Tuning Parameter Selection in Penalized Frailty Models

E. ANDROULAKIS, C. KOUKOUVINOS, AND F. VONTA

National Technical University of Athens, Athens, Greece

The penalized likelihood approach of Fan and Li (2001, 2002) differs from the traditional variable selection procedures in that it deletes the non-significant variables by estimating their coefficients as zero. Nevertheless, the desirable performance of this shrinkage methodology relies heavily on an appropriate selection of the tuning parameter which is involved in the penalty functions. In this work, new estimates of the norm of the error are firstly proposed through the use of Kantorovich inequalities and, subsequently, applied to the frailty models framework. These estimates are used in order to derive a tuning parameter selection procedure for penalized frailty models and clustered data. In contrast with the standard methods, the proposed approach does not depend on resampling and therefore results in a considerable gain in computational time. Moreover, it produces improved results. Simulation studies are presented to support theoretical findings and two real medical data sets are analyzed.

Keywords Clustered data; Error estimation; Generalized cross validation; Penalized frailty model; Penalized likelihood; Tuning parameter; Variable selection

Mathematics Subject Classification Primary: 62-07, 62N01; Secondary: 62J07

1. Introduction

Variable selection is fundamental in high-dimensional statistical modeling and has received much attention in the recent literature. To this end, a family of penalized likelihood approaches has been proposed by Fan and Li (2001). This new methodology was extended further to the case where we have survival data with censored observations (Fan and Li, 2002), where the Cox proportional hazards model and the Gamma frailty model were the two semi-parametric models considered. Compared with traditional methods, like the best subset variable selection, this new approach significantly reduces the computational burden. Moreover, it simultaneously excludes the inactive factors and estimates the regression coefficients of active effects. A noteworthy advantage of this methodology is that the penalized estimates perform as well as the oracle procedure in terms of selecting the correct model. As a result, this framework is applicable to both statistical inference and machine learning problems. Recent related studies include Fan and Li (2006), Leng et al. (2006), Zou and Li (2008), Pötscher and Leeb (2009), and Fan and Lv (2010).

However, this method is based on the use of penalty functions in which a tuning parameter is involved. This parameter influences the performance of penalized methods...
because it controls the extent of penalization. Consequently, the tuning parameter governs the complexity of the resulting model and it must be chosen properly. To this end, Fan and Li (2001, 2002) proposed the application of the cross-validation and generalized cross validation (GCV) statistic (Craven and Wahba, 1979) so as to select the tuning parameter. Moreover, Wang et al. (2007) proposed a tuning parameter selector based on the Bayesian Information Criterion (BIC), concerning the penalized least squares case with the smoothly clipped absolute deviation penalty (SCAD). In addition, Zhang et al. (2010) adopted Nishii’s (1984) generalized information criterion (GIC) in order to select the tuning parameter for the nonconcave penalized likelihood approaches. Recently, Fan and Tang (2013) explored tuning parameter selection for penalized generalized linear regression, where the dimensionality of covariates can increase exponentially with the sample size and examined the GIC systematically.

We must stress that, since parameter tuning in penalized likelihood modeling with standard resampling methods (e.g., cross-validation) can be very time-consuming, the development of error estimate formulas that could be used for parameter tuning without the need for resampling is a very important issue.

In this article, we focus on penalized frailty models for clustered data. In our previous work (Androulakis et al., 2012), we extended the penalized Gamma frailty model methodology of Fan and Li (2002) to a whole class of frailty models, based on a generalized form of the likelihood function designed for clusters (see also Hougaard, 2000), which allows the direct use of many continuous distributions for the frailty parameter. In the likelihood, the Breslow’s estimate of the cumulative hazard function was employed. The aim of this article is firstly to develop new estimates of the norm of the error, related to the estimation of the regression parameters in frailty models with clustered data through the use of Kantorovich inequalities. These estimates where explored in the context of numerical analysis for non-linear systems by Galantai (2001). Based on these estimates a regularization parameter selector for penalized frailty models will also be proposed. It can be considered as an effective alternative approach compared with the commonly used GCV (see also Fan and Li, 2002) as it is highlighted in our simulations. We point out that the proposed procedure has also been used in the case of generalized linear models and more specifically in penalized logistic and poisson regression (Androulakis et al., 2014).

The rest of the article is organized as follows: In Section 2, some details concerning our generalized penalized likelihood for clustered data (Androulakis et al., 2012) are given. In Section 3, the new error estimates are derived via Kantorovich inequalities and, subsequently, employed in the derivation of a tuning parameter selection procedure. In Section 4, a thorough simulation study is provided to illustrate the proposed method for various sample sizes and cluster sizes. The results are discussed and some comments derived from the comparison of the new method with the GCV, are given. In Section 5, an application of the proposed method to two real medical datasets is considered. In the last Section 6, some general conclusions are presented.

2. Penalized Frailty Models for Clustered Data

We consider the frailty model in the case of clusters where the subjects in the same cluster share the same frailty, which is a positive random variable. Let \( n \) be the number of clusters, \( J_i \) the size of the cluster \( i, i = 1, \ldots, n \) and \( u_i \) the shared frailty in the cluster \( i \) with c.d.f. \( F_{u_i} \). The data are described by \((z_{ij}, \delta_{ij}, x_{ij})\), where \( z_{ij} = \min\{T_{ij}, C_{ij}\} \), \( T_{ij} \) is the survival time, \( C_{ij} \) the censoring time, \( \delta_{ij} \) the censoring indicator and \( x_{ij} \) the \( d \)-dimensional vector
of covariates of the \( j \)th individual belonging to the cluster \( i \). We assume that the hazard rate of the \( j \)th subject in the \( i \)th subgroup-cluster is given conditionally on the covariates \( x_{ij} \) and the shared frailty \( u_i \) as

\[
h_{ij}(t|x_{ij}, u_i) = u_i h_0(t)e^{x_{ij}^T \beta},
\]

where \( \beta \in \mathbb{R}^d \) is the parameter of interest and \( h_0(t) \) the baseline hazard function.

In Androulakis et al. (2012), the penalized Gamma frailty model methodology proposed by Fan and Li (2002) was extended to a general class of frailty models in order to be able to use many different continuous distributions for the frailty parameter. We considered the “least informative” nonparametric modeling for \( H_0(t) \) in which \( H_0(t) \) has a possible jump of size \( \lambda_l \) at the observed failure time \( z_l \). Then, 

\[
H_0(t) = \sum_{l=1}^{N_l} \lambda_l I(z_l \leq t),
\]

where \( z_1, \ldots, z_N \) are pooled observed failure times. The resulting penalized profile likelihood has the form

\[
\sum_{i=1}^{n} \sum_{j=1}^{J_i} \delta_{ij} x_{ij}^T \beta + \sum_{l=1}^{N} \ln \lambda_l + \sum_{i=1}^{n} \left( \int_0^\infty \exp \left( -u_i \sum_{j=1}^{J_i} x_{ij}^T \beta \sum_{l=1}^{N} \lambda_l I(z_l \leq z_{ij}) \right) u_i \rho \right) F_u (u_i) \]

\[
- n \sum_{j=1}^{d} p_j(|\beta_j|),
\]

where \( p_j(.) \) is an appropriate penalty function. The most common penalties are the \( L_1 \) penalty which results in the Least Absolute Shrinkage and Selection Operator method (LASSO) (see Tibshirani, 1996), the Hard thresholding penalty (see Antoniadis, 1997), and the SCAD penalty (see Fan and Li, 2001). We will consider for illustrative purposes two continuous frailty models in the simulations and real data, namely, the Gamma and Inverse Gaussian frailty models which are commonly used in survival analysis. For theoretical details see Androulakis et al. (2012).

3. Error Estimates for Frailty Models and Tuning Parameter Selection for the Penalized Case

We now propose our new estimates of the norm of the error in the frailty models framework through the use of the Kantorovich inequality given in the following form (see Greub and Rheinboldt, 1959; Horn and Johnson, 1985; Marcus and Minc, 1992):

\[
\| w \|_2^4 \leq (w^T B w)(w^T B^{-1} w) \leq \frac{1}{4} \left( \frac{\rho_1 + \rho_d}{\rho_1 \rho_d} \right) \| w \|_2^4,
\]

where \( B \in \mathbb{R}^{d \times d} \) is a symmetric positive definite matrix with eigenvalues \( \rho_1 \geq \rho_2 \geq \cdots \geq \rho_d > 0 \) and \( w \in \mathbb{R}^d \) is an arbitrary vector.

Inspired by the work of Galantai (2001), who generalized Auchmuty’s error estimate (Auchmuty, 1992) in numerical analysis to nonlinear systems and based on his steps, we adjust the error estimate that he derived to our situation, that is, the frailty models framework.

Consider the nonlinear score equations that arise from the differentiation of (2) without the penalty term, with respect to \( \beta \), multiplied by \(-1\). We denote them as \( F(\beta) = 0 \) with \( F : \mathbb{R}^d \rightarrow \mathbb{R}^d \) and let \( \beta \) any approximate solution and \( \beta^* \) the exact solution of \( F(\beta) = 0 \).

Let also \( F'(\beta) \) be the Jacobian matrix which coincides with the minus second gradient derivative of (2) without the penalty term. Assume that \( F'(\beta^*) \) is invertible,
$F' \in C^1(S(\beta^*, \delta))$ and $\|F'(\beta_1) - F'(\beta_2)\|_2 \leq M\|\beta_1 - \beta_2\|_2$, $\forall \beta_1, \beta_2 \in S(\beta^*, \delta)$, where

$$S(\beta^*, \delta) = \{ \beta : \|\beta^* - \beta\|_2 \leq \delta \},$$  

(4)

$\delta > 0$ and $\|\cdot\|_2$ denotes the Frobenius norm for matrices and the Euclidean norm for vectors. Now assume that $\beta$ is close to $\beta^*$ and let $B = F'(\beta)F'(\beta)^T$. Assume that $B$ is a symmetric and positive definite matrix. Let also $U\Sigma V^T$ be the Singular Value Decomposition of $F'$, where $\Sigma = diag(\sigma_1, ..., \sigma_d)$ such that $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_d > 0$. If $w \in \mathbb{R}^d$ is an arbitrary vector, we have that

$$w^TBw = \|F'(\beta)^Tw\|_2^2 \quad \text{and} \quad w^TB^{-1}w = \|F'(\beta)^{-1}w\|_2^2$$  

(5)

and $\rho_i = \rho_i(B) = \sigma_i^2(F') = \sigma_i^2$, where $\rho_i$ and $\sigma_i$ are the $i$th eigenvalues and singular values of $B$ and $F'$, respectively. We can now use the Kantorovich inequality (3) which gives

$$\|w\|_2^2 \leq \|F'(\beta)^Tw\|_2\|F'(\beta)^{-1}w\|_2 \leq \frac{1}{2} \frac{\sigma_1^2 + \sigma_d^2}{\sigma_1 \sigma_d} \|w\|_2^2, \quad w \in \mathbb{R}^d.$$  

(6)

Let $w = F(\beta)$. From the Lipschitz continuity, it follows that

$$F(\beta) = F'(\beta)(\beta - \beta^*) + O(\|e\|_2^2)$$  

(7)

and

$$F'(\beta)^{-1}F(\beta) = \beta - \beta^* + O(\|e\|_2^2).$$  

(8)

Hence

$$\|F(\beta)\|_2^2 \leq \|F'(\beta)^TF(\beta)\|_2(\|\beta - \beta^*\|_2 + O(\|e\|_2)) \leq C_2(F'(\beta))\|F(\beta)\|_2^2, \quad \text{and} \quad (9)$$

$$\frac{\|F(\beta)\|_2^2}{\|F'(\beta)^TF(\beta)\|_2} \leq \|\beta - \beta^*\|_2 + O(\|e\|_2^2) \leq C_2(F'(\beta)) \frac{\|F(\beta)\|_2^2}{\|F'(\beta)^TF(\beta)\|_2},$$  

(10)

where $C_2(F'(\beta)) = \frac{1}{2} \frac{\sigma_1^2 + \sigma_d^2}{\sigma_1 \sigma_d}$. Therefore, we have the approximate absolute error estimate

$$\|\beta - \beta^*\|_2 = c \frac{\|F(\beta)\|_2^2}{\|F'(\beta)^TF(\beta)\|_2},$$  

(11)

where $1 \lesssim c \lesssim C_2(F'(\beta))$.

Our aim in this work is to propose an alternative method for estimating the parameter $\lambda$, in the penalized frailty models framework (see Eq. 2), through the proper use of the estimates of the norm of the error (11). In order to apply the Kantorovich inequality to obtain (11) in the frailty models framework, the following assumptions must hold:

A. All clusters are non-empty. The regressors $x$ take finite values, are nondegenerate random variables and are linearly independent within each cluster $i$, $i = 1, \ldots, n$.

B. The baseline cumulative hazard function $H(t) < \infty$ for $t < \infty$.

C. The Laplace transform of the frailty distribution along with its derivatives up to the order $A_i + 3$ for all $i$ exist (but it may not be possible to be given in a closed form). We assume also $\frac{[L^{(A_i+1)}(x)]}{[L^{(A_i)}(x)]}$, $\frac{[L^{(A_i+2)}(x)]}{[L^{(A_i)}(x)]}$, and $\frac{[L^{(A_i+3)}(x)]}{[L^{(A_i)}(x)]}$ are uniformly bounded for all $i$.

Note that assumption C is satisfied for the widely used Gamma and Inverse Gaussian frailty models. Observe that $F'(\beta)$ is continuous with respect to $\beta$. From assumptions
A–C we obtain the Lipschitz continuity of $F'(\beta)$ based on the mean value theorem and
the fact that $F''(\beta)$ is uniformly bounded for $\beta \in S(\beta^*, \delta)$.

Moreover, the matrix $F'(\beta^*)$ is symmetric and non-negative definite as an observed
information matrix. Assumption A excludes degenerate cases that would result in an ob-
served information matrix about $\beta$, with 0 determinant and therefore singular. It is possible
though to encounter this problem in practice especially in high-dimensional data. If such
a problem occurs it affects not only our method for the tuning parameter selection but
the penalized likelihood methodology in general, and in turn the GCV tuning parameter
selection, and even the existence of an MLE of $\beta$. We suppose therefore throughout this
article standard regularity and nondegeneracy conditions about the parameter $\beta$ (Bickel
and Doksum, 2007, Thm. 6.2.2; Van der Vaart, 1998, Thm. 5.39) which imply that the max-
imum likelihood estimator $\hat{\beta}$ of $\beta$ exist and is locally uniquely determined as solution of
the score or likelihood equations and is efficient with variance-covariance matrix $I^{-1}(\beta_0),$ where $I(\beta_0)$ the true Fisher information matrix about $\beta$. If the sample size is large a law
of large numbers along with continuity of the property of positive definiteness in matrix
space is enough to guarantee that the observed information matrix is positive definite if the
true one is.

We should point out here that our estimating error Eq. (11) does not require the actual
calculation of the inverse of the observed information matrix, as compared to the GCV
method in which the trace of the inverse of a matrix, related to the observed information
matrix and the penalty term, needs to be calculated, and therefore, it gains in computational
time and is easily implemented.

The procedure we suggest which can lead to the derivation of the tuning parameter
selector, is described as follows (on the Remarks that follow we will elaborate on the
choices made in our algorithm):

1. Start with a grid containing the initial values of the tuning parameter $\lambda$.
2. For each $\lambda$ in the grid, compute a penalized estimate $\hat{\beta}$.
3. Then evaluate the norm of the error by formula (11), with the constant $c$ taking the
   value of its upper bound, using the appropriate expressions for $F(\hat{\beta})$ and $F'(\hat{\beta}),$
   depending on the frailty distribution considered.
4. The value of the parameter $\lambda$ in the grid which minimizes (11) for the aforementioned
   constant $c$, is the desired tuning parameter and the resulting penalized estimator $\hat{\beta}$
   is our final estimator of $\beta$.

Remark 3.1. It must be noted that the upper bound of constant $c$ in (11), which is
$C_2(F'(\beta))$, is a simple function of the condition number $k_2(F'(\beta)) = \|F'(\beta)\|_2\|F'(\beta)^{-1}\|_2$ of the matrix $F'(\beta)$. Therefore, we stress here that due to the aforementioned assumptions
A–C and the discussion following them, the matrix $F'(\beta)$ is nonsingular. Then the fact
that $C_2(F'(\beta))$ is less than or equal to $k_2(F'(\beta))$ (see also Galantai, 2001), ensures that the
upper bound $C_2(F'(\beta))$ is finite. If the matrix is singular or near singular, we can use the
generalized Moore–Penrose inverse and base our conclusion on the condition number of
the generalized inverse assuming that the latter is finite. This inverse is traditionally also
employed by the GCV method when this problem arises.

Further work is needed, however, for the formal derivation of the average of $c$ in our
framework, especially for high-dimensional settings, and this will be considered in a next
article. However, see some comments in the next section (Remark 4.1) as well as in the
discussion following the applications to the real datasets.
4. Simulations

In this section, a number of simulations is performed in order to demonstrate our new method for the derivation of the tuning parameter. The simulation scheme is based on that of Fan and Li (2002), therefore, we simulated 100 data sets consisting of \( n \) groups and \( J \) subjects in each group from the exponential hazard frailty model

\[
h(t | x, u) = u \exp(x^T \beta),
\]

where the true parameter \( \beta_0 = (0.8, 0, 0, 1, 0, 0, 0.6, 0) \) and the \( x_i \) are marginally standard normal and the correlation between \( x_i \) and \( x_j \) is \( \rho^{ij} \) with \( \rho = 0.5 \). The distribution of the censoring time is an exponential distribution with mean

\[
U \exp(x^T \beta^*)
\]

where \( U \) is randomly generated from the uniform distribution over \([1, 3]\) for each simulated data set so that about 30% data are censored. Here \( \beta^* = \beta_0 \) is regarded as a known constant so that the censoring scheme is noninformative. Concerning the frailty \( u \), we present two cases. First, we consider a Gamma(\( \alpha, 1/\alpha \)) frailty with \( \alpha = 4 \) and, secondly, an Inverse Gaussian frailty \( N^{-1}(b, b) \) with \( b = 2 \) so that in both cases the variance is equal to 1/4.

The performance of the penalized procedures for clustered data, with the SCAD penalty (SCAD), the \( L_1 \) penalty (LASSO), and the hard thresholding penalty (Hard), is compared in terms of their model errors, model complexity, and accuracy, by using the proposed method for the derivation of the tuning parameter and GCV as an alternative. Model errors of the penalized procedures are compared to those of the maximum profile likelihood estimates. Specifically, we present the Median of Relative Model Errors (MRME), see Fan and Li (2002), over the 100 simulated data sets with some combinations of \( n \) and \( J \).

Moreover, the average number of zero coefficients are also reported in the following tables, in which the column labeled “correct” presents the average restricted only to the true zero coefficients, while the column labeled “incorrect” depicts the average of coefficients erroneously set to 0. We also tested the accuracy of the standard error formula proposed by Fan and Li (2002). However, to save space, the results concerning the standard errors can be found in the Web Appendix which is available under the Appendices of Publications link http://www.math.ntua.gr/~ckoukov/. Note that, a good and accurate performance of Fan and Li (2002) formula should result in close values of the true and estimated standard deviations of estimators, while the penalized methods must perform as well as the Oracle estimator, a fact that was successfully achieved.

We have also calculated for each method the mean values of the 100 estimated coefficients of \( X_1, X_4, \) and \( X_7 \) as well as the mean values of \( \alpha \) or \( b \) for which the maximum of the loglikelihood function occurred. Furthermore, the mean values of \( \lambda \) selected by the compared methods are also demonstrated. We also present the results for the Oracle estimator, obtained by fitting the ideal model consisting of the variables \( X_1, X_4, \) and \( X_7 \). It must also be stressed that, due to the fact that GCV also uses some initial values of \( \lambda \) in its routine, the same values were given in both methods, in order to make a strict and fair comparison of them. All computations were conducted using MATLAB codes. The results are given in the next tables.

From Tables 1–4, we extract the following conclusions. We observe that with the use of our method for the derivation of the tuning parameter, there is a significant improvement in the general performance of the penalized methods. Specifically, the MRME is improved for all penalized methods, regardless of the sample sizes, the cluster sizes, or the frailty distribution used. Moreover, in model identifications, the penalized methods and especially the SCAD and Hard, have a higher chance of correctly setting the non-significant coefficients to zero, when using our method, as it is depicted in Tables 1 and 3. The SCAD and Hard methods outperform the LASSO and perform as well as the oracle estimator. However, the performance of the LASSO method is improved with the use of our tuning parameter selector.
**Table 1**
Simulation results for the Gamma frailty

| Method   | GCV: Aver. of 0 coeff. | New Method: Aver. of 0 coeff. |
|----------|------------------------|-------------------------------|
|          | GCV: MRME(%) Correct   | Incorrect                     | New Method: MRME(%) Correct   | Incorrect   |
| $n = 50, J = 2$ |                        |                               |                             |
| SCAD     | 0.7637 4.4300 0.0200   | 0.6302 4.5800 0.0200          |
| LASSO    | 1.1644 3.6200 0.0200   | 0.7597 3.7800 0.0200          |
| Hard     | 0.7430 4.3100 0.1100   | 0.6921 4.6900 0.0700          |
| Oracle   | 0.4206 5             | 0.4206 5                      |
| $n = 50, J = 5$ |                        |                               |                             |
| SCAD     | 0.6231 4.4600 0       | 0.5589 4.8800 0               |
| LASSO    | 0.9445 3.5500 0       | 0.7511 3.8700 0               |
| Hard     | 0.7548 4.1700 0.0400  | 0.6205 4.7300 0               |
| Oracle   | 0.5156 5             | 0.5156 5                      |
| $n = 100, J = 2$ |                        |                               |                             |
| SCAD     | 0.7903 4.4200 0       | 0.6432 4.9900 0               |
| LASSO    | 0.9249 3.7400 0       | 0.7226 3.8600 0               |
| Hard     | 0.8572 4.4800 0.0700  | 0.6797 4.7700 0               |
| Oracle   | 0.6044 5             | 0.6044 5                      |
| $n = 100, J = 5$ |                        |                               |                             |
| SCAD     | 0.6341 4.4500 0       | 0.5309 4.8600 0               |
| LASSO    | 0.7474 3.5200 0       | 0.6743 3.6500 0               |
| Hard     | 0.8115 4.5600 0       | 0.6392 4.7100 0               |
| Oracle   | 0.5872 5             | 0.5872 5                      |

Concerning the mean values of the non-zero coefficients and $\alpha$ or $b$, it can be seen from Tables 2 and 4 that in all cases, the estimated and the real values are very close when implementing the new method, resulting in a similar or even better performance for many cases, as compared to GCV. As far as the mean values of the selected $\lambda$ are concerned, a general remark is that our method yields lower values of $\lambda$. We should stress here that in our simulations, we firstly selected grids for $\lambda$ that did not contain the zero value (see step 1 of the proposed procedure.) The reason is that on one hand it is not our desire to get too close to the MLE, to which a value of $\lambda = 0$ would correspond. On the other hand, we selected the same grids as those used in the GCV method for the purpose of comparison. Note also that the resulting tuning parameter was obtained inside the range of the grid and not in the boundary in all cases considered.

**Remark 4.1.** Galantai (2001) points out that in practice, the constant $c$ seems to take values usually less than 10 in the case of nonrandom matrices. On the other hand, for random matrices, it certainly increases with the value of the dimension $d$ of the parameters. Therefore, further simulation experiments were conducted in order to investigate the possible values of $c$ and especially its upper bound constant $C_2(F'(\beta))$. We note that since our findings, with regard to the MRME and model identifications for these extra setups are
similar to those analyzed before, we do not report the results here for brevity. As far as the values of $C_2(F'(\beta))$ are concerned, we found that for the Gamma and the Inverse Gaussian frailty models, the average value of $C_2(F'(\beta))$ is actually almost always less than 10 for the case of regression parameters with dimension up to 20 and for $n=100$ and $J=5$, for all three penalized methods. At the same time, the minimum upper bound obtained by (11) is for example for $d=20$, around 0.55, for the three penalized methods, while the mean bias of $\hat{\beta}$ is around 0.32.
Table 3

Simulation results for the Inverse Gaussian frailty

| Method | GCV: Aver. of 0 coeff. | New Method: Aver. of 0 coeff. |
|--------|------------------------|-------------------------------|
|        | GCV: MRME(%) Correct Incorrect | New Method: MRME(%) Correct Incorrect |
| $n = 50, J = 2$ |  |  |
| SCAD   | 0.5315 | 4.2700 | 0.0800 | 0.4104 | 4.6600 | 0.0400 |
| LASSO  | 0.5987 | 3.5800 | 0.0900 | 0.5589 | 4.0500 | 0.0700 |
| Hard   | 0.5815 | 4.2600 | 0.2200 | 0.5110 | 4.5500 | 0.1200 |
| Oracle | 0.3979 | 5     | 0     | 0.3979 | 5     | 0     |
| $n = 50, J = 5$ |  |  |
| SCAD   | 0.5978 | 4.3400 | 0 | 0.4332 | 4.7500 | 0 |
| LASSO  | 1.1452 | 3.5100 | 0 | 0.8137 | 3.6400 | 0 |
| Hard   | 0.4778 | 4.3600 | 0.0200 | 0.4395 | 4.7400 | 0 |
| Oracle | 0.4012 | 5 | 0 | 0.4012 | 5 | 0 |
| $n = 100, J = 2$ |  |  |
| SCAD   | 0.7548 | 4.4600 | 0 | 0.6805 | 4.8000 | 0 |
| LASSO  | 1.1272 | 3.5300 | 0 | 0.8299 | 3.7500 | 0 |
| Hard   | 0.9403 | 4.3700 | 0.0700 | 0.8176 | 4.7200 | 0 |
| Oracle | 0.6360 | 5 | 0 | 0.6360 | 5 | 0 |
| $n = 100, J = 5$ |  |  |
| SCAD   | 1.0308 | 4.4800 | 0 | 0.8371 | 4.7600 | 0 |
| LASSO  | 1.8110 | 3.2600 | 0 | 1.1799 | 3.5300 | 0 |
| Hard   | 0.9118 | 4.4700 | 0 | 0.8059 | 4.7200 | 0 |
| Oracle | 0.8438 | 5 | 0 | 0.8438 | 5 | 0 |

Remark 4.2. In an effort to determine whether the choice of the constant $c$, which takes values between 1 and $C_2(F'(β))$, was crucial to the performance of the method, different values of $c$ were considered. For example, we examined the case when $c = 1$ or the case $c = (1 + C_2(F'(β)))/2$. We did not notice improvements in the results. We also examined different choices for $λ$ in step 4 of the algorithm, for example to select the $λ$ which maximizes the criterion (11) with $c$ taking the value of its upper bound. All alternatives led to very high values of MRME. Consequently, the $λ$ that minimizes (11) along with the value $c = C_2(F'(β))$ was our final choice.

Remark 4.3. We also examined the case in which there is a greater correlation between covariates, since independence is a crucial assumption for the application of the Kantorovich inequality, in order to evaluate the effects of a reduced lack of this assumption. Specifically, we examined the case where the correlation between $x_i$ and $x_j$ is $ρ^{|i-j|}$ with $ρ = 0.8$. Our method again outperformed the GCV in terms of MRME and model identifications. To save space, we do not present detailed results here, however, they can be found in the Web Appendix which is available under the Appendices of Publications link http://www.math.ntua.gr/~ckoukouv/.
5. Applications to Real Data

In this section we apply our method for tuning parameter selection in conjunction with the penalized frailty model methodology proposed in Androulakis et al. (2012), to the Non-Hodgkin’s Lymphoma dataset which was also considered in Kosorok et al. (2004) in the context of univariate data, as well as to a high-dimensional dataset by the Hellenic Trauma and Emergency Surgery Society, which was also analyzed in Androulakis et al. (2010) for the case of the penalized Cox’s proportional hazards model.

### Table 4
Mean values of the non-zero coefficients, \( b \) and \( \lambda \), using the New Method and GCV (in parentheses) for the Inverse Gaussian frailty

| Method | \( \beta_1 \) | \( \beta_4 \) | \( \beta_7 \) | \( b \) | \( \lambda \) |
|--------|---------------|---------------|---------------|------|------|
| \( n = 50, J = 2 \) | \( \beta_1 \) | \( \beta_4 \) | \( \beta_7 \) | \( b \) | \( \lambda \) |
| SCAD | 0.7620 (0.7350) | 1.0411 (1.0376) | 0.4570 (0.4552) | 2.4996 (2.6609) | 0.1765 (0.1921) |
| LASSO | 0.7652 (0.5597) | 0.9823 (0.7695) | 0.5676 (0.3784) | 2.5939 (2.7799) | 0.0213 (0.0900) |
| Hard | 0.8393 (0.7843) | 1.0854 (1.0684) | 0.5934 (0.5323) | 2.5482 (2.4978) | 0.4573 (0.5378) |
| Oracle | 0.8267 | 1.0578 | 0.6176 | 2.3931 | — |
| \( n = 50, J = 5 \) | \( \beta_1 \) | \( \beta_4 \) | \( \beta_7 \) | \( b \) | \( \lambda \) |
| SCAD | 0.8197 (0.8167) | 1.0138 (1.013) | 0.6016 (0.5711) | 2.5319 (2.7111) | 0.1222 (0.1719) |
| LASSO | 0.7818 (0.6668) | 0.9696 (0.8364) | 0.5800 (0.4716) | 2.6353 (2.7961) | 0.0123 (0.0484) |
| Hard | 0.8260 (0.8244) | 1.0205 (1.0186) | 0.6204 (0.6105) | 2.5936 (2.7172) | 0.3065 (0.3261) |
| Oracle | 0.8205 | 1.0130 | 0.6142 | 2.4889 | — |
| \( n = 100, J = 2 \) | \( \beta_1 \) | \( \beta_4 \) | \( \beta_7 \) | \( b \) | \( \lambda \) |
| SCAD | 0.7953 (0.7932) | 1.0009 (1.0040) | 0.5775 (0.5537) | 2.3541 (2.3986) | 0.1279 (0.1694) |
| LASSO | 0.7518 (0.6228) | 0.9510 (0.8014) | 0.5604 (0.4355) | 2.7373 (2.7380) | 0.0144 (0.0596) |
| Hard | 0.8083 (0.7971) | 1.0170 (1.0029) | 0.6021 (0.5604) | 2.4550 (2.4638) | 0.3130 (0.3850) |
| Oracle | 0.8000 | 1.0007 | 0.5934 | 2.2354 | — |
| \( n = 100, J = 5 \) | \( \beta_1 \) | \( \beta_4 \) | \( \beta_7 \) | \( b \) | \( \lambda \) |
| SCAD | 0.7583 (0.7526) | 0.9577 (0.9546) | 0.5633 (0.5319) | 2.4547 (2.4959) | 0.0928 (0.1493) |
| LASSO | 0.7326 (0.6556) | 0.9326 (0.8417) | 0.5428 (0.4717) | 2.6742 (2.6285) | 0.0089 (0.0354) |
| Hard | 0.7631 (0.7594) | 0.9667 (0.9578) | 0.5732 (0.5684) | 2.4907 (2.6524) | 0.2062 (0.2202) |
| Oracle | 0.7581 | 0.9563 | 0.5674 | 2.2965 | — |
5.1. Non-Hodgkin’s Lymphoma Data

The data are a subset of 1385 patients with aggressive non-Hodgkin’s lymphoma (NHL) from different institutions and cooperative groups in North America and Europe. These patients were treated with a particular chemotherapy regimen. Survival was documented from the start of treatment until death or loss to follow-up. The data are censored with censoring rate 54.7%. Complete information is available for the following pretreatment covariates for all patients in the subset: age at the diagnosis of NHL (≤ 60 and > 60), performance status (ambulatory or nonambulatory), serum lactate dehydrogenase level (below normal or above normal), number of extranodal disease sites (≤ 1 and > 1), and Ann Arbor classification of tumor stage [stage I or II (localized disease) or stage III or IV (advanced disease)]. Each characteristic is coded 0 for the first group in the parenthesis and 1 for the second. These dichotomous predictors are the basis for the original model (Non-Hodgkin’s Lymphoma Prognostic Factors Project, 1993). The 1385 observations were organized in 19 clusters. The clusters were generated according to survival times. People with earlier survival times share possibly common characteristics or genetic predisposition as well as people that are late survivors. The clusters are almost all of size 50.

We used the maximum likelihood estimation (MLE) method for the Cox’s regression model and the maximum profile likelihood estimation (MPLE) method for the frailty models. Moreover, we used the penalized likelihood methodology with three penalty functions, namely, the SCAD penalty (SCAD), the $L_1$ penalty (LASSO), and the hard thresholding penalty (Hard), along with the GCV and the new approach, as the tuning parameter selection methods. We fitted first the non-penalized and penalized Cox’s regression model on one hand in order to compare these results and on the other hand in order to obtain an initial estimator for the iteration algorithm needed for estimation of parameters in the generalized penalized likelihood (2) including 5 covariates. In the sequel, we considered the Gamma and the Inverse Gaussian frailty models. The results are presented in Table 5 for the Cox model and in Tables 6 and 7 for the Gamma frailty model using the proposed method and the GCV, respectively, for the selection of the tuning parameter. For the Inverse Gaussian frailty model the algorithm diverged and could not produce a maximum, therefore, we conclude that the Inverse Gaussian model is not appropriate to describe the dependency in this set of data. Our conclusion is strengthened by the results in Kosorok et al. (2004) who analyzed the same set of data for the univariate case and concluded that the Gamma frailty model is more appropriate than the Inverse Gaussian model. For the Gamma frailty model and with the use of our method, the maximization of the loglikelihood occurred for $\hat{\alpha} = 0.0403, 0.0533, 0.0533,$ and 0.0403 for the MPLE, the SCAD, LASSO, and HARD penalty cases, respectively, resulting in a frailty term with a high variance ($1/\hat{\alpha} = 24.81$ or 18.76 for the different penalized methods) which suggests that the use of a frailty term is necessary for these data to capture the possible high late dependency present, which is known to be better described by a gamma frailty. For the aforementioned values of $\hat{\alpha}$, the values of $C_2(F'(\beta))$ were 1.6772, 1.6093, and 1.6439 with regard to SCAD, LASSO, and Hard, respectively, while the corresponding values of the error estimate (the minimum upper bound obtained by (11)) were 0.1706, 0.2271, and 0.1320. On the other hand, for the GCV the value of $\hat{\alpha}$ was 0.0403 of all methods except SCAD, for which we had $\hat{\alpha} = 0.0704$. Moreover, the values of $\lambda$ when using our method were 0.0373, 0.0095, and 0.0004 for the SCAD, LASSO, and HARD penalties, while for the GCV the selected values were 0.0430, 0.0161, and 0.0004, respectively.

From Table 5 we observe that all methods keep all factors as statistically significant. Moreover, the MLE and the penalized methods with the SCAD and Hard penalties yield the
Table 5
Estimated coefficients and standard errors (in parentheses) for the penalized Cox’s regression model

| Factors                        | MLE     | SCAD    | LASSO   | Hard    |
|-------------------------------|---------|---------|---------|---------|
| Age at the diagnosis of NHL   | 0.6720  | 0.6720  | 0.5486  | 0.6720  |
|                               | (0.0809)| (0.0809)| (0.0671)| (0.0809)|
| Performance Status            | 0.6134  | 0.6134  | 0.5183  | 0.6134  |
|                               | (0.0884)| (0.0884)| (0.0686)| (0.0884)|
| Serum Lactate Dehydrogenase Level | 0.5850  | 0.5850  | 0.4986  | 0.5850  |
|                               | (0.0875)| (0.0875)| (0.0694)| (0.0875)|
| Number of Extranodal Disease Sites | 0.3950  | 0.3950  | 0.3066  | 0.3950  |
|                               | (0.0873)| (0.0873)| (0.0603)| (0.0873)|
| Tumor Stage                   | 0.3829  | 0.3829  | 0.3155  | 0.3829  |
|                               | (0.1006)| (0.1006)| (0.0628)| (0.1006)|

same model with the same estimates. The loglikelihood values are $-4151.4$ for the MLE, the SCAD, and Hard penalties and $-4155.4$ for the LASSO penalty. On the other hand, for the Gamma frailty model and our method, the loglikelihood values were $-3048.0$, $-3010.5$, $-2927.8$, and $-2934.6$ for the MPLE and the SCAD, LASSO, and HARD penalty cases, respectively. Concerning the GCV method, the corresponding (penalized) loglikelihood values were $-2980.6$ for SCAD, $-2926.4$ for LASSO, and $-2934.6$ for Hard. The loglikelihood values also suggest that the Gamma frailty model is more appropriate than the Cox model.

From Table 6 we observe that the Hard penalty case is more conservative as it keeps all the factors as statistically significant. However, the Serum Lactate Dehydrogenase Level is not significant for the SCAD and LASSO methods. On the other hand, from Table 7 we observe that with the use of GCV, the Performance Status is also not significant for the SCAD and LASSO methods. Note also that the Hard penalty yields the same estimates for both tuning parameter selection methods due to the same value of $\lambda$ that was selected. The attenuation of the regression coefficients under the frailty model as compared to the Cox model is perhaps attributed to the correlation of the estimators of the regression coefficients with the estimator of the frailty variance. Finally, we should stress here that our method took less computational time to be completed, specifically 7655.5 seconds, as compared to the GCV which took 8961.2 seconds.

5.2. Trauma Data

The data were derived from an annual registry conducted during the period 1 January 2005 to 31 December 2005 by the Hellenic Trauma and Emergency Surgery Society involving 30 General Hospitals in Greece. The study was designed to assess the effects of different prognostic factors on the outcome of injured persons. In our dataset, 4800 patients are included of which 13.8% are censored. The set of potential predictor variables was defined according to clinical knowledge and included 111 categorical, binary, and continuous factors, some of which were measured in the clinic and the rest in the emergency room or at the scene of the injury. The response variable $y$ denotes the length of stay for each hospitalized patient, which is the end-point of interest. The observations which were
considered censored, were those that correspond to deaths and transfers to other hospitals (lost to follow-up patients).

The aim of the analysis is to handle a potential hospital heterogeneity, by introducing a Gamma distributed frailty parameter. The 4800 observations were organized in clusters, based on which hospital was each patient admitted to. We applied the same methods as in the Non-Hodgkin’s Lymphoma dataset, using the proposed procedure for the selection of the tuning parameter in the penalized Gamma frailty model. The resulting estimators for the 111 covariates, concerning the MPLE, the SCAD, LASSO and Hard, are available under the Appendices of Publications link at http://www.math.ntua.gr/~ckoukouv/. The results are almost identical, as far as the selection of relevant variables is concerned, for

### Table 6
Estimated coefficients and standard errors (in parentheses) for the Gamma frailty model, using the New Method

| Factors                        | MPLE   | SCAD (New Method) | LASSO (New Method) | Hard (New Method) |
|-------------------------------|--------|-------------------|--------------------|-------------------|
| Age at the diagnosis of NHL   | 0.2997 | 0.2866            | 0.1494             | 0.2583            |
| Performance Status            | 0.1532 | 0.1426            | 0.0055             | 0.1495            |
| Serum Lactate Dehydrogenase Level | 0.0995 | 0(−)             | 0(−)              | 0.0854            |
| Number of Extranodal Disease Sites | 0.1492 | 0.1582            | 0.0769             | 0.1467            |
| Tumor Stage                   | 0.2662 | 0.2562            | 0.1224             | 0.2031            |

| Factors                        | MPLE   | SCAD (GCV)        | LASSO (GCV)        | Hard (GCV)        |
|-------------------------------|--------|------------------|--------------------|------------------|
| Age at the diagnosis of NHL   | 0.2997 | 0.2674           | 0.0811             | 0.2583           |
| Performance Status            | 0.1532 | 0(−)             | 0(−)              | 0.1495           |
| Serum Lactate Dehydrogenase Level | 0.0995 | 0(−)             | 0(−)              | 0.0854           |
| Number of Extranodal Disease Sites | 0.1492 | 0.1585           | 0.0230             | 0.1467           |
| Tumor Stage                   | 0.2662 | 0.2642           | 0.0588             | 0.2031           |

### Table 7
Estimated coefficients and standard errors (in parentheses) for the Gamma frailty model, using GCV

| Factors                        | MPLE   | SCAD (GCV)        | LASSO (GCV)        | Hard (GCV)        |
|-------------------------------|--------|------------------|--------------------|------------------|
| Age at the diagnosis of NHL   | 0.2997 | 0.2674           | 0.0811             | 0.2583           |
| Performance Status            | 0.1532 | 0(−)             | 0(−)              | 0.1495           |
| Serum Lactate Dehydrogenase Level | 0.0995 | 0(−)             | 0(−)              | 0.0854           |
| Number of Extranodal Disease Sites | 0.1492 | 0.1585           | 0.0230             | 0.1467           |
| Tumor Stage                   | 0.2662 | 0.2642           | 0.0588             | 0.2031           |
the three penalized methods resulting in setting 35–37 factors’ coefficients equal to 0. We calculated the standardized coefficients for all the selected factors and we present here the 10 factors having the highest absolute values of standardized coefficients and, therefore, are the most significant (at $\alpha = 5\%$ two-sided). The results are depicted in Table 8.

We also note that for the Gamma frailty model the maximization of the loglikelihood occurred for all methods for $\hat{\alpha} = 20.1217$, resulting in a frailty term with a variance equal to 0.0496 which suggests that a frailty term is probably not necessary for this set of data, therefore, there is no strong evidence of hospital heterogeneity in this dataset. The loglikelihood values were $-33444$, $-31058$, $-29553$, and $-30741$ for the MPLE and the SCAD, LASSO, and HARD penalty cases, respectively. Moreover, the selected values of $\lambda$ were 0.0032, 0.0007, and 0.0114 for the SCAD, LASSO, and HARD penalties.

For the aforementioned value of $\hat{\alpha}$, the values of $C_2(F'(\beta))$ were 114.4288, 129.4753, and 135.7990 with regard to SCAD, LASSO, and Hard, respectively, while the corresponding values of the error estimate [the minimum upper bound obtained by (11)] were 3.9978, 4.5295, and 4.3877. It must be noted that a more extensive analysis is needed in order on one hand to draw general conclusions regarding these data and on the other hand to further reduce the dimension of the relevant factors. However, we do not pursue this here.

### 6. Concluding Remarks

In this article, we propose new estimates of the norm of the error related to the regression parameter in the frailty models framework for clustered data. Based on these estimates we also propose a new procedure for the selection of the thresholding parameter in penalized frailty models. The GCV method is mainly used in the literature for the derivation of this parameter. A thorough simulation study is conducted to compare the new method to GCV and two real medical datasets are also analyzed. Our simulations highlight that the new approach is a quite trustworthy procedure. Moreover, it is easier to implement and
computationally less intensive. Therefore, it could be considered an effective alternative approach for the tuning parameter selection.

Acknowledgments

We would like to thank Professors Dennis K. J. Lin and Runze Li for sending the Matlab codes for the procedures proposed in their articles. We also thank Professor A. Galantai and two anonymous referees for their insightful comments and suggestions that led to a significant improvement of the article.

References

Androulakis, E., Koukouvinos, C., Mylona, K., Vonta, F. (2010). A real survival analysis application via variable selection methods for Cox’s proportional hazards model. *Journal of Applied Statistics* 37:1399–1406.

Androulakis, E., Koukouvinos, C., Vonta, F. (2012). Estimation and variable selection via frailty models with penalized likelihood. *Statistics in Medicine* 31:2223–2239.

Androulakis, E., Koukouvinos, C., Vonta, F. (2014). Tuning parameter selection in penalized generalized linear models for discrete data. *Statistica Neerlandica* 68:276–292.

Antoniadis, A. (1997). Wavelets in statistics: A review (with discussion). *Journal of the Italian Statistical Society* 6:97–144.

Auchmuty, G. (1992). A posteriori error estimates for linear equations. *Numerische Mathematik* 61:1–6.

Bickel, P., Doksum, K. (2007). *Mathematical Statistics*, vol. I, 2nd ed. Upper Saddle River, New Jersey: Pearson Prentice Hall.

Craven, P., Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik* 31:377–403.

Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Association* 96:1348–1360.

Fan, J., Li, R. (2002). Variable selection for Cox’s proportional hazards model and frailty model. *The Annals of Statistics* 30:74–99.

Fan, J., Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. In: Sanz-Sole, M., Soria, J., Varona, J. L., Verdera, J., eds. *Proceedings of the International Congress of Mathematicians*, Vol. 3. pp. 595–622.

Fan, J., Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* 20:101–148.

Fan, Y., Tang, C. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B* 75:531–552.

Galantai, A. (2001). A study of Auchmuty’s error estimate. *Computers and Mathematics with Applications* 42:1093–1102.

Greub, W., Rheinboldt, W. (1959). On a generalization of an inequality of L. V. Kantorovich. *Proceedings of the American Mathematical Society* 10:407–415.

Horn, R. A., Johnson, C. R. (1985). *Matrix Analysis*. Cambridge: Cambridge University Press.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. New York: Springer-Verlag.

Kosorok, M., Lee, B., Fine, J. (2004). Robust inference for univariate proportional hazards frailty regression models. *The Annals of Statistics* 32:1448–1491.

Leng, C., Lin, Y., Wahba, G. (2006). A note on the LASSO and related procedures in model selection. *Statistica Sinica* 16:1273–1284.

Marcus, M., Mine, M. (1992). *A Survey of Matrix Theory and Matrix Inequalities*. New York: Dover.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics* 12:758–765.
Pötscher, B. M., Leeb, H. (2009). On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *Journal of Multivariate Analysis* 100:2065–2082.

The International Non-Hodgkin’s Lymphoma Prognostic Factors Project. (1993). A predictive model for aggressive non-Hodgkin’s lymphoma. *The New England Journal Medicine* 329:987–994.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B* 58:267–288.

Van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Wang, H., Li, R., Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* 94:553–568.

Zhang, Y., Li, R., Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association* 105:312–323.

Zou, H., Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics* 36:1509–1533.