.6) with $F_0 = F_\lambda \psi(t)$. Klaassen (2007) gives results for general transformation models that can be used to check semiparametric asymptotic efficiency of estimates of $\boldsymbol{\beta}$ in the model (2.6).

5.6. Estimation of the Variance of the Profile NPMLE

Hypothesis tests and confidence intervals require standard errors (estimates of the standard deviation) of $\hat{\theta}_A$. An algorithm developed by Tsodikov and Garibotti (2007) combined with the preceding algorithm allows us to compute the profile information matrix which is the observed information matrix derived from the profile likelihood. This provides standard errors $SE(\hat{\theta}_A)$ of $\hat{\theta}_A$.

6. Simulation Results

6.1. PEHR Model Estimates

Monte Carlo (MC) simulation for model (5.13) with F_0 equal to the exponential distribution EXP(1) is based on 1000 Monte Carlo samples with n=100, 200, and 300. $T_{0,i} \sim EXP(1)$, $i = 1, \cdots, n_0$, $T_{1,i} \sim PEHR(\theta = 2, \ or \ 3)$, $i = 1, \cdots, n_1$, $n_0 = n_1 = n/2$, iid. We compute Monte Carlo estimates of the expected values, standard deviations (SD's), and MSE's of $\hat{\theta}_{MOM}$, $\hat{\theta}_A$, and $\hat{\theta}_{PAR}$ where $\hat{\theta}_{PAR}$ is the parametric model MLE, obtained by assuming that F_0 is known and equal to the EXP(1) distribution.

We also compute the Monte Carlo estimates of the expected values $E[SE\ (\hat{\theta}_A)]$ of the standard errors computed as described in Section 5.5. Table 1 summarizes the result.

θ	2			3		
n	100	200	300	100	200	300
$E[\hat{\theta}_{MOM}]$	1.97	2.00	2.02	3.05	3.00	3.01
$E[\hat{\theta}_A]$	1.99	2.00	2.01	3.07	3.00	3.01
$E[\hat{\theta}_{PAR}]$	2.02	2.00	2.01	3.03	3.01	3.00
$SD(\hat{\theta}_{MOM})$	0.819	0.551	0.479	0.972	0.635	0.506
$SD(\hat{\theta}_A)$	0.788	0.541	0.466	0.960	0.618	0.497
$SD(\hat{\theta}_{PAR})$	0.541	0.387	0.317	0.605	0.431	0.335
$E[SE(\hat{\theta}_A)]$	0.752	0.527	0.430	0.837	0.578	0.471
$MSE[\hat{\theta}_{MOM}]$	0.671	0.303	0.230	0.949	0.403	0.256
$MSE[\hat{\theta}_A]$	0.621	0.292	0.217	0.926	0.383	0.248
$MSE[\hat{\theta}_{PAR}]$	0.289	0.150	0.101	0.367	0.186	0.112

TABLE 1

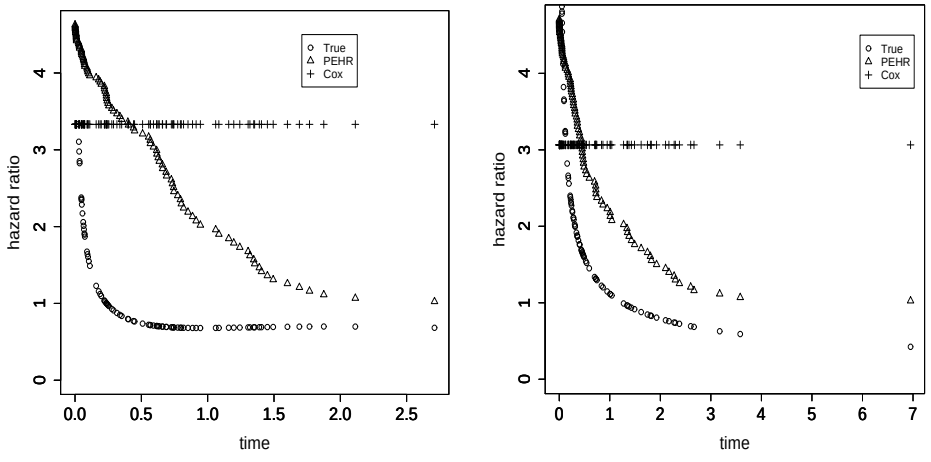
PEHR model simulation estimates (MC=1000, $\theta = 2, 3$)

Overall $\hat{\theta}_{MOM}$, $\hat{\theta}_A$, and $\hat{\theta}_{PAR}$ have almost no bias in the estimation of θ =2 or 3. As expected, the parametric model estimate $\hat{\theta}_{PAR}$ has he smallest MSE. The approximated profile NPMLE $\hat{\theta}_A$ has a smaller MSE than $\hat{\theta}_{MOM}$, but the difference is small. The approximation to SD($\hat{\theta}_A$) is very good and improves as the sample size increases.

6.2. Model Fit for Misspecified Model

Next consider a model that is neither a Cox PH model nor a PEHR model: i.e., $T_{0,i} \sim EXP(1)$ and the true model for $T_{1,i}$ is : Case 1, Gamma(shape=0.5, scale=0.5), and the : Case 2, Weibull(shape=0.5, scale=0.2). Here the target values θ and $\boldsymbol{h}$ are those that minimize the Kullback-Leibler divergence between the true distribution and the model class of distributions (Doksum, Ozeki, Kim and Neto (2007)) .

Fig. 2 shows that PEHR gives better fit than the Cox model.



(*a*) *True model* : Gamma (*shape* = 0.5, *scale* = 0.5).

(*b*) *True model* : Weibull (*shape* = 0.5, *scale* = 0.2).

FIG 2. *Cox and PEHR estimated hazard ratio when $F_0 \sim EXP(1)$ and $F \sim$ Gamma or Weibull.*

7. Estimation in the Normal Copula Model

7.1. The One Covariate Case

Assume that the pair (X,Y) has a joint density f(x,y) with respect to Lebesgue measure on R^2 and a joint distribution function F(x,y). Let F_1 and F_2 be the marginal distribution functions of X and Y, respectively, and let Φ denote the standard normal distribution function. Consider the transformations $X \to Z = \Phi^{-1}(F_1(X))$, $Y \to W = \Phi^{-1}(F_2(Y))$. Then the marginal distributions of Z and W are standard normal. The bivariate normal copula model $\mathcal{F}$ is defined by the assumption that the joint distribution of (Z,W) is bivariate normal with zero mean, unit variance, and correlation coefficient ρ. That is,

$$\mathcal{F} = \{F \ : (\Phi^{-1}(F_1(X)), \Phi^{-1}(F_2(Y))) \sim N(0,0,1,1,\rho)\},$$

where F_1 and F_2 are the marginals of F. Let $\{(X_i, Y_i), i = 1, 2, \cdots, n\}$ be independent and identically distributed with distribution function $F \in \mathcal{F}$, and set $Z_i = \Phi^{-1}(F_1(X_i))$, $W_i = \Phi^{-1}\ (F_2(Y_i))$, $i = 1, 2, \cdots, n$. If we (temporarily) assume that F_1 and F_2 are known, then because E(ZW) $= \rho$, a method of moments "estimate" of ρ, is $r_{MOM} = n^{-1}\sum_{i=1}^{n} Z_i W_i$. The asymptotic distribution of $\sqrt{n}(r_{MOM} - \rho)$ is $N(0, 1 + \rho^2)$ when $F \in \mathcal{F}$. Assuming F_1 and F_2 known, the asymptotic distribution of $\sqrt{n}(r_{MLE} - \rho)$, where r_{MLE} is the maximum likelihood "estimate" of ρ is $N(0, (1-\rho^2)^2/(1+\rho^2))$. The asymptotic variance $(1-\rho^2)^2/(1+\rho^2)$ of r_{MLE} is smaller than the asymptotic variance $(1-\rho^2)^2$ of the usual Pearson correlation coefficient r_P and much smaller than the asymptotic variance $1 + \rho^2$ of r_{MOM}.

Note that $\mathcal{F}$ is invariant under coordinate-wise increasing transformations. That is, if $(X, Y) \sim F \in \mathcal{F}$ and $U = h_1(X), V = h_2(Y)$ with h_1 and h_2 increasing, then the distribution G of (U,V) is in $\mathcal{F}$. If we want methods that are invariant under such transformations, we must use statistics based on the ranks defined in Section 1.

Suppose next that F_1 and F_2 are unknown. It may then make sense to replace the ordered Z's and W's by their expected values. This leads to the Fisher and Yates (1938) or normal scores $E(Z_{(i)})$, $i = 1, \cdots, n$ where $Z_{(1)}, \cdots, Z_{(n)}$ are N(0,1) order statistics. We write $a(i) = E(Z_{(i)})$. An accurate approximation to $E(Z_{(i)})$ is $\Phi^{-1}[(i - 3/8)/(n + 1/4)]$, e.g. Cox (2006).

Let $Z_i^{'} = a(R_i)$, $W_i^{'} = a(S_i)$ where R_i and S_i are the ranks of X_i and Y_i when the X's and Y's are ranked separately. Then we obtain estimates $\hat{\rho}_{MOM}$, $\hat{\rho}_{MLE}$, and $\hat{\rho}_P$ of ρ when F_1 and F_2 are unknown by replacing Z_i and W_i by $Z_i^{'}$ and $W_i^{'}$ in r_{MOM}, r_{MLE} and r_P. In this case $\hat{\rho}_{MOM}$, $\hat{\rho}_P$ are nearly identical and asymptotically equivalent, but they are different from $\hat{\rho}_{MLE}$. We will use $\hat{\rho}_P$ because it is slightly less biased, and denote it by $\hat{\rho}_{NS}$ where NS signifies normal scores. Thus

$$\hat{\rho}_{NS} = \sum Z_i' W_i' / \sum a^2(i). \tag{7.1}$$

It follows from Bhuchongkul (1964) that based on the rank likelihood, $\hat{\rho}_{NS}$ is, uniformly in F_1 and F_2, a locally most powerful test statistics in the bivariate normal copula model. Zou and Hall (2002) gave an asymptotic extension of this result. They also computed the rank likelihood estimate of ρ in the bivariate normal copula model using an improved version of the likelihood sampler in Doksum(1987).

Klaassen and Wellner (1997) found $\sqrt{n}(\hat{\rho}_{NS} - \rho) \to_d N(0, (1-\rho^2)^2)$ in the copula model $\mathcal{F}$ with F_1 and F_2 unknown; the same as for the Pearson correlation in the

bivariate normal model. In fact, in a bivariate normal$(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ model, r_P is the MLE, and r_P and $\hat{\rho}_{NS}$ are asymptotically optimal in the parametric sense.

A fourth possible estimate is the profile NP estimate obtained by fixing ρ and replacing F_1 and F_2 by step functions with jumps $\{p_i\}$ and $\{q_i\}$ at (X_i, Y_i) in the log likelihood for the normal copula model. That is, ignoring constants (ρ is fixed), we maximize

$$
\begin{aligned}
l(\boldsymbol{p}, \boldsymbol{q}) = \sum_i \{ & \log p_i + \log q_i \\
& + \frac{1}{2}((\Phi^{-1}(\sum_{k:X_k \le X_i} p_k))^2 + \frac{1}{2}(\Phi^{-1}(\sum_{k:Y_k \le Y_i} q_k))^2 \\
& + \frac{1}{2}(1-\rho^2)^{-1}[(\Phi^{-1}(\sum_{k:X_k \le X_i} p_k))^2 \\
& - 2\rho\Phi^{-1}(\sum_{k:X_k \le X_i} p_k)\Phi^{-1}(\sum_{k:Y_k \le Y_i} q_k) + (\Phi^{-1}(\sum_{k:Y_k \le Y_i} q_k))^2]\},
\end{aligned} \tag{7.2}
$$

w.r.t. $(\boldsymbol{p}, \boldsymbol{q})$ where $\sum p_i = 1$ and $\sum q_i = 1$. Then given $(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}})$, maximize the log likelihood w.r.t. ρ, which gives $\hat{\rho}_{PROF}$, a profile NPMLE.

***Remark* 7.1 :** An estimate $\hat{\theta}$ of a parameter $\theta \in R$ in a semiparametric model is regular if $\sqrt{n}(\hat{\theta} - \theta) \to_d N(0, V_{\hat{\theta}}(\theta, \eta))$ for some asymptotic variance $V_{\hat{\theta}}(\theta, \eta)$ and if $\hat{\theta}$ satisfies additional regularity conditions given in Bickel et al. (1993, 1998). For $F \in \mathcal{F}$, $V_{r_P}(\rho, F)$ depends on F (e.g. Bickel and Doksum (2007), Example 5.3.6), while $V_{\hat{\rho}_{NS}}(\rho, F)$ does not, as shown by Klaassen and Wellner (1997). Klaassen and Wellner (1997) go on to argue that $(1-\rho^2)^2$ is a semiparametric asymptotic variance lower bound for the class $\mathcal{S}$ of all regular estimates of ρ. Thus $\hat{\rho}_{NS}$ is semiparametrically optimal in the minimax sense:

$$
\sup\{V_{\hat{\rho}_{NS}}(\rho, F) : F \in \mathcal{F}\} = \inf_{\hat{\rho} \in \mathcal{S}} \sup\{V_{\hat{\rho}}(\rho, F) : F \in \mathcal{F}\} \tag{7.3}
$$

***Remark* 7.2 :** Recall that $\hat{\rho}_{NS}$ was obtained by inserting normal scores Z_i' and W_i' in the MOM estimate for the model with F_1 and F_2 known, and that the MLE r_{MLE} for this model has variance $(1-\rho^2)^2/(1+\rho^2)$. Klaassen and Wellner (1997) have shown that the approximate MLE $\hat{\rho}_{MLE}$ obtained from r_{MLE} by replacing (Z_i, W_i) with (Z_i', W_i') is semiparametrically optimal in the same sense as $\hat{\rho}_{NS}$. Because the distribution of the ranks do not depend on F_1 and F_2, this implies that $\hat{\rho}_{NS}$ and $\hat{\rho}_{MLE}$ are asymptotically equivalent for every $F \in \mathcal{F}$. We conjecture that $\hat{\rho}_{PROF}$ is also asymptotically optimal and equivalent to $\hat{\rho}_{NS}$.

***Remark* 7.3 :** The asymptotic distribution of $\hat{\rho}_{NS}$ when the distribution of (X,Y) is not in $\mathcal{F}$ can be obtained from Ruymgaart, Shorack, and Van Zwet (1972) and Ruymgaart (1974) .

7.2. The Multivariate Covariate Case

The normal copula model in the multivariate case is defined as follows: Let $Y \sim G$, $X_j \sim F_j$, $\boldsymbol{h}(\boldsymbol{X}) = (h_1(X_1), \cdots, h_d(X_d))$, where h_j, $j = 0, \cdots, d$ are increasing functions defined by

$$
h_0(Y) = \Phi^{-1}(G(Y)), \tag{7.4}
$$

$$h_j(X_j) = \Phi^{-1}(F_j(X_j)). \tag{7.5}$$

The distribution of the untransformed variables $(\boldsymbol{X}, Y)$ is a *copula model* if we assume that $(\boldsymbol{h}(\boldsymbol{X}), h_0(Y))$ is multivariate normal with 0 means and unit variances.

8. Transformation and NP Models

Consider a regression experiment with response Y and a random covariate vector $\boldsymbol{X} = (X_1, \cdots, X_d)^T$. We will extend the normal scores estimate $\hat{\rho}_{NS}$ of Section 7 to the d dimensioned case and compare it with estimates appropriate for parametric and nonparametric models. In the *copula regression* model of Section 7.2, we can write

$$h_0(Y) = \beta^T \boldsymbol{h}(\boldsymbol{X}) + \epsilon,\ \epsilon \sim N(0, \sigma^2), \tag{8.1}$$

where $\boldsymbol{\beta}$ is the set of regression coefficients when regressing $h_0(Y)$ on $\boldsymbol{h}(\boldsymbol{X})$. The transform both sides Box-Cox model is based on (8.1) with $h_0(Y) = Y^{(\lambda_{d+1})}$, $h_j(X_j) = X_j^{(\lambda_j)}$, $j = 1, \cdots, d$, where $t^{(\lambda)} = (t^\lambda - 1)/\lambda$. Thus for this case, we can write

$$Y^{(\lambda_{d+1})} = \alpha + \boldsymbol{\beta}^T \boldsymbol{X}^{(\boldsymbol{\lambda})} + \epsilon,\ \epsilon \sim N(0, \sigma^2). \tag{8.2}$$

We first consider a procedure for estimating the parameters in model (8.2):

I Profile Likelihood for a multivariate model.

Hernandez and Johnson (1980) considered the one sample multivariate Box-Cox transformation model. This was adopted to regression by Doksum, Ozeki, Kim and Neto (2007). We regard $(Y^{(\lambda_{d+1})}, \boldsymbol{X}^{(\boldsymbol{\lambda})})$ as a d+1 multivariate normal $(\boldsymbol{\mu}, \Sigma)$ vector. Regressing $Y^{(\lambda_{d+1})}$ on $\boldsymbol{X}^{(\boldsymbol{\lambda})}$ leads to (8.2). We fix $\xi \equiv (\boldsymbol{\lambda}, \lambda_{d+1})$ and estimate the parameters in the normal model by maximizing the likelihood thereby obtaining the familiar normal theory estimates $(\hat{\boldsymbol{\mu}}(\xi), \hat{\Sigma}(\xi))$. We plug these into the likelihood and obtain the profile likelihood $l(\xi)$, which we maximize to get $\hat{\xi}$ and the final estimates $(\hat{\boldsymbol{\mu}}(\hat{\xi}), \hat{\Sigma}(\hat{\xi}))$. These are the usual linear model estimates with Y_i, X_{ij} replaced by $Y_i^{(\hat{\lambda}_{d+1})}$, $X_{ij}^{(\hat{\lambda}_j)}$. Similarly, the estimate of $\boldsymbol{\beta}$ in (8.2) is the usual linear model estimate with Y_i, X_{ij} replaced by $Y_i^{(\hat{\lambda}_{d+1})}$, $X_{ij}^{(\hat{\lambda}_j)}$.

***Remark* 8.1 :** We also considered the maximum likelihood estimates of the parameters $\boldsymbol{\beta}, \sigma^2$, and $(\boldsymbol{\lambda}, \lambda_{d+1})$ in model (8.2). This approach has the problem that if we want to test $H_0 : \beta_j = 0$, then λ_j is not identifiable under H_0. Approach I does not have this problem. This is one case where likelihood and profile likelihood are very different. The algorithm for this MLE often failed to converge. When it did converge, it produced results close to those of method I. We omit the details.

***Remark* 8.2 :** As pointed out by Zou and Hall (2002), when d=1, the MLE of ρ in the Box-Cox transformation model with unknown transformation parameters and standardized transformations have the same efficiency as the MLE for the model with known transformation parameters because this Box-Cox model is between the bivariate normal model and the bivariate normal copula model and the MLE's in these models have the same asymptotic variance $(1-\rho^2)^2$. The result that

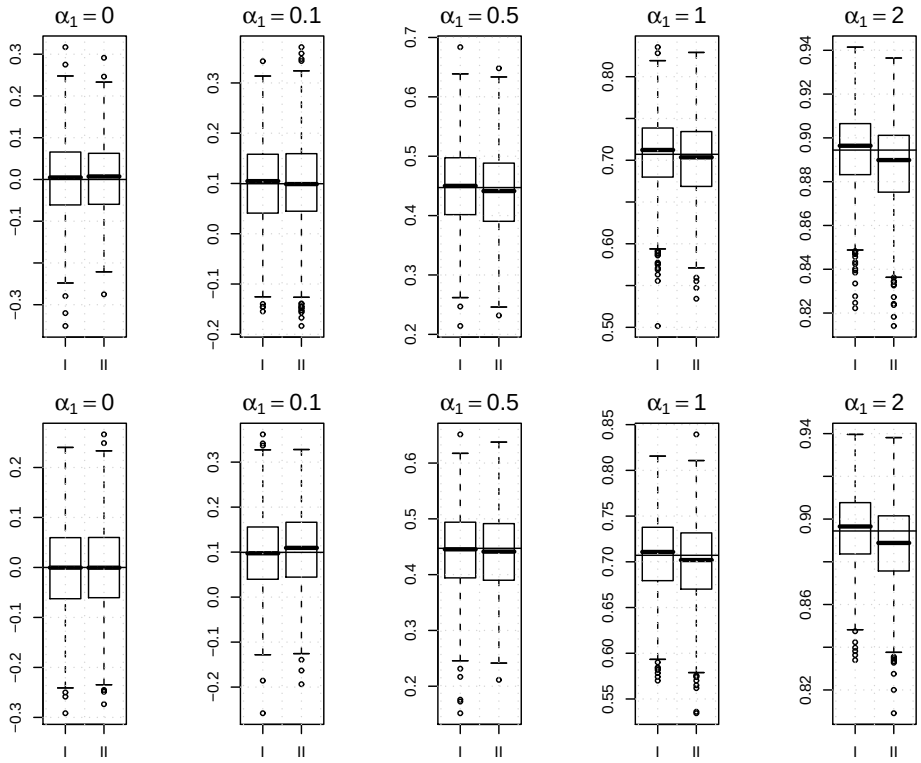


FIG 3. *$\hat{\rho}$ boxplots for correctly specified model (8.7) with n=128. Top 5: $\lambda_1 = \lambda_2 = 0.5$. Bottom 5: $\lambda_1 = \lambda_2 = 1$. The horizontal line gives the true value of ρ.*

the efficiency is the same whether or not the λ's are known in this Box-Cox model was also obtained by Wong (1981). This result is very different from the results of Bickel and Doksum (1981) regarding the estimation of regression coefficients.

***Remark* 8.3 :** Consider the transformation model

$$h_0(Y) = \boldsymbol{\beta}^T \boldsymbol{X} + \epsilon, \ \epsilon \sim N(0, \sigma), \tag{8.3}$$

where $\boldsymbol{X}$ is a vector of random covariates and $h(\cdot)$ is increasing. In this case we can consider the rank estimate $\hat{\boldsymbol{\beta}}_R$ obtained by maximizing the rank likelihood $l_{\boldsymbol{r}}(\boldsymbol{\beta}) = P(\boldsymbol{R} = \boldsymbol{r})$ defined in Section 3. The results of Bickel and Ritov (1997) imply that in a certain sense $\hat{\boldsymbol{\beta}}_R$ is semiparametrically optimal for model (8.3). However the normal scores estimate of $\hat{\boldsymbol{\beta}} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{a}$, where $\boldsymbol{a} = (a(S_1), \cdots, a(S_n))^T$ and $\boldsymbol{x}$ is a vector of nonrandom covariates, is not asymptotically optimal unless $|\boldsymbol{\beta}|/\sigma$ tends to zero at a certain rate as $n \to \infty$ (Doksum (1987)). MC methods for $\hat{\boldsymbol{\beta}}_R$ is introduced in Bickel and Doksum (2009), Section 10.5.

We next introduce a semiparametric approach for the copula regression model and a nonparametric regression approach.

II Normal score substitution.

The model (8.1) with h_j, $j = 0, \cdots, d$, satisfying (7.4) and (7.5) is invariant under increasing transformations. As in the d=1 case, this leads to using the ranks

$\{S_i\}$ of the Y's and the ranks $\{R_{ij} : i = 1, \cdots, n\}$ of X_{ij} among $\{X_{ij} : i = 1, \cdots, n, j = 1, \cdots, d\}$. Because the distribution of the ranks is invariant under increasing transformations, for rank methods, model (8.1) is equivalent to

$$Y' = \boldsymbol{\alpha}^T \boldsymbol{X}' + \epsilon', \tag{8.4}$$

where $X'_j \sim N(0,1)$ and ϵ' are independent, $Y' \sim N(0,1)$ and $\boldsymbol{\alpha}^T = \Sigma^{-1}\boldsymbol{\rho}$ with $\boldsymbol{\rho} = (Corr(X'_1, Y'), \cdots, Corr(X'_d, Y'))^T$ and Σ the correlation matrix of $\boldsymbol{X}' = (X'_j)_{d\times 1}$. Here Σ is assumed to be nonsingular. Based on the distribution of the ranks, $\alpha_1, \cdots, \alpha_d$ are identifiable parameters in model (8.4). These parameters represents the relative importance of the X_j's.

The normal scores $Z'_{ij} = a(R_{ij})$ and $W'_i = a(S_i)$ have approximately the same distribution as the unobservable X'_{ij} and Y'_i in model (8.4). Because $E(Y'|\boldsymbol{x}') = \boldsymbol{\alpha}^T\boldsymbol{x}'$, if we replace X'_{ij} and Y'_i with Z'_{ij} and W'_i, we find that an approximate method of moments estimate of α is

$$\hat{\boldsymbol{\alpha}} = (Z_D^T Z_D)^{-1} Z_D^T \boldsymbol{W'}, \tag{8.5}$$

where Z_D is the no intercept design matrix $(Z'_{ij})_{n\times d'}$, d' is the rank of the matrix (Z'_{ij}), and $\boldsymbol{W'} = (W'_1, \cdots, W'_n)^T$. Any subset of variables $X_j : j \in J$ with the same ranks, say $R_{1J}, \cdots, R_{nJ}$, is collapsed into one variable denoted as X_J with ranks $R_{1J}, \cdots, R_{nJ}$ to avoid singularity. Based on Klaassen and Wellner (1997), we conjecture that $\hat{\boldsymbol{\alpha}}$ is semiparametrically efficient for the multivariate normal copula model.

III Nonparametric estimation.

We next introduce a nonparametric approach. We consider the model

$$Y = m(\boldsymbol{X}) + \epsilon, \tag{8.6}$$

where $m()$ is unknown and ϵ has median zero. To estimate $m()$, we use a cubic B-spline and the R function smooth.spline(). The number of knots are automatically selected (less than the number of observations n.) The smoothing parameter is chosen by generalized cross validation (GCV.)

8.1. Simulation Results

We consider the d=1 case and consider the properties of estimates of $\rho = Corr(h_1(X), h_0(Y))$. In this case, the method II estimate is $\hat{\rho}_{NS}$.

8.1.1. Correctly Specified Model

The true model satisfies

$$Y^{(\lambda_2)} = \alpha_0 + \alpha_1 X^{(\lambda_1)} + \epsilon, \tag{8.7}$$

where $\epsilon \sim N(0, \sigma^2)$ and $X^{(\lambda_1)} \sim N(\mu_1, \sigma_0^2)$ are independent. This model is a subset of the normal copula model $\mathcal{F}$ with

$$F_1(x) = \Phi(\frac{x^{(\lambda_1)} - \mu_1}{\sigma_0}),\ F_2(y) = \Phi(\frac{y^{(\lambda_2)} - \mu_2}{\sigma_2}), \tag{8.8}$$

where $\mu_2 = EY^{(\lambda_2)}$, and $\sigma_2^2 = VarY^{(\lambda_2)}$. We use 1000 MC trials and take $\sigma^2 = 1$, $\sigma_0^2 = 1$, $(\lambda_1,\ \lambda_2) \in \{(0.5, 0.5), (1, 1)\}$, $\alpha_0 = 6$, $\alpha_1 = \in \{0, 0.1, 0.5, 1, 2\}$, and $\mu_1 = 5$.

Fig. 3 shows that methods I and II have similar properties for $\alpha_1 \leq 0.5$. For larger α_1, the normal scores estimate has a downward bias which is negligible for $n \geq 500$ (not shown here.) Method I converges all the time with the constraint $-\mathbf{4} \leq \boldsymbol{\lambda} \leq \mathbf{4}$. Method II does not involve any optimization and hence converges all the time.

8.1.2. Misspecified Model

We simulate the data from

$$Y^{(\lambda_2)} = (1-\gamma)(\alpha_0 + \alpha_1 X^{(\lambda_1)}) + \gamma[L(X)] + \epsilon, \tag{8.9}$$

where $L()$ is a nonlinear function. Thus the model is a Box-Cox model when $\gamma = 0$, but when $\gamma > 0$, we are checking the performance of the methods when the model generating the methods are misspecified.

For comparisons of methods we need a parameter that makes sense for all three methods. One such parameter is

$$m(x) = Median(Y|X = x). \tag{8.10}$$

We consider the 25th, 50th, and 75th population quantiles of X, i.e., our parameters of interest are $m(x_{0.25})$, $m(x_{0.50})$, and $m(x_{0.75})$.

Methods I and II are based on models of the form

$$h_0(Y) = g(\boldsymbol{X}, \boldsymbol{\beta}) + \epsilon, \tag{8.11}$$

where $h_0(\cdot)$ is an increasing function. If X and ϵ are independent and $median(\epsilon) = 0$, then

$$m(x) = h_0^{-1}(g(\boldsymbol{X}, \boldsymbol{\beta})). \tag{8.12}$$

For method I, the MLE of m(x) in model (8.7) is,

$$\hat{m}(x) = (\hat{\lambda}_2[\hat{\beta}_0 + \hat{\beta}_1 \frac{x^{\hat{\lambda}_1} - 1}{\hat{\lambda}_1}] + 1)^{1/\hat{\lambda}_2}. \tag{8.13}$$

For method II, write model (8.1) as

$$h_0(Y) = \rho h_1(X) + \epsilon,\ \epsilon \sim N(0, \sigma^2), \tag{8.14}$$

where $\rho \equiv \rho(h_1(X), h_0(Y))$ is the correlation coefficient. Then by (8.12),

$$m(x) = h_0^{-1}(\rho h_1(x)), \tag{8.15}$$

where

$$h_0^{-1}(t) = F_2^{-1}(\Phi(t)). \tag{8.16}$$

It follows that

$$m(x) = {F_2}^{-1}(\Phi(\rho \Phi^{-1}(F_1(x)))), \tag{8.17}$$

and by replacing F_1 and F_2 by their empiricals, a natural estimate of m(x) is

$$\hat{m}(x) = y_{([n\Phi(\hat{g}(x))])}, \tag{8.18}$$

where $\hat{g}(x) = \hat{\rho}_{NS}\Phi^{-1}(\hat{F}_1(x))$ and [ ] is the greatest integer function.

For method III, we use the smoothing spline estimate of $E(Y|X = x)$ described earlier. In our models with normal errors, $E(Y|X = x)$ coincide with the conditional median m(x).

In the simulation we use model (8.9) with $\epsilon \sim N(0, \sigma^2)$ and $X^{(\lambda_1)} \sim N(\mu_1, \sigma_0^2)$ independent,

$$L(t) = \alpha_0 + \alpha_1\mu_1 - 1.25 + 2.5[1 + \exp(-10(t - \mu_1))]^{-1}, \tag{8.19}$$

$\sigma^2 \in \{0.01, 0.1, 0.5, 1\}$, $\sigma_0^2 = 1$, $(\lambda_1,\ \lambda_2) \in \{(0.5, 0.5), (1, 1)\}$, $\alpha_0 = 6.25$, $\alpha_1 = \in \{0, 0.5, 1, 2\}$, $\mu_1 = 5$, and $\gamma \in \{0, 0.25, 0.5, 0.75, 1\}$. The sample size is n=512. There are 1000 MC trials.

Fig. 4, 5, 6, and 7 are boxplots of $\hat{m}(x_{0.25})$, $\hat{m}(x_{0.50})$, and $\hat{m}(x_{0.75})$ with the setting $(\lambda_1, \lambda_2, \alpha_1, \sigma^2) = (1, 1, 0.5, 0.5)$, $(\lambda_1, \lambda_2, \alpha_1, \sigma^2) = (1, 1, 0.5, 1)$, $(\lambda_1, \lambda_2, \alpha_1, \sigma^2) = (0.5, 0.5, 0.5, 0.5)$, and $(\lambda_1, \lambda_2, \alpha_1, \sigma^2) = (0.5, 0.5, 0.5, 1)$ respectively. Fig.8-11 give MSE's for the estimates $\hat{m}(x_{0.25})$, $\hat{m}(x_{0.50})$, and $\hat{m}(x_{0.75})$.

We see that method I is the best when model (8.2) is correct, that is, $\gamma = 0$. However when the model is increasingly misspecified, i.e., as γ increases, its absolute bias increases which leads to low MSE performance.

Method II is overall best in terms of MSE when $\lambda_1 = \lambda_2 = 0.5$ (Fig 6 and 7). When $\gamma = 0$, it is unbiased and its variance is between Method I and Method III (Fig 4,5,6, and 7).

Method III is overall best in terms of MSE when $\lambda_1 = \lambda_2 = 1$ and the model is badly misspecified. It's smaller bias makes up for its large variance in this case. But its MSE suffers at and near model (8.2) (Fig 6, 7, $\gamma = 0$).

In summary, the normal score procedure performs very well at and close to a copula model. For n large, this is to be expected from the results of Klaassen and Wellner (1997). The normal scores estimate is competitive with the Box-Cox estimate in the transform both sides Box-Cox model.

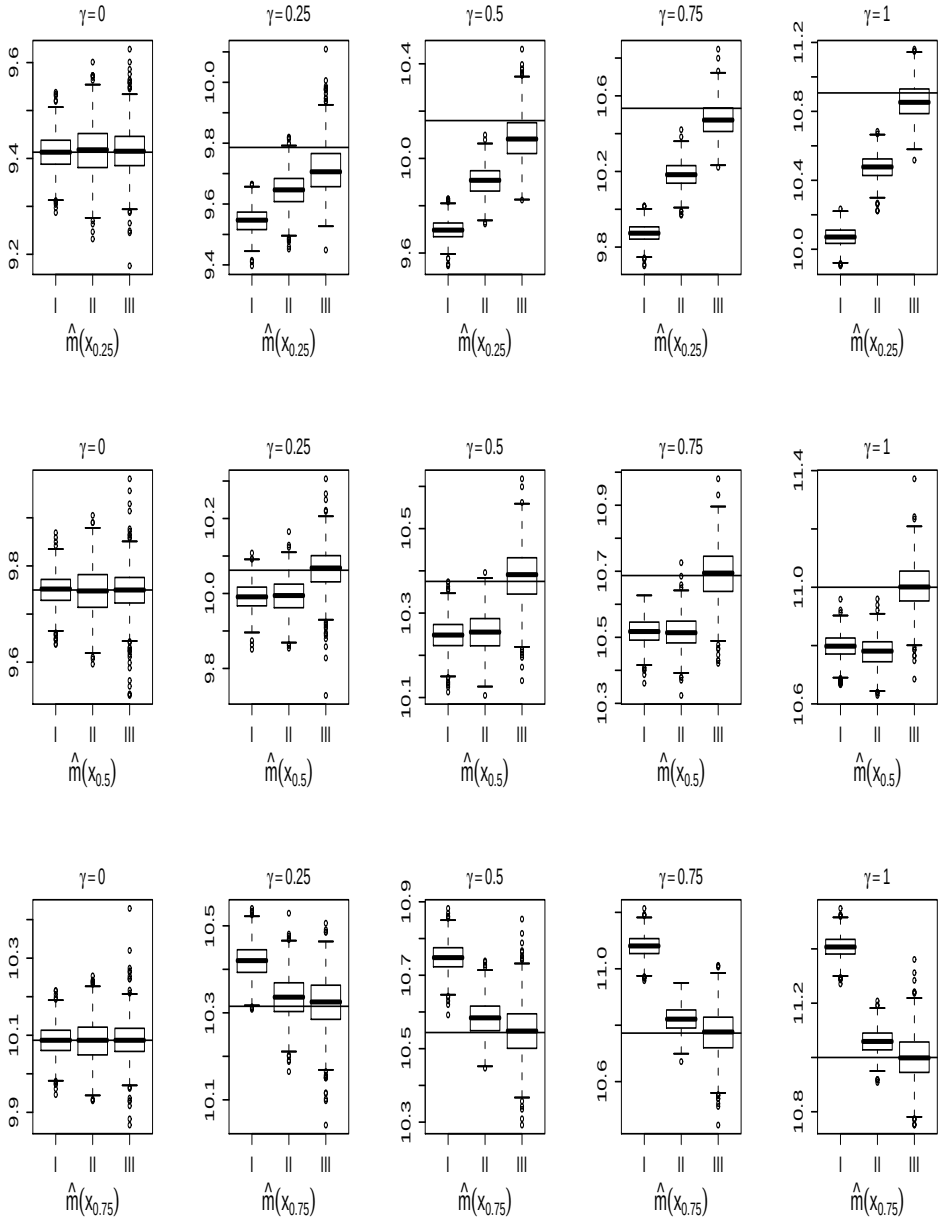


FIG 4. *Boxplots of the three estimates of median regression m(x) for model (8.9) with* $(\lambda_1, \lambda_2, \alpha_1, \sigma^2)$*=(1, 1, 0.5, 0.5). I: profile MLE, II: normal scores, and III: NP, spline. The true value of m(x) is the solid line.*

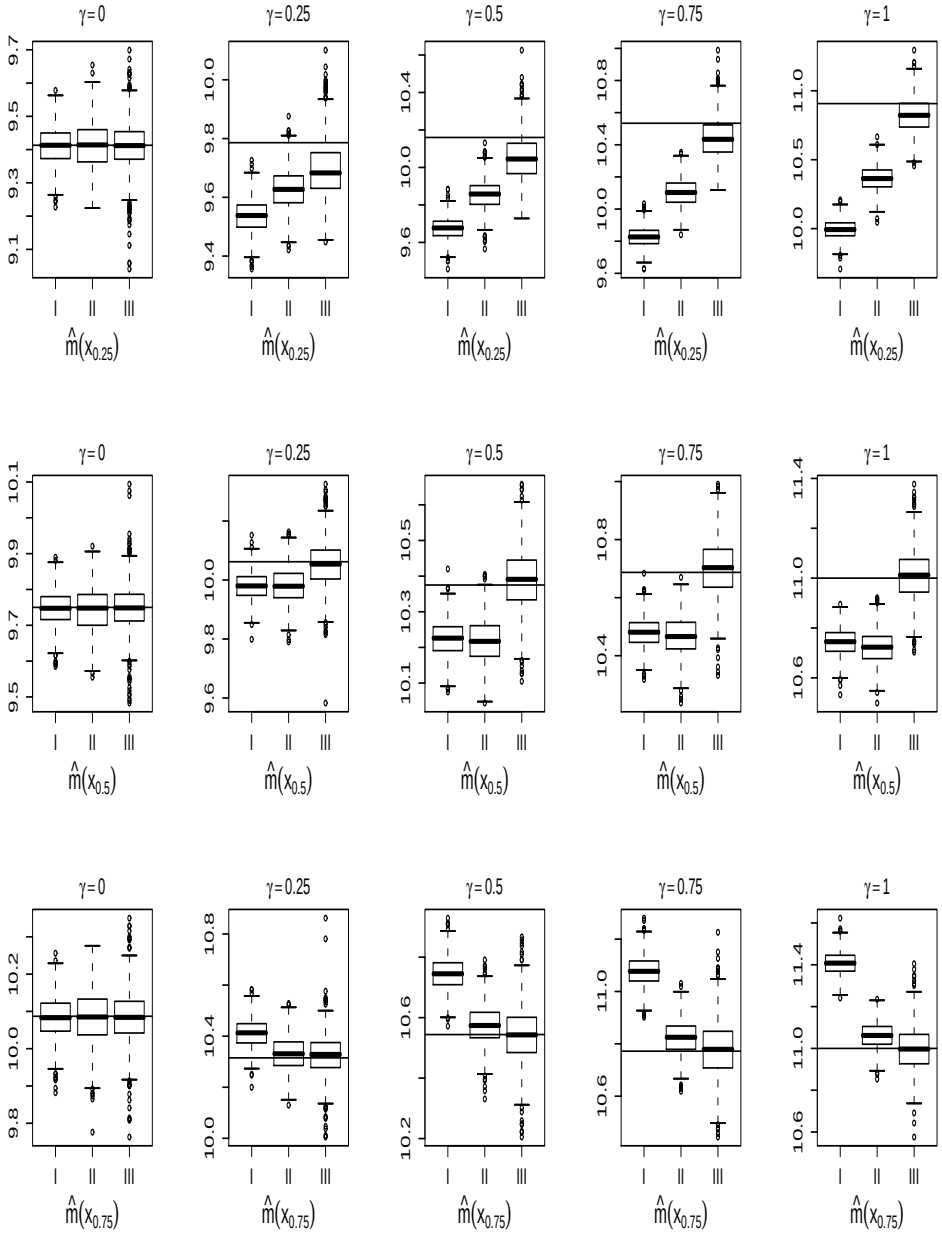


FIG 5. *Boxplots of the three estimates of median regression m(x) for model (8.9) with* $(\lambda_1, \lambda_2, \alpha_1, \sigma^2)$*=(1, 1, 0.5, 1). I: profile MLE, II: normal scores, and III: NP, spline. The true value of m(x) is the solid line.*

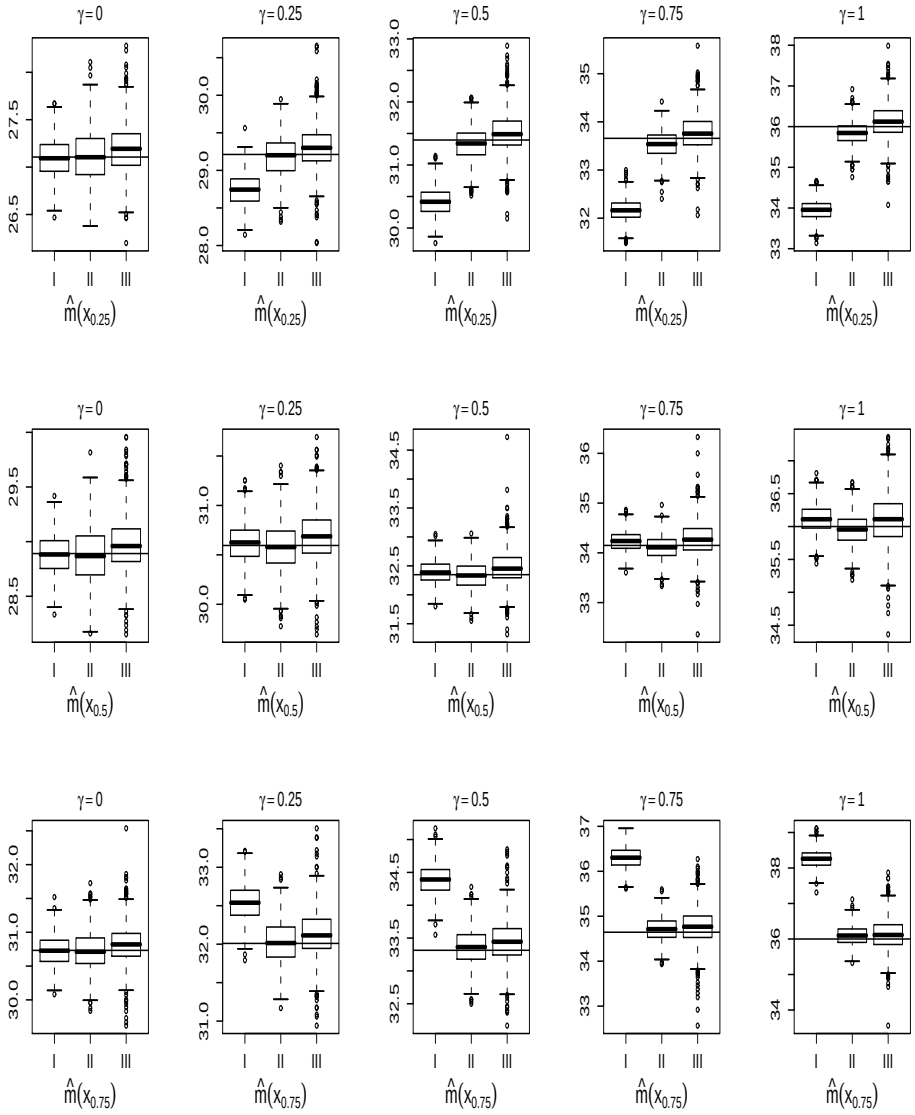


FIG 6. *Boxplots of the three estimates of median regression m(x) for model (8.9) with* $(\lambda_1, \lambda_2, \alpha_1, \sigma^2)$*=(0.5, 0.5, 0.5, 0.5). I: profile MLE, II: normal scores, and III: NP, spline. The true value of m(x) is the solid line.*

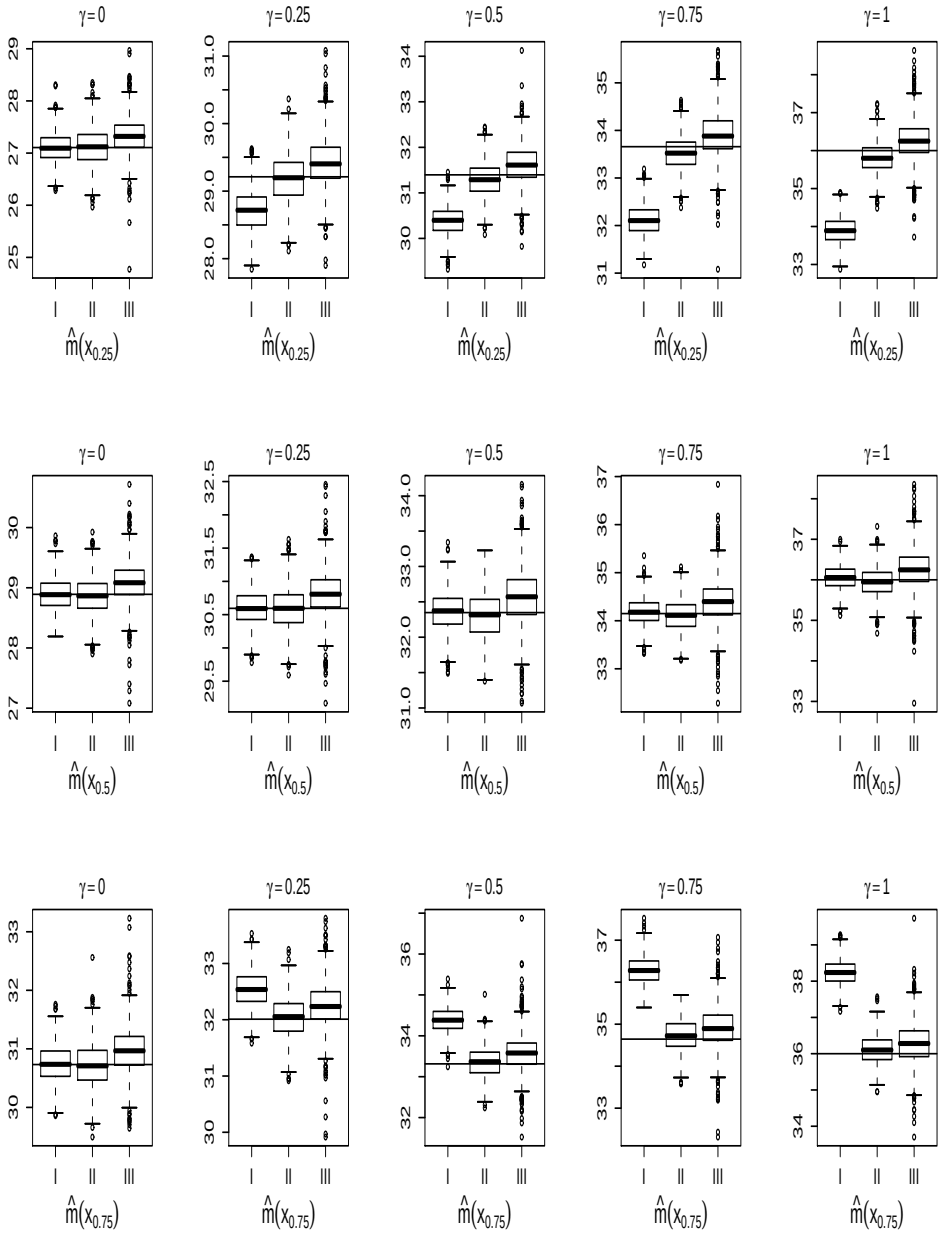


FIG 7. *Boxplots of the three estimates of median regression m(x) for model (8.9) with* $(\lambda_1, \lambda_2, \alpha_1, \sigma^2)$*=(0.5, 0.5, 0.5, 1). I: profile MLE, II: normal scores, and III: NP, spline. The true value of m(x) is the solid line.*

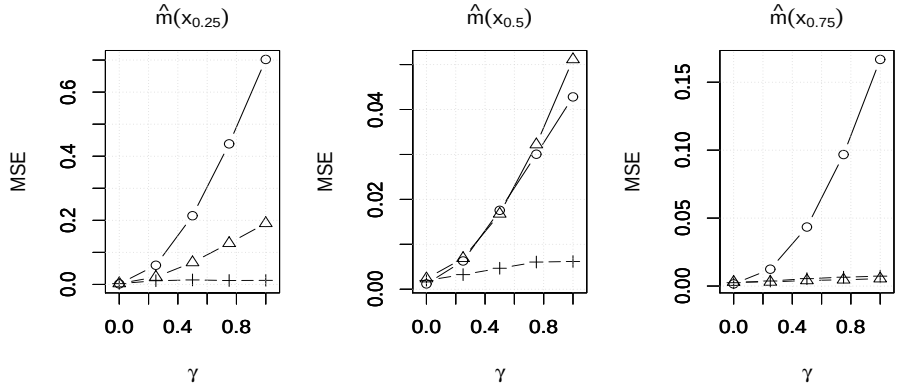


FIG 8. *MSE of the three estimates of m(x) as a function of the misspecification parameter* γ *for* $(\lambda_1, \lambda_2, \alpha_1, \sigma^2)$*=(1, 1, 0.5, 0.5).* $\bigcirc = I$, $\triangle = II$, $+ = III$.

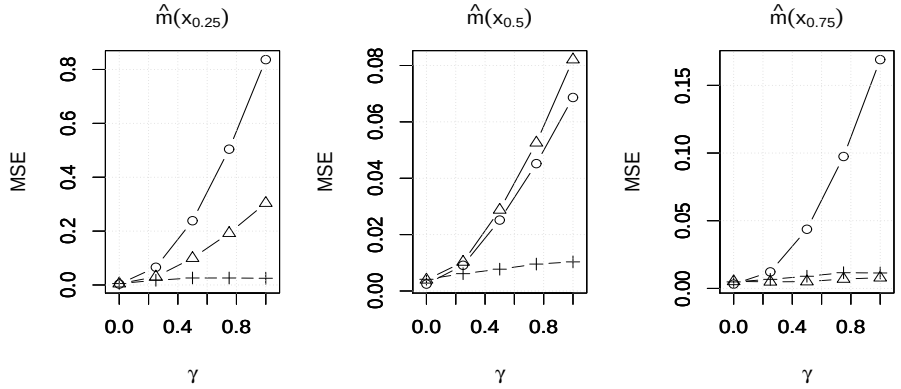


FIG 9. *MSE of the three estimates of m(x) as a function of the misspecification parameter* γ *for* $(\lambda_1, \lambda_2, \alpha_1, \sigma^2)$*=(1, 1, 0.5, 1).* $\bigcirc = I$, $\triangle = II$, $+ = III$.

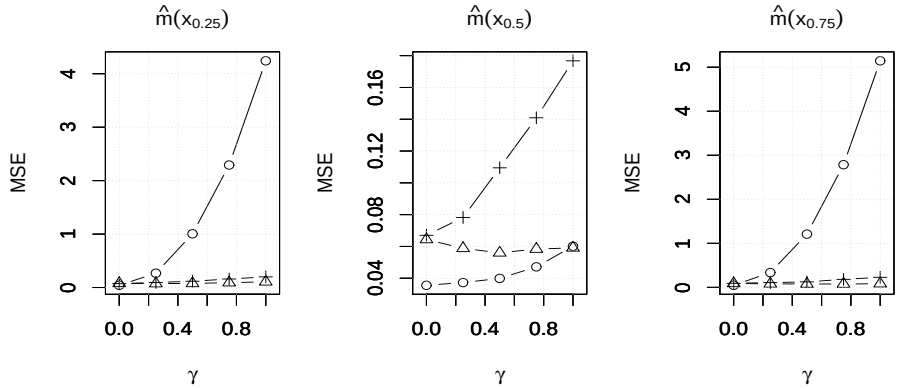


FIG 10. *MSE of the three estimates of m(x) as a function of the misspecification parameter* γ *for* $(\lambda_1, \lambda_2, \alpha_1, \sigma^2)$*=(0.5, 0.5, 0.5, 0.5).* $\bigcirc = I$, $\triangle = II$, $+ = III$.

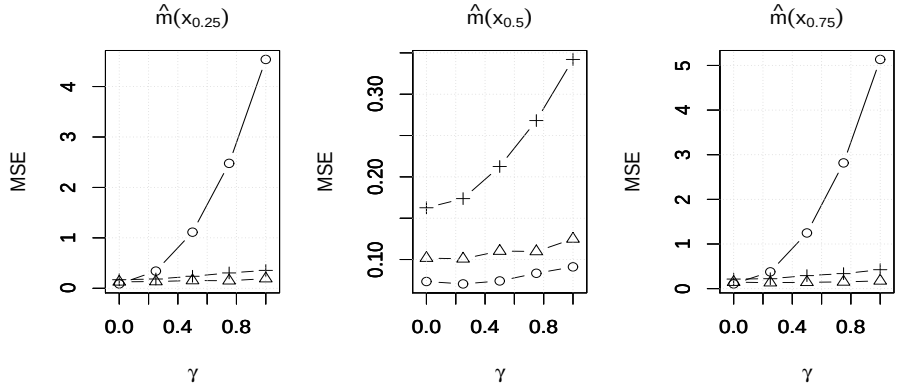


FIG 11. *MSE of the three estimates of m(x) as a function of the misspecification parameter* γ *for* $(\lambda_1, \lambda_2, \alpha_1, \sigma^2)$*=(0.5, 0.5, 0.5, 1).* $\bigcirc = I$, $\triangle = II$, $+ = III$.

Acknowledgements

We thank Peter Bickel, Aad van der Vaart and Jon Wellner for helpful comments.

References

[1] ANDERSEN, P. K., BORGAN, O., GILL, R. D., AND KEIDING N. (1996). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.
[2] BEGUN, J. M., HALL, W. J., HUANG, WEI-MIN, AND WELLNER, J. A. (1983). Information and Asymptotic Efficiency in Parametric-Nonparametric Models. *The Annals of Statistics*, **11**, 2, 432–452.
[3] BELL, C. B., AND DOKSUM, K. A. (1966). "Optimal" One-Sample Distribution-Free Tests and Their Two-Sample Extensions. *The Annals of Mathematical Statistics*, **36**, 1, 120–132.
[4] BHUCHONGKUL, S. (1964). A Class of Nonparametric Tests for Independence in Bivariate Populations. *Ann. Math. Statist.*, **35**,1, 138–149.
[5] BICKEL, P. J., AND DOKSUM, K. A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, **76**, 296–311.
[6] BICKEL, P. J., AND DOKSUM, K. A. (2007). *Mathematical Statistics: Basic Ideas and Selected Topics, 2nd ed. Vol I*, Updated Printing. Pearson Prentice Hall, Upper Saddle River, NJ.
[7] BICKEL, P. J., AND DOKSUM, K. A. To appear (2009). *Mathematical Statistics: Basic Ideas and Selected Topics Vol II*. Pearson Prentice Hall, Upper Saddle River, NJ.
[8] BICKEL, P. J., KLAASSEN, C. A., RITOV, Y., AND WELLNER, J. A. (1993, 1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, New York.
[9] BICKEL, P. J., AND RITOV, Y. (1997). Local asymptotic normality of ranks and covariates in transformation models. In *Festschrift for Lucien Le Cam (D. Pollard and G. L. Yang, eds.)* Springer, New York.
[10] COX, D. R. (1964). Some applications of exponential ordered scores. *J. R. Stat. Soc.*, **B.26**, 103–110.
[11] COX, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc.*, **B.34**, 187–220.
[12] COX, D. R. (1975). Partial likelihood. *Biometrika.*, **62.2**, 269–276.
[13] COX, D. R. (2006). *Principles of Statistical Inference*. Cambridge University Press, Cambridge, 40.
[14] DOKSUM, K. A. (1987). An extension of partial likelihood methods from proportional hazard models to general transformation models. *Annals of Statistics*, **15**, 325–345.
[15] DOKSUM, K. A., OZEKI, A., KIM, J., AND NETO, E. C. (2007). Thinking outside the box: Statistical inference based on Kullback-Leibler Empirical Projections. *Statistics and Probability Letters*. **77**, 1201–1213.
[16] FISHER, R. A. AND YATES, F. (1938). *Statistical Tables for Biological, Agricultural and Medical Research*. (5th edition, 1957) Oliver and Boyd, Edinburgh.
[17] FERGUSON, T. S. (1967). *Mathematical Statistics. A Decision Theoretic Approach*. New York and London, Academic Press.
[18] HAJEK, J. AND SIDAK, Z. (1967). *Theory of Rank Tests*. Academic Press, New York.

[19] Hodges, J. L. and E. L. Lehmann (1963). Estimates of location based on rank tests. *Ann. Math. Stat*, **34**, 598-611.

[20] Hoeffding, W. (1951). 'Optimum' nonparametric tests. *Proc. 2nd Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, Univ. Calif. Press, 83–92.

[21] Hernandez, F. and Johnson, R.A. (1980). The large-sample behavior of transformations to normality. *Journal of the American Statistical Association*, **75**, 855–861.

[22] Kalbfleich, J. D. and Prentice, R. L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika*, **60**, 267–278.

[23] Kalbfleich, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of failure Time Data*, 2nd edition. John Wiley and Sons, Inc., Hoboken, New Jersey.

[24] Klaassen, C. A. J. (2007). A Sturm-Liouville Problem in Semiparametric Transformation Models. In *Advances In Statistical Modeling And Inference. Essays in Honor of Kjell A Doksum*. Editor, Vijay Nair. World Scientific Pub Co Inc., New Jersey.

[25] Kosorok, M. R., Lee B. L., and Fine J.P . (2004). Robust inference for univariate proportional hazards frailty regression models. *The Annals of Statistics*, **32**, No. 4, 1448–1491.

[26] Klaassen, C. A. J. and Wellner, J. A. (1997). Efficient Estimation in the Bivariate Normal Copula Model: Normal Margins Are Least Favourable. *Bernoulli*, **3**, No. 1, 55–77.

[27] Lange, K., Hunter, D. R., and Yang, I. (2000). Optimization Transfer Using Surrogate Objective Functions. *Journal of Computational and Graphical Statistics*, **9**, No. 1, 1–20.

[28] Lehmann, E. L. (1953). The power of rank tests. *Ann. Math. Statist.*, 24, 23–43.

[29] Murphy, S. A. and van der Vaart, A. W. (2000). On Profile Likelihood. *JASA*, **9**, No. 450, 449–465.

[30] Nabeya, S. and Miura, R. (1972). Locally Most Powerful Rank Tests for Lehmann mixture alternatives. *Technical Report*. University of California, Berkeley.

[31] Oakes, D. (1992). Bivariate survival models induced by frailties. *JASA*, **84**, 487–493.

[32] Oakes, D. and Jong-Hyeon Jeong (1998). Frailty Models and Rank Tests. *Lifetime Data Analysis*, **4**, 3, 209–228.

[33] Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, No.2, 237–249.

[34] Owen, A. B. (2001). *Empirical Likelihood*. Chapman and Hall.

[35] Ruymgaart, F. H. (1974). Asymptotic Normality of Nonparametric Tests for Independence. *Ann. Statist.*, **2**, 892–910.

[36] Ruymgaart, F. H., Shorack, G. R., and Van Zwet, W. R. (1972). Asymptotic Normality of Nonparametric Tests for Independence. *Ann. Math. Statist.*, **43**, 1122–1135.

[37] Savage, I. R. (1956). Contributions to the theory of rank orders statistics: the two-sample case. *Ann. Math. Stat.*, **27**, 590–615.

[38] Savage, I. R. (1957). Contributions to the Theory of Rank Order Statistics–The "Trend" Case. *Ann. Math. Stat.*, **28**, 4, 968–977.

[39] Sibuya, M. (1968). Generating Doubly Exponential Random Numbers. *Annals of the Institute of Statistical Mathematics, Tokyo, Supplement V*, 1–7.

[40] Sklar, A. (1959). Fonctions de repartition a n dimensions et leurs marges. *L'Institut de Statistique de L'Universite de Paris*. **8**, 229–231.

[41] Tsodikov, A. (2003). Semiparametric models: a generalized self-consistency approach. *J. R. Statist. Soc.*, **B.65**, 3, 759–774.

[42] Tsodikov, A. and Garibotti, G. (2007). Profile information matrix for nonlinear transformation models. *Lifetime Data Analysis*, **13**, 139–159.

[43] van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge University Press, Cambridge, UK.

[44] Zeng, D. and Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *J. R. Statist.*, **69**, 4, 1–30.

[45] Zou, K.H., and Hall, W.J.(2002) (2002) On estimating a transformation correlation coefficient. *Journal of Applied Statistics*, **29**, 745–760.

[46] Wong, C-W. (1981). Transformation of Independent Variables in Regression Models. *Ph.D. Thesis*, Department of Statistics, University of Berkeley, CA.