Published in final edited form as:

Lifetime Data Anal. 2007 March ; 13(1): 139–159. doi:10.1007/s10985-006-9023-z.

Profile information matrix for nonlinear transformation models

A. Tsodikov and

Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029, USA, tsodikov@umich.edu

G. Garibotti

Centro Regional Universitario Bariloche, Universidad Nacional del Comahue, Quintral 1250, 8400 Bariloche, Argentina, garibott@cab.cnea.gov.ar

Abstract

For semiparametric models, interval estimation and hypothesis testing based on the information matrix for the full model is a challenge because of potentially unlimited dimension. Use of the profile information matrix for a small set of parameters of interest is an appealing alternative. Existing approaches for the estimation of the profile information matrix are either subject to the curse of dimensionality, or are ad-hoc and approximate and can be unstable and numerically inefficient. We propose a numerically stable and efficient algorithm that delivers an exact observed profile information matrix for regression coefficients for the class of Nonlinear Transformation Models [A. Tsodikov (2003) *J R Stat Soc Ser B* 65:759–774]. The algorithm deals with the curse of dimensionality and requires neither large matrix inverses nor explicit expressions for the profile surface.

Keywords

Profile likelihood; Semiparametric models; Information matrix

1 Introduction

In semiparametric models the parameter is partitioned as (β, H) with β a low-dimensional parameter of interest and H a high-dimensional nuisance parameter. For example, in semiparametric regression survival models, β is the vector of regression coefficients and H is the baseline cumulative hazard function estimated as a step-function by the Nonparametric Maximum Likelihood Estimator (NPMLE). The dimension of H is given by the number of distinct failure times and increases with the sample size.

Within the NPMLE framework the following tools are available for interval estimation and hypothesis testing for β .

1. *Likelihood Ratio*. The likelihood ratio statistic for testing $H_0 : \beta = \beta_0$ is defined as,

$$LR(\beta_0) = 2(\ell(\hat{\beta}, \hat{H}) - \ell(\beta_0, \hat{H}(\beta_0))),$$

where ℓ is the log-likelihood function, $(\hat{\beta}, \hat{H})$ is the NPMLE of (β, H) , and $\hat{H}(\beta)$ is the MLE of H given β . Although classical ML theory does not directly apply to

unlimited dimension, for many semiparametric models LR has an asymptotic chi-square distribution with d degrees of freedom, where d is the dimension of β . A $(1 - \alpha)\%$ confidence set for β is given by

$$\{\beta: \text{LR}(\beta) \leq C_{d,\alpha}\},$$

where $C_{d,\alpha}$ is the α percentile of the chi-square distribution with d degrees of freedom. When the asymptotic distribution of LR is unknown, bootstrap can be used to approximate $C_{d,\alpha}$.

The likelihood ratio approach for building confidence regions for β involves inverting the LR surface, which is quite computer intensive as repeated maximizations of the likelihood with respect to H are required.

2. *Wald Statistic.* An alternative method of inference for β is based on the Wald statistic defined as

$$W(\beta) = (\hat{\beta} - \beta)^T \sum_{\beta\beta}^{-1} (\hat{\beta} - \beta),$$

where $\sum_{\beta\beta}$ is the β -submatrix of the inverse of the observed information matrix

$$I = \begin{pmatrix} \frac{-\partial^2 \ell(\beta, H)}{\partial \beta \partial \beta^T} & \frac{-\partial^2 \ell(\beta, H)}{\partial \beta \partial H^T} \\ \frac{-\partial^2 \ell(\beta, H)}{\partial H \partial \beta^T} & \frac{-\partial^2 \ell(\beta, H)}{\partial H \partial H^T} \end{pmatrix} \bigg|_{\beta = \hat{\beta}, H = \hat{H}}$$

Note that in the presence of nuisance parameters the information matrix needs to be inverted twice (Severini 2000, p. 121), the first time in its high-dimensional full model form I , and the second time as a $\dim(\beta)$ -submatrix of $\sum = I^{-1}$.

Under certain conditions, W is asymptotically equivalent to the likelihood ratio and asymptotically has a chi-square distribution with d degrees of freedom. In this case,

$$\{\beta: W(\beta) \leq C_{d,\alpha}\},$$

is a confidence set of approximate coverage probability $1 - \alpha$.

The bottleneck of this procedure is the inversion of a potentially infinitely large matrix I .

The two methods of inference on β described above are based on the full model. An appealing alternative is to consider the so-called profile likelihood (Murphy and van der Vaart 2000)

$$\ell_{\text{pr}}(\beta) = \max_H \ell(\beta, H).$$

The profile likelihood may be used as a likelihood for β . The MLE for β , the first component of the pair $(\hat{\beta}, \hat{H})$ that maximizes $\ell(\beta, H)$, is the maximizer of the profile likelihood function $\ell_{\text{pr}}(\beta)$.

Theoretical justification for the use of the profile likelihood for semiparametric models was given in (Murphy and van der Vaart 1997, 2000; van der Vaart 1998). It was shown that

profile likelihoods with the nuisance parameter estimated out behave like ordinary likelihoods under regularity conditions. These conditions need to be verified on a case-by-case basis as the general theory remains a challenge. Theoretical justification has been obtained for the proportional odds (PO) model (Murphy and van der Vaart 2000; Murphy et al. 1997) and the PH frailty models (Murphy 1994, 1995; Parner 1998; Kosorok et al. 2004).

The observed profile information matrix will be denoted I_{pr} ,

$$I_{\text{pr}} = - \frac{\partial^2 \ell_{\text{pr}}(\beta)}{\partial \beta \partial \beta^T} \Big|_{\beta = \hat{\beta}}.$$

This matrix is asymptotically the same as $\sum_{\beta\beta}^{-1}$ and summarizes partial information on β .

The Likelihood Ratio and Wald statistics based on ℓ_{pr} are easier to obtain than the ones based on the full model provided

- a numerically efficient method is available to profile out the nuisance parameter H , and
- it is possible to derive the exact observed profile information matrix or estimate it in a computationally efficient way.

Fulfilling both conditions is a challenge. First, maximization over H is a problem of potentially very large dimension. Second, in most cases ℓ_{pr} cannot be differentiated analytically. Several alternatives for estimating the profile information matrix have been proposed in the literature. However, they are all approximations, often difficult to calibrate in practice, and algorithms to obtain them are computationally costly.

In this paper we propose a computationally efficient exact solution for the class of semiparametric Nonlinear Transformation Models (NTM) (Tsodikov 2003). The basic assumption that defines this model family is that the survival function at each timepoint t is a *function* of $H(t)$ mapping real numbers $[0, \infty] \rightarrow [0, 1]$ rather than a *functional* mapping a functional space to $[0, 1]$. In other words, the model-based survival function is obtained by plugging a cumulative hazard H or a baseline survival function $F = \exp(-H)$ into a suitably defined parametric function (so-called model-generating function, see Sect.2). Note that a similar assumption underlies the von-Mises Calculus (van der Vaart 1998, p. 291). The NTM class includes the proportional hazards (PH) model, univariate PH frailty models, the PO model, and cure models such as the PHPH model (Tsodikov 2002; Tsodikov et al. 2003). A numerically efficient Quasi-EM algorithm, a subset of the MM family (Lange et al. 2000) was developed to obtain the maximum profile likelihood for NTM models (Tsodikov 2003). The algorithm has since been used in computer intensive settings such as the bootstrap (Dixon et al. 2005).

The algorithm for the exact I_{pr} proposed in this paper works under the following two basic assumptions.

Independence of the future. Independence of the future means that the contribution to the likelihood of an observed event at time t depends on the past $H[0, t]$ of the function H , but not on the future.

Nonlinear Transformation Model Assumption. The survival function given covariates is specified as a parametric transformation of H . A detailed definition is given below.

We compare our method to the following three existing techniques used to estimate the profile information matrix that amount to particular forms of numerical differentiation of the second order.

1. *Discretized second derivative.* Corollary 3 of Murphy and van der Vaart (2000) shows that under certain conditions

$$-2 \frac{\log \ell_{\text{pr}}(\hat{\beta} + h_n \nu_n) - \log \ell_{\text{pr}}(\hat{\beta})}{nh_n^2} \xrightarrow{P} \nu^T I_{\text{pr}} \nu, \quad (1)$$

for all sequences $\nu_n \xrightarrow{P} \nu \in \mathbb{R}$ and $h_n \xrightarrow{P} 0$ such that $(\sqrt{n}h_n)^{-1} = O_p(1)$.

This result can be used to derive an estimate of I_{pr} . Note that this method requires careful maintenance of the speed of convergence of the sequence $\{h_n\}$ as the

condition $(\sqrt{n}h_n)^{-1} = O_p(1)$ implies that the convergence should be neither too slow nor too fast. The reason is that the precision of the discrete differential operator as $n \rightarrow \infty$ on the left side of (1) needs to be measured against the convergence of MLEs to the true value. Indeed, under regularity conditions, the asymptotic expansion of the likelihood ratio statistic about the MLE $\hat{\beta}$ has the form

$n(\hat{\beta} - \beta)^T I_{\text{pr}}(\hat{\beta} - \beta) + o_p(\sqrt{n}\|\hat{\beta} - \beta^*\| + 1)$, where β^* is the true value. The procedure (1) is designed to extract the quadratic term by setting $\beta = \hat{\beta} + h_n \nu_n$, and by ensuring the $1/\sqrt{n}$ rate of convergence of $\hat{\beta}$ to simultaneously $\hat{\beta}$ and β^* so that the quadratic term is indeed the dominant one. Otherwise, the expansion would be dominated by its $o_p(1)$ part if h_n is too fast or by $o_p(\sqrt{n}\|\hat{\beta} - \beta^*\|)$ if h_n is too slow.

See Sect. 4 for further details and implementation of this method.

2. *Fitting a Quadratic Form.* Asymptotically, under regularity conditions, the profile likelihood surface around the true β is quadratic. Nielsen et al. (1992) proposed fitting a quadratic form to $\ell_{\text{pr}}(\hat{\beta})$ in some domain around the maximum likelihood estimator, $\hat{\beta}$, and the derivation of an approximate profile information matrix using the estimated coefficients of the form. Note that globally the likelihood surface is not quadratic. The quadratic approach is difficult to implement as a sufficiently small domain around $\hat{\beta}$ where the likelihood surface can be well approximated by a quadratic form is not well defined. Misspecification of this domain with the quadratic method often leads to estimates of the profile information matrix that are not positive definite, particularly if the number of covariates is large. Yet the domain needs to be large enough to ensure adequate precision and sufficient sample size representing the number of likelihood evaluations within the domain. This balancing act is notoriously difficult as the true variance is unknown and the likelihood surface is specific to the data set being analyzed.
3. *Numerical Differentiation of the Profile Likelihood.* Standard numerical algorithms can be used to numerically differentiate the profile likelihood function. We use Ridder's method (Press et al. 1994) in the examples presented in Sect. 4. The difficulties in the implementation of this idea are similar to the ones with the Quadratic Form approach. Numerical differentiation requires choosing a tolerance for the estimation of the derivatives, and typically involves interpolation of the function. The precision and speed of these methods are inversely related and they vary widely dependent on the tolerance. Since the likelihood surface is dataset-specific, this method may require calibration and tuning for a particular dataset.

The approximating nature of the standard approaches outlined above, the need to balance various tradeoffs in their implementation, and a likely need to tweak implementation based on the dataset at hand, makes it difficult to develop these approaches to the point of automation sufficient for use in standard statistical software.

The algorithm proposed in this paper is exact, automatic and requires no tuning or calibration. This makes it an attractive alternative, particularly with statistical software applications in mind.

The PO model is used in this paper to compare via simulations the performance of the three estimation methods for I_{pr} and the proposed exact algorithm. For different sample sizes, the approaches were compared in terms of the number of operations required to achieve a reasonable standardized precision. Naturally, the exact method outperforms any approximating method if an ever better precision is demanded. In our numerical study we focus on practical precisions where approximating methods could nevertheless represent a viable competition to an exact procedure. Numerical efficiency and precision of the computation of I_{pr} is of great importance for variable selection procedures. In an example involving seven variables, backward variable selection using the Wald statistic based on the exact profile information matrix took less than one third of the time of the quadratic approach. We also compared the estimation methods in terms of relative error. Of all approximating methods, the numerical approach has the smallest relative error.

As a result of these studies we believe that the exact method should be the primary choice for NTM.

2 Nonlinear transformation models

NTM are defined as follows (Tsodikov 2002, 2003).

Definition 1 Let $\gamma(x|\beta, z)$ be a parametrically specified distribution function with x -domain of $[0, 1]$. Let $F(t)$ be a nonparametrically specified baseline survival function. A semiparametric regression survival model is called a Nonlinear Transformation Model if, conditional on the covariates z , its survival function G can be represented in the form

$$G(t|\beta, z) = \gamma(F(t)|\beta, z). \quad (2)$$

The function γ is called the NTM-generating function.

Note that $F(t) = \exp(-H(t))$ where $H(t)$ is the baseline cumulative hazard function. With this in mind we can write the hazard function of the model as

$$\lambda(t|\beta, z) = \frac{\gamma'(F(t)|\beta, z)}{\gamma(F(t)|\beta, z)} F(t)h(t), \quad (3)$$

where $h(t) = H'(t)$ is the baseline hazard function.

In Tsodikov (2003) a Quasi-EM (QEM) point estimation algorithm for the NTM was developed and conditions that ensure its convergence were given.

The algorithm solves a functional self-consistency score equation of the form $H = \psi(\beta, H)$ for H , where ψ is a mapping that generalizes a Nelson–Aalen–Breslow estimator for the proportional hazards model so that its denominator depends on H as well as β . Functional iterations

$$H^{(k+1)} = \psi(\beta, H^{(k)}), k=1, 2, \dots \quad (4)$$

are exercised until $\hat{H}$, the fixed-point of ψ , has been approximated, $H^{(k)} \rightarrow \hat{H}$, as $k \rightarrow \infty$, see Tsodikov (2003) for details.

Although any parameterization of γ in terms of β and z is allowed, in the examples we assume that γ is parameterized through a set of parameters/predictors $\theta, \eta, \dots$, where each predictor is further parameterized using generally different sets of regression coefficients $\beta_1, \beta_2, \dots$, so that $\theta = \exp(\beta_1^T z), \eta = \exp(\beta_2^T z), \dots$

2.1 Profile likelihood approach

Let $t_i, i = 1, \dots, n$ be a set of failure times, arranged in ascending order, $t_{n+1} := \infty$. Associated with each t_i is a set of subjects $\mathcal{D}_i$ with covariates $z_{ij}, j \in \mathcal{D}_i$ who fail at t_i , and a set of subjects $\mathcal{C}_i$ with covariates $z_{ij}, j \in \mathcal{C}_i$ who are censored at time $t \in [t_i, t_{i+1})$. The observed event for the subject ij is a triple (t_i, z_{ij}, c_{ij}) , where c is a censoring indicator, $c = 1$ if failure, $c = 0$ if right censored. Let H be the baseline cumulative hazard, with $H(0) = 0$. We assume that $H(t)$ is a step function with jumps at the failure times $t_i, i = 1, \dots, n$. As a step-function, H can be characterized by the vector $h = (h_1, \dots, h_n)$, where $h_i = \Delta H_i$ is the jump of H at t_i . With this notation, under an NT model (2), (3) and non-informative censoring, the likelihood of survival data takes the form

$$\ell = \sum_{i=1}^n D_i \log(h_i) + \sum_{i=1}^n \sum_{j \in \mathcal{C}_i \cup \mathcal{D}_i} \log \vartheta(F_i | \beta, z_{ij}, c_{ij}),$$

where

$$\vartheta(x | \beta, z, c) = x^c \frac{\partial^c \gamma(x | \beta, z)}{\partial x^c},$$

$\partial^0 \gamma / \partial x^0 = \gamma$, D_i is the number of failures associated with t_i and

$$F_i = F(t_i) = \exp \left(- \sum_{l=1}^i h_l \right).$$

The profile likelihood is defined as a supremum of the full likelihood ℓ taken over the nonparametric part of the model

$$\ell_{\text{pr}}(\beta) = \max_h \ell(\beta, h).$$

The MLE of h for a given β will be denoted $\hat{h}(\beta) = (\hat{h}_1, \dots, \hat{h}_n)$, with $\hat{h}_k = \hat{h}_k(\beta)$, then $\ell_{\text{pr}}(\beta) = \ell(\beta, \hat{h}(\beta))$.

Differentiating ℓ with respect to h and setting the score equal to 0 we obtain $\hat{H}(\beta)$ as the solution of the functional self-consistency equation

$$\widehat{h}_m = \frac{D_m}{\sum_{(i,j) \in \mathcal{R}_m} \Theta(F_i | \beta, z_{ij}, c_{ij})}, m=1, \dots, n, \quad (5)$$

where F_i is a function of $h_1, \dots, h_i$,

$$\Theta(x | \beta, z, c) = - \frac{\partial \log \vartheta(x | \beta, z, c)}{\partial x} = c + x \frac{\gamma^{(c+1)}(x | \beta, z, c)}{\gamma^{(c)}(x | \beta, z, c)}, \quad (6)$$

$\gamma^{(a)} = \partial^a \gamma(x) / \partial x^a$, and $\mathcal{R}_m$ is the set of subjects at risk just prior to t_m , $\mathcal{R}_m = \{(i, j) : i \geq m, j \in \mathcal{C}_i \cup \mathcal{D}_i\}$.

2.2 Point estimation

Point estimation proceeds along the lines of the following nested procedure,

- maximize $\ell_{\text{pr}}(\beta)$ by a conventional nonlinear programming method, for example, the Powell method (Press et al., 1994),
- for each β demanded in the above maximization procedure, find $\max_h \ell(\beta, h)$ as the fixed point of (5).

The QEM algorithm makes use of the straightforward recursion to obtain the profile likelihood,

$$h_m^{(k+1)} = \frac{D_m}{\sum_{(i,j) \in \mathcal{R}_m} \Theta\left(\exp\left(-\sum_{l=1}^i h_l^{(k)}\right) | \beta, z_{ij}, c_{ij}\right)}, m=1, \dots, n, \quad (7)$$

where k counts iterations. Note that an increment of k occurs only once all the parameters h_i , $i = 1, \dots, n$ have been updated.

It can be shown that if Θ is nondecreasing in x , each update of H using (7) strictly improves the likelihood, given β . This guarantees convergence of the sequence of likelihood values $\ell(\beta, h^{(k)})$ to $\ell(\beta, \hat{h}(\beta))$ under fairly general conditions (Tsodikov 2003); that is, $\hat{H}(\beta)$ is the fixed point of the recursion given in (7).

It should be noted that the proposed information matrix algorithm is not contingent on using a specific method for point estimation. Yet it builds on the idea of the self-consistency through implicit differentiation of the self-consistency equation.

3 Profile information matrix

The profile information matrix is the observed information matrix derived from the profile likelihood,

$$I_{\text{pr}}(\beta) = - \frac{\partial^2 \ell_{\text{pr}}(\beta)}{\partial \beta \partial \beta^T}.$$

Implicit differentiation of the profile likelihood yields the following expression for the profile information matrix

$$I_{\text{pr}}(\beta) = I_{\beta\beta} + h_{\beta}^T I_{hh} h_{\beta} + h_{\beta}^T I_{h\beta} + I_{h\beta}^T h_{\beta} + \sum_{m=1}^n \ell_{h_m} h_{m,\beta\beta}, \quad (8)$$

where $h = h(\beta)$ is some function of β , and

$$h_{\beta} = \frac{\partial h(\beta)}{\partial \beta}, \quad I_{ab} = -\frac{\partial^2 \ell}{\partial a \partial b^T}, \quad \ell_{h_m} = \frac{\partial \ell(\beta, h)}{\partial h_m}, \quad \text{and} \quad h_{m,\beta\beta} = \frac{\partial^2 h_m(\beta)}{\partial \beta \partial \beta^T},$$

with a and b equal to β or h .

When evaluated at the MLE $\hat{H}(\beta)$, where $\hat{H}$ is a function defined implicitly as the solution of the score equation

$$\ell_h(\beta, h) = 0 \Rightarrow \hat{h} = \hat{h}(\beta), \quad (9)$$

the information matrix simplifies to

$$I_{\text{pr}} = I_{\beta\beta} + I_{h\beta}^T \hat{h}_{\beta}. \quad (10)$$

Indeed, by virtue of the score equation (9),

$$\ell_h(\beta, \hat{h}(\beta)) \equiv 0. \quad (11)$$

Differentiating (11) with respect to β , we also have

$$\frac{d\ell_h(\beta, \hat{h}(\beta))}{d\beta} = I_{h\beta} + I_{hh} \hat{h}_{\beta} \equiv 0, \quad (12)$$

with (10) now following from (8) on substitution of (11) and (12).

It should be noted, however, that unless the score equation (9) is solved for h *exactly*, the short form of the observed profile information matrix (10) is generally not going to be symmetric. Except in the Cox model, there is no closed form solution to $\hat{H}$, and this function is an output of a numerical algorithm such as (7) converging to $\hat{H}$ with some tolerance. To preserve the symmetry of I_{pr} , we prefer to keep some of the theoretically redundant terms in (8) and use the form

$$I_{\text{pr}}(\beta) = I_{\beta\beta} + \hat{h}_{\beta}^T I_{hh} \hat{h}_{\beta} + \hat{h}_{\beta}^T I_{h\beta} + I_{h\beta}^T \hat{h}_{\beta}. \quad (13)$$

Notice that I_{pr} has dimension $d \times d$, $d = \dim(\beta)$. Therefore only a small matrix needs to be inverted in order to get an estimator of the covariance matrix of regression coefficients.

The difficulty in (13) is that since $\hat{H}(\beta)$ is defined implicitly, so is the potentially large Jacobian matrix $\partial \hat{h} / \partial \beta$. Therefore, the Jacobian is generally unavailable in a closed form. The success in the calculation of the profile information matrix is determined by the

existence of an efficient numerical method to compute $\partial\hat{h}/\partial\beta$. Generally, computation of $\partial\hat{h}/\partial\beta$ is as difficult as taking the inverse of the original full model information matrix ($O(n^3)$ operations required), and this derivation defeats the purpose. However, if the functional $\Theta(H, t|\cdot)$ that defines model contributions to the likelihood depends on (H, t) only through $H(t)$, which is the case for the NT models (2), $\partial\hat{h}/\partial\beta$ can be obtained by solving a system of linear equations with a special structure. This specific structure of the linear system can be exploited to derive an efficient numerical solution given in Proposition 1.

We first show how to obtain $I_{\beta\beta}$, $I_{h\beta}$ and I_{hh} .

The H -score of an NT model is,

$$\frac{\partial\ell}{\partial h_k} = \frac{D_k}{h_k} - \sum_{(i,j) \in \mathcal{R}_k} \Theta(F_i|\beta, z_{ij}, c_{ij}).$$

Differentiating the H -score with respect to β we get,

$$-\frac{\partial^2\ell}{\partial h_k \partial \beta_m} = \sum_{(i,j) \in \mathcal{R}_k} \frac{\partial\Theta}{\partial \beta_m}(F_i|\beta, z_{ij}, c_{ij}).$$

Evaluation of derivatives of Θ or γ with respect to β depends on the parameterization of the model's predictor as a function of explanatory variables z , which is model-specific. Once a model is specified, the calculation of $I_{\beta\beta}$ and $I_{h\beta}$ is straightforward.

Since $F_i = \exp(-\sum_{l=1}^i h_l)$, we have

$$\frac{\partial\Theta(F_i|\cdot)}{\partial h_m} = \begin{cases} Q(F_i|\cdot), & m \leq i, \\ 0, & m > i, \end{cases} \quad (14)$$

where

$$Q(x|\cdot, c) = -x \frac{\partial\Theta(x|\cdot, c)}{\partial x} = -(\Theta(x|\cdot, c) - c)(\Theta(x|\cdot, c+1) - \Theta(x|\cdot, c)), \quad (15)$$

and “ $\cdot$ ” stands for “ β, z ”, and Θ is given by (6) extended to $c = 0, 1, 2$. Note that $\partial\Theta(F_i|\cdot)/\partial h_m$ is a constant in m for $m \leq i$ or $m > i$.

From (14) it follows that,

$$-\frac{\partial^2\ell}{\partial h_k \partial h_m} = \sum_{(i,j) \in \mathcal{R}_{\max\{k,m\}}} Q(F_i|\beta, z_{ij}, c_{ij}) + \frac{D_k}{h_k^2} 1_{\{k=m\}},$$

where

$$1_{\{k=m\}} = \begin{cases} 1, & k=m, \\ 0, & k \neq m. \end{cases}$$

From this we get I_{hh} .

Now we turn our attention to the Jacobian $\partial \hat{h} / \partial \beta$. Proposition 1 gives the main result used to efficiently calculate $\partial \hat{h} / \partial \beta$ in the case of NT models. Its proof is given in the Appendix.

Proposition 1 Let D be an $n \times n$ diagonal matrix with diagonal elements $d_i \neq 0$, $i = 1, \dots, n$.

Let $R = (R_{kl})$ be an $n \times n$ matrix, with $R_{kl} = \sum_{i=\max\{k,l\}}^n a_i$, where a_i , $i = 1, \dots, n$ are real numbers. Let b be an n -dimensional vector.

Define the functions $\varphi_k: \mathbb{R} \rightarrow \mathbb{R}$, $k = 1, \dots, n$ recursively as

$$\begin{aligned} \varphi_n(y) &= \frac{b_n}{d_n} - \frac{a_n}{d_n} y, \\ \varphi_k(y) &= \frac{1}{d_k} \left(b_k - \sum_{i=k}^n a_i y + \sum_{l=k+1}^n \sum_{i=k}^{l-1} a_i \varphi_l(y) \right), \quad k = n-1, \dots, 1, \end{aligned}$$

for y in $\mathbb{R}$. Let $\tilde{\varphi}: \mathbb{R} \rightarrow \mathbb{R}$ be the function given by $\tilde{\varphi}(y) = \sum_{k=1}^n \varphi_k(y)$ and let

$$\tilde{y} = \frac{\tilde{\varphi}(0)}{1 + \tilde{\varphi}(0) - \tilde{\varphi}(1)}.$$

Then the solution to the system of equations $(D + R)x = b$ is the n -dimensional vector $x = (\varphi_1(\tilde{y}), \dots, \varphi_n(\tilde{y}))^T$.

We now show that the Jacobian $\partial \hat{h} / \partial \beta$ satisfies a relationship of the form as discussed in Proposition 1. Differentiating the self-consistency equation (5) implicitly, we get that $\hat{H}$ satisfies the relationship

$$\frac{\partial \hat{h}_m}{\partial \beta_k} = - \frac{\hat{h}_m^2}{D_m} \left(\sum_{l=1}^n \sum_{(i,j) \in \mathcal{R}_{\max\{m,l\}}} Q(F_i | \beta, z_{ij}, c_{ij}) \frac{\partial \hat{h}_l}{\partial \beta_k} + \sum_{(i,j) \in \mathcal{R}_m} \frac{\partial \Theta}{\partial \beta_k}(F_i | \beta, z_{ij}, c_{ij}) \right), \quad (16)$$

where Q is the function given in (15).

Let D be the diagonal matrix with elements

$$d_m = \frac{D_m}{(\hat{h}_m)^2}, \quad m = 1, \dots, d.$$

Let $R = (R_{ml})$ with $R_{ml} = \sum_{i=\max\{m,l\}}^n a_i$, where

$$a_i = \sum_{j \in \mathcal{C}_i \cup \mathcal{D}_i} Q(F_i | \beta, z_{ij}, c_{ij}), \quad i = 1, \dots, n$$

and for $k = 1, \dots, d$ let

$$b^{(k)} = \left(- \sum_{(i,j) \in \mathcal{R}_1} \frac{\partial \Theta(F_i | \beta, z_{ij}, c_{ij})}{\partial \beta_k}, \dots, - \sum_{(i,j) \in \mathcal{R}_n} \frac{\partial \Theta(F_i | \beta, z_{ij}, c_{ij})}{\partial \beta_k} \right)^T.$$

It follows from (16) that

$$\frac{\partial \hat{h}}{\partial \beta_k} = -D^{-1} \left(R \frac{\partial \hat{h}}{\partial \beta_k} - b^{(k)} \right).$$

Hence,

$$(R+D) \frac{\partial \hat{h}}{\partial \beta_k} = b^{(k)}.$$

Therefore, for each $k = 1, \dots, n$ the vector $\partial \hat{h} / \partial \beta_k$ can be obtained from Proposition 1. We now have all the components of (13) defined. This completes the exposition of our method.

4 Examples

In the examples we compare the performance of four methods to compute the observed profile information matrix. A brief explanation of the methods and details on how they were implemented in our examples are given below.

1. *Discretized.* The estimation is based on the result of Corollary 3 in Murphy and van der Vaart (2000). Under certain conditions

$$-2 \frac{\log \ell_{\text{pr}}(\hat{\beta} + h_n v_n) - \log \ell_{\text{pr}}(\hat{\beta})}{nh_n^2} \xrightarrow{P} v^T I_{\text{pr}} v, \quad (17)$$

for all sequences $v_n \xrightarrow{P} v \in \mathbb{R}$ and $h_n \xrightarrow{P} 0$ such that $(\sqrt{n} h_n)^{-1} = O_p(1)$.

In order to estimate all the elements of I_{pr} , we chose $v = e_i$, $i = 1, \dots, d$ and $v = e_i + e_j$, $1 \leq i < j \leq d$, where e_i are Euclidean basis vectors.

We set $v_n \equiv v$, and $h_n = 10 / (\sqrt{n} C^k)$ with $C = 1.4$ and k such that $|f(10 / (\sqrt{n} C^k)) - f(10 / (\sqrt{n} C^{k-1}))| < 0.001$, where $f(h)$ is the left-hand side of Eq. (17) considered as a function of h . This procedure was motivated by Dixon et al. (2005) who considered a choice of h_n in the one dimensional ($d = 1$) situation.

2. *Quadratic.* This approach approximates the profile likelihood surface by a quadratic form and derives the estimate of the information matrix from the coefficients of the form fitted to the surface. Specifically, let $\Delta \beta$ be a vector of deviations of the β values sampled in the vicinity of $\hat{\beta}$, and let $\Delta \ell_{\text{pr}}$ be the induced vector of deviations of the profile likelihood from its maximum value, $\ell_{\text{pr}}(\hat{\beta})$. Then, if $\Delta \beta$ is sufficiently small

$$\Delta \ell_{\text{pr}} \approx \frac{1}{2} \Delta \beta^T I_{\text{pr}} \Delta \beta.$$

Fitting the quadratic form $(1/2)\Delta\beta^T A\Delta\beta$ to points $(\Delta\beta, \Delta\ell_{\text{pr}})$ by least squares produces an estimate, $\hat{A}$, of the profile information matrix I_{pr} .

In our implementation of this method we limit the domain to points that are not rejected at 0.05 significance level by the LR test (applied informally and disregarding the multi-comparison issue). In other words, points β are included if $-2\{\ell_{\text{pr}}(\hat{\beta}) - \ell_{\text{pr}}(\beta)\} \leq C_{d,0.05}$, where $C_{d,0.05}$ is the 0.05th upper tail percentile of the χ^2 distribution with $d = \dim(\beta)$ degrees of freedom. Since the validity of the quadratic approximation is itself a prerequisite for the validity of the likelihood ratio statistic, this choice is far from perfect. Yet this procedure would ensure a desired property of the domain shrinking with sample size, and we know of no better alternative.

3. *Numerical.* The calculation of the observed profile information matrix is carried on using Ridder's numerical differentiation of the profile likelihood function, see Press et al. (1994).

Let $f: \mathbb{R} \rightarrow \mathbb{R}$ be a differentiable function. By definition, the derivative of f is the limit as $h \rightarrow 0$ of the incremental quotient

$$q(h) = \frac{f(x+h) - f(x)}{h}.$$

The basic idea of Ridder's method is to calculate $q(h)$ for several values of h , and then extrapolate the result to the limit $h = 0$. In the case of a function with domain on $\mathbb{R}$, the vector of first derivatives is obtained by applying the algorithm on each coordinate at a time and leaving the other coordinates fixed.

Numerical experimentation showed that this approach gives the second derivatives of $\ell_{\text{pr}}(\beta)$ with very high precision, albeit at a greater computational cost than other methods.

4. *Exact.* This is the method developed in Sect. 3 of this paper for computation of the exact observed profile information matrix for NTM.

PO model will be used as a basis for all our comparisons. The validity of NPMLE and the profile likelihood for this model has been demonstrated elsewhere.

4.1 The PO model

Given covariates z , the survival function $G(t|\beta, z)$ of a PO model can be written in the form,

$$G(t|\beta, z) = G(t|\theta(\beta, z)) = \frac{\theta(\beta, z)}{\theta(\beta, z) + H(t)}, \quad (18)$$

where H is some nonparametrically specified baseline cumulative hazard function, and θ is a predictor. Since $H = -\log F$, the NTM-generating function of the PO model is

$$\gamma(x|\cdot) = \frac{\theta(\cdot)}{\theta(\cdot) - \log x}.$$

A characteristic feature of the PO model is that for any two values, θ_1, θ_2 , of the predictor, the odds ratio

$$\frac{\text{Odds}(G(t|\theta_1))}{\text{Odds}(G(t|\theta_2))} = \frac{\theta_1}{\theta_2}$$

is constant in t .

It follows that

$$\Theta(x|\cdot, c) = \frac{c+1}{\theta(\cdot) - \log x}.$$

We consider an exponential parameterization of the predictor $\theta(\beta, z) = \exp(\beta^T z)$. With this parameterization,

$$\frac{\partial \theta}{\partial \beta} = \theta z, \quad \frac{\partial^2 \theta}{\partial \beta \partial \beta^T} = \theta z z^T.$$

The following derivatives of Θ are necessary to specify the algorithm of Sect.3,

$$\frac{\partial \Theta}{\partial \beta} = \frac{\partial \Theta}{\partial \theta} \theta z, \quad \frac{\partial^2 \Theta}{\partial \beta \partial \beta^T} = \left(\frac{\partial^2 \Theta}{\partial \theta^2} \theta^2 + \frac{\partial \Theta}{\partial \theta} \theta \right) z z^T,$$

where

$$\frac{\partial \Theta(x|\cdot, c)}{\partial \theta} = - \frac{c+1}{(\theta(\cdot) - \log x)^2}, \quad \text{and} \quad \frac{\partial^2 \Theta(x|\cdot, c)}{\partial \theta^2} = \frac{2(c+1)}{(\theta(\cdot) - \log x)^3}.$$

4.2 Real data

As an example, we use data from the National Cancer Institutes Surveillance Epidemiology and End Results (SEER) program. Using the publicly available SEER database, 11,621 cases of primary prostate cancer diagnosed in the state of Utah between 1988 and 1999 were identified. The following selection criteria were applied to a total of 19,819 Utah-cases registered in the database: valid positive survival time, valid stage of the disease, age ≥ 18 years. Prostate cancer specific survival was analyzed by stage of the disease (localized/regional vs. distant). For the definition of stages as well as for other details of the data we refer the reader to SEER documentation <http://www.seer.cancer.gov/>.

The data analysis presented in this paper is a continuation of the one given in Tsodikov (2003). Two groups of patients representing stage at diagnosis of the disease are considered, hence the predictor in the PO model has a single parameter β . The log odds ratio β measures the disadvantage of being in the distant stage relative to local/regional stage. The QEM algorithm was applied to fit the PO model to the data. The maximum likelihood estimate of β was $\hat{\beta} = -3.251$. Confidence intervals for β were obtained using the Wald statistic based on the profile information matrix. The confidence interval based on the quadratic approximation of the profile information matrix was $(-3.416, -3.086)$ and the one obtained through the exact profile information matrix was $(-3.415, -3.086)$. Excellent concordance of the two confidence intervals is due to the large sample size and the small dimension of the regression parameter, a situation when approximating methods tend to be accurate.

In the case of a single parameter, the observed profile information matrix is a scalar. The estimates of the observed profile information matrix were 142.1011, 141.2158 and 141.7424 for the Discretized, Quadratic and Numerical approaches respectively and the Exact value was 141.7423. Although the values were quite similar it is clear that the discretized and quadratic approaches depart from the true value.

4.3 Simulations

4.3.1 Simulations setup—We simulated age at diagnosis for an adult-onset disease using a PO model. The example in its baseline survival is loosely based on incidence of prostate cancer. The baseline survival function was assumed to follow a Weibull distribution with the median of 38 years and shape of 1.8, the risk starting at the age of 18 (a fixed number of 18 years was added to survival time and censoring). With these parameters incidence before the age of 40 is negligible. An independent censoring mechanism was assumed. Censoring times were generated using Weibull distribution with a median of 46 years and shape parameter of 4. Observations in excess of 105 years were type-I censored at 105. Two covariates were introduced, one categorical with 3 levels, and one continuous (a risk factor) with a range between -1 and 1 . Values for both covariates were generated independently. The continuous covariate followed a uniform distribution. The discrete distribution for the categorical covariate assumed the following probabilities, 0.7 (level 1), 0.5 (level 2), and 0.1 (level 3). The following covariate effects were assumed. An effect of the log odds ratio of 2 was assumed for a unit change in the continuous factor. The categorical covariate was assumed to have a progressing effect on the risk of the disease. The log odds ratios comparing level 2 and level 3 to the baseline level 1 were 1.5 and 2.5, respectively.

4.3.2 Speed—To assess the speed of performance of the four methods we calculated the number of operations required to compute the exact information matrix and its approximations. Evaluation of Θ , γ , their analytically specified derivatives or similar comparable procedures were counted as one operation. Figure 1 shows the number of operations by sample size and method. In order to make the performance results comparable, the precision of estimation algorithms was calibrated on an ad-hoc basis so that the relative error of the three methods (Discretized, Quadratic, Numerical) was approximately the same (0.02). For any method A , the relative error was defined as $\|I_{pr}(A) - I_{pr}(\text{Exact})\| / \|I_{pr}(\text{Exact})\|$, where $I_{pr}(A)$ is an estimate of the observed profile information matrix computed using method A , and the norm is defined as the sum of absolute values of all elements of the matrix. Regardless of the sample size, the exact calculation outperformed the approximate methods. Inference based on the discretized second derivative required between 10 and 30 times as many operations as the exact calculation. The quadratic approach required between 60 and 200 times as many operations as the calculation of the exact I_{pr} matrix. The numerical method was computationally very costly requiring between 600 and 7,000 as many operations as the exact approach. However, the numerical approach behaved better than the other two methods in terms of relative error as shown in Sect.4.3.3.

4.3.3 Precision—A sample of size 500 was used to find the smallest possible relative error of the method when adjusting the different parameters involved on an ad-hoc basis. The best relative error achieved by the Discretized method was 0.01 and 8.13×10^5 operations were required. This number was 0.013 for the quadratic approach with 5.32×10^6 operations required, while the numerical approach achieved a relative error of 8×10^{-7} and required 3.87×10^8 operations. This example shows that the numerical approach is perhaps the only one of the approximating methods that can compete with the exact procedure in terms of precision required in real-life analysis. Its high computational cost, however, makes it a poor choice for variable selection and other procedures requiring repeated evaluations of I_{pr} .

4.3.4 Statistical properties—Three sets of experiments were performed with samples of size 100, 500, and 1,000. For each sample size, 1,000 simulated samples were generated. The covariance matrices based on I_{pr} were computed for each sample using the four approaches discussed. The mean and standard deviation of each of the entries of the estimators of covariance matrices under study were estimated from the 1,000 replicates. In addition, point semiparametric MLE estimates of the three parameters entering the profile likelihood (log odds ratios for the continuous factor and level 2 vs. 1 and 3 vs. 1 contrasts) were used to compute the empirical covariance matrix based on 1,000 replicates. A comparison of the estimated means of the entries of the covariance matrices calculated using exact and numerical approaches with the empirical ones were used to evaluate how well these methods estimate the true finite sample variance-covariance. The results are shown in Table 1. Two factors contributed to the distance between the exact and numerical approaches and the empirical one: the finite-sample bias of covariance estimates based on I_{pr} , and the bias in the estimate of I_{pr} by an approximating method (this latter bias does not pertain to the exact method). The following basic conclusions are evident from the Table 1.

1. All methods are much better at estimating variances (left half of the table) than covariances (right half of the table);
2. The precision of estimation of covariance improves rapidly with sample size;
3. Under all sample sizes the numerical approach showed excellent concordance with the exact method. This is in agreement with our earlier observation that of all approximating methods, the numerical approach is most precise albeit computationally costly. The quadratic method was the least precise in this analysis;
4. For reasonably large sample sizes all methods are in good agreement with the empirical estimate.

5 Discussion

In this paper we have proposed a method to compute the profile information matrix based on implicit differentiation of the self-consistency equation. Computationally the method outperformed all existing approaches to the best of our knowledge. An attractive property of the procedure is that it is exact contingent upon point estimates. Even though exact point estimates are hardly ever available, the precision of variance-covariance estimation is improved as the method does not add any error to the one associated with imprecision of point estimates. Numerically efficient and stable procedures for point estimates have been developed earlier and provide a good complement to this methodology. We recommend the Exact method as a preferred choice with NTM.

Since derivatives of the profile likelihood are defined implicitly, applying the Newton–Raphson method to the profile likelihood for point estimation is a challenge. The Newton–Raphson method typically requires exact inverse Hessian matrix and is not guaranteed to converge if this matrix is approximated. The results of this paper can be used to provide an exact inverse Hessian matrix at any point in the parameter space and thus enable the Newton–Raphson method for use with the profile likelihood.

Acknowledgments

This research is supported by National Cancer Institute grant U01 CA97414, and Department of Defence grant DAMD17-03-1-0034.

Appendix

6.1 Proof of Proposition 1

The equation $(D + R)x = b$ implies

$$d_k x_k + \sum_{l=1}^n R_{kl} x_l = b_k, k=1, \dots, n. \quad (19)$$

Since $R_{kl} = \sum_{i=\max\{k,l\}}^n a_i$, it follows from (19) that for $k = 1, \dots, n$,

$$\begin{aligned} b_k &= d_k x_k + \sum_{i=k}^n a_i \sum_{l=1}^k x_l + \sum_{l=k+1}^n \sum_{i=l}^n a_i x_l \\ &= d_k x_k + \sum_{i=k}^n a_i \sum_{l=1}^k x_l + \sum_{i=k+1}^n \sum_{l=k+1}^i a_i x_l \\ &= d_k x_k + \sum_{i=k}^n a_i \left(\sum_{l=1}^n x_l - \sum_{l=i+1}^n x_l \right). \end{aligned}$$

The second equality above is a consequence of a change of summation order.

Hence, solving the system of equations $(D + R)x = b$ is equivalent to solving the system

$$\begin{aligned} x_k &= \frac{1}{d_k} \left(b_k - \sum_{i=k}^n a_i y + \sum_{l=k+1}^n \sum_{i=k}^{l-1} a_i x_l \right), k=1, \dots, n \\ y &= \sum_{l=1}^n x_l. \end{aligned}$$

Notice that $\{x_k\}$ are in fact functions of y obtained recursively, $x_k = \varphi_k(y)$, and

$y = \sum_{k=1}^n \varphi_k(y) = \tilde{\varphi}(y)$. The solution of this system of equations is the vector $(\varphi_1(\tilde{y}), \dots, \varphi_n(\tilde{y}))^T$ where $\tilde{y}$ satisfies $\varphi(\tilde{y}) = \tilde{y}$. Since $\varphi_k, k = 1, \dots, n$ are linear functions of y , so is φ . Hence, $\varphi(y) = ay + b$ with

$$a = \tilde{\varphi}(1) - 1 - \tilde{\varphi}(0) \quad \text{and} \quad b = \tilde{\varphi}(0).$$

Therefore,

$$\tilde{y} = \frac{\tilde{\varphi}(0)}{1 + \tilde{\varphi}(0) - \tilde{\varphi}(1)}.$$

References

- Dixon JR, Kosorok MR, Lee BL. Functional inference in semiparametric models using the piggyback bootstrap. *Ann Inst Statist Math* 2005;57:255–277.
- Kosorok MR, Lee BL, Fine JP. Robust inference for univariate proportional hazards frailty regression models. *Ann Statist* 2004;32:1448–1491.

- Lange K, Hunter DR, Yang I. Optimization transfer using surrogate objective functions (with discussion). *J Comput Graph Statist* 2000;9:1–59.
- Murphy SA. Consistency in a proportional hazards model incorporating a random effect. *Ann Statist* 1994;22(2):712–731.
- Murphy SA. Asymptotic theory for the frailty model. *Ann Statist* 1995;23(1):182–198.
- Murphy SA, van der Vaart AW. Semiparametric likelihood ratio inference. *Ann Statist* 1997;25:1471–1509.
- Murphy SA, van der Vaart AW. On profile likelihood. *J Am Statist Assoc* 2000;95:449–485.
- Murphy SA, Rossini AJ, van der Vaart AW. Maximum likelihood estimation in the proportional odds model. *J Am Statist Assoc* 1997;92(439):968–976.
- Nielsen GG, Gill RD, Andersen PK, Sorensen TI. A counting process approach to maximum likelihood estimation in frailty models. *Scand J Statist* 1992;19:25–43.
- Parner E. Asymptotic theory for the correlated Gamma-frailty model. *Ann Statist* 1998;26:183–214.
- Press, WH.; Flannery, BP.; Teukolsky, SA.; Vetterling, WT. The art of scientific computing. New York: Cambridge University Press; 1994. Numerical recipes in Pascal.
- Severini, TA. Likelihood methods in statistics. New York: Oxford University Press; 2000.
- Tsodikov A. Semiparametric models of long- and short-term survival: an application to the analysis of breast cancer survival in Utah by age and stage. *Statist Med* 2002;21:895–920.
- Tsodikov A. Semiparametric models: a generalized self-consistency approach. *J R Statist Soc, Ser B* 2003;65:759–774.
- Tsodikov A, Ibrahim JG, Yakovlev AY. Estimating cure rates from survival data: an alternative to two-component mixture models. *J Am Statist Assoc* 2003;98:1063–1078.
- van der Vaart, AW. Cambridge series in statistical and probabilistic mathematics. Cambridge UK: Cambridge University Press; 1998. Asymptotic statistics.

Fig. 1. Operations by sample size characteristics of four methods of computation of the observed profile information matrix. The Exact method developed in this paper shows the highest numerical efficiency

Table 1

Empirical covariance matrix based on 1,000 replicates and mean covariance matrix and sd based on 1,000 replicates for the four methods studied

Method	Mean (sd)					
	σ_{11}	σ_{22}	σ_{33}	σ_{12}	σ_{13}	σ_{23}
Sample size: 100						
empirical	0.405	0.597	0.207	0.102	0.044	0.045
exact	0.426 (0.170)	1.020 (0.347)	0.192 (0.041)	0.080 (0.022)	0.045 (0.035)	0.062 (0.046)
num	0.426 (0.170)	1.019 (0.346)	0.192 (0.041)	0.080 (0.022)	0.044 (0.035)	0.062 (0.046)
quad	1.009 (21.986)	0.103 (21.374)	1.139 (31.762)	- 0.087 (9.674)	- 0.188 (5.822)	- 0.364 (7.674)
Mur	0.464 (0.203)	1.155 (0.411)	0.206 (0.046)	0.077 (0.025)	0.046 (0.042)	0.067 (0.054)
Sample size: 500						
empirical	0.082	0.259	0.037	0.014	0.011	0.011
exact	0.074 (0.010)	0.245 (0.117)	0.036 (0.003)	0.016 (0.002)	0.009 (0.003)	0.013 (0.004)
num	0.074 (0.010)	0.245 (0.117)	0.036 (0.003)	0.016 (0.002)	0.009 (0.003)	0.013 (0.004)
quad	0.074 (0.034)	0.347 (0.490)	0.040 (0.039)	- 0.003 (0.067)	0.002 (0.022)	0.032 (0.053)
Mur	0.076 (0.011)	0.260 (0.128)	0.036 (0.003)	0.014 (0.002)	0.008 (0.003)	0.013 (0.004)
Sample size: 1,000						
empirical	0.037	0.116	0.018	0.006	0.005	0.006
exact	0.037 (0.003)	0.112 (0.028)	0.018 (0.001)	0.008 (0.001)	0.004 (0.001)	0.007 (0.001)
num	0.037 (0.003)	0.112 (0.028)	0.018 (0.001)	0.008 (0.001)	0.004 (0.001)	0.007 (0.001)
quad	0.038 (0.066)	0.114 (0.052)	0.018 (0.020)	0.002 (0.044)	0.003 (0.036)	0.008 (0.024)
Mur	0.037 (0.004)	0.116 (0.030)	0.018 (0.001)	0.007 (0.001)	0.004 (0.001)	0.006 (0.001)