Published in final edited form as:
J R Stat Soc Series B Stat Methodol. 2003 August 1; 65(3): 759–774. doi:10.1111/1467-9868.00414.

Semiparametric models: a generalized self-consistency approach

A. Tsodikov
University of Utah, Salt Lake City, USA

Summary

In semiparametric models, the dimension *d* of the maximum likelihood problem is potentially unlimited. Conventional estimation methods generally behave like $O(d^3)$. A new $O(d)$ estimation procedure is proposed for a large class of semiparametric models. Potentially unlimited dimension is handled in a numerically efficient way through a Nelson–Aalen-like estimator. Discussion of the new method is put in the context of recently developed minorization–maximization algorithms based on surrogate objective functions. The procedure for semiparametric models is used to demonstrate three methods to construct a surrogate objective function: using the difference of two concave functions, the EM way and the new quasi-EM (QEM) approach. The QEM approach is based on a generalization of the EM-like construction of the surrogate objective function so it does not depend on the missing data representation of the model. Like the EM algorithm, the QEM method has a dual interpretation, a result of merging the idea of surrogate maximization with the idea of imputation and self-consistency. The new approach is compared with other possible approaches by using simulations and analysis of real data. The proportional odds model is used as an example throughout the paper.

Keywords

EM algorithm; Frailty; Nonparametric maximum likelihood estimation; Profile likelihood; Semiparametric models

1. Introduction

Potentially unlimited dimension has been the most critical deterrent to the use of maximum likelihood estimation (MLE) in semiparametric regression models. In survival analysis, methods based on the partial likelihood (Cox, 1972) are specific to the proportional hazards (PH) model and do not extend to other models. Straightforward Newton-type methods of maximizing the likelihood for the full model generally require $O(d^3)$ operations to solve the system of score equations, where *d* is the number of model parameters. The principal part of the set of *d* parameters in a semiparametric model is used to specify a stepwise function *H* which approaches a continuous 'true' *H* in probability, as $d \to \infty$. Although theoretically almost any likelihood can be maximized by a Newton-type method, its high complexity makes the problem computationally difficult for large *d*. The development of general, stable and numerically efficient algorithms for semiparametric MLE has been a long-standing problem (Fleming and Lin, 2000). Such algorithms are the subject of this paper. The argument goes as follows. The bottle-neck of a maximization algorithm for a semiparametric

Address for correspondence: A. Tsodikov, Division of Biostatistics, Huntsman Cancer Institute, University of Utah, 2000 Circle of Hope, Salt Lake City, UT 84112-5550, USA. atsodiko@hci.utah.edu.

likelihood is the estimation of H. Let l be the log-likelihood of a semiparametric model, treated as a functional of H. Consider a class of continuous semiparametric models with the log-likelihood of the form (informally)

$$l=\sum_t D_t \log\{\mathrm{d}H(t)\}+\sum_t \log\{\vartheta(H,t|z)\}, \tag{1}$$

where D_t is the number of exact observations at t (failures), z is a vector of covariates and $\vartheta > 0$ is some functional of H. The basic assumption that contributes to equation (1) is that the probability of failure in $[t, t + \mathrm{d}t]$ is proportional to $\mathrm{d}H(t)$, which is differentiability. To obtain an estimator for H, we differentiate l with respect to the set of $\{\mathrm{d}H(\tau)\}$. Informally, we arrive at the so-called self-consistency equation

$$\mathrm{d}H(\tau)=D_\tau/\sum_t \Theta(H,t|z), \tag{2}$$

where Θ is a functional representing a negative ‘derivative’ of $\log(\vartheta)$. Since both sides of equation (2) depend on H, an iterative procedure is required to make the equation self-consistent,

$$\mathrm{d}H^{(k+1)}(\tau)=D_\tau/\sum_t \Theta(H^{(k)},t|z), \tag{3}$$

where k counts iterations. Iterative updating of H by using equation (3) is the basic idea behind the algorithm. As we shall see, the above procedure is intimately linked to the EM algorithm as used to fit certain PH frailty models in survival analysis (Oakes, 1989; Klein, 1992; Nielsen *et al.*, 1992). The EM algorithm handles H in an $O(d)$ way through the use of the Nelson–Aalen–Breslow estimator (Andersen *et al.*, 1993) for the cumulative hazard H. This is made possible as the M-step reduces to the PH model. However, a large amount of analytic work would be required to specify an estimation procedure for a new non-PH model. Expectation at the E-step may prove to be inaccessible in a closed form, and Monte Carlo extensions of the EM approach are much less computationally attractive. Recently, an optimization transfer approach (Lange *et al.*, 2000) was proposed that allows us to construct EM-like procedures without the use of missing data. For a target function $l(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^n$, the minorization–maximization (MM) algorithm (Lange *et al.*, 2000) proceeds by construction of the so-called surrogate objective function $Q(\mathbf{x}|\mathbf{y})$ such that $Q(\mathbf{y}|\mathbf{y}) = l(\mathbf{y})$, and $Q(\mathbf{x}|\mathbf{y}) \leq l(\mathbf{y})$, for any $\mathbf{x}$, to ensure monotonicity of the procedure. Maximization of the target function l proceeds iteratively as

$$\mathbf{x}^{(k+1)}=\arg\max_{\mathbf{x}}\{Q(\mathbf{x}|\mathbf{x}^{(k)})\}. \tag{4}$$

The MM algorithm converges in l and in $\mathbf{x}$ under fairly general conditions (Lange *et al.*, 2000). In the likelihood interpretation, the EM algorithm is a particular case of the MM algorithm. Unfortunately, there is no automatic way to construct Q. The procedure (3) interpreted as an MM algorithm is used to highlight three methods to construct a surrogate objective function: using the difference of two concave functions, the EM way and a new quasi-EM (QEM) approach. These methods link the EM algorithm for frailty models and its modifications with the MM algorithms. In the QEM approach, ‘E’ in the EM is replaced by the quasi-expectation operator QE, which is not based on the concept of a random variable.

The result is the so-called QEM algorithm, which presents us with a recipe of generalizing an EM procedure into a 'distribution-free' one, representing a particular MM algorithm.

2. Profile likelihood approach

The problem of nonparametric maximum likelihood estimation (NPMLE) with the semiparametric model is to find estimates of regression coefficients **β** and an NPMLE estimate of *H* such that they deliver the maximum of a suitably defined likelihood function *l* = *l*(**β**, *H*). In this paper we use a profile likelihood approach to maximize *l*. The profile likelihood is defined as a supremum of the full likelihood taken over the nonparametric part of the model

$$l_{\mathrm{pr}}(\beta) = \max_{H} \{l(\beta, H)\}. \tag{5}$$

Assuming that we can find the global maximum of *l* with respect to *H*, given **β**, we may write the profile likelihood as an implicit function of **β**

$$l_{\mathrm{pr}}(\beta) = l\{\beta, H(\beta)\}, \tag{6}$$

where *H*(**β**) is the solution of equation (5). Our algorithms will be designed following a straightforward nested procedure:

a. maximize $l_{\mathrm{pr}}(\boldsymbol{\beta})$ by a conventional non-linear programming method (e.g. a directions set method);

b. for any **β** as demanded in the above maximization procedure, solve problem (5).

As the number of parameters of a semiparametric model is potentially unlimited, obtaining the inverse of the full information matrix can be computationally prohibitive. Therefore, we use the profile information matrix

$$\mathbf{I}^{\mathrm{P}}_{\beta,\beta} = -\frac{\partial^2 l_{\mathrm{pr}}\{\beta l_{\mathrm{pr}}(\beta)\}}{\partial\beta\partial\beta^T} \tag{7}$$

to derive a standard error estimator for **β**. In this paper we adopt a pragmatic numerical approach. In the course of maximization of the profile likelihood with respect to **β**, a dense sample of the profile likelihood surface is generated near a stationary point. The curvature of the profile likelihood surface at the stationary point can be estimated by fitting a quadratic function to some domain around the point by using least squares. For example, the domain can be limited to points that cannot be rejected by using the likelihood ratio test (applied informally). Alternatively, a more sophisticated approach can be used based on implicit differentiation of l_{pr}.

The rest of the paper will be devoted to constructing efficient NPMLE methods for obtaining l_{pr}, i.e. for maximizing *l* with respect to *H*, given **β**, as this is the crux of the matter.

Practically, inference based on the profile likelihood is similar to that based on the partial likelihood for the PH model, which is quite convenient. Theoretically, however, inference based on the profile likelihood is not straightforward, as the usual theory of MLE does not apply to unlimited dimension. Important results have been obtained regarding a theoretical justification for the NPMLE method and the profile likelihood for semiparametric models

(Murphy, 2000; van der Vaart, 1998; Murphy and van der Vaart, 1997). It was shown that profile likelihoods with nuisance parameters estimated out behave like ordinary likelihoods under some conditions. In particular, these results apply to the PH model, the proportional odds (PO) model (Murphy, 2000; Murphy *et al.*, 1997) and the PH frailty model (Murphy, 1994, 1995), and presumably to most other models.

Let t_i, $i = 1,\ldots,n$, be a set of times, arranged in increasing order, and define $t_{n+1} \coloneqq \infty$. Associated with each t_i is a set of individuals $\mathscr{D}_i$ with time-independent covariates $\mathbf{z}_{ij}$, $j \in \mathscr{D}_i$, who fail at t_i, and a similar set of individuals $\mathscr{C}_i$ with covariates $\mathbf{z}_{ij}$, $j \in \mathscr{C}_i$, who are censored at t_i. The observed event $\mathscr{E}_{ij}$ for the subject ij is a triple $(t_i, \mathbf{z}_{ij}, c_{ij})$, where c is a censoring indicator: $c = 1$ if failure; $c = 0$ if right censored. For any function $A(t)$, let $A_i = A(t_i)$, $\Delta A_i = |A(t_i) - A(t_i - 0)|$. A stepwise function H can be characterized by two vectors $\Delta\mathbf{H} = (\Delta H_1,\ldots,\Delta H_n)^T$ and $\mathbf{t} = (t_1,\ldots,t_n)^T$. With this notation, the likelihood of survival data under non-informative censoring takes the form

$$l=\sum_{i=1}^{n} D_i \log(\Delta H_i)+\sum_{i=1}^{n}\sum_{j\in\mathscr{C}_i\cup\mathscr{D}_i}\log\{\vartheta(\Delta\mathbf{H}, t_i|\beta, \mathbf{z}_{ij}, c_{ij})\}, \tag{8}$$

where D_i is the number of failures that are associated with t_i, and the function ϑ will be specified later for the class of non-linear transformation models (NTMs).

3. EM algorithm for a semiparametric model

For example, consider a PO model for the survival function G, given covariates $\mathbf{z}$,

$$G(t|\beta, \mathbf{z})=G\{t|\theta(\beta, \mathbf{z})\}=\frac{\theta(\beta, \mathbf{z})}{\theta(\beta, \mathbf{z})+H(t)}, \tag{9}$$

where θ is a predictor and H is some nonparametrically specified base-line cumulative hazard. The model is named after the PO property that for any two values of the predictor, θ_1 and θ_2, with corresponding survival functions $G_i(t) = G(t|\theta_i)$, $i = 1, 2$, the odds ratio

$$\frac{\text{odds}\{G_1(t)\}}{\text{odds}\{G_2(t)\}}=\frac{\theta_1}{\theta_2}$$

is a constant in t, where odds$(a) = a/(1 - a)$.

This paper was inspired by the idea of representing a semiparametric model as a mixture (frailty) model, and to use the EM algorithm to fit it. With this idea in mind, consider a PH mixture model

$$G(t|\beta, \mathbf{z})=E\{F(t)^{U(\beta,\mathbf{z})}|\mathbf{z}\}, \tag{10}$$

where $F = \exp(-H)$ is the base-line survival function corresponding to H, and $U = U(\boldsymbol{\beta}, \mathbf{z})$ is used to indicate that the distribution of random variable U depends on covariates and regression coefficients. This model can be considered a compact expression for a family of so-called PH frailty models, or PH models with random effects considered by Hougaard (1984), Klein (1992), Nielsen *et al.* (1992), Wassel and Moeschberger (1993), Clayton and

Cuzick (1985) and many others, for different distributions of U, possibly dependent on covariates.

To construct the EM algorithm for a particular model (PO in the example), we represent it as a PH mixture model (inverse transform), and then follow the usual logic of the EM algorithm construction for frailty models, as for example in Nielsen *et al.* (1992).

3.1. Inverse transform

We note that $\mathscr{L}(s|\cdot) = E[\exp\{-s\,U(\cdot)\}]$ is the Laplace transform of $U(\cdot)$, and that for the PH mixture model (10)

$$G(t|\cdot)=\mathscr{L}\{H(t)|\cdot\}=\mathscr{L}[-\log\{F(t)\}|\cdot].$$

From the latter equation and equation (9), we conclude that U for the PO model represents exponential regression, as $\mathscr{L}(s|\cdot) = \theta(\cdot)/\{\theta(\cdot) + s\}$ is the Laplace transform of an exponential distribution with mean θ^{-1}.

3.2. Complete-data likelihood

With the PH mixture model (10), pretend that U is known for each subject ij, continuing the notation of Section 2. The complete-data likelihood under non-informative right censoring corresponds to the PH model with predictors U_{ij}

$$l_{\mathrm{cd}}=\sum_{i=1}^{n}\left\{D_i\log(\Delta H_i)-\sum_{j\in\mathscr{C}_i\cup\mathscr{D}_i}U_{ij}H_i\right\}. \qquad (11)$$

3.3. E-step

Since the complete-data likelihood (11) is linear in missing data U_{ij}, the E-step reduces to imputation of each U by the corresponding $\hat{U}$, the conditional expectation of U, given the observed event. Using the exponential distribution of U with mean θ^{-1}, after a little algebra, we obtain

$$\widehat{U}=\frac{\int uF^{u}(uh)^{c}\theta\ \exp(-\theta u)\,\mathrm{d}u}{\int F^{u}(uh)^{c}\theta\ \exp(-\theta u)\,\mathrm{d}u}=\frac{\Gamma(c+2)\theta/(\theta+H)^{c+2}}{\Gamma(c+1)\theta/(\theta+H)^{c+1}}=\frac{c+1}{\theta(\beta,\mathbf{z})+H(t)}, \qquad (12)$$

where h is the hazard function corresponding to H. A similar derivation of $\hat{U}$ for a gamma frailty model can be found, for example, in Parner (1998).

3.4. M-step

Maximization of the complete-data likelihood (11) with respect to H, and with U_{ij} imputed by $\hat{U}_{ij}$, results in the Nelson–Aalen estimator

$$\Delta H_m=D_m/\sum_{ij\in\mathscr{R}_m}\widehat{U}_{ij}, \qquad m=1,\ldots,n,$$

where $\mathscr{R}_m = \{ij : j \in \mathscr{D}_i \cup \mathscr{C}_i, i \geq m\}$ is the set of subjects at risk just before t_m.

3.5. EM procedure for the proportional odds model

Finally, for the PO model we have the iterative EM procedure

$$\Delta H_m^{(k+1)}=D_m\left\{\sum_{ij\in\mathcal{R}_m}\frac{c_{ij}+1}{\theta(\beta,\mathbf{z}_{ij})+H_i^{(k)}}\right\}^{-1},\qquad m=1,\ldots,n, \tag{13}$$

where k counts iterations.

3.6. Alternative derivation of procedure (13)

It is intriguing that we can formally derive procedure (13) as an immediate corollary of the argument presented in Section 1. Indeed, using equation (9), we write the likelihood for the PO model as

$$l=\sum_{i=1}^{n}D_i\ \log(\Delta H_i)+\sum_{j\in\mathcal{C}_i\cup\mathcal{D}_i}\log\left[\frac{\theta(\beta,\mathbf{z}_{ij})}{\{\theta(\beta,\mathbf{z}_{ij})+H_i\}^{c_{ij}+1}}\right], \tag{14}$$

On differentiating equation (14) with respect to ΔH_m, and assigning the iteration index k as in equations (1)–(3), we obtain expression (13).

This observation deserves discussion. The EM derivation presented above for the PO model is model specific and its feasibility depends on the success and simplicity of the inverse Laplace transformand the integrals that are evaluated at the E-step (12). The PH mixture representation of a semiparametric model may not exist, in which case the EM derivation ultimately fails. Necessary and sufficient conditions for this representation to exist are a corollary of the Bernstein theorem (Feller, 1971): the survival function must be a completely monotonic function of H. A function $\psi(H)$ is called completely monotonic if all its derivatives $\psi^{(i)}$ exist, $i = 1, 2,\ldots,$ and $(-1)^i\ \psi^{(i)}(H) \geq 0$, $H > 0$. In particular, the survival function (10) of the PH mixture model is an infinitely differentiable function of F. The alternative derivation of procedure (13) bypasses all the above-mentioned difficulty and formally works for any model in a straightforward and simple fashion. This raises a series of questions. Does the procedure of Section 1 work for any model? What is its relationship to the EM algorithm? Does it inherit the monotonicity, stability and convergence of the EM algorithm?

A clue to generalizing the EM algorithm described above is the observation that the derivation of the E-step (12) does not require knowledge of the distribution of U. Indeed, denote by $\gamma(x)$ the moment-generating function of U (other arguments are omitted), so that $\gamma(x) = E(x^U) = \mathcal{L}\{-\log(x)\}$. Observe that the first equation in expression (12) can be written as

$$\widehat{U}=\frac{E(F^U U^{c+1})}{E(F^U U^c)}=c+F\frac{\gamma^{(c+1)}(F)}{\gamma^{(c)}(F)}, \tag{15}$$

where $\gamma^{(c)}$ denotes the derivative of order c; $\gamma^{(0)} \coloneqq \gamma$, $c = 0, 1$. Expression (15) represents a variation on the topic of the derivation of moments from the transform of a distribution. The consequence of equation (15) is a straightforward and general specification of the *E*-step for any mixture model formulated in terms of the moment-generating function. In fact, it is even more general as will be shown in what follows. To elaborate further on the issues raised above, we need to make the few theoretical observations considered in the next section.

4. General concepts

4.1. Construction using the difference of two concave functions

For studying procedure (3), the following MM construction (Lange *et al.*, 2000) is useful. Let $l(\mathbf{x}) = B(\mathbf{x}) - A(\mathbf{x})$, where A and B are differentiable concave functions. The iterative maximization procedure,

$$\nabla B(\mathbf{x}^{(k+1)})=\nabla A(\mathbf{x}^{(k)}), \tag{16}$$

where $\nabla A(\mathbf{x}) = \partial A/\partial \mathbf{x}$ is the gradient of A, represents an MM algorithm, as follows from convexity arguments. The surrogate objective function for the above construction has the form

$$\mathscr{Q}(\mathbf{x}|\mathbf{x}^{(k)})=B(\mathbf{x}^{(k)})-A(\mathbf{x})+\nabla^{\mathrm{T}}A(\mathbf{x}^{(k)})(\mathbf{x}-\mathbf{x}^{(k)}). \tag{17}$$

4.2. EM construction

Let $\mathscr{E}$ be the observed event, and U be a random variable (vector) representing missing data. The EM algorithm is a method to maximize the log-likelihood function $l(\mathbf{x}) = \log\{L(\mathbf{x})\}$ of the form

$$L(\mathbf{x})=E\{L_0(\mathbf{x}|U)\}, \tag{18}$$

where $L_0(\mathbf{x}|U)$ is the conditional likelihood, given missing data (the likelihood constructed to estimate $\mathbf{x}$ as if U were a covariate), and $\mathbf{x}$ does not include parameters of the distribution of U. To facilitate ‘distribution-free’ generalizations, we intentionally avoid explicit expressions involving the distribution of U, and we use the conditional rather than the joint likelihood of U and $\mathscr{E}$ to represent the EM procedure. In this construction, unknown parameters entering the distribution of U do not participate in the procedure, and the maximization is considered with respect to parameters $\mathbf{x}$ only. For any function $f(U)$, the conditional expectation of $f(U)$, given observed event $\mathscr{E}$ and $\mathbf{x}$, is represented as

$$E\{f(U)|\mathscr{E},\mathbf{x}\}=\frac{E\{f(U)L_0(\mathbf{x}|U)\}}{E\{L_0(\mathbf{x}|U)\}}.$$

This suggests the following explicit functional notation for the conditional expectation operator

$$E(f|g):=\frac{E(fg)}{E(g)}, \tag{19}$$

for any functions f and g of U, where U is a random variable, and $E(g)$ is the probability of the condition. A standard Jensen inequality argument shows that, with this notation,

$$Q(\mathbf{x}|\mathbf{y})=l(\mathbf{y})+E\{l_0(\mathbf{x}|U)-l_0(\mathbf{y}|U)|L_0(\mathbf{y}|U)\}, \qquad l_0=\log(L_0), \tag{20}$$

is a surrogate objective function for the target function $l(\mathbf{x})$. The operation of finding $\hat{U}$ such that $l_0(\mathbf{x}|\hat{U}) = E\{l_0(\mathbf{x}|U)|L_0(\mathbf{y}|U)\}$ is referred to as missing data imputation. If imputation is easy (E-step), maximization of Q with respect to $\mathbf{x}$ reduces to that of $l_0(\mathbf{x}|\hat{U})$, a complete-data problem.

To prove that any converging sequence $\mathbf{x}^{(k)} \to \mathbf{x}^*$, designed according to equation (4), gives us a stationary point in the limit, we follow the argument at $\mathbf{x} = \mathbf{y} = \mathbf{x}^*$

$$\frac{\partial Q(\mathbf{x}|\mathbf{y})}{\partial \mathbf{x}}=\frac{E\{\partial L_0(\mathbf{x}|U)/\partial \mathbf{x}\}}{L(\mathbf{y})}=\frac{1}{L(\mathbf{y})}\frac{\partial L(\mathbf{x})}{\partial \mathbf{x}}=0, \tag{21}$$

which implies that the score equation $\partial L(\mathbf{x})/\partial \mathbf{x} = 0$ is satisfied in the limit.

The EM algorithm proceeds by iterations $E\{l_0'(\mathbf{x}^{(k+1)}|U)\, L_0(\mathbf{x}^{(k)}|U)\}=0$, where at the kth iteration this equation is solved for $x^{(k+1)}$.

4.3. Quasi-EM construction

Let us revisit the EM construction under the question what properties of the E-operator did we actually use in the derivation of Section 4.2? They are conditional expectation performed according to expression (19), linearity and the Jensen inequality in equation (20) and interchangeability of E and $\partial/\partial\mathbf{x}$ in equation (21). Operators satisfying these properties will be called QE. As soon as a QE operator satisfying these requirements and such that $L(\mathbf{x}) = \text{QE}\{L_0(\mathbf{x}|U)\}$ is constructed, the likelihood function L can be maximized by an MM algorithm with the surrogate objective function as in equation (20) with E replaced by QE. The rationale behind this substitution is that the evaluation of E requires that U be a random variable, and that we know its distribution, whereas that of QE does not.

Formally, let $\mathscr{B}$ be some set of basis functions (including the function $f(u) \equiv 1$), and $\mathscr{S}$ be a set of all admissible functions stretched on $\mathscr{B}$ using linear combinations. In other words $\mathscr{S}$ consists of functions f such that $f = \Sigma_i a_i f_i$ for any sequence (finite or infinite) of functions $\{f_i\}$, $f_i \in \mathscr{B}$, and real numbers $\{a_i\}$.

Define QE as a linear functional mapping $\mathscr{S}$ to real numbers such that

a. $\text{QE}(\mathbf{1}) \coloneqq 1$, where $\mathbf{1}$ means a function that is equivalent to 1 (*normalization*),

b. for any $f = \Sigma_i a_i f_i \in \mathscr{S}\ f_i \in \mathscr{B}$, $\text{QE}(f) \coloneqq \Sigma_i a_i\, \text{QE}(f_i)$ (*linearity*),

c. for any function $f(u, a) \in \mathscr{S}$ and such that $\partial f(u, a)/\partial a \in \mathscr{S}$

$$\frac{\partial \text{QE}\{f(U,a)}{\partial a}=\text{QE}\left\{\frac{\partial f(U,a)}{\partial a}\right\}$$

(*interchangeability*),

d. as in expression (19), for any functions g and $fg \in \mathscr{S}$ and such that $\text{QE}(g) \neq 0$,

$$\text{QE}(f|g) := \frac{\text{QE}(fg)}{\text{QE}(g)} \tag{22}$$

(*conditional QE*) and

e. given the functions g, fg and $g\log(f) \in \mathscr{S}$

$$\mathrm{QE}\{\log(f)|g\} \le \log\{\mathrm{QE}(f|g)\} \tag{23}$$

(*Jensen inequality*).

Let us consider the QE requirements more closely. We start by postulating one basis function $f_0(U, a)$ and the value of QE on that basis function $\mathrm{QE}(f_0) \coloneqq \gamma(a)$, where γ is some function of a. Dependent on how many times we are allowed to differentiate under the QE symbol, the derivatives

$$f^{(i)}(U,a)=\partial^{(i)}f_0(U,a)/\partial a^i$$

must also be included in the set of admissible functions $\mathscr{S}$ so that derivatives of $\mathrm{QE}(f_0)$ can be defined. Moreover, QE on $f^{(i)}$, $i = 1, 2,\ldots$, are automatically defined by the interchangeability property through derivatives of γ, as $\mathrm{QE}\{f^{(i)}\} = \gamma^{(i)}$. As we can see, QE construction is cloned from f_0 and γ. Mathematical expectation $E\{f_0(U, a)\}$ is an integral transform of U, where a is an argument of the transform. Dependent on the choice of the function f_0, QE mimics differential properties of the corresponding transform, whereas the function γ is not necessarily an integral transform.

To study procedure (3), a QE construction based on moment-generating functions is useful. This construction is cloned from the basis function $f_0(U, a) = a^U$, $0 \le a \le 1$, $U \ge 0$. For a mathematical expectation, $E(f_0) = \gamma$ is the moment-generating function of U. If we want to be able to differentiate twice under the QE symbol, we need two more basis functions Ua^U and U^2a^U, so that

$$\left.\begin{array}{r}\mathrm{QE}(a^U)=\gamma(a),\\ \mathrm{QE}(Ua^U)=a\gamma'(a),\\ \mathrm{QE}(U^2a^U)=a\gamma'(a)+a^2\gamma''(a),\end{array}\right\} \tag{24}$$

by the interchangeability property with two derivatives of γ allowed. We now derive the Jensen inequality for this construction. First, noting equation (15), introduce the conditional QE

$$\Theta(x|c) := \mathrm{QE}(U|U^c x^U)=c+x\frac{\gamma^{(c+1)}(x)}{\gamma^{(c)}(x)}, \tag{25}$$

where we used expressions (22) and (24) to obtain the right-hand part of expression (25), c = 0, 1. The function Θ is a surrogate of conditional expectation of U, given observed data, and it serves as an imputation operator in the QEM construction. Next, let $\mathscr{B}_k$ be a family of functions $\mathscr{B}_k = \{U^k a^U, 0 \le a \le 1, U \ge 0\}$. Then by using the fact that, for $f = f(U, a) = a^U$ and $g = g(U, b) = U^c b^U$,

$$\frac{\partial}{\partial a}[\mathrm{QE}\{\log(f)|g\} - \log\{\mathrm{QE}(f|g)\}]=\frac{1}{a}\{\Theta(b|c) - \Theta(ab|c)\},$$

the following can be proved.

> *Theorem 1* (Jensen inequality for QE). Let $\Theta(x|c)$ as defined by expression (25) be a non-decreasing function of x, $c = 0, 1$. Then inequality (23) holds true for any $f \in \mathcal{B}_0$ and $g \in \mathcal{B}_0 \cup \mathcal{B}_1$.

It is interesting that mathematical expectation satisfies the assumption of non-decreasing Θ, which we call the *convexity assumption* for reasons that will become clear later as we discuss an application of the above theory to semiparametric likelihood.

> *Theorem 2* (convexity for mathematical expectation). Let γ be a function defined by using the mathematical expectation operator E as $\gamma(x) \coloneqq E(x^U)$, where U is a non-negative random variable. Then $\Theta(x|c)$ as defined in expression (25) is non-decreasing in x for any $c = 0, 1$.

Proof. The Cauchy–Schwartz inequality with functions $\xi(U, x) = U^{1+c/2}x^{U/2}$ and $\zeta(U, x) = U^{c/2}x^{U/2}$ can be used to show that $\Theta'(x|c) \geq 0$.

5. Non-linear transformation models

5.1. Definition

To study procedure (1)–(3) in more detail using the developments of Section 4, we need to specify a certain structure of the likelihood function to be optimized. To do this, let us confine ourselves to a large, yet specific, class of semiparametric survival models. Consider a parametric regression model with support on [0, 1]. Let $\gamma(x|\boldsymbol{\beta}, \mathbf{z})$, $x \in [0, 1]$, be a parametrically specified distribution function in x, conditional on covariates $\mathbf{z}$. We require that γ be twice differentiable with respect to x and regression coefficients $\boldsymbol{\beta}$.

We can nowdefine a semiparametrically specified survival function $G(t|\boldsymbol{\beta}, \mathbf{z})$, given covariates, as

$$G(t|\beta, \mathbf{z})=\gamma\{F(t)|\beta, \mathbf{z}\}, \tag{26}$$

where the base-line survival function F is specified nonparametrically. The class of models (26) will be called NTMs, to give it a name. Functions like γ will be called NTM-generating functions. An NTM is obtained by plugging a nonparametrically specified survival function F into a parametric distribution function γ with the support compatible with the range of F. One important subclass of NTMs is the family of PH mixture models (10), for which γ is a moment-generating function of U,

$$\gamma(x|\beta, \mathbf{z})=E(x^{U(\beta,\mathbf{z})}|\mathbf{z}).$$

To represent a semiparametric model in the NTM form, we need to express its survival function G as a function γ of a base-line survival function F (this representation is not unique and is not always possible). For example, from equation (9) we obtain the PO model in the form $G(t|\cdot) = \theta(\cdot)/[\theta(\cdot) - \log\{F(t)\}]$, which gives

$$\gamma(x|\cdot)=\frac{\theta(\cdot)}{\theta(\cdot)-\log(x)}. \tag{27}$$

The class of NTMs includes the class of linear transformation models (Cheng *et al.*, 1995, 1997). It is easy to show that a linear transformation model can be represented as $\gamma(x|\boldsymbol{\beta}, \mathbf{z}) = p[\log\{\theta(\boldsymbol{\beta}, \mathbf{z})\} + q(x)]$, where p is a parametrically specified tail function, q is an inverse tail function and θ is a predictor (it is convenient to specify *q* as the inverse of *p*; then θ = 1 corresponds to the base-line $\gamma(x|\cdot) = x$).

5.2. Algorithm

In the survival analysis formulation, under non-informative right censoring, the contribution of observations sampled from an NTM (26) to the likelihood are $\log(-\mathrm{d}G)$ and $\log(G)$, for a failure and censored observation respectively. We have

$$-\mathrm{d}G(t|\beta,\mathbf{z})=\gamma'\{F(t)|\beta,\mathbf{z}\}\,F(t)\,\mathrm{d}H(t),$$

where $\gamma'(x|\cdot) = \partial\gamma(x|\cdot)/\partial x$, differentials are taken with respect to *t* under a continuous model and we recall that $F = \exp(-H)$. We may now rewrite the likelihood (8) for an NTM as

$$l=\sum_{i=1}^{n}D_i\,\log(\Delta H_i)+\sum_{i=1}^{n}\sum_{j\in\mathscr{C}_i\cup\mathscr{D}_i}\log\{\vartheta(F_i|\beta,\mathbf{z}_{ij},c_{ij})\},\tag{28}$$

where

$$\vartheta(x|\cdot,c)=x^{c}\,\gamma^{(c)}(x|\cdot)\ ,$$

and ΔH_i is substituted for $\mathrm{d}H(t_i)$. It is easy to check that a negative derivative of $\log\{\vartheta(F_i|\cdot, c)\}$ with respect to ΔH_m is represented by $\Theta(F_i|\cdot, c)$, if $m \le i$, and is equal to 0 otherwise, where the function Θ is defined by expression (25). Therefore, the construction (1)–(3) of Section 1 leads us to the iteration scheme

$$\Delta H_m^{(k+1)}=D_m/\sum_{ij\in\mathscr{R}_m}\Theta(F_i^{(k)}|\beta,\mathbf{z}_{ij},c_{ij}).\tag{29}$$

This procedure is a generalization of procedure (13) to the NTM family. For the PO model, substituting equation (27) into expression (25), we obtain

$$\Theta(x|\cdot,c)=\frac{c+1}{\theta(\cdot)-\ \log(x)}.\tag{30}$$

It is clear that, with Θ given by equation (30), the general procedure (29) turns into the procedure (13) derived for the PO model in Section 3.

5.3. Quasi-expectation form of a non-linear transformation model

We now make use of the QE theory of Section 4.3 to provide a link between NTMs and the QE operator. Equations (24) summarizing second-order differential properties of the QE operator will be the main instrument of this section.

First, let us synchronize the development of Section 4.3 and the definition of NTM (26) in Section 5.1 by assuming that the function γ that is used in both sections is the same function. In fact, we already used this synchronization when we noticed in the previous section that Θ in equation (29) and in expression (25) is the same function. Now, from the first line of expression (24), with $F(t)$ instead of *a*, we obtain the QE form of the NT model

$$G(t|\beta,\mathbf{z})=\mathrm{QE}_{\beta,\mathbf{z}}\{F(t)^{U}\}\,, \tag{31}$$

where the subscript **β**, **z** to the QE operator indicates that QE is defined by using the function γ(*x*|**β**, **z**). Equation (31) is a postulate in the definition of the QE operator, and its link to the NTMs is established as we assume that QE is defined by using an NTM-generating function γ.

Now, consider the likelihood function *l* (28). Given an observation (*t*, **z**, *c*), *c* = 0, 1, its contribution υ{*F*(*t*)|**β**, **z**, *c*} to the likelihood *L* = exp(*l*) can be written as

$$\upsilon(F|\cdot,c)=\vartheta(F|\cdot,c)\Delta H^{c}=\Delta H^{c}F^{c}\gamma^{(c)}(F|\cdot,c)=\mathrm{QE}(\Delta H^{c}U^{c}F^{U})\,,$$

where the last equation follows from the first two lines of expression (24) and linearity of QE. As a result, the likelihood of an NTM mimics that of a mixture model

$$L=\prod\mathrm{QE}(\Delta H^{c}U^{c}\,F^{U}).$$

Consider the hazard function λ(*t*|**z**), corresponding to the survival function *G*(*t*|**z**). Differentiating the survival function (26), and using expression (24), we have

$$\lambda(t|\cdot)=-\frac{1}{G(t|\cdot)}\frac{\partial G(t|\cdot)}{\partial t}=\frac{\gamma'\{F(t)|\cdot\}\,F(t)}{\gamma\{F(t)|\cdot\}}h(t)=\frac{\mathrm{QE}\{U\,F(t)^{U}\}}{\mathrm{QE}\{F(t)^{U}\}}h(t),$$

where *h* is the hazard function corresponding to *F*. Applying the definition of conditional QE (22) to this expression, and using expression (25), we obtain

$$\lambda(t|\mathbf{z})=\mathrm{QE}\{U|F(t)^{U}\}\,h(t)=\Theta(F|\cdot,0)\,h(t).$$

This is a generalization of the fact that the population hazard function at time *t* in a heterogeneous population is represented as a conditional average, given survival up to time *t*.

Bringing these derivations together with expression (25), we have the following theorem.

> *Theorem 3* (QEM construction). Consider a survival analysis problem for an NTM generated by the function γ(*x*|**β**, **z**), with fixed covariates. With the QE operator as defined in Section 4.3, and using the same NTM-generating function γ in its definition, the following representations are valid: *survival function*,
>
> $$G(t|\beta,\mathbf{z})=\mathrm{QE}_{\beta,\mathbf{z}}\{F(t)^{U}\};$$
>
> *hazard function*,
>
> $$\lambda(t|\mathbf{z})=\mathrm{QE}\{U|F(t)^{U}\}\,h(t)=\Theta(F|\cdot,0)\,h(t),$$

where λ and *h* are hazards functions corresponding to *G* and *F* respectively; *likelihood function*,

$$l=\sum_{i=1}^{n}\left(\sum_{j\in\mathscr{C}_i\cup\mathscr{D}_i}\log[\,\mathrm{QE}_{\beta,\mathbf{z}_{ij}}\{(U\Delta H_i)^{c_{ij}}F_i^U\}]\right);$$

imputation operator,

$$\widehat{U}=\mathrm{QE}(U|U^c\,F^U)=\Theta(F|\cdot,c),\qquad c=0,1,$$

where $\hat{U}$ denotes *U*, imputed by using the conditional QE operator.

6. Summary and justification of the procedure

Let us now go back to the EM algorithm of Section 3 and see how the results obtained since then allow us to streamline and justify our algorithm construction, using the PO model as an example. In summary, we now have the following procedure.

a. Obtain the NTM-generating function, representing the model survival function as a function of *F*. For the PO model (9) $G(t|\cdot)=\theta(\cdot)[\theta(\cdot)-\log\{F(t)\}]^{-1}$, we have equation (27),

$$\gamma(x|\cdot)=\theta(\cdot)\{\theta(\cdot)-\log(x)\}^{-1}.$$

b. Obtain the imputation operator (25)

$$\Theta(x|\cdot)=c+x\,\gamma^{(c+1)}(x|\cdot)\gamma^{(c)}(x|\cdot)^{-1}.$$

For the PO model, this results in equation (30),

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

Check that $\Theta(x|\cdot)$ is a non-decreasing function of *x* (see the justification below).

c. Obtain the profile likelihood by iterations (29),

$$\Delta H_m^{(k+1)}=D_m\left\{\sum_{ij\in\mathscr{R}_m}\Theta(F_i^{(k)}|\beta,\mathbf{z}_{ij},c_{ij})\right\}^{-1}.$$

d. Maximize the profile likelihood with respect to **β** as in Section 2.

For the PH mixture model, QE is equivalent to E (compare equations (31) and (10)), which makes Θ the conditional expectation of *U*, given observed data (compare expressions (25) and (15)). In this case, the above procedure is an EM algorithm.

Justification of this procedure works through the proof of monotonicity (i.e. the likelihood is improved at each step) under the following assumption.

6.1. Convexity assumption

Consider an NTM with the NTM-generating function γ. Assume that

$$\Theta(x|\beta, \mathbf{z}, c) \text{ is a non−decreasing function of } x, \text{ for any } \beta, \mathbf{z}, c. \tag{32}$$

We have two ways to show monotonicity.

a. Observe that, under assumption (32), the likelihood (28), as a function of the vector $\Delta\mathbf{H}$, represents a difference between two concave functions $\Sigma_i D_i \log(\Delta H_i)$ and $-\Sigma_{ij} \log\{\vartheta(F_i|\cdot)\}$. This follows from the fact that Θ is the negative derivative of log(ϑ) with respect to H. Therefore, monotonicity follows from the results of Section 4.1.

b. Observe that the likelihood is represented as a QE $L = \Pi\ \mathrm{QE}(\Delta H^c U^c F^U)$ (Section 5.3), and that under assumption (32) the QE operator satisfies the Jensen inequality (Section 4.3). Therefore, the EM proof of monotonicity works.

Convergence of the algorithm under monotonicity follows from the results of Lange *et al.* (2000) and Wu (1983) under fairly general conditions.

7. Real data example

As an example we use data from the National Cancer Institute’s ‘Surveillance epidemiology and end results’ programme. Using the publicly available database for the programme, 11621 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 19819 Utah cases registered in the database: valid positive survival time, valid stage of the disease and age 18 years or more. Prostate cancer specific survival was analysed by the stage of the disease (localized or regional *versus* distant). For the definition of stages as well as for other details of the data we refer the reader to the documentation for the programme at http://seer.cancer.gov/.

The PH and the PO models with **z** representing two groups corresponding to the localized or regional stage (10765 cases) and distant stage (856 cases) respectively were fitted by using the profile MM algorithm. The log-odds-ratio was estimated as $\hat{\beta} = -3{:}251$ with 95% asymptotic confidence interval (−3:416, −3:086).A likelihood ratio test showed that the difference between groups is highly significant ($p < 0{:}0001$). Observed (Kaplan–Meier) and expected model-based estimates of the survival functions by group are shown in Fig. 1. The PO model showed a superior fit to the data.

On the basis of the PO model, four different approaches to model fitting are compared.

a. *MM or QEM*: the method is described in Section 6. Maximization of the profile likelihood is performed by the Powell method (Press *et al.*, 1994).

b. *EM*: the EM method is similar to the EM algorithm as used to fit frailty models with predictor θ(**β**, **z**). Using the QEM formulation, the procedure is as follows.

  i. With the current iteration $\boldsymbol{\beta}^{(k)}$ and $F^{(k)}$ compute $V_{ij}^{(k)}=\theta^{-1}(\beta^{(k)}, \mathbf{z}_{ij})\ \Theta(F_i^{(k)}|\beta^{(k)}, \mathbf{z}_{ij})$ for each subject ij.

  ii. Maximize the partial likelihood for a PH model with (imputed) predictor $V_{ij}^{(k)}\theta(\beta, \mathbf{z}_{ij})$ with respect to **β**. Set $\boldsymbol{\beta}^{(k+1)}$ at the solution.

  iii. Update the function F by using the Nelson–Aalen estimator for the PH model fitted at the previous step. Denote the solution by $F^{(k+1)}$.

   - **iv.** Set $k = k + 1$. Continue iterations until convergence.

- **c.** *Parametric*: in the parametric method, the function F is specified as a Weibull survival function. The parametric regression model is fitted by using the Powell method applied to all model parameters.
- **d.** *Full model (Powell)*: apply the Powell method to maximize the log-likelihood of the full semiparametric model with respect to the joint vector of regression coefficients $\boldsymbol{\beta}$ and the base-line hazard $\Delta\mathbf{H}$.

Computation of θ, Θ or γ is counted as one operation. For a given tolerance ε, the stopping rule is defined as $l^{(k+1)} - l^{(k)} < \varepsilon$. If the method required solving a nested numerical problem (MM or EM), the tolerance for the nested problem is specified as ε/100.

First, we evaluate the precision by operations characteristics of the above numerical methods. The precision is measured by $l^* - \mathrm{l}^{(k)}$, where the exact solution l^* was approximated by the solution obtained for $\varepsilon = 10^{-20}$. Shown in Fig. 2 are the precision by operations curves for the four methods, obtained by varying ε. It is clear from Fig. 2 that the profile MM algorithm outperforms the other approaches in the number of operations that are required to reach a given precision. The profile MM method is closely followed by the frailty EM algorithm. Fitting the full semiparametric model by the Powell method shows the worst performance. The advantage of the EM-like approaches compared with methods that invoke the function F into a conventional maximization is explained by the utilization of a closed form solution for F in the form of the Nelson–Aalen–Breslow estimator. For the same reason, the MM and the EM procedures show a linear trend with increasing dimension, given fixed precision as shown in Fig. 3. To obtain Fig. 3, samples of size 100–1000 were generated from the parametric (Weibull) PO model fitted to the same data. The MM, EM and full model (Powell) procedures were applied to each such sample. The tolerance $\varepsilon = 10^{-3}$ was used for the MM algorithm. The tolerance for the other two methods was tuned to give a likelihood that was as close as possible yet smaller than the likelihood achieved by the MM method (to keep the comparison conservative). The profile MM algorithm shows the most favourable behaviour with increasing dimension, followed by the EM procedure. It comes as no surprise that the full model (Powell) method shows the greatest complexity.

8. Conclusion

We presented an application of the general MM principle to a class of semiparametric models. Three methods of specifying the surrogate objective function were demonstrated. In particular, we clarified the connection between the likelihood-based MM principle and the imputation-based self-consistency principle that is used in EM algorithms for semiparametric models. To study this connection, we built an EM-like world behind the MM algorithm by using the QEM construction. The approaches were illustrated by using continuous NTMs in a survival analysis context. This is just one example of how these constructions can be used. Discrete survival models, cure models, multivariate semiparametric models, models with time-dependent covariates and many other statistical models can be treated by application of the principles presented in this paper. Construction of surrogate objective functions is not straightforward. Having an option to work a particular problem from both ends (likelihood or convexity *versus* imputation or self-consistency) may increase the chance of finding efficient and general procedures that are applicable to large classes of models.

Acknowledgments

The author thanks Dr Ken Boucher, Dr Marvin Zelen, the referees and the Joint Editor for many helpful comments. This research is partially supported by National Cancer Institute grant 1U01 CA97414-01 (Cancer Intervention and

Surveillance Modeling Network), Department of Defence grant DAMD17-03-1-0034 and National Institute of Ageing grant 1RO1 AG14650-01A2.

References

Andersen, P.; Borgan, Ø.; Gill, R.; Keiding, N. Statistical Models based on Counting Processes. New York: Springer; 1993.

Cheng S, Wei L, Ying Z. Analysis of transformation models with censored data. Biometrika 1995;82:835–845.

Cheng S, Wei L, Ying Z. Predicting survival probabilities with semiparametric transformation models. J. Am. Statist. Ass 1997;92:227–235.

Clayton D, Cuzick J. Multivariate generalizations of the proportional hazards model (with discussion). J. R. Statist. Soc. A 1985;148:82–117.

Cox D. Regression models and life-tables (with discussion). J. R. Statist. Soc. B 1972;34:187–220.

Feller, W. An Introduction to Probability Theory and Its Applications. New York: Wiley; 1971.

Fleming T, Lin D. Survival analysis in clinical trials: past developments and future directions. Biometrics 2000;56:971–983. [PubMed: 11129494]

Hougaard P. Life table methods for heterogeneous populations: distributions describing the heterogeneity. Biometrika 1984;71:75–83.

Klein J. Semiparametric estimation of random effects using the Cox model based on the EM algorithm. Biometrics 1992;48:795–806. [PubMed: 1420842]

Lange K, Hunter D, Yang I. Optimization transfer using surrogate objective functions (with discussion). J. Comput. Graph. Statist 2000;9:1–59.

Murphy S. Consistency in a proportional hazards model incorporating a random effect. Ann. Statist 1994;22:712–731.

Murphy S. Asymptotic theory for the frailty model. Ann. Statist 1995;23:182–198.

Murphy S. On profile likelihood. J. Am. Statist. Ass 2000;95:449–485.

Murphy S, Rossini A, van der Vaart A. Maximum likelihood estimation in the proportional odds model. J. Am. Statist. Ass 1997;92:968–976.

Murphy S, van der Vaart A. Semiparametric likelihood ratio inference. Ann. Statist 1997;25:1471–1509.

Nielsen G, Gill R, Andersen P, Sorensen T. A counting process approach to maximum likelihood estimation in frailty models. Scand. J. Statist 1992;19:25–43.

Oakes D. Bivariate survival models induced by frailties. J. Am. Statist. Ass 1989;84:487–493.

Parner E. Asymptotic theory for the correlated gamma frailty model. Ann. Statist 1998;26:183–214.

Press, W.; Flannery, B.; Teukolsky, S.; Vetterling, W. Numerical Recipes in Pascal: the Art of Scientific Computing. New York: Cambridge University Press; 1994.

van der Vaart, A. Asymptotic Statistics. Cambridge: Cambridge University Press; 1998.

Wassel J, Moeschberger M. A bivariate survival model with modified gamma frailty for assessing the impact of interventions. Statist. Med 1993;12:241–248.

Wu C. On the convergence properties of the EM algorithm. Ann. Statist 1983;11:95–103.

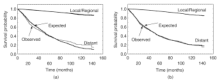

Fig. 1.

Observed (————) *versus* expected (————) plots corresponding to (a) the PH and (b) the PO model fitted to the prostate cancer data

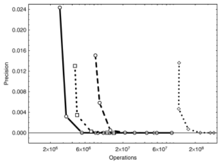

Fig. 2.

Precision of likelihood maximization by the number of operations (precision is measured as the difference between the limiting value of the likelihood as operations tend to ∞ and the maximal likelihood value achieved under a stopping rule; curves closer to the *y*-axis correspond to more efficient numerical methods): ————, MM; -------, EM;– – –, parametric; ·······, full model

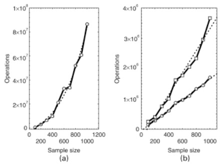

Fig. 3.

Numerical efficiency by sample size for (a) the full model (·······, polynomial fit) and (b) the EM (□) and MM (○) methods (·······, linear fits): ————, number of operations needed to achieve a fixed precision by sample size (each point corresponds to a sample generated from a parametric PO model with a Weibull base-line survival function with parameters specified by using the model fit to data described in Section 7)