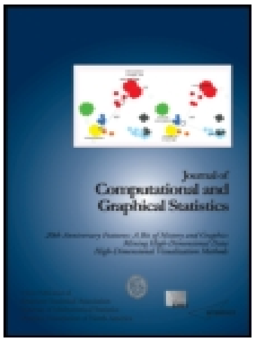

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Penalized Survival Models and Frailty

Terry M Therneau, Patricia M Grambsch & V. Shane Pankratz

Penalized Survival Models and Frailty

Terry M. THERNEAU, Patricia M. GRAMBSCH, and V. Shane PANKRATZ

Interest in the use of random effects in the survival analysis setting has been increasing. However, the computational complexity of such frailty models has limited their general use. Although fitting frailty models has traditionally been difficult, standard algorithms for fitting Cox semiparametric and parametric regression models can be readily extended to include penalized regression. We demonstrate that solutions for gamma shared frailty models can be obtained exactly via penalized estimation. Similarly, Gaussian frailty models are closely linked to penalized models. Fitting frailty models with penalized likelihoods can be made quite efficient by taking advantage of computational methods available for penalized models. We have implemented penalized regression for the coxph function of S-Plus and illustrate the algorithms with examples using the Cox model.

Key Words: Cox model; Penalized likelihood; Proportional hazards; Random effects.

1. INTRODUCTION

In the last several years there has been significant and active research concerning the addition of random effects to survival models. In this setting, a random effect is a continuous variable that describes excess risk or *frailty* for distinct categories, such as individuals or families, over and above any measured covariates. The idea is that individuals have different frailties, and that those who are most frail will die earlier than the others. Aalen (1988) provided theoretical and practical motivation for frailty models by discussing the impact of heterogeneity on analyses, and by illustrating how random effects can deal with it.

Frailties are useful in modeling correlations in multivariate survival and event history data. Examples include recurrent events such as epileptic seizures or depressive episodes, where an individual's frailty influences the occurrence of events, and community trials,

Terry M. Therneau is Division Chair, Division of Biostatistics, Mayo Clinic, Harwick 7, 200 First Street SW, Rochester, MN 55905 (E-mail: therneau@mayo.edu). Patricia M. Grambsch is Associate Professor, Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Building, MMC 303, 420 Delaware Street SE, Minneapolis, MN 55455 (E-mail: pat@biostat.umn.edu). V. Shane Pankratz is Senior Associate Consultant, Division of Biostatistics, Mayo Clinic, Harwick 7, 200 First Street SW, Rochester, MN 55905 (E-mail: pankratz.vernon@mayo.edu).

Journal of Computational and Graphical Statistics, Volume 12, Number 1, Pages 156–175
DOI: 10.1198/1061860031365

where the different events within each community share a common frailty. The simplest model, implicit in these examples, is the shared frailty model. In this model, all the units within each category share a common frailty, each unit belongs to precisely one category, and frailties of different categories are independent. More complex models are possible. Frailties can be nested; individuals within a family may share a common frailty, while families within communities share another common frailty. Frailties can also be correlated, as in studies of pedigrees. Due to its simplicity, we emphasize the shared frailty model here.

Frailties are usually viewed as unobserved covariates. This has led to the use of the EM algorithm as an estimation tool. However, the algorithm is slow, variance estimates require further computation, and no implementation has appeared in any of the more widely available packages.

Penalized models provide an alternate approach. The frailty terms are treated as additional regression coefficients which are constrained by a penalty function added to the log-likelihood. They are computationally similar to other shrinkage methods for penalized regression such as ridge regression, the lasso, and smoothing splines. Standard algorithms for fitting Cox semiparametric and parametric models can be simply extended to include penalty functions. These methods usually converge quickly and produce both point and variance estimates for model parameters.

The next section discusses the link between penalized estimation and frailty models. In particular, we demonstrate that if the frailty has a gamma distribution, then the shared frailty model can be written exactly as a penalized likelihood. We also show that Gaussian frailty models are closely linked to penalized models. We then turn to computational issues in implementing penalized techniques for fitting proportional hazard frailty models. We describe our S-Plus implementation and illustrate the algorithms with several examples.

2. FRAILTY MODELS

Assume that the data for subject i, who is a member of the jth of q families, follows a proportional hazards shared frailty model. The hazard can be written as

$$\lambda_i(t) = \lambda_0(t)\varpi_{j(i)} e^{\mathbf{X}_i \boldsymbol{\beta}}, \tag{2.1}$$

where $j(i)$ denotes that individual i belongs to family j, $\varpi_{j(i)} = \varpi_j$ is the frailty for family j, $\mathbf{X}$ is the covariate matrix of dimension n by p, and $\boldsymbol{\beta}$ is a vector of regression coefficients. The ϖ's are independent and identically distributed from some positive scale family with density function $f(\varpi; \theta)$, having variance θ and mean 1 for identifiability.

If the ϖ's are known, the complete data log-likelihood is

$$\sum_{i=1}^{n} \left[\int_0^\infty Y_i(t)[\log(\lambda_0(t)) + \log(\varpi_{j(i)}) + \mathbf{X}_i\boldsymbol{\beta}] dN_i(t) \right.$$
$$\left. - \int_0^\infty Y_i(t)\varpi_{j(i)} \exp(\mathbf{X}_i\boldsymbol{\beta})\lambda_0(t)dt + \log f(\varpi_{j(i)}; \theta) \right].$$

If the ϖ are viewed as missing data, the problem can be approached using the EM algorithm. Parner (1997) laid out a general framework. Let $\phi(s) = \phi(s, \theta)$ be the Laplace transform of the distribution of ϖ, and let $\phi^{(n)}(s)$ be its nth derivative with respect to s. Let $A_j = A_j(\boldsymbol{\beta}, \lambda_0) = \sum \int_0^\infty Y_i(s)\exp(\mathbf{X}_i\boldsymbol{\beta})d\Lambda_0(s)$, where the sum is over the members of family j, and let d_j be the number of events in the jth family. The log-likelihood of the observed data,

$$L_m(\boldsymbol{\beta}, \lambda_0; \theta) = \sum_{i=1}^{n} \delta_i \log\left(\int_0^\infty Y_i(t)e^{\mathbf{X}_i\boldsymbol{\beta}}\lambda_0(t)dt\right) + \sum_{j=1}^{q} \log[(-1)^{d_j}\phi^{(d_j)}(A_j)], \quad (2.2)$$

is found by integrating over the distribution of ϖ. For any fixed value of θ, Parner suggested maximizing this likelihood for $\boldsymbol{\beta}$ and λ_0 by an EM algorithm, which alternates between the following steps.

1. M-step. Treat the current estimate of ϖ as a fixed value or *offset*, and update $\boldsymbol{\beta}$ and λ_0 as in usual Cox regression. Note that for given $\boldsymbol{\beta}$ and ϖ,

$$d\hat{\Lambda}_0(t; \boldsymbol{\beta}, \varpi) = \sum dN_i(t) / \sum Y_i(t)\varpi_{j(i)}\exp(\mathbf{X}_i\boldsymbol{\beta}). \quad (2.3)$$

2. E-step. Compute ϖ as the expected value given the current values $\boldsymbol{\beta}$ and λ_0 and the data;

$$\varpi_j = -\frac{\phi^{(d_j+1)}(\hat{A}_j)}{\phi^{(d_j)}(\hat{A}_j)}, \quad (2.4)$$

where $\hat{A}_j = A_j(\boldsymbol{\beta}, \hat{\lambda}_0(\boldsymbol{\beta}, \omega))$.

Equations (2.2) and (2.4) are derived assuming a shared frailty model and unfortunately do not hold for more complex models. Parner suggested that estimation of θ be done by maximizing the profile log-likelihood

$$L_m(\theta) = L_m(\hat{\boldsymbol{\beta}}(\theta), \hat{\lambda}_0(\theta), \theta). \quad (2.5)$$

Although ϖ is not an explicit parameter of the observed log-likelihood, the EM algorithm provides an estimate of this vector.

The penalized regression formulation for the shared frailty model is most easily developed from an alternative version of the hazard,

$$\lambda_i(t) = \lambda_0(t)e^{\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\omega}, \quad (2.6)$$

which is equivalent to Equation (2.1). In this case, $\varpi_j = \exp(\omega_j)$, $\mathbf{Z}$ is matrix of q indicator variables such that $\mathbf{Z}_{ij} = 1$ when subject i is a member of family j and 0 otherwise, and each individual belongs to only one family. Estimation under this model is done by maximizing a penalized partial log-likelihood

$$\text{PPL} = \text{PL}(\boldsymbol{\beta}, \omega; \text{data}) - g(\omega; \theta)$$

over both $\boldsymbol{\beta}$ and ω. Here PL is the log of the usual Cox partial likelihood,

$$\text{PL}(\boldsymbol{\beta}, \omega) = \sum_{i=1}^{n} \int_0^\infty \left[Y_i(t)(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\omega) - \log\left\{\sum_k Y_k(t)\exp(\mathbf{X}_k\boldsymbol{\beta} + \mathbf{Z}_k\omega)\right\}\right] dN_i(t), \quad (2.7)$$

and g is a penalty function chosen by the investigatorto restrict the values of ω. The parameter θ is a tuning constant which may be prespecified or adapted to the data. Typically, one would choose the penalty function to "shrink" ω toward zero and use θ to control the amount of shrinkage.

To estimate $\boldsymbol{\beta}$ and ω, one solves the score equations. Because the penalty function does not involve $\boldsymbol{\beta}$, $\partial\text{PPL}/\partial\boldsymbol{\beta} = \partial\text{PL}/\partial\boldsymbol{\beta}$. Therefore, the score equations for $\boldsymbol{\beta}$ are identical to those for an ordinary Cox model treating $\mathbf{Z}\omega$ as an offset term. If we define

$$\mathbf{Z}_j(t) = \mathbf{Z}_j(\boldsymbol{\beta}, \omega, t) = \frac{\sum \mathbf{Z}_{ij} Y_i(s) \exp[\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\omega]}{\sum Y_i(s) \exp[\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\omega]}, \tag{2.8}$$

then

$$\frac{\partial\text{PPL}}{\partial\omega_j} = \sum_{i=1}^{n} \int_0^\infty (\mathbf{Z}_{ij} - \mathbf{Z}_j(t)) dN_i(t) - \frac{\partial g(\omega;\theta)}{\partial\omega_j}. \tag{2.9}$$

Recall that for given $\boldsymbol{\beta}$ and ω, the Breslow estimator of the underlying hazard is

$$d\hat{\Lambda}_0(t;\boldsymbol{\beta},\omega) = \sum dN_i(t) / \sum Y_i(t) \exp(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\omega),$$

which is just Equation (2.3) in different notation. Let $\hat{\lambda}_i = \hat{\lambda}_i(\boldsymbol{\beta},\omega) = \int_0^\infty Y_i(s) d\hat{\Lambda}_0(s;\boldsymbol{\beta},\omega)$. Simple algebra shows that the score equation for ω_j is

$$\frac{\partial\text{PPL}}{\partial\omega_j} = \sum_{i=1}^{n} \left[\mathbf{Z}_{ij}\delta_i - \mathbf{Z}_{ij}\hat{\lambda}_i e^{\mathbf{X}_i\boldsymbol{\beta}+\mathbf{Z}_i\omega}\right] - \frac{\partial g(\omega;\theta)}{\partial\omega_j} = 0. \tag{2.10}$$

Because of the structure of the matrix $\mathbf{Z}$, this equation simplifies to

$$\frac{\partial\text{PPL}}{\partial\omega_j} = \left[d_j - \hat{A}_j e^{\omega_j}\right] - \frac{\partial g(\omega;\theta)}{\partial\omega_j} = 0, \tag{2.11}$$

where d_j and $\hat{A}_j$ are as defined earlier.

The penalized likelihood can be fitted with the Newton–Raphson algorithm. In addition to the score vectors $\partial\text{PPL}/\partial\boldsymbol{\beta}$ and $\partial\text{PPL}/\partial\omega$, this requires the Hessian of the penalized partial log-likelihood:

$$\mathbf{H} = \mathbf{H}(\boldsymbol{\beta},\omega) = \mathcal{I} + \begin{pmatrix} 0 & 0 \\ 0 & g'' \end{pmatrix}, \tag{2.12}$$

where $\mathcal{I} = \mathcal{I}(\boldsymbol{\beta},\omega)$ is the usual Cox model information matrix, the second derivative matrix of the PL with respect to $\boldsymbol{\beta}$ and ω.

2.1 Gamma Frailty

Details of the EM approach for the shared gamma frailty model can be found in Nielsen, Gill, Andersen, and Srensen (1992) and Klein (1992). Equations (2.4) and (2.2) can be used to rederive their results, and help make the connection to penalized methods. Here we demonstrate that for any fixed θ, the penalized log-likelihood with appropriate choice of penalty function and the observed-data log-likelihood in Equation (2.2) have the same solution.

Let the frailty have a gamma distribution with mean 1 and variance $\theta = 1/\nu$. The density of ϖ can be written as

$$\log[f(\varpi;\nu)] = (\nu - 1)\log(\varpi) - \nu\varpi + \nu\log(\nu) - \log\Gamma(\nu).$$

This has a Laplace transform of $\phi(s) = (1 + s/\nu)^{-\nu}$. The derivatives of $\phi(s)$ are

$$\phi^{(d)}(s) \;=\; \left(-\frac{1}{\nu}\right)^{d}\left(1+\frac{s}{\nu}\right)^{-(\nu+d)}\prod_{i=0}^{d-1}(\nu+i)\,,$$

and Equation (2.4) reduces to

$$e^{\omega_j} \;=\; \frac{d_j+\nu}{\hat{A}_j+\nu}\,. \tag{2.13}$$

Lemma 1. *The solution to the penalized partial likelihood model, with penalty function* $g(\omega;\theta) = -1/\theta\sum_{j=1}^{q}[\omega_j - \exp(\omega_j)]$, *coincides with the EM solution for any fixed value of* θ.

Proof: For $\boldsymbol{\beta}$, the EM and penalized methods have the same score equation, which includes $\mathbf{Z}\omega$ as a fixed offset. Thus, if the solutions for ω are the same, those for $\boldsymbol{\beta}$ will be also. Let $(\hat{\boldsymbol{\beta}},\hat{\omega})$ be a solution to the the EM process. Then $\hat{\omega}$ must satisfy Equation (2.13) exactly, not just as an update step. Rearranging terms, we see that $\hat{A}_j = \exp(-\hat{\omega}_j)(d_j+\nu) - \nu$. Substituting this into the penalized score equation and simplifying with $\nu = 1/\theta$ a fixed quantity, we see that

$$\begin{aligned}
\frac{\partial \mathrm{PPL}(\hat{\boldsymbol{\beta}},\hat{\omega})}{\partial\hat{\omega}_j} &= \left[d_j - \hat{A}_j e^{\hat{\omega}_j}\right] - \frac{\partial g(\hat{\omega};\theta)}{\partial\hat{\omega}_j}\\
&= \left[d_j - e^{-\hat{\omega}_j}\left(d_j + \tfrac{1}{\theta} - \tfrac{1}{\theta}e^{\hat{\omega}_j}\right)e^{\hat{\omega}_j}\right] + \tfrac{1}{\theta}(1 - e^{\hat{\omega}_j})\\
&= 0.
\end{aligned}$$

This shows that the solution to the EM algorithm is also a solution to the penalized score equations. Therefore, for any fixed θ, the penalized log-likelihood and the observed-data log-likelihood in Equation (2.2) have the same solution, although these two equations are *not* equal to one another. □

Furthermore, if we let $\mathrm{PPL}(\theta) = \mathrm{PPL}(\hat{\boldsymbol{\beta}}(\theta),\hat{\omega}(\theta),\theta)$, then we can write Equation (2.5), the profile log-likelihood for θ, as $\mathrm{PPL}(\theta)$ plus a correction that only involves θ and the d_j's. Using the fact that each row of $\mathbf{Z}$ has exactly one 1 and $q-1$ 0's, we see that the Cox PL for $(\hat{\boldsymbol{\beta}},\hat{\omega})$ must be the same as that for $(\hat{\boldsymbol{\beta}},\hat{\omega}+c)$ for any constant c. Simple algebra shows that the value of c which minimizes the penalty portion of the PPL is such that

$$\sum_{i=1}^{q} e^{\hat{\omega}_j} = q\,. \tag{2.14}$$

Using the identities in Equations (2.13) and (2.14), recalling that they hold only at the solution point, we show in the appendix that

$$L_m(\theta) = \text{PPL}(\theta) + \sum_{j=1}^{q} \nu - (\nu + d_j)\log(\nu + d_j) + \nu \log \nu + \log\left(\frac{\Gamma(\nu + d_j)}{\Gamma(\nu)}\right). \quad (2.15)$$

It is useful to consider $L_m(\theta) + \sum_{j=1}^{q} d_j$, rather than $L_m(\theta)$, because the profile log-likelihood converges to $\text{PL}(\hat{\beta}) - \sum d_j$ as the variance of the random effect goes to zero. Adding $\sum d_j$ to $L_m(\theta)$ makes the maximized marginal likelihood from a frailty model with small θ comparable to the maximized likelihood from a nonfrailty model.

The fitting program for a shared gamma frailty consists of an inner and outer loop. For any fixed θ, Newton–Raphson iteration is used to solve the penalized model in a few (usually 3–5) steps, and return the corresponding value of the PPL. The outer loop chooses θ to maximize the profile likelihood in Equation (2.15), which is easily done as it is a unimodal function of one parameter.

All of the results presented in this section were dependent on the correct choice of a penalty function. For gamma frailties, the penalty function that links the penalized and EM results is directly related to the density of the random effect; the log of the density for ω, where $\exp(\omega)$ has a gamma distribution, is equal to $[\omega - \exp(\omega)]/\theta$ plus additional terms not involving ω. Similarly, the penalty we use for a Gaussian frailty is related to a log-density, as discussed in the next section.

2.2 GAUSSIAN FRAILTY

McGilchrist and Aisbett (1991, 1993), suggested a Gaussian density for ω in a shared frailty model. This leads to the penalized partial likelihood

$$\text{PPL} = \text{PL} - (1/2\theta) \sum_{j=1}^{q} \omega_j^2, \quad (2.16)$$

where θ is the variance of the random effect.

The authors did not provide an exact connection to the marginal likelihood that can be used to choose the variance parameter θ. Instead, they noted the similarity of the Cox model's Newton–Raphson step to an iteratively reweighted least-squares calculation. Using this observation, they proposed using standard estimators from Gaussian problems. This leads to choosing θ such that it satisfies

$$\theta = \frac{\sum_{j=1}^{q} \omega_j^2 + r}{q}. \quad (2.17)$$

The value of r varies depending on the estimation technique used. For BLUP, $r = 0$; for MLE, $r = \text{trace}[(\mathbf{H}_{22})^{-1}]$; and for REML, $r = \text{trace}[(\mathbf{H}^{-1})_{22}]$, where $\mathbf{H}$ is the Hessian of the penalized partial log likelihood in Equation (2.12) and $\mathbf{H}_{22}$ is the lower right $q \times q$ submatrix corresponding to the random effects.

The Gaussian approach was justified and expanded by Ripatti and Palmgren (2000). Let the random effects have a positive definite covariance matrix $\mathbf{D} = \mathbf{D}(\theta)$. This provides a rich class of models for the random effects; for example, setting $\mathbf{D} = \theta I$ results in a shared frailty model. The marginal log-likelihood is

$$L_m(\boldsymbol{\beta}, \theta) = -1/2 \log |\mathbf{D}| + \log \left\{ \int \exp[\mathrm{PL}(\boldsymbol{\beta}, \omega) - 1/2 \omega' \mathbf{D}^{-1/2} \omega] d\omega \right\}.$$

Note that, unlike Parner's approach, this marginal log-likelihood does not involve $\lambda_0(t)$; that has already been partialed out to give the Cox partial log-likelihood. Following the methods of Breslow and Clayton (1993), Ripatti and Palmgren (2000) used a Laplace approximation to the above integral to get an approximate marginal log-likelihood.

$$L_m(\boldsymbol{\beta}, \theta) \approx \mathrm{PL}(\boldsymbol{\beta}, \tilde{\omega}) - 1/2 \left(\tilde{\omega}' \mathbf{D}^{-1} \tilde{\omega} + \log |\mathbf{D}| + \log |\mathbf{H}(\boldsymbol{\beta}, \tilde{\omega})_{22}| \right) , \qquad (2.18)$$

where $\tilde{\omega} = \tilde{\omega}(\boldsymbol{\beta}, \theta)$ solves

$$\sum_{i=1}^{n} \int_0^\infty (\mathbf{Z}_{ij} - \mathbf{Z}_j(t)) dN_i(t) - D(\theta)^{-1} \tilde{\omega} = 0$$

which is comparable to Equation (2.9). The first two terms of the approximate marginal log-likelihood correspond to a penalized partial likelihood with $g(\omega; \theta) = 1/2 \tilde{\omega}' \mathbf{D}(\theta)^{-1} \tilde{\omega}$. This reduces to Equation (2.16) in the case of a shared frailty model. We can ignore the third term of Equation (2.18) as $\mathbf{D}$ is constant for fixed θ. Ignoring the fourth term can influence the estimates. However, Ripatti and Palmgren explored the performance of the estimators via simulation for models with and without fixed effects terms. The simulation results suggest that the loss of information is slight.

As shown by Ripatti and Palmgren (2000), the estimating equation for θ_j is

$$\mathrm{trace} \left[\mathbf{D}^{-1} \frac{\partial \mathbf{D}}{\partial \theta_j} \right] + \mathrm{trace} \left[(\mathbf{H}_{22})^{-1} \frac{\partial \mathbf{D}^{-1}}{\partial \theta_j} \right] - \omega' \mathbf{D}^{-1} \frac{\partial \mathbf{D}}{\partial \theta_j} \mathbf{D}^{-1} \omega = 0 \,. \qquad (2.19)$$

The Fisher information matrix, obtained by taking the expectation with respect to ω, has a jk element of

$$\begin{aligned} &(1/2) \quad \mathrm{trace} \left[\mathbf{D}^{-1} \frac{\partial \mathbf{D}}{\partial \theta_j} \mathbf{D}^{-1} \partial \theta_k + \mathbf{D}^{-1} \frac{\partial^2 \mathbf{D}}{\partial \theta_j \partial \theta_k} \right] \\ +&(1/2) \quad \mathrm{trace} \left[(\mathbf{H}_{22})^{-1} \frac{\partial \mathbf{D}}{\partial \theta_j} (\mathbf{H}_{22})^{-1} \frac{\partial \mathbf{D}}{\partial \theta_k} - (\mathbf{H}_{22})^{-1} \frac{\partial^2 \mathbf{D}^{-1}}{\partial \theta_j \partial \theta_k} \right] . \qquad (2.20) \end{aligned}$$

For the shared frailty model the estimating equation reduces to

$$\hat{\theta} = \frac{\omega' \omega + \mathrm{trace}[(\mathbf{H}_{22})^{-1}]}{q} ,$$

which is equivalent to the MLE formula of McGilchrist (1993).

Yau and McGilchrist (1998) displayed a similar formula for the ML estimate for an arbitrary correlation matrix $\mathbf{D}$, and apply the results to the CGD dataset using an AR(1) structure for the multiple infections within subject. (Unfortunately, differences in how ties are handled make it impossible to replicate their fits). In that article, they also defined an REML estimate, which is identical to Equations (2.19) and (2.20) above, but with $(\mathbf{H}^{-1})_{22}$ replacing $(\mathbf{H}_{22})^{-1}$. Additionally, their simulations show Equation (2.20) to be an overestimate of the actual standard error.

3. COMPUTATIONAL ISSUES

Thus far, we have discussed the relationship between frailty models and penalized likelihood estimation. This section describes several issues important to the computational implementation of penalized likelihood methods for Cox models with random effects.

3.1 PENALIZED LIKELIHOOD INFERENCE

Consider a Cox model with both constrained and unconstrained effects, as shown in Equation (2.6). The model is fit by maximizing the penalized partial log-likelihood (PPL). We assume that θ is fixed. Consider testing the set of hypotheses $\mathbf{z} = \mathbf{C}(\boldsymbol{\beta}', \boldsymbol{\omega}')' = 0$, where $(\boldsymbol{\beta}', \boldsymbol{\omega}')'$ is the combined vector of $p + q$ parameters, and $\mathbf{C}$ is a $k \times p + q$ matrix of full row rank k, $k \leq p + q$. Gray (1992) suggested that

$$\mathbf{V} = \mathbf{H}^{-1}\mathcal{I}\mathbf{H}^{-1} \tag{3.1}$$

be used as the covariance estimate of the parameter estimates. He recommends a Wald type test statistic, $\mathbf{z}'(\mathbf{C}\mathbf{H}^{-1}\mathbf{C}')^{-1}\mathbf{z}$, with generalized degrees of freedom

$$\text{df} = \text{trace}[(\mathbf{C}\mathbf{H}^{-1}\mathbf{C}')^{-1}\,(\mathbf{C}\mathbf{V}\mathbf{C}')]\,.$$

The total degrees of freedom for the model ($\mathbf{C} = I$) simplifies to

$$\begin{aligned}\text{df} &= \text{trace}[\mathbf{H}\mathbf{V}] \\ &= \text{trace}[\mathbf{H}(\mathbf{H}^{-1}(\mathbf{H} - \mathbf{G})\mathbf{H}^{-1})] \\ &= (p + q) - \text{trace}[\mathbf{G}\mathbf{H}^{-1}].\end{aligned} \tag{3.2}$$

Under H_0, the distribution of the test statistic is asymptotically the same as $\sum e_i \mathbf{X}_i^2$, where the e_i are the k eigenvalues of the matrix $(\mathbf{C}\mathbf{H}^{-1}\mathbf{C}')^{-1}\,(\mathbf{C}\mathbf{V}\mathbf{C}')$ and the $\mathbf{X}_i$ are iid standard Gaussian random variables. In nonpenalized models, the e_i are all either 0 or 1, and the test statistic has an asymptotic chi-square distribution on $\sum e_i$ degrees of freedom. In penalized models, the test statistic has mean $\sum e_i$ and variance $2\sum e_i^2 < 2\sum e_i$ because $0 \leq e_i \leq 1$. Using a reference chi-square distribution with df $= \sum e_i$ will tend to be conservative.

Verweij and Van Houwelingen (1994) discussed penalized Cox models in the context of restricting the parameter estimates. They use $\mathbf{H}^{-1}$ as a "pseudo standard error," and an "effective degrees of freedom" identical to Equation (3.2). With this variance matrix, the test statistic $\mathbf{z}'(\mathbf{C}\mathbf{H}^{-1}\mathbf{C}')^{-1}\mathbf{z}$ is a usual Wald test. To choose an optimal model they recommended either the Akaike information criterion (AIC) which uses the degrees of freedom described above or the cross-validated (partial) log-likelihood CVL, which uses a degrees of freedom estimate based on a robust variance estimator.

Our algorithm makes both $\mathbf{H}^{-1}$ and $\mathbf{H}^{-1}\mathcal{I}\mathbf{H}^{-1}$ available. Significance tests are based on $\mathbf{H}^{-1}$ as the more conservative choice. Simulation experiments for the related problem of penalized smoothing splines in Cox regression (not shown) suggest that this is the more reliable choice for tests, but we do not have more definitive results to support this.

In our implementation, the computation of the degrees of freedom and variance matrices are specialized to avoid any intermediate steps that would give a q by q result, where q is the number of constrained coefficients.

3.2 SPARSE COMPUTATION

When performing estimation with frailty models, memory and time considerations can become an issue. For instance, if there are 300 families, each with a frailty term, and 4 other variables, then the full information matrix has $304^2 = 92{,}416$ elements. The Cholesky decomposition must be applied to this matrix within each Newton–Raphson iteration. In our S-Plus implementation, we have applied a technique that can provide significant savings in space and time.

If we partition the information matrix of a Cox shared frailty model according to the rows of $\mathbf{X}$ and $\mathbf{Z}$, and arrange the matrix as

$$\mathcal{I} = \begin{pmatrix} \mathcal{I}_{\mathbf{ZZ}} & \mathcal{I}_{\mathbf{ZX}} \\ \mathcal{I}_{\mathbf{XZ}} & \mathcal{I}_{\mathbf{XX}} \end{pmatrix},$$

then the upper left corner will be a diagonally dominant matrix, having almost the form of the variance matrix for a multinomial distribution. Adding the penalty further increases the dominance of the diagonal. Therefore, using a *sparse* computation option, where only the diagonal of $\mathcal{I}_{\mathbf{ZZ}}$ is retained, should not have a large impact on the estimation procedure.

Ignoring a piece of the full information matrix has a number of implications. First, the speed of the Cholesky factorization is increased dramatically. Second, the savings in space can be considerable. If we use the sparse option with the example above, the information matrix consists of only $\mathcal{I}_{\mathbf{XZ}}$ and $\mathcal{I}_{\mathbf{XX}}$, with $304 * 4 = 1{,}216$ elements, along with the 300 element diagonal of $\mathcal{I}_{\mathbf{ZZ}}$, a savings of over 95% in memory space. Third, because the score vector and likelihood are not changed, the solution point is identical to the one obtained in the nonsparse case, discounting trivial differences due to distinct iteration paths. Fourth, the Newton–Raphson iteration may undergo a slight loss of efficiency so that 1–2 more iterations are required. However, because each N–R iteration requires the Cholesky decomposition of the information matrix, the sparse problem is much faster per-iteration than the full matrix version. Finally, the full information matrix is a part of the formulas for the post-fit estimates of degrees of freedom and standard error. In a small number of simple examples, the effect of the sparse approximation on these estimates has been surprisingly small.

We have found two cases where our sparse method does not perform acceptably. The first is if the variance of the random effect is quite large (>5). In this case, each N–R step may require a large number (>15) of iterations. The second is if one group contains a majority of the observations. The off diagonal terms are too important to ignore in this case, and the approximate N–R iteration does not converge.

4. EXAMPLES

We now present two examples where we use our S-Plus functions to obtain estimates from frailty models. The first deals with the survival of kidney catheters. The second examines the effect of UDCA in patients with primary biliary cirrhosis.

4.1 Survival of Kidney Catheters

The following dataset was presented by McGilchrist and Aisbett (1991). Each observation is the time to infection, at the point of insertion of the catheter, for kidney patients using portable dialysis equipment. Catheters may be removed for reasons other than infection, in which case the observation is censored. There are 38 patients, each with exactly 2 observations. Variables are the subject id, age, sex ($1 =$ male, $2 =$ female), disease type (glomerulo nephritis, acute nephritis, polycystic kidney disease, and other), and the time to infection or censoring for each insertion. We first fit two ordinary Cox models, followed by a gamma frailty fit.

```
> kfit1 <- coxph(Surv(time, status) ~ age + sex, data=kidney)
> kfit2 <- coxph(Surv(time, status) ~ age + sex + disease,
                data=kidney)
> kfit3 <-  coxph(Surv(time, status) ~ age + sex + disease +
                                      frailty(id), data=kidney)
> kfit3
               coef se(coef)    se2 Chisq DF       p
       age  0.00318 0.0111   0.0111  0.08 1   7.8e-01
       sex -1.48314 0.3582   0.3582 17.14 1   3.5e-05
 diseaseGN  0.08796 0.4064   0.4064  0.05 1   8.3e-01
 diseaseAN  0.35079 0.3997   0.3997  0.77 1   3.8e-01
diseasePKD -1.43111 0.6311   0.6311  5.14 1   2.3e-02
frailty(id)                          0.00 0   9.5e-01

Iterations: 6 outer, 29 Newton-Raphson
Penalized terms:
     Variance of random effect= 1.47e-07   M-likelihood = -179.1
Degrees of freedom for terms= 1 1 3 0
Likelihood ratio test=17.6  on 5 df, p=0.00342  n= 76
```

Many of the labels in this output are self-explanatory. The `se(coef)` estimates are from the diagonal elements of $\mathbf{H}^{-1}$, and `se2` uses the diagonal entries in $\mathbf{H}^{-1}\mathcal{I}\mathbf{H}^{-1}$. In this particular dataset, they are identical, but that is not always the case. The `M-likelihood` is the marginal likelihood in Equation (2.15), evaluated at the MLE.

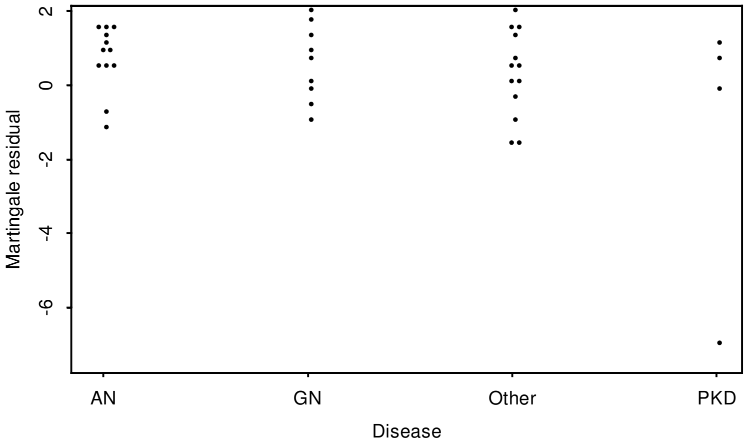

Figure 1. *Residuals for the kidney data from model kfit1.*

The partial log-likelihood values for first two models are -184.3 and -179.1, with 2 and 5 degrees of freedom respectively. Hence, the disease variable is a significant addition. In the third fit, the program provided an estimate of the MLE of θ, the variance of the random effect, that was essentially 0.

When the disease variable is left out of the random effects model, however, we get a quite different result.

```
> kfit4 <- coxph(Surv(time, status) ~ age + sex + frailty(id),
        data=kidney)
> kfit4
                coef se(coef)    se2 Chisq   DF       p
        age  0.00522 0.0119   0.0088  0.19  1.0 0.66000
        sex -1.58335 0.4594   0.3515 11.88  1.0 0.00057
frailty(id)                          22.97 12.9 0.04100

Iterations: 7 outer, 49 Newton-Raphson
     Variance of random effect= 0.408   M-likelihood = -181.6
Degrees of freedom for terms=  0.6  0.6 12.9
Likelihood ratio test=46.6  on 14.06 df, p=2.36e-05  n= 76
```

In this case, both the approximate Wald test and the likelihood ratio test indicate that the variance of the random effect is greater than zero. The Wald test shown in the printout is not as accurate as the the comparison of the marginal likelihood to that from `kfit1` (-184.3 vs. -181.6), which gives a chi-square statistic of 5.4 on 1 degree of freedom for a p value of 0.02. As discussed by Nielsen et al. (1992), this chi-square test for θ is not affected by the boundary at zero.

Figure 1 shows the reason for the discrepancy of the results between the two models. The graph shows the martingale residuals for each subject (the sum of the residuals from the two observations), based on the simplest model, `kfit1`. Note the outlier in the lower right, corresponding to a 46-year-old male whose age was quite close to the median for the study (45.5 years). There were 10 males and most had early failures: 2 observations were censored at 4 and 8 days, respectively, and the remaining 16 male kidneys had a median time to infection of 19 days. Subject 21, however, had failures at 154 and 562 days. With this subject removed, neither the disease ($p = 0.53$) nor the frailty ($p > 0.9$) are important. With this subject in the model, it is a toss-up whether the disease or the frailty term will be credited with `significance'. Using a Gaussian frailty with REML gives partial importance to each.

```
> mfit1 <- coxph(Surv(time,status) ~ age + sex + disease +
                  frailty(id, dist=´gauss´, sparse=F), data=kidney)
> mfit1
                               coef se(coef)     se2 Chisq   DF       p
                        age  0.00492 0.0149   0.0108  0.11  1.0 0.74000
                        sex -1.70204 0.4631   0.3613 13.51  1.0 0.00024
                  diseaseAN  0.39442 0.5428   0.4052  0.53  1.0 0.47000
                  diseaseGN  0.18173 0.5413   0.4017  0.11  1.0 0.74000
                 diseasePKD -1.13160 0.8175   0.6298  1.92  1.0 0.17000
frailty(id, dist = "gauss                             18.13 12.3 0.12000

Iterations: 6 outer, 17 Newton-Raphson
     Variance of random effect= 0.509   M-likelihood = -171.9
Degrees of freedom for terms=  0.5  0.6  1.7 12.3
Likelihood ratio test=118  on 15.14 df, p=0  n= 76
```

This dataset has been used as an example in number of articles, unfortunately not always with an acknowledgment of the influence of this outlier. Even so, it remains useful as an example of large correlation between fixed (sex) and random effects, and as an illustration of the potential impact of an outlier on the various distributions for θ. Although it is useful for illustrating the application of frailty models, it is not a good test case for making a general comparison of the gamma and Gaussian frailty models.

The sparse routines have some impact on the solution for a Gaussian model, since the REML estimate depends on the matrix $\mathbf{H}$. Using the `sparse=T` option in the frailty function, the routine required 32 Newton–Raphson iterations and gave a solution of $\theta = 0.493$, but with about one third the total computing time.

The standard error estimates reported by a penalized coxph model in S-Plus are computed under the assumption that θ is fixed. For some models, such as a smoothing spline with user specified degrees of freedom, this assumption is correct. For the above frailty models it clearly is not and the standard errors are an underestimate. Using the bootstrap, we found the standard error to be *much* higher for this dataset, which is not surprising given the inordinate influence of a single subject. More useful bootstrap results appear in the second example.

These answers differ slightly from the original author's (McGilchrist 1993) results.

Table 1. Total Number of Events in the UDCA Trial

	UDCA	*Placebo*
Death	6	10
Transplant	6	6
Drug toxicity	0	0
Voluntary withdrawal	11	18
Histologic progression	8	12
Development of varices	8	17
Development of ascites	1	5
Development of encephalopathy	3	1
Doubling of bilirubin	2	15
Worsening of symptoms	7	9

That article presented formulas that are completely valid only for untied data, and this dataset has five tied pairs and one quadruple. This is a small proportion of the data, and in a standard Cox model the ties would barely perturb the answers. Unfortunately, the REML solution for θ can be very sensitive to small changes in the data.

4.2 UDCA in Patients With PBC

Primary biliary cirrhosis (PBC) is a chronic cholestatic liver disease characterized by progressive destruction of the bile ducts. PBC frequently progresses to cirrhosis, which may lead to death from liver failure unless liver transplant is offered—an extensive and costly procedure. Trials have been held for several promising agents, but an effective therapy remains elusive. Although progression of disease is inexorable the time course can be very long; many patients survive 10 or more years from their initial diagnosis before requiring a transplant.

A randomized double-blind trial of a new agent, ursodeoxycholic acid (UDCA), was conducted at the Mayo Clinic from 1988 to 1992 and enrolled 180 patients. The data were reported in Lindor et al. (1994); the analysis shown here has slightly longer follow-up. The endpoints of the study were predefined and are shown in Table 1. While the event endpoints are all unique, for example, no single patient had more than one instance of death, patients could experience more than one of the different endpoints, for example, both transplant and death. The primary report was based on an analysis of time to the first event; 58/84 placebo and 34/86 UDCA patients have at least one event. An analysis that used all of the events data would seem to be more complete, being based on all 93 placebo and 52 UDCA events. Three possible methods of analysis present themselves to deal with these correlated outcomes. The simplest is to consider time to the first adverse event. In this case, each patient has a single observation and correlation is not an issue. The second is a marginal analysis in the manner of Lin (1994). The third involves the use of a frailty model. The dataset for the latter two options is essentially a concatenation of the nine individual datasets that would be created for an analysis of time to death (censoring all other causes), time to transplant, time to withdrawal, and so on, with the event type as a stratification variable.

Table 2. Results of Three Models for the UDCA Data

	β	*se(β)*	*robust se*
		First event	
treatment	−0.94	0.22	0.22
bilirubin	0.74	0.24	0.23
stage	−0.02	0.25	0.25
		Marginal model	
treatment	−0.80	0.17	0.23
bilirubin	0.77	0.18	0.25
stage	0.05	0.21	0.28
		Frailty model	
treatment	−0.96	0.28	0.32
bilirubin	0.74	0.31	0.32
stage	0.31	0.32	0.35

Three covariates were used in each of these models. The first was an indicator for treatment by UDCA, and the other two were indicators for the presence of two of the stratification factors used in treatment assignment. The resulting parameter estimates and their standard errors are shown in Table 2. The robust standard error estimates for the frailty model were obtained from 1,000 bootstrap realizations where θ was fixed at the original model's estimate. They show the ordinary standard error to be quite reliable as an estimate when θ is incorrectly fixed in advance. The bootstrap standard error estimates obtained when θ was estimated were higher than those shown in Table 2, being 0.42, 0.51, and 0.50 for treatment, bilirubin, and stage, respectively.

Upon examination of Table 2, two outcomes are immediately obvious. First, the naive variance is an underestimate in the multiple event model; accounting for the within-patient correlation is important. Second, the multiple-event robust variances and the frailty variance estimates are slightly larger than the variances for first events only. The use of multiple events added no information to the analysis!

A closer look at the data reveals the cause of the difficulty. Patients participating in the study returned for evaluation once a year, which is the point at which most of the outcomes were measured. For instance, one patient had five events, four of which were recorded on July 20, 1990. The fifth, death, occurred on July 22. Similar outcomes are seen for many others. Figure 2 shows the event times for the 31 subjects with multiple adverse outcomes, with a circle marking each event. The data has been jittered slightly to avoid overlap. It appears that the use of multiple event types was useful in this study only to make the detection of "liver failure" more sensitive. Given that failure has occurred, the number of positive markers for failure was irrelevant.

In this situation we expect the frailty model to show significant within patient correlation, and indeed this is the case. The variance of the random effect is estimated as 1.47 in a shared gamma frailty model and is highly significant ($\chi^2 = 31.6$ on 1 df). The estimated value of Kendall's τ is $\theta/(2+\theta) = 0.42$.

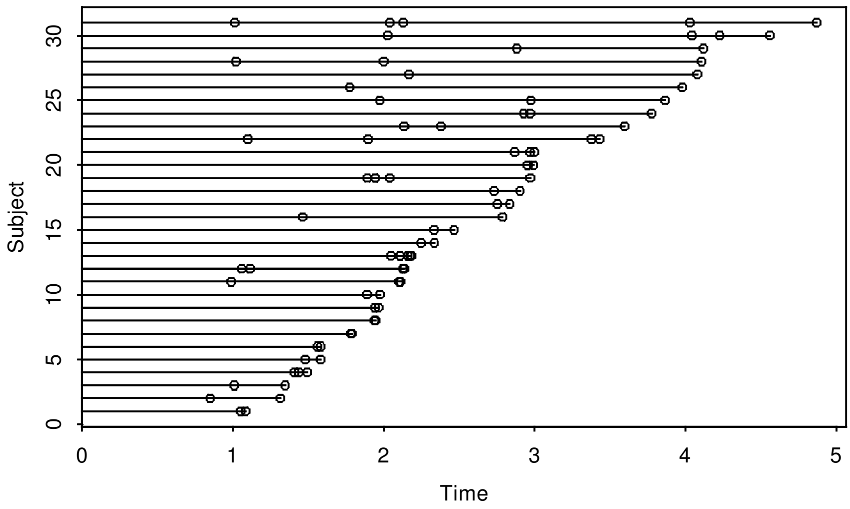

Figure 2. *Multiple failure times for the UDCA data.*

4.3 COMPUTATION TIME

The penalized likelihood approach provides a number of programming efficiencies. It allows a single central S language function to handle gamma and Gaussian frailty models, as well as other regression methodologies such as smoothing splines and ridge regression. It also allows for the incorporation of frailty into the existing S code for survival models in a fairly natural way. Although the programming benefits are evident, the potential gains in actual computational time have not yet been completely addressed. In this section, we compare the results from the penalized models to those from a pure EM approach for the computation of $\boldsymbol{\beta}$ given θ, and then for the dual problem of jointly estimating θ with $\boldsymbol{\beta}$.

4.3.1 θ Fixed

When θ is known for the gamma frailty model, substituting the derivative of g into Equation (2.11) and solving for exp(ω_j) leads directly to Equation (2.13). That is, the E-M update step is similar to a steepest descent algorithm in the sense that it uses only first derivative information. Our S code uses second derivative information in a Newton–Raphson step, or when using the sparse computation option, an approximate N–R step. In the situation when θ is fixed, we might then expect the E-M approach to be competitive with the N–R update if the random effect coefficients ω_j are essentially independent of $\boldsymbol{\beta}$, but less efficient than N–R if this is not the case. The kidney data, with and without inclusion of the disease covariate, provides a good illustration of both. Figure 3 shows a trace of the E-M iteration path for a subset of the frailty coefficients ω, with θ set to an arbitrary value of 1. The N–R algorithm declared convergence after 5 iterations for this scenario, and the sparse variant after 10. Their iteration paths for subject 21, the lowest curve on the plot, are shown

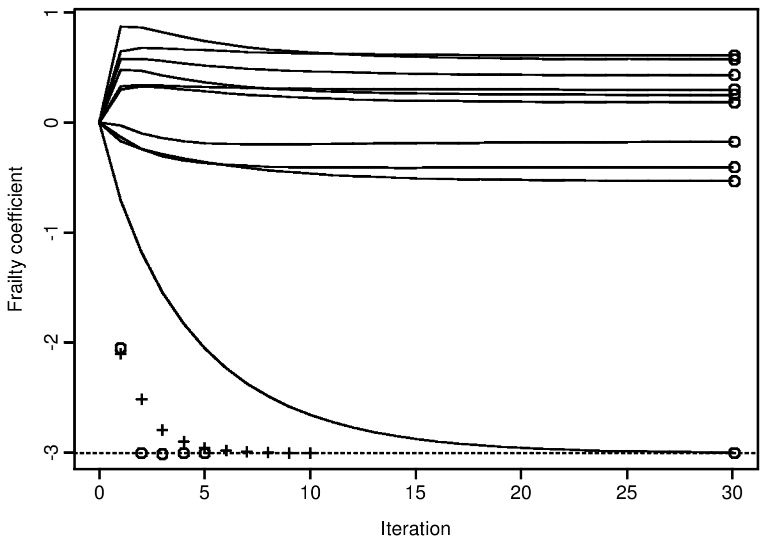

Figure 3. Coefficient traces for the E-M alogorithmforfrailty coefficients 1, 5, 9, ..., 37, for the kidney data with age, sex and disease also in the model. The lowest trace is subject 21. The iteration paths for subject 21's frailty under the full and sparse penalized likelihood approach are shown with a ○ and +, respectively; final values for the penalized result, for each curve, are also shown on the right.

as "○" and "+", respectively. The E-M approach requires 30 iterations for this coefficient. This is essentially a comparison of the efficiency of using first derivative, derivative + variance, and derivative + variance + covariance terms. Other work (not shown) indicates that the sparse computation scheme becomes more competitive with the full N-R approach, in terms of iteration counts, as the number of frailty coefficients increases.

A repeat of the exercise with only age as a covariate in the model is shown in Figure 4. In this case, there is far less correlation between the covariate and the frailty terms, and the E-M suffers only about a 2:1 disadvantage compared to the approximate N–R. The EM code we used for these computations is an S-Plus function, and the penalized models use both S-Plus and compiled C language code. Therefore, a comparison of these runs based on CPU time would not be fair to the EM approach.

4.3.2 θ Estimated

When estimating the value of θ from the data, the difference between the profile likelihood approach used here and an E-M update is more profound. The E-M update equation for $\hat{\theta}$ is the solution to

$$\sum_j \left[\left(\nu \log(\nu) - \log\Gamma(\nu)\right) + \gamma(\nu + d_j) - \left(\log(\nu + \hat{A}_j) + (\nu + d_j)/(\nu + \hat{A}_j)\right)\right] = 0$$

(Klein 1992), where γ is the digamma function. This update must itself be solved iteratively, whereas the profile likelihood is a simple correction of the PPL. For the kidney data and a model with only age and frailty, the E-M update requires approximately 520 iterations for

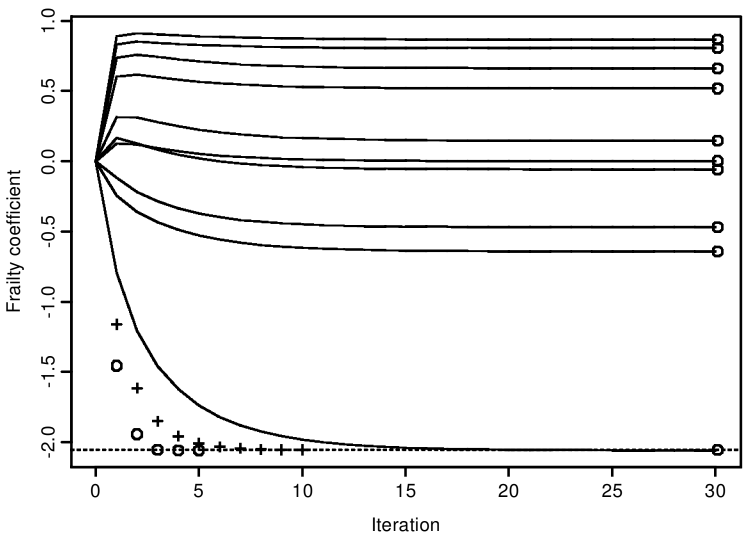

Figure 4. Coefficient traces for the E-M algorithm for frailty coefficients 1, 5, 9, . . ., 37, for the kidney data with only age as a covariate. The lowest trace is subject 21. The iteration paths for subject 21's frailty under the full and sparse penalized likelihood approach are shown with a ○ and +, respectively.

adequate convergence. Final convergence of the E-M required 760 iterations, but the change in loglik over the last 200 is in the seventh significant digit. The profile search required only 10 iterations to converge on the estimate of θ. The authors have observed a similar ratio of iteration counts across a range of datasets.

5. CONCLUDING REMARKS

Penalized estimation techniques are useful estimation tools. We have now shown that estimation using shared Gamma frailty models can be performed exactly with penalized likelihood methods. This is true for models with time-dependent covariates as well as for models with time-independent covariates, which we focused on in an attempt to keep the notation simple. We have yet to find such a correspondence for more general Gamma frailty models, such as the nested frailty model of Guo and Rodrguez (1992). However, more general Gaussian frailty models can be approximately estimated using penalized estimation procedures. Also, the programs support the use of AIC or corrected AIC (Hurvich, Simonoff, and Tsai 1998) as a selection criteria. With this approach, models can be fit beyond those for which a formal ML-penalized correspondence has been worked out. Some examples are models with multiple frailty terms or models with other frailty distributions. Using AIC as the optimization criteria for θ and the log of a t-distribution density as the penalty term, for instance, appears to give similar results to more formal MCMC methods on two (small) local examples. Wider experience and/or formal results are needed to understand the relative merits of various approaches: likelihood versus degrees-of-freedom based estimates of the frailty variance, and choice of frailty distribution.

We outline several important issues regarding the variance of the random effect, θ, below.

- The software does not print an estimate of the variance of θ. However, a plot of the profile likelihood can easily be obtained by fitting a sequence of models with fixed θ. This profile likelihood is often seriously asymmetric with a longer right tail, raising concerns about the utility of se(θ) for either confidence intervals or tests. The current computer code does iteration on the $\sqrt{\theta}$ scale. Although this seems to speed convergence, other scales may be more appropriate for the creation of confidence intervals.
- The estimate of the random effect is often much less precise than the estimates of the coefficients $\hat{\boldsymbol{\beta}}$. It is unclear how large a sample size is needed for reliable estimation.
- The software prints out an approximate Wald test, $\omega'(\mathbf{H}^{-1})_{22}\omega$, based on the fitted frailty coefficients. Since the number of frailty coefficents often grows with sample size, while the *effective* number might not, the statistical properties of the test are unknown. The printed test seems to be successful as a first "very significant/not at all significant" approximation, but final judgement should be based on the likelihood ratio test derived by comparing the printed `M-likelihood` value to the fit without the frailty term.
- The standard errors of the estimate are calculated as though θ were fixed. This is true for some penalized problems, but false for the two examples given here. A bootstrap evaluation with θ fixed at $\hat{\theta}$ gives standard errors for the other parameters that agree with our asymptotic formula, but with θ free the standard errors are larger by 30 to 60%.

Beyond its extendability, an important benefit of the penalized approach is speed. The computer code is fast enough that we can use it with computationally intensive secondary techniques such as the bootstrap. For instance, it took just over 11 minutes to perform 1,000 bootstrap realizations of the kidney data holding θ fixed, and 35 minutes when θ was allowed to vary.

In summary, certain classes of frailty models can be formulated as penalized likelihoods. Because of its connection to other work in penalized regression, computational improvements are possible for selected models. For shared frailty models, use of a sparse Cholesky factorization provides significant computational advantages. Other, similar, gains can be made with other frailty models. As an example, genetic family studies can be cast as a frailty model with one random effect per subject, and correlations among random effects that are block diagonal with one block per family. This can be efficiently handled using a more general sparse Cholesky algorithm.

APPENDIX: CORRESPONDENCE OF MARGINAL LOG-LIKELIHOODS AT THE SOLUTION POINT

Here, we obtain the realized value of the marginal log-likelihood at the solution point in terms of the penalized likelihood for the gamma shared frailty model. This justifies Equation (2.15).

Expanding Equation (2.2) gives

$$L_m(\boldsymbol{\beta}, \lambda_0; \theta) = \sum_{i=1}^{n} \delta_i \log \left(\int Y_i(t) e^{\mathbf{X}_i \boldsymbol{\beta}} d\Lambda_0(t) \right) + \sum_{j=1}^{q} [-d_j \log \nu - (\nu + d_j) \log \left(1 + A_j/\nu\right) + \log\{\Gamma(\nu + d_j)/\Gamma(\nu)\}].$$

The log profile likelihood for θ is just this function restricted to the one-dimensional curve defined by the maximizing values of $\hat{\boldsymbol{\beta}}(\theta), \hat{\omega}(\theta), \hat{\lambda}_0(\theta)$. On that curve $\hat{A}_j = (d_j + \nu - \nu e^{\hat{\omega}_j})/e^{\hat{\omega}_j}$ (see Equation (2.13)). With this substitution, after some rearrangement we get

$$L_m(\theta) = \sum_{i=1}^{n} \delta_i \log \left(\hat{\lambda}_i e^{\mathbf{X}_i \hat{\boldsymbol{\beta}} + \mathbf{Z}_i \hat{\omega}} \right) + \sum_{j=1}^{q} \left[-(\nu + d_j) \log(\nu + d_j) + \nu \log(\nu e^{\hat{\omega}_j}) + \log \Gamma(\nu + d_j) - \log \Gamma(\nu) \right],$$

where δ_i is a 0/1 indicator for an event for individual i.

Subtracting and adding the penalty function $g(\omega; \theta) = -1/\theta \sum_{j=1}^{q} \omega_j - \exp(\omega_j)$, evaluated at $\hat{\omega}$ results in

$$\begin{aligned} L_m(\theta) &= \sum_{i=1}^{n} \delta_i \log(\hat{\lambda}_i e^{\mathbf{X}_i \hat{\boldsymbol{\beta}} + \mathbf{Z}_i \hat{\omega}}) - g(\hat{\omega}; \theta) \\ &\quad + \sum_{j=1}^{q} [-\nu \hat{\omega}_j + \nu e^{\hat{\omega}_j} - (\nu + d_j) \log(\nu + d_j) + \nu \log(\nu e^{\hat{\omega}_j}) \\ &\quad + \log \Gamma(\nu + d_j) - \log \Gamma(\nu)] \\ &= \text{PPL}(\theta) + \sum_{j=1}^{q} \left[\nu - (\nu + d_j) \log(\nu + d_j) + \nu \log \nu + \log \left(\frac{\Gamma(\nu + d_j)}{\Gamma(\nu)} \right) \right], \end{aligned}$$

where the last step follows from Equation (2.14).

Note that, because considerable loss of accuracy can occur if one subtracts values of the log-gamma function, it is computationally advantageous to use

$$\log \left(\frac{\Gamma(\nu + d_j)}{\Gamma(\nu)} \right) = \sum_{i=0}^{d_j - 1} \log \left(\frac{\nu + i}{\nu + d_j} \right)$$

rather than

$$\log \left(\frac{\Gamma(\nu + d_j)}{\Gamma(\nu)} \right) = \log(\Gamma(\nu + d_j)) - \log(\Gamma(\nu)).$$

[Received May 2000. Revised October 2001.]

REFERENCES

Aalen, O. O. (1988), "Heterogeneity in Survival Analysis," *Statistics in Medicine*, 7, 1121–1137.

Breslow, N. E., and Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88, 9–25.

Gray, R. J. (1992), "Flexible Methods for Analyzing Survival Data Using Splines, With Applications to Breast Cancer Prognosis," *Journal of the American Statistical Association*, 87, 942–951.

Guo, G., and Rodrguez, G. (1992), "Estimating a Multivariate Proportional Hazards Model for Clustered Data Using the EM Algorithm, With an Application to Child Survival in Guatemala," *Journal of the American Statistical Association*, 87, 969–976.

Hurvich, C. M., Simonoff, J. S., and Tsai, Chih-Ling (1998), "Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion," *Journal of the Royal Statistical Society*, Ser. B, 60, 271–293.

Klein, J. P. (1992), "Semiparametric Estimation of Random Effects Using the Cox Model Based on the EM Algorithm," *Biometrics*, 48, 795–806.

Lin, D. Y. (1994), "Cox Regression Analysis of Multivariate Failure Time Data, the Marginal Approach," *Statistics in Medicine*, 13, 2233–2247.

Lindor, K. D., Dickson, E. R., Baldus, W. P., Jorgensen, R. A., Ludwig, J., Murtaugh, P. A., Harrison, J. M., Wiesner, R. H., Anderson, M. L., Lange, S. M., LeSage, G., Rossi, S. S., and Hofman, A. F. (1994), "Ursodeoxycholic Acid in the Treatment of Primary Biliary Cirrhosis," *Gastroenterology*, 106, 1284–1290.

McGilchrist, C. A. (1993), "REML eEstimation for Survival Models With Frailty," *Biometrics*, 49, 221–225.

McGilchrist, C. A., and Aisbett, C.W. (1991), "Regression With Frailty in Survival Analysis," *Biometrics*, 47, 461–466.

Nielsen, G. G., Gill, R. D., Andersen, P. K., and Srensen, T. I. (1992), "A Counting Process Approach to Maximum Likelihood Estimation of Frailty Models," *Scandinavian Journal of Statistics*, 19, 25–43.

Parner, E. (1997), "Inference in Semiparametric Frailty Models," Technical report, Ph.D. dissertation, University of Aarhus, Denmark.

Ripatti, S., and Palmgren, J. (2000), "Estimation of Multivariate Frailty Models Using Penalized Partial Likelihood," *Biometrics*, 56, 1016–1022.

Verweij, J. M., and Van Houwelingen, H. C. (1994), "Penalized Likelihood in Cox Regression," *Statistics in Medicine*, 13, 2427–2436.

Yau, K. K. W., and McGilchrist, C.A. (1998), "ML and REML Estimation in Survival Analysis with Time Dependent Correlated Frailty," *Statistics in Medicine*, 17, 1201–1213.