Journal of the Korean Statistical Society (2004), 33: 1, pp 25-34

REGRESSION WITH CENSORED DATA BY LEAST SQUARES SUPPORT VECTOR MACHINE

DAEHAK KIM[1], JOOYONG SHIM[1] AND KWANGSIK OH[1]

ABSTRACT

In this paper we propose a prediction method on the regression model with randomly censored observations of the training data set. The least squares support vector machine regression is applied for the regression function prediction by incorporating the weights assessed upon each observation in the optimization problem. Numerical examples are given to show the performance of the proposed prediction method.

AMS 2000 subject classifications. Primary 62N01; Secondary 62J02.
Keywords. Regression models, randomly censored data, least squares support vector machines.

1. INTRODUCTION

The least squares support vector machine (LS-SVM), a modified version of support vector machine introduced by Vapnik (1995, 1998) in a least squares sense, has been proposed for classification and regression by Suykens and Vandewalle (1999). In LS-SVM the solution is given by a linear system instead of a quadratic program problem. The fact that LS-SVM has an explicit primal-dual formulations has a number of advantages. Taking account of the fact that the computational complexity increases strongly as the number of training data becomes larger, LS-SVM regression can be estimated efficiently for the huge data set by using iterative methods.

The accelerated failure time model (AFT) and the least squares method to accommodate the censored data seem appealing since they are familiar and well understood. Koul *et al.* (1981) gave a simple least squares type estimation procedure in the censored regression model with the weighted observations and also

Received April 2003; accepted September 2003.
[1]Department of Statistical Information, Catholic University of Daegu, Gyengsan 712-702, Korea (e-mail : dhkim@cu.ac.kr)

showed the consistency and asymptotic normality of the estimator. Zhou (1992) proposed an M-estimator of the regression parameter based on the censored data using the weights Koul *et al.* (1981) proposed.

In this paper we obtain the predicted regression function by LS-SVM based on the censored observations of the training data set. The similar weighting scheme as Zhou (1992) used and the squared error loss function are included in the optimization problem of LS-SVM. In Section 2 we give an overview of LS-SVM regression. In Section 3 we suggest a prediction method on the regression model with randomly censored data by LS-SVM with the weighting scheme as Zhou (1992) used. Numerical studies with simulated data sets were performed in Section 4. Finally, Section 5 has concluding remarks.

2. Least Squares Support Vector Machines

Let the training data set be denoted by $\{\mathbf{x}_i, y_i\}_{i=1}^n$, with each input $\mathbf{x}_i \in \mathbb{R}^d$ and the output $y_i \in \mathbb{R}$. For this kind of data set, we can consider the two types of regression, linear and nonlinear regression based on least squares support vector machines. In this section we give an overview of LS-SVM regression for linear and nonlinear cases, respectively.

2.1. Linear regression

For the case of well known linear regression, we can assume the functional form of unknown regression function f for given input vector $\mathbf{x}$ by

$$f(\mathbf{x}) = \mathbf{w}'\mathbf{x} + b \tag{2.1}$$

where b is a bias term and $\mathbf{w}$ is an appropriate weight vector. The least squares support vector approach to minimizing the guaranteed risk bound for linear model leads to the optimization problem defined with a regularization parameter γ as

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}'\mathbf{w} + \frac{\gamma}{2}\sum_{i=1}^{n} e_i^2 \tag{2.2}$$

over $(\mathbf{w}, b, \mathbf{e})$ subject to equality constraints

$$y_i - \mathbf{w}'\mathbf{x}_i - b = e_i, \quad i = 1, \ldots, n \tag{2.3}$$

where $\mathbf{e} = (e_1, \ldots, e_n)$. The Lagrangian function can be constructed as

$$L(\mathbf{w}, b, \mathbf{e} : \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}'\mathbf{w} + \frac{\gamma}{2}\sum_{i=1}^{n} e_i^2 - \sum_{i=1}^{n} \alpha_i(\mathbf{w}'\mathbf{x}_i + b + e_i - y_i) \tag{2.4}$$

where α_i's are the Lagrange multipliers. The conditions for optimality are given by

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{w}} &= \mathbf{0} \rightarrow \mathbf{w} = \sum_{i=1}^{n} \alpha_i \mathbf{x}_i, \\
\frac{\partial L}{\partial b} &= 0 \rightarrow \sum_{i=1}^{n} \alpha_i = 0, \\
\frac{\partial L}{\partial e_i} &= 0 \rightarrow \alpha_i = \gamma e_i, \quad i = 1, \ldots, n, \\
\frac{\partial L}{\partial \alpha_i} &= 0 \rightarrow \mathbf{w}'\mathbf{x}_i + b + e_i - y_i = 0, \quad i = 1, \ldots, n
\end{aligned}$$

with solution

$$\begin{bmatrix} 0 & \mathbf{1}' \\ \mathbf{1} & \boldsymbol{\Omega}_L + \gamma^{-1}\mathbf{I} \end{bmatrix} \begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{y} \end{bmatrix} \tag{2.5}$$

with $\mathbf{y} = (y_1, \ldots, y_n)'$, $\mathbf{1} = (1, \ldots, 1)'$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)'$ and $\boldsymbol{\Omega}_L = \{\mathbf{x}_k'\mathbf{x}_l\}$, $k, l = 1, \ldots, n$. Solving the linear equation (2.5) the estimators of the optimal bias and Lagrange multipliers, $\widehat{b}$ and $\widehat{\alpha}_i$'s can be obtained. And then the optimal regression function for the given $\mathbf{x}$ is obtained as

$$\widehat{f}(\mathbf{x}) = \sum_{i=1}^{n} \widehat{\alpha}_i \mathbf{x}_i'\mathbf{x} + \widehat{b}. \tag{2.6}$$

2.2. Nonlinear regression

So far we have explained the case of linear regression, which is not always appropriate for all tasks. To allow for the case of nonlinear regression, the input vectors are nonlinearly transformed into a potentially higher dimensional feature space by a nonlinear mapping function ϕ and then a linear regression is performed there. Nonlinear regression function can be written as

$$f(\mathbf{x}) = \mathbf{w}'\phi(\mathbf{x}) + b \tag{2.7}$$

where b is a bias term, $\mathbf{w}$ is an appropriate weight vector and $\phi(\cdot)$ is a nonlinear mapping function. The least squares support vector approach for nonlinear model leads to the optimization problem defined with a regularization parameter γ as

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}'\mathbf{w} + \frac{\gamma}{2}\sum_{i=1}^{n} e_i^2 \tag{2.8}$$

over $(\mathbf{w}, b, \mathbf{e})$ subject to equality constraints

$$y_i - \mathbf{w}'\phi(\mathbf{x}_i) - b = e_i, \quad i = 1, \ldots, n. \tag{2.9}$$

The Lagrangian function can be constructed as

$$L(\mathbf{w}, b, \mathbf{e} : \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}'\mathbf{w} + \frac{\gamma}{2}\sum_{i=1}^{n} e_i^2 - \sum_{i=1}^{n} \alpha_i\{\mathbf{w}'\phi(\mathbf{x}_i) + b + e_i - y_i\} \tag{2.10}$$

where α_i's are the Lagrange multipliers. The conditions for optimality are given by

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{w}} = \mathbf{0} &\rightarrow \mathbf{w} = \sum_{i=1}^{n} \alpha_i \phi(\mathbf{x}_i), \\
\frac{\partial L}{\partial b} = 0 &\rightarrow \sum_{i=1}^{n} \alpha_i = 0, \\
\frac{\partial L}{\partial e_i} = 0 &\rightarrow \alpha_i = \gamma e_i, \quad i = 1, \ldots, n, \\
\frac{\partial L}{\partial \alpha_i} = 0 &\rightarrow \mathbf{w}'\phi(\mathbf{x}_i) + b + e_i - y_i = 0, \quad i = 1, \ldots, n.
\end{aligned}$$

Thus for the case of nonlinear regression, the estimators of the optimal bias and Lagrange multipliers can be obtained by solving the linear equation

$$\begin{bmatrix} 0 & \mathbf{1}' \\ \mathbf{1} & \boldsymbol{\Omega}_N + \gamma^{-1}\mathbf{I} \end{bmatrix} \begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{y} \end{bmatrix} \tag{2.11}$$

with $\mathbf{y} = (y_1, \ldots, y_n)'$, $\mathbf{1} = (1, \ldots, 1)'$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)'$, and $\boldsymbol{\Omega}_N = \{K_{kl}\}, k, l = 1, 2, \ldots, n$, where

$$K_{kl} = \phi(\mathbf{x}_k)'\phi(\mathbf{x}_l).$$

For this nonlinear regression, solution of (2.11) requires the computations of dot products $\phi(\mathbf{x}_k)'\phi(\mathbf{x}_l)$, $k, l = 1, \ldots, n$, in a potentially higher dimensional feature space. Under certain conditions (Mercer, 1909), these demanding computations can be reduced significantly by introducing a kernel function K such that

$$\phi(\mathbf{x}_k)'\phi(\mathbf{x}_l) = K(\mathbf{x}_k, \mathbf{x}_l).$$

Several choices of kernel functions are possible. RBF (Radial Basis Function) is the most frequently used kernel function. The optimal nonlinear regression function for the given $\mathbf{x}$ is obtained as

$$\widehat{f}(\mathbf{x}) = \sum_{i=1}^{n} \widehat{\alpha}_i K(\mathbf{x}_i, \mathbf{x}) + \widehat{b}. \tag{2.12}$$

The linear regression model (2.1) can be regarded as a special case of nonlinear regression model (2.7). By using an identity feature mapping function ϕ in nonlinear regression model, that is, $K(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1'\mathbf{x}_2$, it reduces to linear regression model.

3. REGRESSION WITH CENSORED DATA BY LS-SVM

In this section we suggest a prediction method on the regression model with randomly censored data by LS-SVM. For the suggestion, we consider the censored linear regression model first and then extend the result of censored linear regression model to censored nonlinear regression model.

Consider the censored linear regression model (AFT model) for the response variables T_i's,

$$T_i = \boldsymbol{\beta}'\mathbf{x}_i + b + \epsilon_i, \quad i = 1, \ldots, n,$$

where $(\boldsymbol{\beta}', b)'$ is the regression parameter vector of the model and ϵ_i's are unobservable errors assumed to be independent with zero means and bounded variances. Let C_i's be the censoring variables assumed to be independent and identically distributed having a cumulative distribution function $G(y) = P(C_i \leq y)$. The parameter vector of interest is $(\boldsymbol{\beta}', b)'$, and T_i is not observed but

$$\delta_i = I_{(T_i < C_i)} \quad \text{and} \quad Y_i = \min(T_i, C_i),$$

where $I_{(\cdot)}$ denotes the indicator function. In most practical cases $G(\cdot)$ is not known and needs to be estimated by the Kaplan-Meier estimator or its variation, $\widehat{G}(\cdot)$. The problem considered here is that of the estimation of $(\boldsymbol{\beta}', b)'$ based on $(\delta_1, Y_1, \mathbf{x}_1), \ldots, (\delta_n, Y_n, \mathbf{x}_n)$. Koul *et al.* (1981) defined a new observable response Y_i^* with weights ζ_i as

$$Y_i^* = \zeta_i Y_i \quad \text{where} \quad \zeta_i = \frac{\delta_i}{1 - G(Y_i)} \tag{3.1}$$

and showed Y_i^* has the same mean as T_i and thus follows the same linear model as T_i does. And the estimator of $(\boldsymbol{\beta}', b)'$ is obtained from

$$(\widehat{\boldsymbol{\beta}}', \widehat{b})' = \underset{(\boldsymbol{\beta}', b)'}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i^* - \boldsymbol{\beta}'\mathbf{x}_i - b)^2.$$

Zhou (1992) proposed an M-estimator of the regression parameter with a general

loss function $\rho(\cdot)$ using the weights ζ_i,

$$(\widehat{\beta}', \widehat{b})' = \underset{(\beta', b)'}{\operatorname{argmin}} \sum_{i=1}^{n} \zeta_i \rho(Y_i - \beta' \mathbf{x}_i - b).$$

We apply the weighting scheme (3.1) to (2.2) with squared error loss function. Then we can construct the optimal problem

$$\min_{\beta} \frac{1}{2}\beta'\beta + \frac{\gamma}{2}\sum_{i=1}^{n} e_i^2. \tag{3.2}$$

Since the second term in the equation (3.2) controls the empirical risk as Zhou (1992) proposed with squared error loss function, we modify the equality constraints in (2.3) to

$$\sqrt{\zeta_i}\left(Y_i - \beta' \mathbf{x}_i - b\right) = e_i, \quad i = 1, \ldots, n.$$

Thus the estimators of the optimal bias and Lagrange multipliers, $\widehat{b}$ and $\widehat{\alpha}_i$'s, can be obtained from the linear equation

$$\begin{bmatrix} 0 & \sqrt{\widehat{\zeta}}' \\ \sqrt{\widehat{\zeta}} & \mathbf{\Omega}_L^* + \gamma^{-1}\mathrm{I} \end{bmatrix} \begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{y}^* \end{bmatrix} \tag{3.3}$$

where

$$\mathbf{y}^* = \left(\sqrt{\widehat{\zeta}_1} y_1, \ldots, \sqrt{\widehat{\zeta}_n} y_n\right)', \ \sqrt{\widehat{\zeta}} = \left(\sqrt{\widehat{\zeta}_1}, \ldots, \sqrt{\widehat{\zeta}_n}\right)', \ \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)'$$

and

$$\mathbf{\Omega}_L^* = \left\{ \sqrt{\widehat{\zeta}_k}\sqrt{\widehat{\zeta}_l}\mathbf{x}_k'\mathbf{x}_l \right\}, \quad \widehat{\zeta}_i = \frac{\delta_i}{1 - \widehat{G}(y_i)}, \quad k, l = 1, \ldots, n,$$

with $\widehat{G}(\cdot)$ as an empirical distribution function of Y. Solving the above linear equation the optimal bias and Lagrange multipliers, $\widehat{b}$ and $\widehat{\alpha}_i$'s can be obtained. And then the optimal linear regression function for given $\mathbf{x}$ is predicted as

$$\widehat{f}(\mathbf{x}) = \widehat{\beta}'\mathbf{x} + \widehat{b} = \sum_{i=1}^{n} \sqrt{\widehat{\zeta}_i}\widehat{\alpha}_i \mathbf{x}_i'\mathbf{x} + \widehat{b}. \tag{3.4}$$

For the censored nonlinear regression using kernel function and feature mapping function mentioned in Section 2.2 we can construct the optimal problem

$$\min_{\beta} \frac{1}{2}\beta'\beta + \frac{\gamma}{2}\sum_{i=1}^{n} e_i^2 \tag{3.5}$$

with the equality constraints as

$$\sqrt{\zeta_i}\{Y_i - \beta'\phi(\mathbf{x}_i) - b\} = e_i, \quad i = 1, \ldots, n.$$

Thus the estimators of the optimal bias and Lagrange multipliers, $\widehat{b}$ and $\widehat{\alpha}_i$'s, are obtained by solving the following linear equation

$$\begin{bmatrix} 0 & \sqrt{\widehat{\zeta}}' \\ \sqrt{\widehat{\zeta}} & \mathbf{\Omega}_N^* + \gamma^{-1}\mathbf{I} \end{bmatrix} \begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{y}^* \end{bmatrix} \tag{3.6}$$

where

$$\mathbf{y}^* = \left(\sqrt{\widehat{\zeta}_1}y_1, \ldots, \sqrt{\widehat{\zeta}_n}y_n\right)', \ \sqrt{\widehat{\zeta}} = \left(\sqrt{\widehat{\zeta}_1}, \ldots, \sqrt{\widehat{\zeta}_n}\right)', \ \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)'$$

and

$$\mathbf{\Omega}_N^* = \left\{\sqrt{\widehat{\zeta}_k}\sqrt{\widehat{\zeta}_l}K(\mathbf{x}_k, \mathbf{x}_l)\right\}, \quad \widehat{\zeta}_i = \frac{\delta_i}{1 - \widehat{G}(y_i)}, \quad k, l = 1, \ldots, n,$$

with $\widehat{G}(\cdot)$ as an empirical distribution function of Y. Then the optimal nonlinear regression function for given $\mathbf{x}$ is predicted as

$$\widehat{f}(\mathbf{x}) = \sum_{i=1}^{n} \sqrt{\widehat{\zeta}_i}\widehat{\alpha}_i K(\mathbf{x}_i, \mathbf{x}) + \widehat{b}. \tag{3.7}$$

4. NUMERICAL STUDIES

We illustrate the performance of the proposed prediction method for the regression function by LS-SVM using the weights defined in (3.1). The training data sets were generated on the linear and the nonlinear regression model, respectively, which include the censored observations.

For the censored linear regression model the response variables T_i's can be expressed as

$$T_i = \beta x_i + b + \epsilon_i, \quad i = 1, \ldots, n. \tag{4.1}$$

For the generation of data, we choose $(\beta, b) = (1, 1)$ without loss of generality. Then the true value of the regression function of response variable given the covariate x is $f(x) = x + 1$. For training data set, 200 of x's are generated from a uniform distribution, $U(0, 1)$, and 200 of (t, c)'s are generated from logistic distributions, $L(x + 1, 10)$ and $L(x + 1 + cc, 10)$, respectively, where cc is chosen

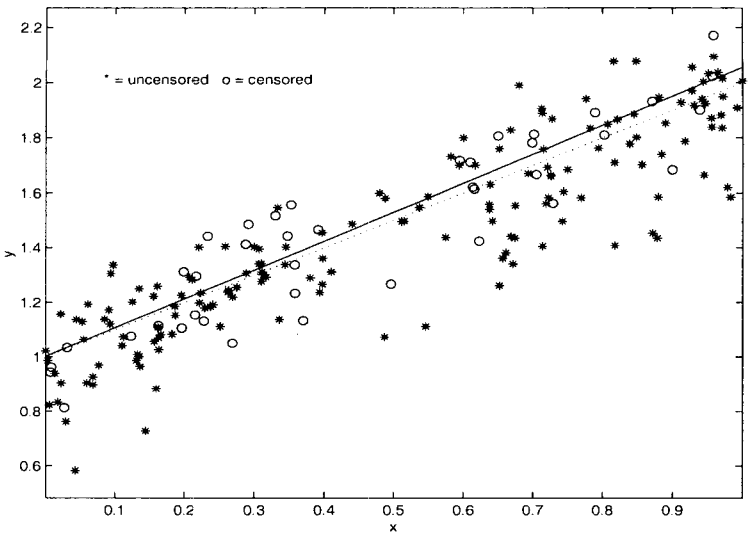

FIGURE 1 *True regression functions given x's and their predicted values for the linear regression model.*

for 20% censoring proportion. For the test data set, 200 of (x, t, c)'s are generated by the same way as for the training data set. For the optimization problem in (3.1), the value of regularization parameter γ is chosen as 500 by the cross-validation method with uncensored observations in the training data set. Solving the linear equation (3.3) with the value of regularization parameter γ chosen as 500, the estimators of the optimal Lagrange multipliers and bias, $\widehat{\alpha}_i$'s and $\widehat{b}$, can be obtained. Then by the equation (3.4) the regression parameter estimators are obtained as $(\widehat{\beta}, \widehat{b}) = (1.0548, 1.0020)$. Therefore we can write the predicted regression function given x as

$$\widehat{f}(x) = 1.0548x + 1.0020.$$

In Figure 1 we represent the true regression functions (dotted line) and their predicted values (solid line), respectively. From this figure we can see the predicted values look close to the true linear regression functions for x's from the test data set even though under 20% censorship.

The result shows that the proposed method works well. For the comparison of our proposed method with existing method, Zhou (1992)'s method were considered. Zhou (1992) proposed an M-estimator of the regression parameter with a general loss function. But for the comparison we used a squared error loss function on the same data set. As a result we have a predicted regression

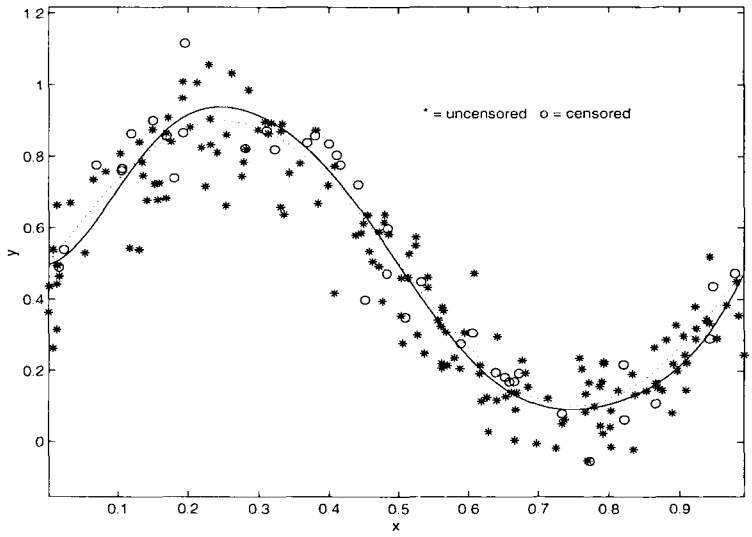

FIGURE 2 *True regression functions given x's and their predicted values for the nonlinear regression model.*

function given x by

$$\widehat{f}(x) = 1.0549x + 1.0019.$$

We cannot find a significant difference between the proposed method and Zhou (1992)'s method in this linear regression model on the given data set.

Now consider the censored nonlinear regression model for the response variables T_i's of the form,

$$T_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n. \tag{4.2}$$

For the training data set, 200 of x's are generated from a uniform distribution $U(0.1)$ and 200 of (t, c)'s are generated from the following normal distributions,

$$N(0.5 + 0.4\sin(2\pi x),\ 0.01) \quad \text{and} \quad N(0.5 + 0.4\sin(2\pi x) + cc,\ 0.01),$$

respectively. cc is chosen for 20% censoring proportion. 200 of (x, t, c)'s are also generated for the test data set. The radial basis function (RBF) kernel is used for the numerical studies, which is defined as

$$K(x_1, x_2) = \exp\left\{-\frac{1}{2\sigma^2}(x_1 - x_2)^2\right\}.$$

The values of γ and σ in RBF kernel are chosen as 500 and 0.2, respectively, by the cross-validation method with uncensored observations in the training data

set. Solving the linear equation (3.6) with the training data set, the estimators of the optimal Lagrange multipliers and bias, $\widehat{\alpha}_i$'s and $\widehat{b}$, can be obtained. Then by the equation (3.7) the predicted regression function given x in the test data set is obtained. Figure 2 shows true regression functions (dotted line) and their predicted values (solid line), respectively. The predicted values look close to the true regression functions in this nonlinear model as in linear model for x's from the test data set even though under 20% censorship. For the case of nonlinear regression model, it is hard to find a regression method on the censored data set to compare with the proposed method.

5. Concluding Remarks

Through the numerical studies, we showed that the proposed prediction method by LS-SVM provides a satisfying solutions to the censored linear regression model and the censored nonlinear regression model respectively. Particularly for the censored nonlinear regression model, the proposed method can be used without heavy computations and shows a satisfying result. In future work, we intend to devise algorithms for predicting intervals of regression parameter based on the training data set which might be randomly censored, by using LS-SVM or the other efficient machine learning methods.

Acknowledgements

The authors wish to thank to the referees for many helpful comments and careful readings of this paper, which have improved the presentation of this paper considerably.

References

Koul, H., Susarla, V. and van Ryzin, J. (1981). "Regression analysis with randomly right censored data", *The Annals of Statistics*, **9**, 1276–1288.

Mercer, J. (1909). "Functions of positive and negative type, and their connection with the theory of integral equations", *Philosophical Transactions of the Royal Society of London*, **A209**, 415–446.

Suykens, J. A. K. and Vandewalle, J. (1999). "Least square support vector machine classifier", *Neural Processing Letters*, **9**, 293–300.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*, Springer, New York.

Vapnik, V. N. (1998). *Statistical Learning Theory*, Wiley, New York.

Zhou, M. (1992). "M-estimation in censored linear models", *Biometrika*, **79**, 837–841.