Statistica Neerlandica (2001) Vol. 55, nr. 1, pp. 17–34

Shrinkage and penalized likelihood as methods to improve predictive accuracy

J. C. van Houwelingen*

Department of Medical Statistics, Leiden, The Netherlands P.O. Box 9604 2300 RC Leiden, The Netherlands

A review is given of shrinkage and penalization as tools to improve predictive accuracy of regression models. The James-Stein estimator is taken as starting point. Procedures covered are Pre-test Estimation, the Ridge Regression of Hoerl and Kennard, the Shrinkage Estimators of Copas and Van Houwelingen and Le Cessie, the LASSO of Tibshirani and the Garotte of Breiman. An attempt is made to place all these procedures in a unifying framework of semi-Bayesian methodology. Applications are briefly mentioned, but not amply discussed.

Key words and phrases: Pre-test Estimation, Ridge Regression, LASSO, Garotte

1 Introduction

In the setting of regression and prediction shrinkage estimators have a long history. Some well-known papers are HOERL and KENNARD (1970), COPAS (1983) and VAN HOUWELINGEN and LE CESSIE (1990). These estimators are related to the James-Stein (J-S) estimator (JAMES and STEIN, 1962) and can all be understood from an empirical Bayes point of view as pointed out by EFRON and MORRIS (1972). A modern formulation is by penalized likelihood (HASTIE and TIBSHIRANI, 1986 and EILERS and MARX, 1996) with all kind of new fancy penalties, such as the LASSO of TIBSHIRANI (1996) and the Garotte of BREIMAN (1995). The purpose of this paper is to give a unifying review of all these methods. Practical applications are given in some of the other papers in this special issue.

The starting point is the James-Stein shrinkage estimator that is useful in estimating many similar quantities as in disease mapping, comparison of institutions, meta-analysis etcetera. Section 2 discusses the J-S estimator, and the intuition behind it. Section 3 discusses classical and modern alternatives to James-Stein estimation. From section 4 on, attention is focused on regression problems. Section 4 discusses penalized regression in general terms and introduces the distinction between problems with natural penalties such as fitting splines to be discussed in section 5 and

* email: jcvanhouwelingen@lumc.nl

problems without natural penalties such as multivariate regression, to be discussed in section 7. In between, standard errors and confidence intervals for penalized regression are discussed in section 6. Section 7 also contains a small simulation study. The last section briefly discusses software for penalized regression.

2 James-Stein revisited

The idea of shrinkage estimators goes back to the famous James-Stein estimator (JAMES and STEIN, 1962). In a nutshell the main result is that for the case of k independent normals with given standard deviation equal to one and unknown means, that is $X_1, \ldots, X_k \sim N(\mu_i, 1)$, the shrinkage estimator

$$\tilde{\mu}_i = \overline{X} + \left(1 - \frac{k-3}{\sum(X_i - \overline{X})^2}\right)^+ (X_i - \overline{X})$$

has uniformly smaller mean squared error (MSE) than the Maximum Likelihood Estimator (MLE) $\hat{\mu}_i = X_i$, that is

$$MSE(\tilde{\mu}) = \sum_{i=1}^{k} E(\tilde{\mu}_i - \mu_i)^2 < MSE(\hat{\mu}) = \sum_{i=1}^{k} E(\hat{\mu}_i - \mu_i)^2 \text{ for all } (\mu_1, \ldots, \mu_k).$$

The factor $(1 - (k-3)/\Sigma(X_i - \overline{X})^2)^+$ is known as the shrinkage factor and often denoted by c.

The proof requires some hard mathematics and will not be discussed here. It is more important to get some intuitive feeling why shrinkage might work.

2.1 *Graphical explanation*

The simplest way of developing intuition is by making some graphs in a simulation experiment where the value of the parameter μ is known to the experimenter.

Figure 1 shows some data with arbitrary μ_is and X_is simulated according to the model $X_i = \mu_i + e_i$ with $e_i \sim N(0, 1)$. The regression line of X on μ has slope ≈ 1 as could be expected. However, the regression of μ on X has slope < 1 as can be seen in Figure 2. That phenomenon (attenuation) is well-known in errors-in-measurement models. It is also seen in repeated measures where the best prediction of the next observation (BLUP) shows similar shrinkage towards the mean. Figures 1 and 2 are purely hypothetical, because we cannot observe μ. So we cannot estimate the slope from the graphs and we require a more formal approach to obtain an estimate of the ideal regression of μ on X.

2.2 *Empirical Bayes explanation*

EFRON and MORRIS (1972) give an empirical Bayes explanation of the James-Stein phenomenon.

The key step is to assume that the μ_is are i.i.d with some normal distribution,

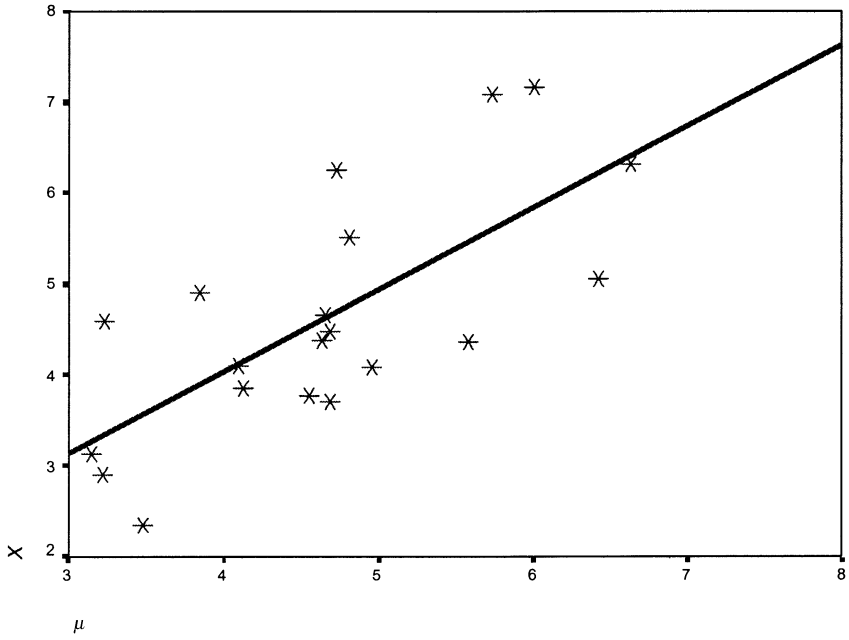

Fig. 1. Relation between X and μ. The graph shows the true μs, the observed Xs and the regression of X on μ.

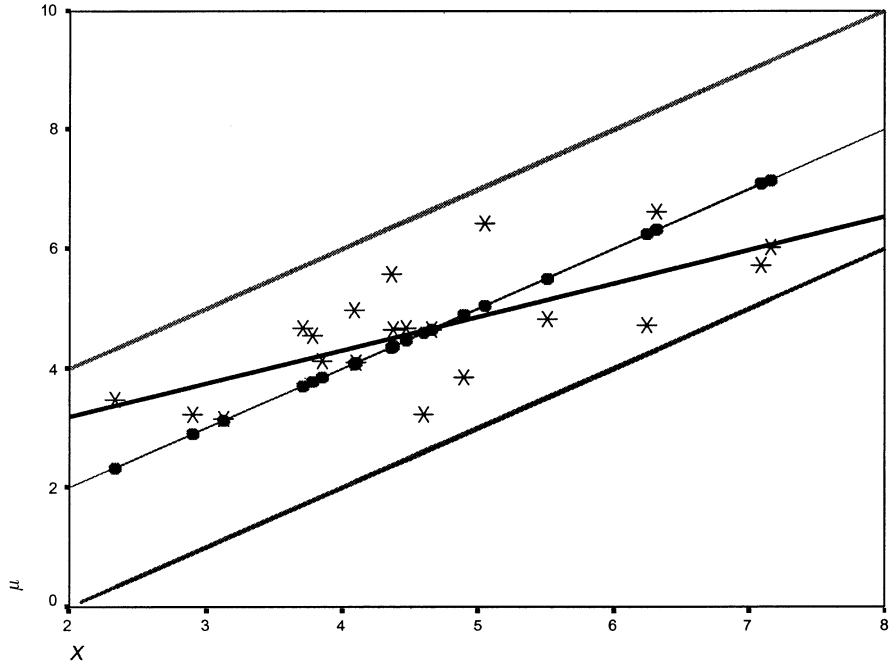

Fig. 2. Relation between μ and X. The graph shows the point estimates, the 95%-confidence intervals for μ, the true μs and the regression of μ on X.

$N(\mu, \tau^2)$ say. Then, the posterior distribution of each μ_i given X_i is again normal with

$$E[\mu_i|X_i] = \mu + \frac{\tau^2}{\tau^2+1}(X_i - \mu) \text{ and } \qquad \text{var}(\mu_i|X_i) = \frac{\tau^2}{\tau^2+1}$$

This formula nicely shows the shrinkage towards the mean. In this model, the parameters of the mixing distribution μ and τ^2 can estimated from the marginal distribution $X_i \sim N(\mu, \tau^2+1)$. Everybody will agree on $\hat{\mu} = \overline{X}$, but there are different ways to estimate τ^2. Actually, the parameter to estimate is $\tau^2/(\tau^2+1)$ and just plugging in an estimate for τ^2 might not be the best idea. (The James-Stein estimator is based on the observation that $(k-3)/\Sigma(X_i - \overline{X})^2$ is the unbiased estimator of $1/(\tau^2+1)$.)

Since the parameters μ and τ^2 have to be estimated, it is not guaranteed that the empirical Bayes estimator works better than the classical MLE. As shown by James and Stein, it only gives an improvement if $k > 3$.

So far we have assumed that the (conditional) variance of the Xs was given and constant (homo-skedastic). It is not hard to generalize the idea to hetero-skedastic models.

2.3 *Non-Bayesian explanation*

Not everybody likes the Bayesian idea of modelling the unknown μ_is as being drawn from some distribution. The original paper by James and Stein does not have this Bayesian flavour, so shrinkage could well be explained without being (empirical) Bayesian. The simplest idea is to find out if, for some reason, we decide to use a shrinkage estimator, what the best choice would be. That is the reasoning behind the ridge estimator of HOERL and KENNARD (1970).

Let us take an estimator that shrinks towards some value μ_0, that is

$$\hat{\mu}_i = \mu_0 + c(X_i - \mu_0)$$

Such a linear shrinkage is suggested by the regression plot of Figure 2. Componentwise, the bias is $(c-1)(\mu_i - \mu_0)$ and the variance c^2. Hence, the mean squared error is given by

$$MSE = \Sigma E[(\hat{\mu}_i - \mu_i)^2] = (c-1)^2\Sigma(\mu_i - \mu_0)^2 + kc^2.$$

First, we observe (with Hoerl and Kennard) that $\partial MSE/\partial c > 0$ at $c = 1$, so it is always profitable to shrink a little. The optimal parameters minimizing the MSE are given by

$$\mu_0 = \bar{\mu} \text{ and } \qquad c = 1 - \frac{k}{\Sigma(\mu_i - \bar{\mu})^2 + k} = \frac{\text{var}(\mu)}{\text{var}(\mu)+1}$$

These can be estimated from the data and yield the same estimator as the empirical Bayes approach.

A different non-Bayesian motivation of shrinkage is given by penalization. The idea is that we do not think that the μ_is are very different and, therefore, we penalize

the log-likelihood or the sum-of-squares by a penalty that increase with increasing differences between the μ_is. That leads to estimation of the μ_is by minimizing

$$\Sigma(X_i - \mu_i)^2 + \lambda\Sigma(\mu_i - \bar{\mu})^2.$$

The solution of this minimization problem is just the shrinkage estimator with $c = 1/(\lambda + 1)$. The optimal λ can again be obtained from the data by estimating the Mean Squared Error of the procedure. It should be emphasized that we do not need anything like cross-validation to estimate the optimal shrinkage/penalty parameter. Even stronger, it is not possible to perform cross-validation by lack of repeated measures.

3 Alternatives to the James-Stein estimator

In the literature other procedures have been suggested that also aim at improving over the maximum likelihood estimator. We discuss the classical pre-test estimators and the 'modern' LASSO and Garotte.

3.1 *Pre-test estimators*

The classical procedure is to test some (linear) null-hypothesis about the μ_is by means of Analysis of Variance and estimate the μ_is under that restriction if the null-hypothesis is not rejected. (BANCROFT, 1964) The simplest case is to take H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$. If H_0 is not rejected, the μ_is are estimated by the common mean. If H_0 is rejected, the μ_is are estimated by the individual observations. All kinds of extensions are possible that allow for the grouping of observations and the search for outliers. What is needed is a formal description of how to proceed. The behaviour of the procedure depends on the significance level α used in the test phase and the configurations of the μ_is. The MSE is hard to obtain and it is not easy to obtain practical rules which significance level α should be used. There is some old theory (STEIN, 1981) about how to estimate the MSE that could be useful, but it is far from straight forward how it should be implemented. The choice $\alpha = 0.05$, used as a rule of thumb by all applied statisticians, lacks a clear motivation. A Bayesian formulation of the testing problem does not seem to be very helpful either.

3.2 *LASSO*

This procedure is due to TIBSHIRANI (1996) inspired by BREIMAN (1995). It is usually introduced as a penalized likelihood method with a penalty based on the sum of absolute deviations $\lambda\Sigma|\mu_i - \mu|$, but we think that it is better understood from the empirical Bayes point of view. Instead of the normal mixing distribution that leads to J-S shrinkage, we take the double exponential prior with density $1/(2\tau)\exp(-|\mu_i - \mu|/\tau)$ and estimate the parameter μ_i by the posterior mode. Given μ and τ the minimization of $\Sigma(X_i - \mu_i)^2 + \Sigma|\mu_i - \mu|/\tau$ leads to the interesting estimator

$$\hat{\mu}_i = \mu \qquad \text{if } |X_i - \mu| \leqslant 1/\tau$$
$$\hat{\mu}_i = X_i - 1/\tau \qquad \text{if } X_i > \mu + 1/\tau$$
$$\hat{\mu}_i = X_i + 1/\tau \qquad \text{if } X_i < \mu - 1/\tau.$$

The resulting estimator is shown in Figure 3.

The procedure behaves like correcting the outliers a little bit and pooling the rest of the observations. It has much of the flavour of the pre-test estimator. However, it is far from obvious what values of μ and τ should be used. If you take the empirical Bayesian point of view, it is not clear how to estimate μ and τ from the marginal distribution of the X_is. If you view it from the penalized likelihood point of view, it is far from easy to estimate the MSE and to find the values with the lowest estimated MSE.

3.3 *Garotte*

This procedure is one of the many data-analytic proposals of Leo Breiman (BREIMAN, 1995). It defines an estimator by means of individual shrinkage factors, that is

$$\hat{\mu}_i = \overline{X} + c_i(X_i - \overline{X})(c_i > 0)$$

The amount of shrinkage is controlled by putting a penalty on the shrinkage factors. The quadratic penalty $\lambda \Sigma c_i^2$ leads to $c_i = (X_i - \overline{X})^2/(\lambda + (X_i - \overline{X})^2)$ and the linear penalty $2\lambda \Sigma c_i$ leads to $c_i = (1 - \lambda/(X_i - \overline{X})^2)^+$. The latter is shown in Figure 3 as well.

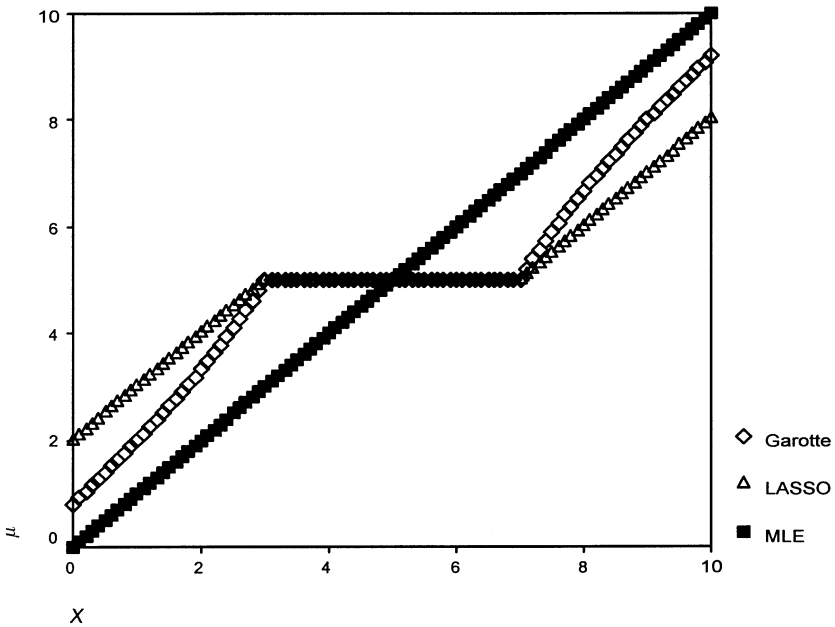

Fig. 3. Comparison of the MLE, the LASSO with $\mu = 5$ and $\tau = 0.5$ and the linear penalty Garotte with $\lambda = 4$.

The intuitive advantage is that extreme outliers are not corrected at all. However, the behaviour of this estimator is very hard to determine analytically and the penalty also lacks any Bayesian interpretation.

3.4 *Comparing alternatives*

The performance of any of these shrinkage estimators depends on the true value of $\mu_1, \ldots, \mu_k$. Simulation studies can never cover all possibilities. For given $\mu_1, \ldots, \mu_k$ that procedure performs best for which the corresponding prior model gives the best fit. Therefore, simulation studies that prove superiority of a particular procedure should always be mistrusted. They only show that the author managed to find some μ-configuration for which his procedures works best. In a purely Bayesian setting these alternatives could be combined by defining mixture models for the distribution of the μ_is.

The pre-test estimators and the related LASSO and Garotte seem to be most appealing to the practitioner. However, the procedures based on quadratic penalties (or normal mixtures if you are a Bayesian) have the most transparent behaviour of the estimators and are best controlled. Therefore, we will focus in the sequel on shrinkage implied by quadratic penalties.

4 Penalized regression

In this section we switch from the unstructured ANOVA-type of situation with k different μ_is to the regression situation, where there is more structure in the μ_is. We consider the classical regression model

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + error$$

with n observations, p covariates, an intercept and homo-skedastic error with unknown variance σ^2. In matrix notation we have

$$Y = X\beta + e$$

The classic Ordinary Least Squares (OLS)-estimator is given by

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$$

The question is if we can improve upon the OLS-estimator by shrinking the β_is somehow.

The answer is: Yes we can, if we are careful.

The oldest procedure is the Ridge Regression that was introduced by HOERL and KENNARD (1970), mainly to fight multicollinearity in linear regression. To define the ridge estimator, we first centre (and standardize) all variables. The intercept is taken as $\hat{\beta}_0 = \overline{Y}$ and further ignored. The regression coefficients are estimated by

$$\hat{\beta}_{Ridge} = (X'X + kI)^{-1}X'Y \quad (k > 0)$$

Adding the 'ridge' kI to $X'X$ has two effects. First, the regression coefficients behave

less wildly if there is multicollinearity. Secondly, the $MSE = E\|Y - X\hat{\beta}\|^2$ gets smaller by virtue of the trade off between bias and variance, if k is not too big. The question is what the best choice is for k. Before answering that, we put the ridge regression estimator in the more general setting of the Penalized Sum of Squares methodology.

Let P be any non-negative definite matrix for which $\beta' P\beta$ serves as a reasonable penalty on the vector of regression coefficients β. The Penalized Sum of Squares estimator is defined by

$$\hat{\beta}_{Pen} \text{ minimizes } \|Y - X\beta\|^2 + \lambda\beta' P\beta$$

It is not hard to check that it is given by

$$\hat{\beta}_{Pen} = (X'X + \lambda P)^{-1} X'Y$$

(Ridge regression is obtained by taking $P_{ii} = 1$ if $i > 0$ and $P_{ij} = 0$ elsewhere.)

The criterion for the choice of optimal λ (and β) is the actual (ACT) expected prediction error (PE) defined by

$$PE_{ACT} = \frac{1}{n}\Sigma E_{future}(Y_{i,new} - X_i\hat{\beta}_{Pen})^2$$

We use the terminology of VAN HOUWELINGEN and LE CESSIE (1990). In this expression, expectation is only with respect to the future (new observations), not with respect to the past (observations that were used to estimate β). It measures the performance of the actually obtained predictor if new observations are to be predicted in the same design points. PE_{ACT} depends on the unknown parameters β and σ. It can be estimated by comparing it with the apparent (APP) prediction error on the current data set

$$PE_{APP} = \frac{1}{n}\Sigma(Y_i - X_i\hat{\beta}_{Pen})^2$$

In the spirit of AKAIKE (1973) it can be shown that

$$E_{past}(PE_{ACT} - PE_{APP}) = \frac{2\sigma^2}{n} dim_{eff}$$

where

$$dim_{eff} = trace((X'X + \lambda P)^{-1} X'X)$$

Therefore, the 'optimal' λ can be obtained by minimizing

$$\widehat{PE}_{ACT} = PE_{APP} + \frac{2}{n}\hat{\sigma}^2 dim_{eff},$$

with

$$\hat{\sigma}^2 = \Sigma(Y_i - X_i\hat{\beta}_{Pen})^2/(n - dim_{eff}).$$

This is closely related to Akaike's AIC and Mallow's C_P (MALLOWS, 1973). However, it should be kept in mind that *the optimal λ is only estimated, it is not known*. It is not

clear what the price is of estimating λ. It might be quite high. This issue is related to the dimension problem of the James-Stein estimator, that only works for at least four 'observations'. It might be expected that shrinkage in the regression setting only works if there are at least three covariates yielding four unknown parameters.

An alternative approach to obtain the optimal λ is by means of cross-validation. The procedure is well-known: leave the i-th observation out, fit the model to the remaining observations (result denoted by $(-i)$) and see how well the i-th observation is predicted. That leads to minimizing the cross-validated PE

$$PE_{CV} = \frac{1}{n}\sum (Y_i - X_i\hat{\beta}^{(-i)}_{Penal})^2$$

Cross-validation is intuitively appealing, but it is not 100% clear what actually is estimated by PE_{CV} and it might be impossible to compute $\hat{\beta}^{(-i)}_{Penal}$ for all i, if leaving out some observation leads to collinearity of the design matrix. On the other hand, it nicely mimics the situation of random X_is in contrast to 'Akaike' that is based on fixed X_is. The practical conclusion is that cross-validation is fine in most cases. Cross-validation can even be replaced by bootstrapping (EFRON and TIBSHIRANI, 1986). That gives a better estimate of PE_{ACT} but is more computer-intensive.

The main issue is that there is a single parameter λ that controls the whole process and that it is attempted to find the value of λ that gives the smallest prediction error. The true prediction error is unknown, but we can try to estimate it.

It is also possible to extend the system to generalized penalties of the structure $P = \Sigma\lambda_i P_i$, but the more parameters have to be estimated, the less stable the estimator becomes. The main question is what penalty to use. We can distinguish between situations where natural penalties exist such as in the case of smoothing problems and situations without natural penalties.

5 Regression problems with natural penalties

The most outspoken example is that of fitting a smoothing spline. The simple model is

$$Y = f(X) + error$$

No model for $f(x)$ is assumed but 'smoothness' in a very vague sense.

Let the observations be $(Y_1, X_1), \ldots, (Y_n, X_n)$. The function f is estimated by minimizing

$$\sum_{i=1}^{n}(Y_i - f(X_i))^2 + \lambda\int f^{(k)}(x)^2\,\mathrm{d}x \qquad (f^{(k)}(x) \text{ is the } k\text{-th derivative})$$

Generally, the problem can be solved in two steps by first minimizing the integral in the expression above for fixed values of $f(X_i)$, $f(X_i) = \mu_i$, say. The minimum is a quadratic function of the μ_is, $\mu' P\mu$ say, that serves as the penalty in a reduced problem where not the whole function, but only the $\mu_i = f(X_i)$s have to be estimated.

Once the μ_is are estimated, the complete function is estimated by the k-th order interpolation spline.

For equidistant X, one can simplify the problem to minimizing

$$\Sigma(Y_i - \mu_i)^2 + \lambda\Sigma[\Delta^{(k)}(\mu_i)]^2$$

with $\Delta^{(k)}$: k-th order difference.

The effect of enhanced smoothing caused by larger values of λ is demonstrated in Table 1 and Figure 4. In this example 2nd-order differences are used.

Table 1. Effect of different penalty weights λ when smoothing the data of Figure 5

λ	SS	dim_{eff}	$\hat{\sigma}^2$	$SS + 2dim_{eff}\hat{\sigma}^2$
1	43.357	9.001	3.613	108.406
10	62.746	5.287	3.993	104.974
100	70.504	3.365	3.998	97.411
1000	92.186	2.353	4.944	115.452
10000	109.125	2.045	5.757	132.670

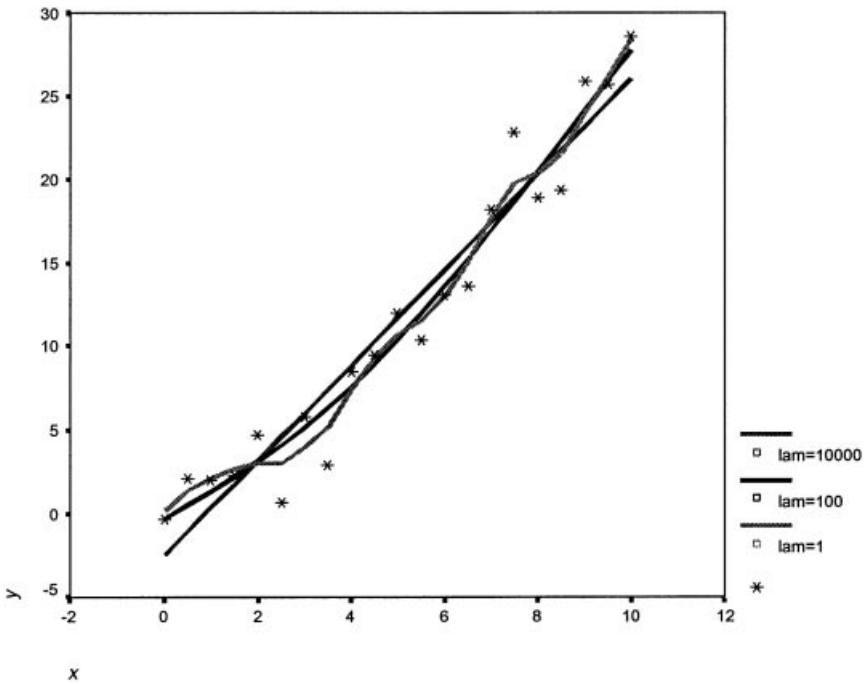

Fig. 4. Smoothing splines for different values of λ. The straight line corresponds with $\lambda = 10\,000$, the parabola-like line with $\lambda = 100$ and the wavy line with $\lambda = 1$.

This approach can be extended to the generalized additive models in more covariates X_1, X_2 ... of HASTIE and TIBSHIRANI (1986), that is

$$Y = f_1(X_1) + f_2(X_2) + \ldots + error$$

However, it is not clear how the penalties for the different covariates should be combined. Having separate penalties with independent λ-weights leads to intractable optimization problems.

If there are many observations, fitting a completely unspecified function leads to a very large number of free parameters and the danger of numerical problems. A practical simplification that makes the problem more tractable is to categorize X first in k categories $C_1, \ldots, C_k$ that are represented by $(k-1)$ dummy covariates and corresponding parameters γ_i. It should be realized that the definition of a reasonable smoothness penalty depends on the definition of dummies. For example, if the dummies are defined as

$$D_i = [X \in C_{i+1}] \qquad (i = 1, \ldots, k-1) \text{ and } \qquad Y = \gamma_0 + \sum_{1}^{k-1} \gamma_i \mathrm{D}_i + error,$$

that is C_1 is taken as baseline, the penalty based on first order differences is

$$Penalty = \gamma_1^2 + \sum_{i=2}^{k-1} (\gamma_i - \gamma_{i-1})^2$$

A definition of the first-order difference penalty that does not depend on the definition of the dummies is given by

$$Penalty = \sum_{i=2}^{k} (E[Y|X \in C_i] - E[Y|X \in C_{i-1}])^2$$

This penalty reflects the ordinal character of the categorization. A purely nominal penalty would be given by

$$Penalty = \sum_{i=1}^{k} (E[Y|X \in C_i] - \overline{E[Y|X \in C_i]})^2.$$

A nice compromise between splines and categorization is given by the penalized B-splines of EILERS and MARX (1996).

The model can simply be extended to more ordinal categorical covariates and interactions between them. An elaborated example is in the field of survival analysis is given by VERWEIJ and VAN HOUWELINGEN (1994). For survival data the objective function to minimize is $-2 * \ln(\textit{partial likelihood})$ (we will come back to that later), but the penalty functions for lack of smoothness of the regression coefficients are the same. In that example, penalized categorical systems gave a slightly better fit than simple ordinal 1, 2, 3, ... coding combined with linear regression. The reader is referred to the paper for more details.

Another example of natural penalization is where the (continuous) covariates

$X_1, \ldots, X_k$ form a 'signal'. In that case it is natural to assume that the β_is should depend in a smooth way on the index i.

An example discussed in LE CESSIE and VAN HOUWELINGEN (1992) is a study on the survival of ovarian cancer patients. The prognostic signal is the histogram of the DNA-content (C) of a sample of a patient's tumour cells. The histogram is produced by DNA cytometry. The content C is standardized in such a way that the modal value of C equals two for normal cells. The histogram consists of a fixed number of k classes. So, X_i is the relative frequency of class i. Using a first order difference penalty, reasonably interpretable results were obtained: histograms that peak around C-values of 2 or 4 are associated with a good survival prognosis, histograms that have peaks at other C-values are associated with poor prognosis, especially if very large C-values (8 or higher) are very frequent in the histogram.

Similar regression on signals arises when $X_1, \ldots, X_k$ represents a physical, chemical or DNA-marker spectrum or a time series of bio-marker values. For a further discussion see EILERS and MARX (1999).

Finally, similar smooth sequences of parameters appear in modelling time-dependent effects as in VERWEIJ and VAN HOUWELINGEN (1995), see also HASTIE and TIBSHIRANI (1993).

6 Standard errors and confidence intervals

Penalized estimators do not fit well into classical theory. The standard errors could be obtained by sandwich estimators as used by Generalized Estimating Equations (LIANG, ZEGER and QAQISH, 1992), but the big problem is the unknown bias induced by the bias–variance trade-off.

A practical solution is to act as if the estimators are unbiased with covariance matrix $\hat{\sigma}^2(X'X + \hat{\lambda}P)^{-1}$. Actually, this can be seen as a 'Bayesian' interpretation of the penalty function as sketched in section 2.2. If P is non-singular, one acts as if the prior distribution of β is $N(0, (\lambda P)^{-1})$ and compute posterior means, (co)variances and posterior credibility intervals. In this line of reasoning λ and σ^2 can be estimated from the marginal distribution of the random effect model: $Y \sim N(0, \sigma^2 I + (\lambda P)^{-1})$. The corresponding criterion is known as ABIC (Akaike's Bayesian Information Criterion) (which does not coincide with SCHWARZ (1978)'s BIC criterion). The question that is still to be studied is how much the estimated λ differs from the one based on AIC. If P is singular, things are more complicated, but the main line of reasoning is maintained. See HEISTERKAMP *et al.* (1999) for an application of ABIC in smoothing AIDS incidence figures. The confidence intervals described above do not take into account that λ (and σ^2) are estimated as well. That can be handled by using 'hyperprior Bayes' methdology, i.e. by putting a prior on λ, σ^2 and the part of the regression parameter β that is not controlled by the penalty. Analytically, things become intractable, but MCMC might be helpful here. More research is needed. Our fear is that effect of not knowing λ (and σ^2) could be disappointingly large.

7 Regression problems without natural penalties

In general parametric regression models with many covariates such as used in building prognostic models, there is no natural penalty on the vector of regression coefficients. The penalty used in classical ridge regression comes more or less out of the blue. The obtained solution is not invariant under scale transformations. A solution is either to scale Xs first or (equivalently) take a diagonal penalty matrix with $P_{ii} = \Sigma_j (X_{ij} - X_{i.})^2$. This could be described as ridge regression on the correlation matrix instead of covariance matrix. This estimator is still not invariant under linear transformation of the covariates. The only invariant penalty matrix is the cross-product-matrix of the covariates $P_{ij} = \Sigma_l (X_{il} - X_{i.})(X_{jl} - X_{j.})$. That could be described as transforming the covariates to an orthonormal basis with the constant vector as first component and taking $P_{ii} = 1$ for $i = 2, \ldots$ and all other $P_{ij} = 0$. The optimal estimator under penalty λP shows uniform shrinkage, that is

$$\hat{\beta}_{shrink} = c\hat{\beta}_{OLS} \quad \text{(not for the constant) with } c = 1/(1+\lambda)$$

This is the shrinkage estimator of COPAS (1983) and VAN HOUWELINGEN and LE CESSIE (1990). It can also be seen as a Bayesian solution with respect to a prior with a covariance matrix $\tau^2 (X'X)^{-1}$.

There are several methods to estimate the optimal choice for c. The first is by cross-validation. That leads in a natural way to *calibration regression*, that is regression of Y_i on $X_i \hat{\beta}_{OLS}^{(-i)}$ leading to $\hat{c}_{CV}$. Another way is by optimization of AIC, which leads to

$$\hat{c}_{AIC} = 1 - p/(\hat{\beta}(X - \overline{X})'(X - \overline{X})\hat{\beta}/\hat{\sigma}^2) = 1 - p/(model\ \chi^2),$$

where p is the number of covariates. See also VAN HOUWELINGEN and LE CESSIE (1992). A third way is by bootstrapping leading to $\hat{c}_B$. The precise way of application of the bootstrap is to estimate the prediction error and to minimize that. A practical, but not quite equivalent method is to compute an estimate of the shrinkage factor in each bootstrap by regression of the original data on the predictor estimated in the bootstrap and to average the shrinkage factor over the bootstrap samples. The latter procedure will be applied in section 7.3. Once the shrinkage factor c is estimated, the shrunken predictor $\overline{Y} + \hat{c}(X - \overline{X})\hat{\beta}_{OLS}$ is used, hoping that it leads to smaller prediction error.

The condition of invariance under linear combination of covariates that leads to the uniform shrinkage factor is not always natural. Many people have a desire for individual shrinkage. That desire is met by the Garotte of BREIMAN (1995) that is defined as the minimizer of

$$\sum \left(Y_i - \hat{\beta}_0 - \sum_j c_j (X_{ij} - X_{i.})\hat{\beta}_j \right)^2 + \lambda P(c)$$

where $\hat{\beta}$ is the OLS estimator and $P(c)$ some penalty on the cs. There are several choices for the penalty as discussed in section 3.3 . Intuitively it is all great, but the

method lacks a theoretical foundation. An alternative is to group the covariates in subgroups of supposedly 'weak' predictors and supposedly 'strong' predictors and to have penalties and free λ per subgroup, provided the subgroups are not too small. The optimal λs can be estimated by minimizing the estimated prediction error, using either cross-validation, approximate AIC or bootstrapping. Generally speaking the resulting predictor has a complicated structure, is not equivalent with group-wise shrinkage and can not be obtained by calibration regression.

7.1 *Non-quadratic penalties*

The intuitive desire to have a shrinkage procedure that produces about the same answer as pre-test estimators can be satisfied by taking non-quadratic penalties. As we have seen in section 3.2 the LASSO of TIBSHIRANI (1996) that is obtained by scaling the Xs and taking the penalty to be $\lambda\Sigma|\beta_i|$ leads to shrinkage patterns as sketched in Figure 3, that is βs close to zero are put at zero and larger βs are virtually unchanged. However, finding the optimal βs is far from straightforward.

A solution close to classical best subset selection is obtained by taking the penalty simply to be equal to $\lambda \times$ the number of non-zero βs. The advantage is again that many $\hat{\beta}_i = 0$ are obtained. However, the disadvantage of all these 'intuitive' penalties is that the optimal solution for given λ is hard to obtain and that finding the optimal λ is even harder. Tibshirani gives some approximation to the prediction error of his LASSO, but the motivation is not very rigorous. Furthermore, it is far from clear how some kind of confidence intervals can be obtained. Proof of the nice results obtained for some type of penalty is often given by simulation. However, it should be realized that penalties have some Bayesian flavour as $-\ln$*(prior density)* and that it is very easy to manipulate simulations by taking an example that is more likely under the prior corresponding to the penalty studied. So, the LASSO will behave well if the βs look like coming from a double exponential distribution, that is many close to zero and some very big. The classic Ridge Regression will behave better if the histogram of the βs look more normal.

The problem of estimating the prediction error for the non-quadratic penalties can be overcome by taking the 'hyperprior Bayes' view by interpreting the penalty as $-\ln$*(prior density)* and putting a vague prior on the remaining free parameters β_0, σ_2 and λ. In such a model the hard optimization problems can be replaced by relatively easy but time-consuming Markov Chain Monte Carlo simulation.

7.2 *Non-normal outcomes*

Everything discussed so far easily carries over to non-normal models, such as logistic regression, Poisson regression, Cox regression and the like. Generally speaking we have independent observations $Y_1, \ldots, Y_n$ with a distribution depending on $\theta_i = X_i\beta$, for example logistic regression with $P(Y = 1) = 1 - P(Y = 0) = \exp(\theta)/(1 + \exp(\theta))$, yielding a log-likelihood $l(\beta) = \Sigma l_i(\beta|Y_i, X_i)$. (In Cox regression with its partial likelihood things are slightly more complicated.) As a criterion for predictive value we take the expected log-likelihood

$(1/n)E_{future}[\Sigma l_i(\hat{\beta}|Y_{new,i}, X_i)]$, sometimes also known as the Kullback–Leibler criterion.

Now, the penalized likelihood estimator maximizes $l(\beta) - \frac{1}{2}\lambda\beta' P\beta$. An AIC type of criterion can be based on the relation

$$E_{past+future}[\Sigma l_i(\hat{\beta}_{Penal}|Y_i, X_i) - \Sigma l_i(\hat{\beta}_{Penal}|Y_{new,i}, X_i) \approx dim_{eff}$$

with

$$dim_{eff} = trace((H + \lambda P)^{-1} H) \text{ and } H = -\frac{\partial^2 l(\hat{\beta}_{Penal})}{\partial \beta^2}$$

The approximated AIC is given by

$$AIC = l(\hat{\beta}) - dim_{eff}$$

(Similar approximations can also be obtained for ABIC.)

Observe that things are slightly easier here than in the normal case, because we do not have to estimate the scale factor σ.

Of course, it is again possible to estimate the 'prediction error' by cross-validation or bootstrapping. It is not so hard to show that CV and AIC are nearly identical. In the case of the normal distribution uniform shrinkage was obtained from the requirement that the penalty was invariant under linear transformations of the covariates. In the non-normal case that is less obvious. The penalized estimator is getting close to uniform shrinkage if the penalty is chosen proportional to information matrix H. However, the latter depends on β and it seems dangerous to let the penalty depend on an estimated value of β. We conjecture that uniform shrinkage is alright if the MLE is reasonably stable, but that it is wiser to use ridge regression type penalties (as used in LE CESSIE and VAN HOUWELINGEN 1992) that do not depend on the estimated parameters. More research is needed about the applicability of uniform shrinkage and cross-validation based calibration regression in the non-normal case.

7.3 *A simulation study*

As stated before, it is not warranted that the predictive performance is actually improved if penalty weight λ or the shrinkage factor c are estimated from the data. To see whether any gain could be achieved a simulation experiment was performed, simular to the one reported in VAN HOUWELINGEN (1997).

We take the model

$$Y = 0.2X_1 + 0.15X_2 + 0.1X_3 + 0.05X_4 + error$$

with X_1, X_2, X_3, X_4 and *error* all independent $N(0, 1)$. In each simulation experiment we take a sample size of $n = 100$ and we perform 250 of these experiments.

In each simulation we compute the Least Squares estimator $\hat{\beta}_{OLS}$ and the shrinkage factors $\hat{c}_{AIC}$, $\hat{c}_{CV}$ and $\hat{c}_B(= \hat{c}_{Bootstrap})$. Since we know the 'future' in such a simulation experiment, we can also compute the optimal shrinkage factor c_{opt} and the

MSE of $X\hat{\beta}_{OLS}$ and all shrinkage estimators. The results are presented in Table 2 and Figure 5.

The first conclusion is that there is indeed some gain from shrinkage and that bootstrapping works best. This is in line with the general experience that boot-

Table 2. Comparing the efficacy of different shrinkage estimators in the simulation study. The table indicates the fraction of simulation experiments in which the shrinkage estimator has small prediction error than the OLS-predictor

shrinkage method	better than $X\hat{\beta}_{OLS}$
$\hat{c}_{AIC}$	77%
$\hat{c}_{CV}$	75%
$\hat{c}_{B}$	90%

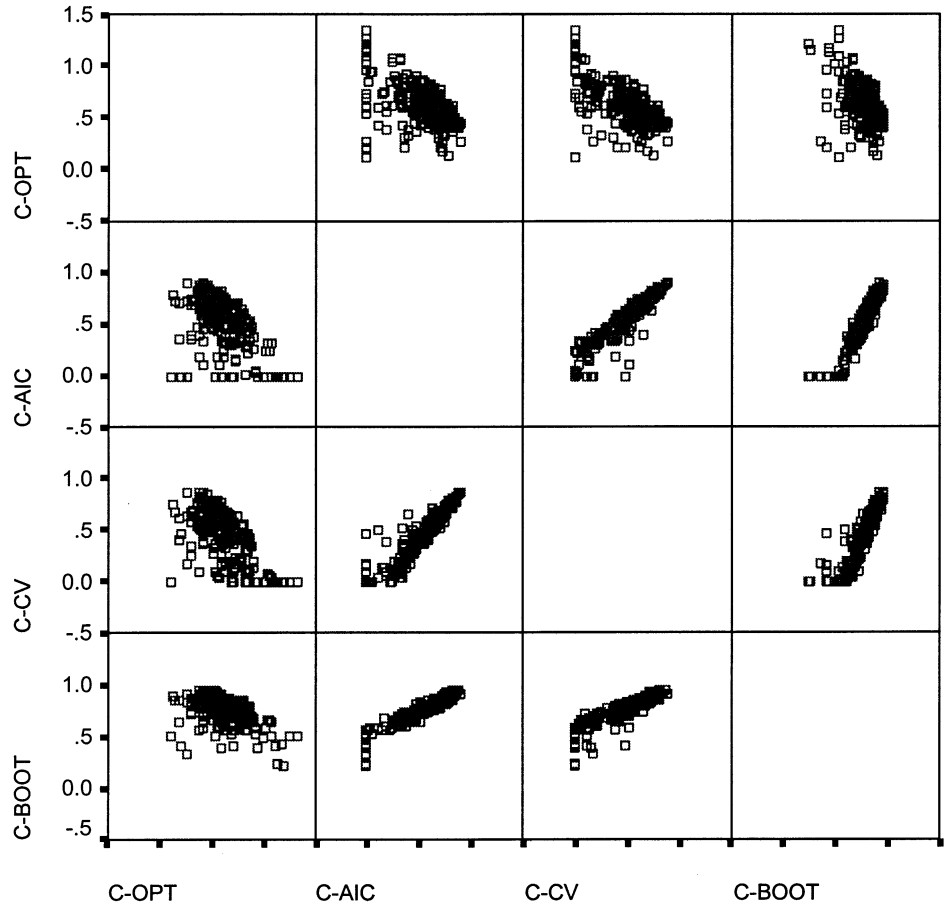

Fig. 5. Cross-scatter of different shrinkage estimators in the simulation study.

strapping gives better estimates of the prediction error than cross-validation or AIC-type methods.

It is surprising to observe that the estimated shrinkage factors can be quite off the mark and are **negatively correlated** with optimal shrinkage factor. This can be best understood from the AIC-based shrinkage factor $\hat{c}_{AIC} = 1 - p/(\hat{\beta}(X - \overline{X})'(X - \overline{X})\hat{\beta}/\hat{\sigma}^2) = 1 - p/(model\ \chi^2)$. If β is 'large' by random fluctuation, the model χ^2 gets large and $\hat{c}_{AIC}$ stays close to one, while the optimal solution (that knows the true value of β) corrects for the 'large' β by taking small value $\hat{c}_{opt}$.

This shows once again that intuition has to be mistrusted and that shrinkage works on the average but may fail in the particular unique problem on which the statistician is working.

8 Software

It is not so hard for an experienced programmer to write his own penalized likelihood procedures and to find the optimal penalty weight λ as long as they use quadratic penalties. Absolute value penalties as used in the LASSO is much harder to implement.

There are all kind of software modules available through the Internet. For example, Frank Harrell has incorporated shrinkage and penalization in his Design software for S-plus that can be approached through

http://hesweb1.med.virginia.edu/biostat/s/index.html

The recent paper by VERBYLA *et al*. (1999) shows how software for Random Effect Models (SAS Proc Mixed, S-plus LME, etc.) can be used to obtain smoothing splines. We think these ideas can be generalized to penalized regression, but we do not have practical experience so far in forcing existing standard software to carry out penalized regression.

References

AKAIKE, H. (1973), Information theory and an extension of the maximum likelihood principle, in: B. N. PETROV and F. CSAKI, 2nd International Symposium *on Information Theory*, 268–281, Akademiai Kiado, Budapest.

BANCROFT, T. A. (1964), Analysis and inference for incompletely specified models involving the use of preliminary test(s) of significance, *Biometrics* **20**, 427–442.

BREIMAN, L. (1995), Better subset selection using the nonnegative Garotte, *Technometrics* **37**, 373–383.

COPAS, J. B. (1983), Regression, prediction and shrinkage, *Journal of the Royal Statistical Society, Series B* **45**, 311–354.

EFRON, B. and C. MORRIS (1972), Limiting risk of Bayes and empirical Bayes estimators, *Journal of the American Statistical Association* **67**, 130–139.

EFRON, B. and R. TIBSHIRANI (1986), Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy, *Statistical Science* **1**, 54–74.

EILERS, P. H. C. and B. D. MARX (1996), Flexible smoothing using B-splines and penalized likelihood, *Statistical Science* **11**, 89–121.

HARRELL Jr, F. E., K. L. LEE and D. B. MARK (1996), Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacym and measuring and reducing errors, *Statistics in Medicine* **15**, 361–387.

HARRELL Jr, F. E., P. A. MARGOLIS, S. GOVE, K. E. MASON, E. K. MULHOLLAND, D. LEHMANN, L. MUHE, S. GATCHALIAN and H. F. EICHENWALD (1998), Development of a clinical prediction model for an ordinal outcome: the World Health Organization multicentre study of clinical signs and etiological agents of pneumonia, sepsis and meningitis in young infants, *Statistics in Medicine* **17**, 939–944.

HASTIE, T. J. and R. J. TIBSHIRANI (1986), Generalized additive models, *Statistical Science* **1**, 297–318.

HASTIE, T. and R. TIBSHIRANI (1993), Varying-coefficient models, *Journal of the Royal Statistical Society*, *Series B* **55**, 757–796.

HEISTERKAMP, S. H., J. C. VAN HOUWELINGEN and A. M. DOWNS (1999), Empirical Bayesian estimators for a Poisson process propagated in time, *Biometrical Journal* **41**, 385–400.

HOERL, A. E. and R. W. KENNARD (1970), Ridge regression, biased estimation for nonorthogonal problems, *Technometrics* **12**, 55–67.

JAMES, W. and C. STEIN (1962), Estimation with quadratic loss, *Proceeding of the Fourth Berkeley Symposium* **1**, 361–373.

LE CESSIE, S. and J. C. VAN HOUWELINGEN (1992), Ridge Estimators in logistic regression, *Applied Statistics* **41**, 191–201.

LIANG, K. Y., S. L. ZEGER and B. QAQISH (1992), Multivariate regression-analyses for categorical-data, *Journal of the Royal Statistical Society*, *Series B* **54**, 3–40.

MALLOWS, C. L. (1973), Some comments on C-P, *Technometrics* **15**, 661–675.

MARX, B. D. and P. H. C. EILERS (1999), Generalized linear regression on sampled signals and curves: A P-spline approach, *Technometrics* **41**, 1–13.

SCHWARZ, G. (1978), Estimating the dimension of a model, *Annals of Statistics* **6**, 461–464.

STEIN, C. M. (1981), Estimating the mean of a multivariate normal distribution, *Annals of Statistics* **9**, 1135–1151.

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

VAN HOUWELINGEN, H. C. (1995), Validation, calibration and updating of prognostic models, *Biocybernetics and Biomedical Engineering* **17**, 127–137.

VAN HOUWELINGEN, J. C. and S. LE CESSIE (1990), Predictive value of statistical models, *Statistics in Medicine* **9**, 1303–1325.

VERBYLA, A. P., B. R. CULLIS, M. G. KENWARD and S. J. WELHAM (1999), The analysis of designed experiments and longitudinal data by using smoothing splines, *Journal of the Royal Statistical Society*, *Series C-Applied Statistics* **48**, 269–300.

VERWEIJ, P. J. M. and J. C. VAN HOUWELINGEN (1994), Penalized likelihood in Cox regression, *Statistics in Medicine* **13**, 2427–2436.

VERWEIJ, P. J. M., and H. C. VAN HOUWELINGEN (1995), Time-dependent effects of fixed covariates in Cox regression, *Biometrics* **51**, 1550–1556.

Received: January 2000. Revised: July 2000.