Extending the Iterative Convex Minorant Algorithm to the Cox Model for Interval-Censored Data

Wei Pan

The iterative convex minorant (ICM) algorithm (Groeneboom and Wellner, 1992) is fast in computing the NPMLE of the distribution function for interval censored data without covariates. We reformulate the ICM as a generalized gradient projection method (GGP), which leads to a natural extension to the Cox model. It is also easily extended to support the Lasso (Tibshirani, 1996). Some simulation results are also shown. For illustration we reanalyze two real datasets.

Key Words: Cross-validation; EM; Generalized gradient projection; ICM; Interval censoring; Isotonic regression; Lasso; NPMLE.

1. INTRODUCTION

There has been an increasing research interest in interval censored data recently, largely due to its popularity in studies of AIDS and other chronic diseases, in which the event of interest is never observed exactly but only known to occur within some time intervals. In these studies a baseline examination was conducted, then after some fixed time the first followup took place, then second and third, etc. followups (if any) after intervals of several years each. If the event happens, we can only be aware of its happening between two examinations. Otherwise, it may only happen after the end of the last followup. This produces interval censoring. One fundamental statistical problem in survival analysis is to estimate the distribution function of the event time from censored data.

In a more general framework Turnbull (1976) presents an EM algorithm (Dempster, Laird and Rubin, 1977) to compute the nonparametric maximum likelihood estimator (NPMLE) of the distribution function for this kind of data without covariates. Groeneboom and Wellner (1992, shortened as G&W in the sequel) propose an iterative convex minorant (ICM) algorithm, which they suggest is much faster than the EM. In the Cox model with interval censored data, Finkelstein (1986) gives a modified Newton-Raphson method after reparameterization. This approach uses the

1Division of Biostatistics, University of Minnesota, A460 Mayo Building (Box 303), Minneapolis, MN 55455. Email: weip@biostat.umn.edu
Extending the ICM

inverse Hessian matrix at each iteration. However, for continuous time model without grouping, the dimension of the unknown parameters is in the order of the number of observations, which is often more than several hundreds. Many of today's ordinary workstations cannot afford this storage space. Moreover, the Hessian matrix is sometimes near singular, which may cause some numerical problems when inverting it. In contrast, neither the EM nor the ICM needs to store and invert the Hessian matrix. In fact, as analyzed in the next section, the ICM only uses the diagonal elements of the Hessian. On the other hand, in spite of its slowness the EM is hard to be used in the Cox model for continuous survival times (Alioum and Commenges, 1996). Hence an extension of the ICM to the Cox model has a practical importance.

For low-dimensional regression coefficients, a profile likelihood approach has been suggested (Huang 1996; Huang and Wellner 1995). It works by fixing the regression coefficient at some possible values, then for each fixed value of regression coefficient maximizing the likelihood with respect to the baseline distribution by a modified ICM algorithm. Then the point achieving the maximum of this profile likelihood (possibly after smoothing) is taken as the final estimate. Just recently Huang and Wellner (1997) suggest a more general method based on the back-fitting scheme (Hastie and Tibshirani 1990), though no simulation results are shown. We will compare our approach with theirs later.

Originally the ICM was formulated in terms of some special stochastic processes and isotonic regression algorithms (G&W; Huang and Wellner 1997). In Section 2 we identify it as a special case of the generalized gradient projection method (GGP) in the optimization literature (Mangasarian, 1996; Bertsekas, 1982). From this connection a natural extension of the ICM to the Cox model is almost trivial. We further extend it to support the Lasso, a biased regression scheme proposed by Tibshirani (1996, 1997). In Section 3 simulations were conducted to assess the performance of the extended ICM methods. Finally two real world problems are used to illustrate the methods, followed by a brief discussion.

2. COMPUTATIONAL PROCEDURES

2.1 GGP

We first briefly review the GGP, a general optimization scheme (Mangasarian, 1996; Bertsekas, 1982). It is helpful to regard it as a Newton-type iterative algorithm in constrained optimization.
Extending the ICM

Specifically, suppose we want to maximize a function $f(x)$ in a closed convex set $\mathcal{X}$. Let $\nabla f$ denote the first derivative of $f$, $H$ a symmetric positive definite matrix, which can be often taken as the negative Hessian matrix of $f$ if $f$ is strictly concave. The GGP updates its current estimate $x^{(m)}$ by

$$
x^{(m+1)} = \text{Proj}[x^{(m)} + \alpha^{(m)} H^{(m)-1} \nabla f(x^{(m)}), H^{(m)}, \mathcal{X}],
$$

where $\alpha^{(m)} > 0$ is a suitably chosen stepsize, and $\text{Proj}$ is the projection operation defined by:

$$
\text{Proj} [x_0, H, \mathcal{X}] := \arg \min_x \{(x - x_0)' H (x - x_0) : x \in \mathcal{X}\}.
$$

The iterate $x^{(m)}$ at convergence is taken as a solution of the maximization problem. The rationale behind the GGP is intuitively explained by the following result:

1) $x^* \in \arg \max_{x \in \mathcal{X}} f(x)$;
2) $x^* = \text{Proj} [x^* + \alpha H^{-1} \nabla f(x^*), H, \mathcal{X}]$ for any $\alpha > 0$;
3) $x^* = \text{Proj} [x^* + \alpha H^{-1} \nabla f(x^*), H, \mathcal{X}]$ for some $\alpha > 0$;

If $f$ is differentiable, then 1) implies 2); On the other hand, if $f$ is convex, then 3) implies 1).

We notice by passing that taking $H$ as an identity matrix (i.e. completely ignoring the second order information of $f$) results in the gradient projection method (Polak 1971).

Some theoretical properties of the GGP, such as its superlinear or linear convergence rate (depending on the choice of $H$), have been established in Bertsekas (1982). Furthermore, Bertsekas has shown its effectiveness via computational examples involving as many as 10000 variables. This feature is especially attractive in our current setting (often involving more than several hundred parameters for medium to large samples).

2.2 ICM for Interval Censored Data without Covariates

For interval censored data, the time of the event of interest, say survival time $X_i$, is not observed exactly. Observations available are censoring times, $(U_i, V_i)$ for $i = 1, ..., n$. We only know that $0 \leq U_i \leq X_i \leq V_i \leq \infty$. As usual we assume that $X_i$ and $(U_i, V_i)$ are independent, and $\{X_i\}$ and $\{(U_i, V_i)\}$ are respectively two iid samples. We are interested in estimating the distribution function of $X_i$, $F_0$, from $\{(U_i, V_i)\}$.

The (conditional) log-likelihood can be written down as (Turnbull, 1976):

$$
L(F|\{(U_i, V_i)\}) = \sum_{i=1}^{n} \log (F(V_i) - F(U_i-)).
$$
Notice that $L$ can be further expressed in terms of the cumulative hazard function $\Lambda$. This is actually what we did in implementing our algorithms. There are two advantages with this parametrization:
i) concavity of the likelihood function is readily available, even in the Cox model (Huang and Wellner, 1997); ii) in contrast to restricting $F \leq 1$ there is no upper bound on $\Lambda$. But for simplicity of notation we still use the distribution function throughout.

The nonparametric maximum likelihood estimator (NPMLE) $\hat{F}$ of $F_0$ is a right-continuous function maximizing $L(F)$. Turnbull (1976) shows that $\hat{F}$ can only have jumps between the order statistics $\tau_i$, $i = 1, \ldots, k$, of $U_i$ and $V_i$, $1 \leq i \leq n$. However, the probability distribution of $\hat{F}$ within some intervals is nonidentifiable. This nonidentifiability will not influence the computation of $\hat{F}$ by the EM algorithm. To facilitate the computation of $\hat{F}$ by the ICM, G&W define the NPMLE $\hat{F}$ as piecewise constant with possible discontinuities only at the order statistics $\{\tau_i\}$, and moreover $\hat{F}$ may be less than 1 at the largest order statistic $\tau_k$. These two different definitions seem to be non-critical. In the sequel, we adopt the second one.

Turnbull (1976) also suggests how to reduce the number of the order statistics $\{\tau_i\}$ on which $\hat{F}$ may have jumps. We denote the possible support of the NPMLE as a disjoint union of intervals $[a_i, b_i)$, $i = 1, \ldots, m$. (It is likely that the NPMLE has no probability mass in some of these intervals.)

Generally we use $v$ to denote a column vector with components $v_i$, and similarly the $(i, j)$th element of a matrix $M$ with $M_{ij}$.

So far the ICM is only applicable to interval censored data without covariates. The original exposition of the ICM was in terms of stochastic processes (Groeneboom and Wellner, 1992). Now we reformulate it as a GGP method (Mangasarian 1996; Bertsekas 1982) so that its connection with the Newton-Raphson algorithm and gradient projection method is obvious.

Let the first derivative (i.e. gradient) and second derivative (i.e. the Hessian) of $L$ with respect to $F = (F(\tau_1), \ldots, F(\tau_k))^t$ be $\nabla L$ and $\nabla^2 L$ respectively. $G$ is a $k \times k$ diagonal matrix with the same diagonal elements as $-\nabla^2 L$. To have a proper (sub)distribution function we need to restrict $F \in \mathcal{R} = \{F : 0 \leq F(\tau_1) \leq \ldots \leq F(\tau_k) \leq 1\}$. Denote the estimates from the $i$th iteration as $F^{(i)}$. Then the ICM iterates as

$$F^{(m+1)} = \text{Proj}[F^{(m)} + G(F^{(m)})^{-1}\nabla L(F^{(m)}), G(F^{(m)}), \mathcal{R}],$$
Extending the ICM

where the projection by definition is

\[
\text{Proj}[y, G, R] := \arg \min_x \left\{ \sum_{i=1}^k (y_i - x_i)^2 G_{ii} : 0 \leq x_1 \leq x_2 \leq \ldots \leq x_k \leq 1 \right\},
\]

which is just an isotonic least squares regression problem and hence can be efficiently accomplished by some well-known algorithms such as the pool-adjacent-violators algorithm (PAVA) (Robertson et al., 1988). The time complexity of the PAVA is \(O(k^2)\) and easy to program. Graphically the solution of the isotonic regression (with a simple order as \(R\)) can be represented by a greatest convex minorant of the plot \(\{(0, 0), (\sum_{i=1}^j G_{ii}, \sum_{i=1}^j G_{ii} x_i) \text{ for } i = 1, \ldots, k\}\). This may be the reason why the above (iterative) algorithm is named ICM. The projection guarantees that the new \(F^{(m+1)}\) is still a proper (sub)distribution function. Moreover, the projection seems to speed up convergence as suggested in G&W. Evidently this is a special case of the GGP method. Without the projection, the above iteration is just a Newton-like step where only the diagonal elements of the Hessian matrix are used. The estimate \(F^{(m)}\) at convergence is taken as the NPMLE.

We follow Aragon and Eberly (1992) to further define the damped iterative convex minorant algorithm (DICM) as

\[
F^{(m+1)} = \text{Proj}[F^{(m)} + \alpha_j G(F^{(m)})^{-1} \nabla L(F^{(m)}), G(F^{(m)}), R],
\]

where \(\alpha_k\) is an Armijo-like stepsize (Polak 1971) or can be simply chosen by

\[
\alpha_j = \max \left\{1/2^i : L(F^{(m+1)}) > L(F^{(m)}), i = 0, 1, 2, \ldots \right\},
\]

which tries to take a large stepsize while increases the likelihood.

The appeal of the DICM is its global convergence and increasing likelihood in each step, though Zhan and Wellner (1995) point out some possible problems in the proof of the global convergence in Aragon and Eberly (1992). On the other hand, G&W (page 70) suggest using “buffer” to keep the estimate as a proper distribution and to avoid the likelihood becoming 0. However, they do not say what to be done if the new estimate does make the likelihood be 0. Damping is a natural way to avoid this problem. Hence in our discussion we refer to ICM and DICM equivalently.

2.3 Extending ICM to the Cox Model

The Cox proportional hazards model is probably the most widely used one in survival analysis. It assumes that the hazard function \(h(x|Z)\) with covariate (vector) \(Z\) is proportional to an unknown
Extending the ICM

baseline hazard \( h_0(x) \): \( h(x|Z) = h_0(x) \exp(Z'\beta) \), or formulated in terms of distribution functions:

\[
1 - F(x|Z) = [1 - F_0(x)]^{\exp(Z'\beta)},
\]

where \( F_0 \) is the corresponding baseline distribution (unknown), and \( \beta \) is the regression coefficient (vector). Let the \( n \) observations be labelled \((U_1, V_1, Z_1), \ldots,(U_n, V_n, Z_n)\), where \((U_i, V_i)\) as before is the lower and upper boundaries of the censoring interval, and \( Z_i \) is the covariate (vector). The survival time \( X_i \) is only observed to be inside \((U_i, V_i)\). As usual we assume that \( X_i \) and \((U_i, V_i)\) are mutually independent. The joint log-likelihood can be written down (up to a constant) as (Finkelstein 1986):

\[
L(F_0, \beta) = \sum_{i=1}^{n} \log \left\{ (1 - F_0(U_i-))^{\exp(Z_i'\beta)} - (1 - F_0(V_i))^{\exp(Z_i'\beta)} \right\}.
\]

\( L \) is maximized to obtain the NPMLE \( \hat{\beta} \) of the regression coefficient along with that of the baseline distribution, \( \hat{F}_0 \).

Here we approach this maximization problem by extending the ICM. Computationally this is different from Finkelstein (1986), where the Newton-Raphson method is adopted after reparametrization. As before we would like to avoid using the full Hessian matrix due to the possible difficulties mentioned earlier. Our extended ICM is simple (without inverting a high-dimensional matrix) and reasonably fast when applied in this setting.

Extending the ICM is straightforward by considering the regression coefficient as another component of parameters in addition to the baseline distribution. Now let \( \nabla_1 L(F_0, \beta) = \partial L(F_0, \beta)/\partial F_0 \) and \( \nabla_2 L(F_0, \beta) = \partial L(F_0, \beta)/\partial \beta \) denote the first derivatives; \( G_1(F_0, \beta) \) and \( G_2(F_0, \beta) \) the corresponding diagonal matrices of the negative second derivatives. Then our extended ICM iterates as

\[
F_0^{(m+1)} = \text{Proj}[F_0^{(m)} + \alpha_j G_1(F_0^{(m)}, \beta^{(m)})^{-1} \nabla_1 L(F_0^{(m)}, \beta^{(m)}), G_1(F_0^{(m)}, \beta^{(m)}), \mathcal{R}],
\]

\[
\beta^{(m+1)} = \beta^{(m)} + \alpha_j G_2(F_0^{(m)}, \beta^{(m)})^{-1} \nabla_2 L(F_0^{(m)}, \beta^{(m)}),
\]

where \( \alpha_j \) and \( \text{Proj} \) are defined as before. Again the projection step is taken to ensure that the estimate of \( F_0 \) is a proper (sub)distribution, which requires that \( \hat{F}_0 \) is nondecreasing and between 0 and 1. Since no constraint is on \( \beta \), no projection is performed on it. We will discuss another application of the GP to constrained \( \beta \) next.
Extending the ICM

Notice that inverting the diagonal matrices \( G_1 \) and \( G_2 \) is trivial. But to ensure numerical stability of inverting \( G_1 \) and \( G_2 \) (and positive-definiteness of \( G_1 \) and \( G_2 \) for truncated data as mentioned at the end of the paper), the Levenberg-Marquardt adjustment is adopted (Thisted, 1988). Specifically, \( G_1^{-1} \) and \( G_2^{-1} \) are respectively replaced by \( (G_1 + \epsilon_0 I)^{-1} \) and \( (G_2 + \epsilon_0 I)^{-1} \) with a suitably chosen small \( \epsilon_0 > 0 \) when they are near singular (or non-positive-definite for truncated data). We note that our extended ICM is different from that outlined by Huang and Wellner (1997), which is basically a back-fitting algorithm (Hastie and Tibshirani 1990): \( L(F_0, \beta^{(m)}) \) is first maximized with respect to \( F_0 \) at the fixed \( \beta^{(m)} \) (by an ICM-like algorithm), which leads to an \( F_0^{(m+1)} \); then \( L(F_0^{(m+1)}, \beta) \) is maximized with respect to \( \beta \) at the fixed \( F_0^{(m+1)} \) to get another estimate \( \beta^{(m+1)} \); the above two steps are repeated until convergence. In contrast, we propose to maximize the likelihood jointly with \( F_0 \) and \( \beta \), which seems to be more common and likely more efficient. Huang and Wellner (1997) did not give any specific numerical results. Some numerical comparison is presented in Section 3.

Intuitively the (extended) ICM will converge, since the likelihood function is upper-bounded and each step of the (extended) ICM increases the likelihood. When we don’t use the second order information (i.e. \( G \) is an identity matrix), the GGP reduces to the GP, which only has a linear convergence rate. On the other hand, the GGP can have a superlinear convergence rate if \( G \) is suitably chosen (part of which is diagonal and the remaining part equals the corresponding Hessian) (Propositions 3 and 4, Bertsekas 1982). Since the used information \( G \) in our (extended) ICM is between them, we can expect the convergence rate of the (extended) ICM is also between linear and superlinear; furthermore, if the diagonal elements of the Hessian dominate the non-diagonals, the convergence rate of the (extended) ICM will at least be close to superlinearity. Rigorous derivations are not pursued here, though our limited experience seems to be positive.

2.4 Extending ICM to Support Lasso

Tibshirani (1996) proposes a new estimation procedure called Lasso in linear regression and then extends it to the Cox model for right-censored data (Tibshirani, 1997). The Lasso works by constraining the sum of the absolute values of the regression coefficients being less than a constant. This immediately reminds one the ridge regression, where the constraint is replaced by the sum of the squared regression coefficients. As in ridge regression, the Lasso is a biased estimation
method which nonetheless may lead to smaller mean squared errors by reducing the variability of the estimates.

However, to our knowledge, there is still no application of the Lasso method to the Cox model with interval censored data in the literature. Unlike in linear regression or in the Cox model with right censored data, here we have a high-dimensional (and ordered) nuisance parameter, the baseline distribution. (The baseline distribution is eliminated from the partial likelihood in the Cox model for right censored data; but we do not have an explicit partial likelihood for interval censored data.) The ICM is readily adapted to support the Lasso by using the ideas of the GGP. Now we need to constrain \( \beta \) in \( C = \{ \beta : \sum_{i=1}^{r} |\beta_i| \leq \gamma \} \) (where \( \gamma \) is a specified constant). This only needs to modify our previous algorithm by

\[
\beta^{(m+1)} = \text{Proj}[\beta^{(m)} + \alpha_j G_2(F_0^{(m)}, \beta^{(m)} )^{-1} \nabla_2 L(F_0^{(m)}, \beta^{(m)} ), G_2(F_0^{(m)}, \beta^{(m)} )] \cup C, 
\]

The projection of a point \( \beta \) into \( C \) is a standard quadratic programming problem. In particular, Tibshirani (1996) gives two efficient algorithms. We term the above \( \beta^{(m)} \) at convergence as the Lasso estimate.

The \( \gamma \) in the constraint can be chosen by generalized cross-validation for right censored data (Tibshirani 1997). Here we use a 10-fold likelihood cross-validation (Silverman 1986).

### 3. SIMULATION

In our simulations, the baseline distribution \( F_0 \) is Weibull with shape parameter 2 and scale parameter 1. Only binary covariates are used, each component of which takes 0 or 1 with an equal probability. The regression coefficient is also a vector of 0 and 1. The first examination time \( T \) is \( \text{Uniform}(0, \theta) \). To mimic many panel studies we take the length of the time interval between two follow-up examinations to be constant, \( len = 0.5 \). Generally if we have \( k+1 \) examinations, survival time \( X_i \) is accordingly censored in one of \( (0, T_i], (T_i, T_i + len], ..., (T_i + k*len, \infty) \). \( k \) and \( \theta \) jointly determine the censoring pattern. In our simulation studies there are four configurations:

1. \( k = 1, \theta = 2, \beta = (1, 0, 0) \);
2. \( k = 1, \theta = 2, \beta = (0.333, 0.333, 0.333) \);
3. \( k = 2, \theta = 1, \beta = (1, 0, 0) \);
4. \( k = 2, \theta = 1, \beta = (0.333, 0.333, 0.333) \);
It is evident that in the first two cases the average length of censoring intervals tends to be longer than that in the other two, which implies more information loss from censoring. The first and third cases represent one large covariate effect while the other two many smaller covariate effects. They are used to compare the NPMLE with the Lasso estimate.

The sample size is always 100.

All the programs are implemented in C on a SunSparc 10. The starting value for the baseline distribution is the discrete uniform in the possible support of the NPMLE, whereas it is zero for the regression coefficient. An algorithm is stopped when both the log-likelihood increment and the change of the regression coefficient from two consecutive iterations are less than $10^{-5}$.

3.1 NPMLE and Bootstrap

For cases 1 and 3, suppose we know the true model; in other words, we know there is only one non-zero coefficient $\beta_1$. The ICM is so fast that often it converges within several seconds. The fast speed of the ICM facilitates using some computing-intensive methods, such as bootstrap and cross-validation. The cross-validation is used in choosing the Lasso constraint $\gamma$. Now we apply bootstrap to estimate the variability of the NPMLE $\hat{\beta}$. For related work in the context of right-censored data see Burr (1992).

As reviewed in Huang and Wellner (1997), in practice there are two approaches to estimate the variance of $\hat{\beta}$: one is by using the observed information matrix; another is through the profile likelihood for low-dimensional $\beta$. As we pointed out earlier, for continuous time model without grouping the observed information matrix is often of too high-dimension, and sometimes is near singular. There is also no theoretical justification for their use. Bootstrapping seems a natural alternative, though we are not aware of any empirical studies yet, which probably results from lack of faster algorithms.

Our bootstrapping scheme is as discussed by Efron and Tibshirani (1986), where the same number of observations are randomly taken with replacement from the original sample to form a bootstrap sample. We used 1000 bootstrap samples. The results are reported in Table 1. Noticeably the NPMLE is biased upwards and with a large variability for case 1. The reason is too much information loss from censoring, particularly with an overwhelming 60% left-censoring (and 10% right-censoring approximately). For a left-censored observation $(0, V_i, Z_i)$ its contribution to the
Table 1: NPMLE of the regression coefficient, $\hat{\beta}$, and the bootstrap estimate of its standard error, $se$. Their means and Monte Carlo standard errors were calculated from 500 replications. $\beta = 1$.

<table>
<thead>
<tr>
<th>Case</th>
<th>mean($\hat{\beta}$)</th>
<th>se($\hat{\beta}$)</th>
<th>mean($se$)</th>
<th>se($se$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.58</td>
<td>0.89</td>
<td>1.01</td>
<td>0.66</td>
</tr>
<tr>
<td>3</td>
<td>1.12</td>
<td>0.29</td>
<td>0.33</td>
<td>0.13</td>
</tr>
</tbody>
</table>

log-likelihood is

$$L_i = \log \left\{ 1 - (1 - F(V_i))^{\exp(Z_i \beta)} \right\},$$

which will be maximized when $\beta$ tends to the infinity for a positive $Z_i$. Nevertheless, the bootstrap estimate of the variance of the NPMLE is still sensible. In case 3, the left-censoring percentage is reduced to 30% (and there is about 10% right-censoring). The bootstrap estimate improves along with the NPMLE. In overall we can see that the bootstrap works reasonably well.

3.2 NPMLE and Lasso Estimate

In this subsection we compare the performance of the NPMLE and the Lasso estimate of the regression coefficient. Unlike in the last subsection, now all three covariates are used; in other words, we do not know in advance which $\beta_i$ is zero.

Our results are presented in Table 2, which are pretty similar to that of Tibshirani (1996) for ordinary least squares regression. The Lasso estimates tend to have larger biases (down towards zero) but much smaller variances, hence smaller mean squared errors (MSE), than the NPMLE. However, as the performance of the NPMLE improves in cases 3 and 4, the Lasso estimate is only on a slight edge.

3.3 A Comparison with Other Algorithms

As pointed out earlier, Finkelstein’s Newton-Raphson algorithm is not practical to use when the sample size is large, whereas the profile likelihood approach (Huang 1996; Huang and Wellner 1995) is only applicable with low-dimensional regression coefficients. Hence the only left competitor is the algorithm outlined in Huang and Wellner (1997). Essentially their method is to alternate
Table 2: NPMLEs and Lasso estimates of regression coefficients (with their standard errors in parentheses) from 500 replications. \( \beta = (1, 0, 0) \) for Cases 1 and 3; \( \beta = (0.333, 0.333, 0.333) \) for other two.

<table>
<thead>
<tr>
<th>Case</th>
<th>MSE</th>
<th>( \hat{\beta}_1 )</th>
<th>( \hat{\beta}_2 )</th>
<th>( \hat{\beta}_3 )</th>
<th>MSE</th>
<th>( \hat{\beta}_1 )</th>
<th>( \hat{\beta}_2 )</th>
<th>( \hat{\beta}_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.82</td>
<td>1.31</td>
<td>0.00</td>
<td>0.01</td>
<td>0.30</td>
<td>0.71</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>(0.29)</td>
<td>(0.63)</td>
<td>(0.40)</td>
<td>(0.41)</td>
<td>(0.29)</td>
<td>(0.40)</td>
<td>(0.17)</td>
<td>(0.18)</td>
</tr>
<tr>
<td>2</td>
<td>0.53</td>
<td>0.44</td>
<td>0.43</td>
<td>0.44</td>
<td>0.27</td>
<td>0.16</td>
<td>0.15</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>(0.65)</td>
<td>(0.41)</td>
<td>(0.40)</td>
<td>(0.44)</td>
<td>(0.17)</td>
<td>(0.25)</td>
<td>(0.22)</td>
<td>(0.25)</td>
</tr>
<tr>
<td>3</td>
<td>0.25</td>
<td>1.14</td>
<td>0.00</td>
<td>0.00</td>
<td>0.21</td>
<td>0.68</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>(0.28)</td>
<td>(0.30)</td>
<td>(0.26)</td>
<td>(0.27)</td>
<td>(0.20)</td>
<td>(0.30)</td>
<td>(0.09)</td>
<td>(0.11)</td>
</tr>
<tr>
<td>4</td>
<td>0.23</td>
<td>0.39</td>
<td>0.38</td>
<td>0.39</td>
<td>0.21</td>
<td>0.17</td>
<td>0.17</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>(0.21)</td>
<td>(0.27)</td>
<td>(0.26)</td>
<td>(0.27)</td>
<td>(0.09)</td>
<td>(0.21)</td>
<td>(0.20)</td>
<td>(0.21)</td>
</tr>
</tbody>
</table>

maximizing the likelihood with respect to the baseline distribution and with the regression coefficients. In contrast, we propose to maximize the likelihood with respect to the baseline distribution and the regression coefficients simultaneously. The advantage of our algorithm can be explained intuitively as follows. Suppose we want to maximize a bivariate function, say \( f(x, y) \), with respect to \( x \) and \( y \). If possible, we would like to maximize \( f(x, y) \) with respect to \( x \) and \( y \) jointly, rather than maximize \( f(x, y) \) with respect to \( x \), then with \( y \) and repeat this process. We note that due to this connection between two algorithms it may appear that our proposal is a trivial extension of Huang and Wellner's. However, this connection is established largely by our formulation of the ICM as a GGP.

We compared these two algorithms in the configuration of Case 3. Both algorithms used the same starting values and convergence criteria as before. The results are shown in Table 3. It is apparent that our algorithm is over two times faster than Huang and Wellner's. In particular, the log-likelihoods at the convergence of the two algorithms are very close to each other; in fact, all the log-likelihood values at the convergence of our algorithm are slightly larger than their corresponding parts from Huang and Wellner's. This guarantees that our algorithm was not favorably stopped.
Table 3: Speed of our new and Huang and Wellner’s algorithm in computing the NPMLE in Case 3, with 500 replications. The CPU time is in seconds; Log-likelihood is that at convergence. All the corresponding standard errors are in parentheses.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>CPU time</th>
<th>Log-likelihood</th>
<th>$\hat{\beta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Huang-Wellner</td>
<td>1.36 (0.24)</td>
<td>-84.67098 (6.78609)</td>
<td>1.1154 (0.2903)</td>
</tr>
<tr>
<td>New</td>
<td>0.44 (0.08)</td>
<td>-84.67097 (6.786011)</td>
<td>1.1163 (0.2905)</td>
</tr>
</tbody>
</table>

earlier.

To investigate our algorithm’s (in)dependence on starting values, we used some random starting values for the first ten datasets in Case 3; each one is replicated for 100 times. The initial baseline hazard increment in each of its support intervals was taken as a uniform random number between 0.00001 and 0.1, while the starting regression coefficients were from a uniform distribution between -2 and 2. For each dataset, the algorithm all converged to the neighborhood of the same point: the estimated regression coefficients agreed up to the third decimal digit while the log-likelihood value was the same. However, more work is needed to clarify whether there are any local maxima to which the algorithm may converge, as one referee pointed out.

4. REAL EXAMPLES

Now we apply our extended ICM to reanalyze two real datasets via the NPMLE. They have already been discussed among others in Finkelstein (1986), and Huang (1994) and Huang and Wellner (1995). Finkelstein focuses on hypothesis testing while Huang (1994) and Huang and Wellner (1995) establish the asymptotic normality and asymptotic efficiency of the NPMLE of the regression coefficient.

The first dataset is from a tumorigenicity study, where 144 mice were assigned to either a germfree or a conventional environment. The survival time here is the tumor onset time, which cannot be observed directly. The presence of the tumor can only be detected at the time of the sacrifice or death of the animal. Hence the survival time is either left censored or right censored (i.e. interval-censoring case 1). If we take the germfree group as the baseline with covariate $Z = 0$ and another group with $Z = 1$, then the NPMLE $\hat{\beta}$ of the regression coefficient from the Cox model
Extending the ICM

is $\hat{\beta} = -0.68$ with estimated standard error 0.41 by bootstrap (with 1000 bootstrap replications). They are different from Huang’s results: -0.55 and 0.29. Huang uses the profile likelihood to compute the NPMLE and approximates its variance by asymptotics, which involves some density estimation. Our 95% bootstrap percentile confidence interval for $\beta$ is $[-1.26, 0.34]$, compared with Huang’s asymptotic normal approximation $[-1.12, 0.01]$.

The second example considered is the Breast Cosmesis Study dataset (Finkelstein 1986; Finkelstein and Wolfe 1985). Two medical treatments are given to 90 early breast cancer patients after tumorectomy: one is a mix of primary radiotherapy and adjuvant chemotherapy, and another is radiotherapy only. The interest of this clinical trial is to investigate which treatment has better long-term cosmetic effects. Interval-censoring results since the patients could be visited every 4 to 6 months and those living farther away from the clinic had even longer followup intervals. If we take the radiotherapy group as the baseline group, the NPMLE of the regression coefficient from the Cox model is $\hat{\beta} = 0.735$ with estimated standard error 0.35 by bootstrap. These results are not much different from the corresponding 0.795 and 0.29 from Huang and Wellner (1995). But the bootstrap inference seems more conservative than theirs. The 95% bootstrap percentile and asymptotic confidence intervals are respectively $[0.11, 1.49]$ and $[0.23, 1.36]$.

Further studies are needed to clarify these discrepancies.

5. SUMMARY

In computing the NPMLE of the distribution function for interval censored data, the dimension of the parameters is in the order of the sample size, which is often more than several hundreds. Hence the space and time overheads incurred by using the Hessian (or its approximation by a full matrix) are prohibitive. This motivated us to investigate algorithms that avoid using these large matrices. Clearly we are trading time for space. In particular, we do not claim that our extended ICM is faster than Finkelstein’s Newton-Raphson method.

The ICM is fast but still simple without using a high-dimensional matrix, such as the Hessian, for interval censored data with no covariates. Our formulation of the ICM as a GGP helps not only to establish its connection with the Newton-Raphson algorithm, but also to ease its extension to the Cox model and to support the Lasso. Its fast speed also facilitates combining its use with some computing-intensive methods, such as bootstrap and cross-validation. In particular, our limited
simulation studies seem to favor the Lasso estimates of the regression coefficients in terms of mean squared errors as in linear regression setting (Tibshirani, 1996).

Finally, we note that our extended ICM (with Levenberg-Marquardt adjustment) is readily applicable to truncated and interval censored data (Pan and Chappell 1998).

Acknowledgements

The author would like to thank Olvi Mangasarian for introducing the GGP; thanks also go to Rick Chappell for stimulating discussions and support. Many insightful comments from the Associate Editor and three anonymous referees dramatically improved the presentation.
References


