OUKA
Osaka University Knowledge Archive

Title	Penalized Likelihood Approaches to Nonparametric Regression Problems
Author(s)	坂本, 亘
Citation	大阪大学, 1998, 博士論文
Version Type	VoR
URL	https://doi.org/10.11501/3144026
rights	
Note	

Osaka University Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

Osaka University

Penalized Likelihood Approaches to Nonparametric Regression Problems

Wataru Sakamoto

Department of Mathematical Science,
Graduate School of Engineering Science,
Osaka University

1998

Abstract

Maximum penalized likelihood estimation is applied in non(semi)-parametric regression problems, and enables us exploratory identification and diagnostics of nonlinear regression relationships. The smoothing parameter λ controls the trade-off between the smoothness and the goodness-of-fit of a function. The method of cross-validation is used for selecting λ, but the generalized cross-validation based on the squared error criterion shows bad behavior in non-normal case and often can not select reasonable λ. The main purpose of this thesis is to propose a method which gives more suitable λ and to evaluate the performance of it.

This thesis consists of three parts below. Firstly the method of maximum penalized likelihood estimation in non(semi)-parametric regression problems is summarized. It is described as an extension of penalized least squares and generalized linear models, and is applied to logistic regression, Poisson regression and density smoothing for classified data. Secondly the likelihood-based cross-validation score is described as a tool to select λ adaptively, and a method of simple calculation for the delete-one estimate is proposed. A score of similar form to the Akaike information criterion (AIC) is also derived. The scores by the simple calculation are compared with the one by the exact calculation and the performance of the approximation is evaluated. Thirdly the proposed scores are compared with the ones of standard procedures. A variety of data in literature are examined. Simulations are performed to compare the patterns of selecting λ and overall goodness-of-fit and to evaluate the effects of factors.

The simple calculation of the delete-one estimate coincides with the one-step approximation based on the Newton–Raphson method in the case of canonical link, and the cross-validation scores by the simple calculation provide good approximations to the one by the exact one if λ is not extremely small. Furthermore the cross-validation scores by the simple calculation have little risk of choosing extremely small λ and make it possible to select λ adaptively. They have the effect of reducing the bias of estimates and provide better performance in the sense of overall goodness-of-fit. These scores are useful especially in the case of small sample size and in the case of binary logistic regression.

Acknowledgments

The author got a lot of benefit from many people for completing this dissertation. His counselor, Professor Shingo Shirahata provided various guidance and suggestion through the whole of this dissertation. Professors Masashi Goto and Nobuo Inagaki of Osaka University reviewed the first draft of the dissertation carefully and provided useful comments and encouragement. The author expresses his sincere appreciation for their efforts. Dr. In-sun Chu provided a suggestion leading to the study of nonparametric regression. Professor Megu Ohtaki of Hiroshima University provided helpful advice on the problem of selecting the smoothing parameter. The author is grateful for their assistance.

The author would like to thank all of the teachers, his seniors, contemporaries and juniors and the clerks of the Department of Mathematical Science for their support during his being at the graduate school. He also wishes to express his gratitude to the members of the Biostatistics Research Association (BRA), colleagues of Professor Masashi Goto, for their kindnesses.

The author is grateful for the support by a scholarship granted from the *Nihon Ikuei-kai.*

Finally the author would like to give special thank to his parents and sister for their encouragement, for their economical and mental support and for their taking care of his health.

Contents

Notations

$\mathrm{E}(\cdot)$	Expectation
$\mathrm{var}(\cdot)$	Variance
$\mathrm{P}(\cdot)$	Probability
$\mathrm{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$\mathrm{B}(m, p)$	Binomial distribution with m trials and probability of success p
$\mathrm{Po}(\mu)$	Poisson distribution with mean μ
$\mathrm{Ga}(\alpha, \beta)$	Gamma distribution with shape parameter α and scale β
Φ	Cumulative distribution function of the standard normal distribution
$\mathrm{tr}A$	Trace of a matrix A
$\mathrm{diag}(\cdots)$	Diagonal matrix
$\hat{}$	Estimates (least squares, or maximum likelihood)
$(-i)$	Estimate when the ith observation is deleted
(i)	Vector or matrix from which the ith component is removed
i	Observation number $(i = 1, \ldots, n)$
j	Variable number $(j = 1, \ldots, p)$ (sometimes used in place of i)
k	Basis number $(k = 1, \ldots, q)$
y_i	Response variable
t_i, t_{ij}	Explanatory variables (nonparametric term)
$\boldsymbol{x}_i$	Vector of explanatory variables (parametric term) $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$
α	Constant in a regression model
$\boldsymbol{\beta}$	Vector of regression coefficient to be estimated $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^T$
f, f_j	Regression functions to be estimated
θ_i	Natural parameter of the exponential family
μ_i	Expectation of y_i, $\mu_i = b'(\theta_i)$
ϕ	Scale (nuisance) parameter
η_i	Predictor or predicted value
a_i, b, c	Functions employed in the density (probability) function of some distribution belonging to the exponential family
m_i	Prior weight, $a_i(\phi) = \phi / m_i$ (especially, the number of trials in the binomial distribution)
$V(\mu)$	Variance function of the exponential family
$\mathcal{S}$	Penalized (weighted) sum of squares
l	Log-likelihood
Π	Penalized log-likelihood
$J(f)$	Roughness penalty for f
λ, λ_j	Smoothing parameters (positive number)
w_i	Weight in weighted least squares, or working weight
z_i	Working response
φ_k	Basis function for the functional space of f (e.g., B-spline basis)
ξ_k	Coefficient associated with f, $f(t_i) = \sum_{k=1}^{n} \xi_k \varphi_k(t_i)$

$\boldsymbol{\xi}$	Coefficient vector, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_q)^T$
$\boldsymbol{\eta}$	Predictor vector, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$
$\boldsymbol{z}$	Working response vector, $\boldsymbol{z} = (z_1, \ldots, z_n)^T$
X	Design matrix $(n \times p)$, $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T = \{x_{ij}\}$
W	Weight matrix $(n \times n)$, $W = \mathrm{diag}(w_1, \ldots, w_n)$
B	Basis matrix $(n \times q)$, $B = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n]^T = \{\varphi_k(t_i)\}$
K, K_j	Penalty matrix $(q \times q$, non-negative definite), $J(f) = \boldsymbol{\xi}^T K \boldsymbol{\xi}$
S_λ, $S_{\lambda_j}^{(j)}$	Smoother matrix $(n \times n)$, $S_\lambda = B(B^T W B + \lambda K)^{-1} B^T W$
A (A_λ), A_{ij}	Hat matrix and its component (A_{ii} : Leverage value)
ν	Equivalent degrees of freedom for a model, $\nu = \mathrm{tr}A$
σ^2	Error variance
r_i	Residual
D, D^*, d_i	Deviance, scaled deviance and deviance increment
χ^2	Pearson chi-square statistic
p_i	Probability of success in binomial distribution
q_i	Person-time, or the population size
x_r	Raw data in the context of density estimation $(r = 1, \ldots, N)$
C_i	Class
g	Density function
h	Bin width

Abbreviations

MLE	Maximum likelihood estimate
MPLE	Maximum penalized likelihood estimate
GLM	Generalized linear model
GAM	Generalized additive model
EDF	Equivalent degrees of freedom
OCV	Ordinary cross-validation score (based on squared-error criterion)
GCV	Generalized cross-validation
UBR	Unbiased risk estimate
AIC	Akaike information criterion
LCV	Likelihood-based cross-validation score (by exact calculation)
LCV_1	Likelihood-based cross-validation score by simple calculation
LCV_2	AIC-like form of the likelihood-based cross-validation score
ACV	Approximate cross-validation score
GACV	Generalized approximate cross-validation
AMSE	Averaged mean squared error
ASB	Averaged squared bias
AV	Averaged variance
KL	Kullback–Leibler distance

Chapter 1

Introduction

1.1 Penalized Likelihood Approach

Nonparametric regression, which has been recently developed and is also known as smoothing, is a statistical technique to estimate regression functions without specifying the functional form of them in a parametric way. In order to estimate them under non-normal distributions such as logistic regression, Poisson regression and so on, the approach of maximum penalized likelihood is often useful. Maximizing ordinary likelihood yields a regression function that shows rapid variation, and hence a penalty term for the roughness of the function is added to the likelihood. The approach enables us exploratory identification and diagnostics of nonlinear regression relationships.

A simple formulation of the penalized likelihood approach is as follows. Denote the log-likelihood function by $l(\boldsymbol{\theta})$, and let the natural parameter $\boldsymbol{\theta}$ be linked to an unknown function f of some explanatory variable(s). Then we maximize the penalized log-likelihood

$$l(\boldsymbol{\theta}) - \frac{\lambda}{2} J(f)$$

to obtain an estimate of f. The estimate of f is called the *maximum penalized likelihood estimate* (MPLE). Here $J(f)$ is a roughness penalty for f. The positive number λ is called the *smoothing parameter*, which controls the trade-off between the smoothness and the goodness-of-fit of f. As λ tends to infinity $\hat{f}$ becomes smoother and tends to a parametric function (linear, quadratic, and so on according to the form of the penalty $J(f)$), while as λ tends to zero $\hat{f}$ becomes rougher and tends to an interpolating function.

The maximum penalized likelihood estimation is a natural extension of the penalized least squares. Hence the algorithm of the maximum penalized likelihood estimation is easily constructed as the iteration of the penalized least squares algorithm (Green and Silverman, 1994). Moreover, just as in ordinary maximum likelihood estimation, the Fisher scoring algorithm can be also applied to the penalized likelihood approach.

The maximum penalized likelihood estimation is originally proposed by Good and Gaskins (1971) in the context of density estimation. Silverman (1982)

improved the method of density estimation by the maximum penalized likelihood approach, and O'Sullivan (1988) developed an algorithm of Silverman's (1982) method using B-spline approximation. In the context of nonparametric regression, Anderson and Blair (1982) applied the maximum penalized likelihood estimation to logistic regression models. O'Sullivan, Yandell and Raynor (1986) considered estimating nonparametric regression functions in generalized linear models (GLMs) (Nelder and Wedderburn, 1972). Green and Yandell (1985) proposed semiparametric generalized linear models, in which predictors were the sum of nonparametric function and parametric functions. Hastie and Tibshirani (1986) proposed generalized additive models, the combination of additive models and GLMs. The idea of maximizing penalized likelihood is also contained in these models, and it has been extended for various purposes.

1.2 Selecting the Smoothing Parameter

As described in the previous section, the smoothing parameter λ controls the smoothness of the estimated function $\hat{f}$. If we have some criterion for selecting the optimal value of λ adaptively from data, diagnostics on nonlinear regression structure will be possible.

Some criterion for selecting the smoothing parameter have been proposed, mainly from the viewpoint of prediction. By analogy with the selection of λ in penalized least squares problem, the method of cross-validation has been most commonly used. O'Sullivan, Yandell and Raynor (1986) suggested that each iteration of maximum penalized likelihood algorithm should be equivalent to penalized weighted least squares algorithm, and that the generalized cross-validation (GCV) score (Wahba, 1977; Craven and Wahba, 1979) could be evaluated on the final iteration.

Green and Yandell (1985) used the GCV method to determine the smoothing parameter in applying semiparametric logistic regression to bioassay data. They pointed out that the GCV score might be badly behaved, that is, the score might have no global minimum and tend to zero as $\lambda \to 0$. It has been shown by many authors that the GCV score in penalized least squares has many optimality (Craven and Wahba, 1979), but in maximum penalized likelihood estimation, as Green and Silverman (1994) states, it seems natural to construct the cross-validation score in terms of likelihood.

The likelihood-based cross-validation (LCV) score requires the delete-one estimates of the natural parameters, but calculating the exact values of them is expensive. Sakamoto and Shirahata (1997b) proposed a method for simple calculation of the delete-one estimates and the approximate LCV score. The method coincides with the one-step approximation to the delete-one estimates in Newton–Raphson algorithm.

1.3 Purpose and Composition

The purpose of this thesis is described as follows. The first is to summarize the subject on maximum penalized likelihood estimation in non(semi)-parametric

regression problems. Penalized likelihood is formulated in the framework of GLMs, and the algorithm is constructed using the methods of Fisher scoring. Semiparametric GLMs and generalized additive models are also stated. It is illustrated that the approach of penalized likelihood can be applied in various situations — binary or binomial logistic regression models, Poisson regression models, density smoothing for classified data and so on.

The second purpose is to describe the likelihood-based cross-validation score as a tool to select the smoothing parameter adaptively, and to propose a method of simple calculation for the delete-one estimate. Another score of similar form to the Akaike information criterion (AIC) is also derived. Some theoretical grounds on these scores are discussed. The scores by the simple calculation are compared with the one by the exact calculation and the performance of approximation is evaluated.

The third purpose is to compare the scores proposed with the ones of standard procedures (GCV, AIC, etc.) and to show usefulness of them. A variety of data in literature are examined, and some simulations are performed to compare the patterns of selecting the smoothing parameter and overall goodness-of-fit and to evaluate the effects of factors.

This thesis is composed as follows.

Chapter 2 describes maximum penalized likelihood estimation. For preparation, penalized least squares and generalized linear models are briefly summarized in Section 2.1 and Section 2.2, respectively. Maximum penalized likelihood estimation is then introduced in Section 2.3 in the framework of GLM, divided into the case of nonparametric, semiparametric and additive models. Inference and diagnostics in non(semi)-parametric GLMs are also discussed there. Fitting non(semi)-parametric GLMs is illustrated in some case studies in Section 2.4.

Chapter 3 discusses the methods for adaptive selection of the smoothing parameter. At first standard procedures are described in Section 3.1. The likelihood-based cross-validation score by simple calculation and the AIC-like score are proposed in Section 3.2. The comparison of the simple calculation with the exact one is performed in Section 3.3. Some of other selection procedures proposed are taken up and compared with our scores in Section 3.4.

Chapter 4 reports the comparison of our likelihood-based score with the standard selection procedure. Logistic regression models, Poisson regression models and density smoothing for classified data are stated. Data sets in literature are examined in Section 4.1, and simulation is performed in Section 4.2.

In the last chapter our results are summarized and our future prospects are discussed. In Appendices, the splines often used as the basis of smoothed functions are briefly described, and data used in this thesis are also listed.

Chapter 2

Maximum Penalized Likelihood Estimation

2.1 Penalized Least Squares

In this section we briefly discuss penalized least squares, a special case of penalized likelihood approaches.

2.1.1 Nonparametric Regression Models

Suppose that each of the responses y_i, $i = 1, \ldots, n$, is observed at an explanatory variable t_i. We fit a *nonparametric regression model*

$$y_i = f(t_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

where $f(t)$ is a smooth function, to be estimated, and ϵ_i are independent errors with mean zero.

Applying the ordinary least squares to the nonparametric regression model yields the estimate of f that interpolates data. So we consider minimizing the penalized (weighted) sum of squares

$$\mathcal{S}(f) = \sum_{i=1}^{n} w_i\{y_i - f(t_i)\}^2 + \lambda J(f), \tag{2.1}$$

where w_i are appropriate weights (we discuss the weighted version because of the generalization in the later sections), λ is the smoothing parameter (a positive number, specified in this chapter), and $J(f)$ is a roughness penalty. The method is called *penalized least squares.*

The roughness penalty functional $J(f)$ is designed so that a rougher function f gives a larger value of $J(f)$. For example, if we use $J(f) = \int\{f''(t)\}^2 dt$, the integral of the squared second derivative of f, the estimated function $\hat{f}$ that minimizes $\mathcal{S}(f)$ in the class of twice differentiable functions is necessarily a natural cubic spline with knots at $t_1, \ldots, t_n$, called the smoothing spline. Hence the class of smoothing splines is often used when minimizing $\mathcal{S}(f)$. In addition, $\hat{f}$ tends to the ordinary least squares estimate as λ tends to infinity,

and $\hat{f}$ tends to the function that interpolates the points (t_i, y_i) as λ tends to zero. See Eubank (1988), Green and Silverman (1994) and Appendix.

Assume that the space of f is spanned by q smooth basis functions φ_k, $k = 1, \ldots, q$. For example, B-splines, which are orthogonal basis for splines and are constructed from $t_1, \ldots, t_n$, are often used as $\{\varphi_k\}$. In some cases such as natural splines, q is equal to n. Then f is written as $f(t) = \sum_{k=1}^{q} \xi_k \varphi_k(t)$. Moreover, using the coefficients $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_q)^T$, assume that $J(f) = \boldsymbol{\xi}^T K \boldsymbol{\xi}$, where K is a $q \times q$ non-negative definite symmetric matrix. If the penalty $J(f) = \int \{f''(t)\}^2 dt$ is used, $J(f)$ can be written in such a quadratic form using $t_1, \ldots, t_n$ (see Appendix). Let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, $\boldsymbol{f} = (f_1, \ldots, f_n)^T$, $W = \mathrm{diag}(w_1, \ldots, w_n)$ and denote the $n \times q$ basis matrix $\{\varphi_k(t_i)\}_{i=1,\ldots,n;k=1,\ldots,q}$ by B. The penalized sum of squares (2.1) is then rewritten as

$$\begin{aligned} \mathcal{S}(\boldsymbol{f}) &= (\boldsymbol{y} - \boldsymbol{f})^T W (\boldsymbol{y} - \boldsymbol{f}) + \lambda \boldsymbol{\xi}^T K \boldsymbol{\xi} \\ &= (\boldsymbol{y} - B\boldsymbol{\xi})^T W (\boldsymbol{y} - B\boldsymbol{\xi}) + \lambda \boldsymbol{\xi}^T K \boldsymbol{\xi}. \end{aligned} \tag{2.2}$$

The coefficient vector $\boldsymbol{\xi}$ that minimizes (2.2), denoted by $\hat{\boldsymbol{\xi}}$, is found by differencing (2.2) on $\boldsymbol{\xi}$ and setting $\partial \mathcal{S}/\partial \boldsymbol{\xi} = \mathbf{0}$, and we obtain $\hat{\boldsymbol{\xi}} = (B^T W B + \lambda K)^{-1} B^T W \boldsymbol{y}$. Therefore the values of estimated function f becomes

$$\hat{\boldsymbol{f}} = (\hat{f}(t_1), \ldots, \hat{f}(t_n)) = B(B^T W B + \lambda K)^{-1} B^T W \boldsymbol{y}. \tag{2.3}$$

The $n \times n$ matrix $S_\lambda = B(B^T W B + \lambda K)^{-1} B^T W$ is called the smoother matrix.

2.1.2 Semiparametric Regression Models

Next, we consider the case that responses are observed with several explanatory variables. However, it is difficult to estimate a function of several variables. Semiparametric regression models, and additive models discussed in the next section, are simple approximations to such a surface. Especially, semiparametric regression models that contain one nonparametric function of a single variable are the simplest extension, and provides an algorithm without any iteration.

Suppose that each of the responses y_i, $i = 1, \ldots, n$, is observed with a vector of explanatory variables $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ and another variable t_i (assumed to be one-dimensional for simplicity). A *semiparametric regression model* is formulated as

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + f(t_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\boldsymbol{\beta}$ is a p-vector of regression coefficients and $f(t)$ is a smooth function, both of which are to be estimated. The vector $\boldsymbol{\beta}$ and the function $f(t)$ are estimated by minimizing the penalized weighted sum of squares

$$\mathcal{S}(\boldsymbol{\beta}, f) = \sum_{i=1}^{n} w_i \{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta} - f(t_i)\}^2 + \lambda J(f). \tag{2.4}$$

Assume that $f(t) = \sum_{k=1}^{q} \xi_k \varphi_k(t)$ (i.e., $\boldsymbol{f} = B\boldsymbol{\xi}$) and $J(f) = \boldsymbol{\xi}^T K \boldsymbol{\xi}$ as in the previous subsection, and let the design matrix $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T$. Then (2.4) is rewritten as

$$\mathcal{S}(\boldsymbol{\beta}, \boldsymbol{f}) = (\boldsymbol{y} - X\boldsymbol{\beta} - B\boldsymbol{\xi})^T W (\boldsymbol{y} - X\boldsymbol{\beta} - B\boldsymbol{\xi}) + \lambda \boldsymbol{\xi}^T K \boldsymbol{\xi}.$$

Moreover, assume that each column of X is independent of the null space of K. Setting $\partial \mathcal{S}/\partial \boldsymbol{\beta} = \mathbf{0}$ and $\partial \mathcal{S}/\partial \boldsymbol{f} = \mathbf{0}$, we have a system of equations

$$X^T W X \boldsymbol{\beta} = X^T W (\boldsymbol{y} - B\boldsymbol{\xi}) \tag{2.5a}$$

and

$$(B^T W B + \lambda K)\boldsymbol{\xi} = B^T W (\boldsymbol{y} - X\boldsymbol{\beta}). \tag{2.5b}$$

The equation (2.5a) is the ordinary least squares equation to the response vector $\boldsymbol{y} - B\boldsymbol{\xi}$, and (2.5b) is the penalized least squares algorithm to the response vector $\boldsymbol{y} - X\boldsymbol{\beta}$. The vectors $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ can be also found using (2.5a) and (2.5b) alternatively. Under the condition described above, this iterative scheme always converges (Green and Silverman, 1994, Theorem 4.2).

If we denote the smoother matrix by $S_\lambda = B(B^T W B + \lambda K)^{-1} B^T W$, we have $\boldsymbol{f} = B\boldsymbol{\xi} = S_\lambda(\boldsymbol{y} - X\boldsymbol{\beta})$ from (2.5b). By substituting it into (2.5a) and rearranging it, the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{f}$ is represented as

$$\hat{\boldsymbol{\beta}} = \{X^T W (I - S_\lambda) X\}^{-1} X^T W (I - S_\lambda)\boldsymbol{y} \tag{2.6a}$$

and

$$\hat{\boldsymbol{f}} = S_\lambda(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}). \tag{2.6b}$$

Therefore, when the nonparametric term is a univariate function, once computing $S_\lambda X$ and $S_\lambda \boldsymbol{y}$ by ordinary smoothing operations, $\hat{\boldsymbol{\beta}}$ can be obtained using a common linear equation algorithm without any iteration (see Green and Silverman, 1994).

Rice (1986) pointed out that, if there exists a regression dependence of $\{\boldsymbol{x}_i\}$ on $\{t_i\}$, such as

$$x_{ij} = g_j(t_i) + \delta_{ij}, \quad i = 1, \ldots, n, \; j = 1, \ldots, p,$$

where g_j is a smooth function and δ_{ij} is an error, the estimate $\hat{\boldsymbol{\beta}}$ might contain bias. Speckman's (1988) alternative approach diminishes some of the bias of $\hat{\boldsymbol{\beta}}$. Let $\tilde{X} = S_\lambda X$ and $\tilde{\boldsymbol{y}} = S_\lambda \boldsymbol{y}$, which can be considered to contain no error components. Estimate $\boldsymbol{\beta}$ and $\boldsymbol{f}$ by

$$\hat{\boldsymbol{\beta}} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{\boldsymbol{y}}$$

and

$$\hat{\boldsymbol{f}} = S_\lambda(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}) = \tilde{\boldsymbol{y}} - \tilde{X}\hat{\boldsymbol{\beta}}.$$

They are called partial regression estimates. Chen and Shiau (1991) considered a two-stage spline smoothing method. Sakamoto and Shirahata (1997a) investigated these three types of estimates and compared the bias of $\hat{\boldsymbol{\beta}}$ by a simulation study.

2.1.3 Additive Models and Other Models

Suppose that each of the responses y_i, $i = 1, \ldots, n$ is observed with p explanatory variables $t_{i1}, \ldots, t_{ip}$. As a natural generalization of multiple linear regression model, an additive relationship on $t_{i1}, \ldots, t_{ip}$ such as

$$y_i = \alpha + \sum_{j=1}^{p} f_j(t_{ij}) + \epsilon_i, \quad i = 1, \ldots, n,$$

is often assumed, where a constant α and smooth functions f_j, $j = 1, \ldots, p$, are estimated. Buja, Hastie and Tibshirani's paper (1989) and Hastie and Tibshirani's monograph (1990) discuss the *additive models* in detail.

The additive model can be also estimated using penalized least squares by minimizing

$$\mathcal{S}(f_1, \ldots, f_p) = \sum_{i=1}^{n} w_i \left\{ y_i - \alpha - \sum_{j=1}^{p} f_j(t_{ij}) \right\}^2 + \sum_{j=1}^{p} \lambda_j J(f_j), \tag{2.7}$$

where λ_j's are smoothing parameters.

Let B_j and K_j be the basis matrix and the penalty matrix, respectively, constructed from the values of the jth explanatory variable, $t_{1j}, \ldots, t_{nj}$. The smoother matrix on the jth explanatory variable is then denoted by $S_{\lambda_j}^{(j)} = B_j(B_j^T W B_j + \lambda_j K_j)^{-1} B_j^T W$. In the same way as in the previous subsections, minimizing (2.7) yields a system of equations

$$\boldsymbol{f}_j = S_{\lambda_j}^{(j)} \left(\boldsymbol{y} - \boldsymbol{\alpha} - \sum_{k \neq j} \boldsymbol{f}_k \right), \quad j = 1, \ldots, p, \tag{2.8}$$

where $\boldsymbol{\alpha} = (\alpha, \ldots, \alpha)^T$ and $\boldsymbol{f}_j = (f_j(t_{1j}), \ldots, f_j(t_{nj}))^T$. Initially α is set to be $\bar{y}$, the sample mean of $y_1, \ldots, y_n$, and $\boldsymbol{f}_j$'s are all set to be zero. Given current values of $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_p$, a new value of $\boldsymbol{f}_j$ is obtained from equation (2.8) in the order of $j = 1, \ldots, p$, and the cycle is repeated until all $\boldsymbol{f}_j$'s converge. This procedure is known as the *backfitting algorithm* (Friedman and Stuetzle, 1981; Hastie and Tibshirani, 1986, 1990).

Other methods for non(semi)-parametric regression analysis have been proposed. Projection pursuit regression (PPR) (Friedman and Stuetzle, 1981) assumes a model of the form

$$y_i = \sum_{j=1}^{p} f_j(\boldsymbol{\alpha}_j^T \boldsymbol{x}_i) + \epsilon_i,$$

where $\boldsymbol{\alpha}_j^T \boldsymbol{x}_i$ is a one-dimensional projection of the vector of explanatory variables $\boldsymbol{x}_i$, and f_j is a smooth function. On the other hand, the approach of alternative conditional expectations (ACE) (Breiman and Friedman, 1985) supposes a model that contains a transformation of a response variable, that is,

$$\mathrm{E}[\theta(Y)] = \sum_{j=1}^{p} \mathrm{E}[\phi_j(X_j)],$$

and selects the transformations θ^* and $\phi_1^*, \ldots, \phi_p^*$ that maximize the correlation between both sides. The backfitting approach is also adopted in these regression methods. Recently, non(semi)-parametric regression models that take interaction into account are proposed: multivariate adaptive regression splines (MARS) (Friedman, 1991), interaction splines (Gu *et al.*, 1989), and so on. Gu (1989) provided the FORTRAN program RKPACK for fitting the interaction spline models. Nagai *et al.* (1994) gives a genealogy of computer-intensive tools for non(semi)-parametric regression analysis.

2.1.4 Inference and Diagnostics in Non(semi)-parametric Regression

The *equivalent degrees of freedom* (EDF) indicates the effective number of parameters. When the vector of predicted values $\hat{\boldsymbol{y}} = (\hat{y}_1, \ldots, \hat{y}_n)^T$ is denoted by $\hat{\boldsymbol{y}} = A\boldsymbol{y}$ for some matrix A, called the hat matrix, the overall EDF for the model is defined as the trace of A. For the univariate nonparametric regression model in Section 2.1.1, the EDF becomes $\nu = \mathrm{tr}S_\lambda$. Green and Silverman (1994) defined the equivalent degrees of freedom for *noise*. It takes the value more than 2, and the value becomes larger as λ tends to zero, that is, the estimated function $\hat{f}$ becomes rougher. For the semiparametric regression model in Section 2.1.2, the overall EDF becomes

$$\nu = \mathrm{tr}S_\lambda + \mathrm{tr}[\{X^T W(I - S_\lambda)X\}^{-1} X^T W(I - S_\lambda)^2 X]$$

(Green and Yandell, 1985), and hence the EDF for the nonparametric term $\hat{f}(t_i)$ becomes $\nu - p$. For the additive model in Section 2.1.3, the EDF is difficult to compute. The EDF for the term $\hat{f}_j(t_{ij})$ is often approximated by $\mathrm{tr}S_{\lambda_j}^{(j)} - 1$, where one is subtracted since each of the term contains one redundant constant, and hence the overall EDF for the additive model is approximated by $\nu = \sum_{j=1}^{p} \mathrm{tr}S_{\lambda_j}^{(j)} - p + 1$.

Each component of the hat matrix A, denoted by A_{ij}, determines how y_j affects $\hat{y}_i$. Especially, the diagonal components of A are called the leverage values. They satisfies $0 \leq A_{ii} \leq 1$. Properties of the leverage values are investigated by Eubank (1984). The leverage values are useful for a diagnostic tool of explanatory variables.

Diagnostics for residuals have also been developed. One of the diagnostic measures is the studentized residual

$$r_i^* = r_i / \sqrt{\hat{\sigma}^2(1 - A_{ii})},$$

which is the ordinary $r_i = y_i - \hat{y}_i$ divided by its standard error, where $\hat{\sigma}^2$ is an estimate of the error variance σ^2. Silverman (1985) suggested the studentized residual as an analogue of the one proposed by Cook and Weisberg (1982) in ordinary least squares regression problems. Eubank (1985) and Eubank and Gunst (1986) derived it from the Bayesian viewpoint. Moreover, Eubank (1985) and Eubank and Gunst (1986) proposed the studentized deleted residuals

$$r_{(i)}^* = r_i / \sqrt{\hat{\sigma}_{(i)}^2(1 - A_{ii})},$$

where $\hat{\sigma}^2_{(i)}$ is an estimate of σ^2 after removing the influence of the observation y_i from s^2. These measures can evaluate the goodness-of-fit to each observation. Eubank (1985) and Eubank and Gunst (1986) also derived some influential measures called DFIT and DFITS, which allow to detect the existence of influential observations.

Estimating the error variance is required in various cases, such as diagnosing influential observations, constructing confidence intervals, and so on. The first approach is the one called local differencing, suggested by Rice (1984) and Gasser, Sroka and Jennen-Steinmetz (1986). Their estimators are constructed on the basis of the variance of the first (or second) differences of the adjacent data. The second approach is based on the residual sum of squares. The errors ϵ_i are assumed to have a common variance σ^2. The natural analogue of the ordinary least squares yields the estimate of σ^2,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - \nu},$$

where the denominator $n - \nu$ is the EDF for *noise*, and the numerator is the residual sum of squares. If the true function f (or all f_j's) are in the null space of $J(f)$ (or $J(f_j)$), $\hat{\sigma}^2$ is an unbiased estimator of σ^2 (Buckley, Eagleson and Silverman, 1988).

One common way of constructing confidence intervals is to base on Bayesian approaches, proposed by Wahba (1983). For simplicity, the univariate nonparametric regression model in Section 2.1.1 is considered. Assuming some specific form of (prior) stochastic Gaussian process to $f(t)$, the posterior distribution of $\boldsymbol{f} = (f(t_1), \ldots, f(t_n))^T$ for given $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is the multivariate normal distribution with mean $\hat{\boldsymbol{f}} = S_\lambda \boldsymbol{y}$ and variance matrix $\sigma^2 S_\lambda$, where σ^2 is the variance of ϵ_i. Hence a $100 \times (1 - \alpha)$ % Bayesian posterior probability interval for each $f(t_i)$ is given by $\hat{f}(t_i) \pm \mathrm{z}(\alpha/2)\hat{\sigma}(S_\lambda)_{ii}$, where $\mathrm{z}(\alpha/2)$ is the upper $100 \times \alpha/2$ % point of the standard normal distribution, $\hat{\sigma}$ is some estimate of σ as defined in the last paragraph, and $(S_\lambda)_{ii}$ is the ith diagonal matrix of S_λ. For details about Bayesian approaches, see Eubank (1988) and Wahba (1990). Some authors discuss bootstrap approaches (e.g., Hastie and Tibshirani, 1990).

2.2 Generalized Linear Models

In this section we briefly describe generalized linear models (GLMs), proposed by Nelder and Wedderburn (1972).

2.2.1 Generalized Linear Models

Generalized linear models have two components below. For the stochastic component, suppose that responses y_i, $i = 1, \ldots, n$, are independent and have distributions belonging to the *exponential family*, which has the density (or probability) function of such a form as

$$p(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i; \phi)\right\}, \tag{2.9}$$

where a_i $(i = 1, \ldots, n)$, b and c are functions specific to each distribution. We assume that the function a_i is of the form $a_i(\phi) = \phi/m_i$ for some known weight m_i. In (2.9), θ_i, $i = 1, \ldots, n$, are called the natural parameters, and are related to mean structure since they satisfy

$$\mu_i \equiv \mathrm{E}(y_i) = b'(\theta_i) \tag{2.10}$$

under some regularity conditions. Also, ϕ is called the scale (or nuisance) parameter. The variance of y_i becomes

$$\mathrm{var}(y_i) = b''(\theta_i)a_i(\phi) = b''(\theta_i)\phi/m_i.$$

The factor $b''(\theta_i)$ depends on μ_i by (2.10), so $b''(\theta_i)$ is called the variance function, and is denoted by $V(\mu_i)$.

For the systematic component, assume that there exists a link function G such that

$$G(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta} \tag{2.11}$$

for given p-vector of explanatory variables $\boldsymbol{x}_i$, and the p-vector $\boldsymbol{\beta}$ is to be estimated. The right-hand side of (2.11) is called the linear predictor, denoted by η_i.

Estimating $\boldsymbol{\beta}$ can be based on maximum likelihood method. Denote the log-likelihood function by l, that is,

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n}\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi/m_i} + c(y_i;\phi)\right\}. \tag{2.12}$$

Notice that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ is linked to $\boldsymbol{\beta}$ by (2.10) and (2.11). The log-likelihood (2.12) is maximized over $\boldsymbol{\beta}$. Nelder and Wedderburn (1972) proposed Fisher scoring for the numerical evaluation of the maximum likelihood estimate (MLE) $\hat{\boldsymbol{\beta}}$ in GLMs. The Fisher scoring algorithm has the following form of the updating equation

$$\boldsymbol{\beta}^{new} = \boldsymbol{\beta}^{old} + \left\{\mathrm{E}\left(-\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right)\right\}^{-1}\frac{\partial l}{\partial\boldsymbol{\beta}},$$

where both derivatives are evaluated at $\boldsymbol{\beta}^{old}$. The updating is repeated until $\boldsymbol{\beta}$ converges.

In GLMs, the updating equation is written explicitly. Let $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$ and $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T$. From (2.10) and (2.11), some evaluation of derivatives yields

$$\frac{\partial l}{\partial\boldsymbol{\beta}} = \left(\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\beta}}\right)^T\frac{\partial l}{\partial\boldsymbol{\eta}} = X^T\frac{\partial l}{\partial\boldsymbol{\eta}} = \frac{1}{\phi}X^TW(\boldsymbol{z} - X\boldsymbol{\beta}^{old})$$

and

$$\mathrm{E}\left(-\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right) = \left(\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\beta}}\right)^T\mathrm{E}\left(-\frac{\partial^2 l}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T}\right)\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\beta}} = \frac{1}{\phi}X^TWX,$$

where $\boldsymbol{z}$ is the n-vector called the working response vector with ith component

$$z_i = (y_i - \mu_i)G'(\mu_i) + \boldsymbol{x}_i^T\boldsymbol{\beta}^{old}, \tag{2.13}$$

and W is the $n \times n$ diagonal matrix called the working weight matrix with ith component

$$w_i = \{G'(\mu_i)^2 b''(\theta_i)\}^{-1} m_i. \tag{2.14}$$

Therefore the updating equation becomes

$$\boldsymbol{\beta}^{new} = (X^T W X)^{-1} X^T W \boldsymbol{z}. \tag{2.15}$$

Notice that, for each iteration, $\boldsymbol{z}$ and W are reevaluated by (2.13) and (2.14), respectively, using μ_i and θ_i that are reevaluated by (2.11) and (2.10) from $\boldsymbol{\beta}^{old}$. The updating equation (2.15) has the form of a weighted least squares equation, so the algorithm is one of the iterative reweighted least squares (IRLS) (Green, 1984). Many statistical packages and libraries provide the routines for the IRLS.

For details on Fisher scoring algorithm in GLMs, see also McCullagh and Nelder (1989) and Green and Silverman (1994).

2.2.2 Models included in GLMs

We describe three examples which are representative of GLMs.

(a) *Normal linear regression model.* Suppose that each of responses y_i has the normal distribution $\mathrm{N}(\mu_i, \sigma^2)$, that is, $a_i(\phi) = \sigma^2$, $b(\theta_i) = \frac{1}{2}\theta_i^2$ and $c(y_i, \phi) = -\frac{1}{2}(y/\sigma)^2 - \log \sigma\sqrt{2\pi}$ in (2.9). So μ_i coincides with θ_i in (2.10), and the variance function is $V(\mu) = 1$. If G is the identical function in (2.11), the normal linear regression model

$$y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i, \quad \epsilon \sim \mathrm{N}(0, \sigma^2)$$

is obtained.

(b) *Binomial distribution (binary response) case.* Suppose that $\tilde{y}_i$, $i = 1, \ldots, n$, has the binomial distribution $\mathrm{B}(m_i, \mu_i)$, and then $y_i = \tilde{y}_i/m_i$ takes discrete values in [0,1]. Especially, when all the m_i's are 1, y_i become binary random variables, each of which takes only two values 0 and 1 with the probability $\mathrm{P}(y_i = 1) = \mu_i$. Put $a_i(\phi) = 1/m_i$, $b(\theta_i) = \log(1+e^{\theta_i})$ and $c(y_i; \phi) = \log \binom{m_i}{m_i y_i}$ in (2.9). The mean of y_i is written as $\mu = e^{\theta_i}/(1+e^{\theta_i})$ from (2.10), and the variance function is $V(\mu) = \mu(1-\mu)$. If we put G as the logit function of μ_i:

$$G(\mu_i) = \log \frac{\mu_i}{1-\mu_i},$$

then θ_i coincides with $G(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$, and the logistic regression model

$$\log \frac{\mu_i}{1-\mu_i} = \boldsymbol{x}_i^T\boldsymbol{\beta}$$

is obtained. As the link function G, the probit function $G(\mu_i) = \Phi^{-1}(\mu_i)$ is also often used, where Φ is the cumulative distribution function of the standard normal distribution.

(c) *Poisson distribution case.* Suppose that each of y_i has the Poisson distribution with the mean parameter μ_i, that is, $a_i(\phi) \equiv 1$, $b(\theta_i) = e^{\theta_i}$ and $c(y_i, \phi) = -\log(y_i!)$ in (2.9). Then the mean becomes $\mu_i = e_i^{\theta}$ from (2.10), and the variance function is $V(\mu) = \mu$. If we put G as $G(\mu_i) = \log \mu_i$, then θ_i equals to $\boldsymbol{x}_i^T \boldsymbol{\beta}$, and the Poisson regression model

$$\log(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

is obtained. This model is also known as the log-linear model, which is often used in analysis of contingency tables.

If G is the inverse function of $b'(\theta_i)$, then θ_i coincides with the linear predictor $\boldsymbol{x}_i^T \boldsymbol{\beta}$. In this case G is called the canonical link function. In each of the three examples described above (in (b), if G is the logit function), G is the canonical link function.

2.2.3 Influence and Diagnostics in GLMs

The deviance is one of the measures of goodness-of-fit of a model to data. The scaled deviance D^* is defined as the log-likelihood ratio statistic $D^* = 2[l_{\max} - l(\hat{\boldsymbol{\theta}})]$, where $l_{\max}$ is the maximum log-likelihood for the saturated model that allows one parameter for each observation. The vector $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_n)$ is linked to the MLE $\hat{\boldsymbol{\beta}}$. The nuisance parameter ϕ is multiplied so that the deviance is independent of ϕ, and hence the deviance D becomes

$$D = \phi D^* = 2 \sum_{i=1}^{n} m_i [\{y_i \tilde{\theta}_i - b(\tilde{\theta}_i)\} - \{y_i \hat{\theta}_i - b(\hat{\theta}_i)\}],$$

where $\tilde{\theta}_i$ satisfies $b'(\tilde{\theta}_i) = y_i$, which gives $l_{\max}$. In the normal distribution case, the deviance D results in the residual sum of squares. So, in the way similar to analysis of variance, analysis of deviance (ANODEV) is often implemented for model selection in hierarchical models, where the scaled deviance D^* is compared to the chi-squared distribution with the degrees of freedom $n - p$. The contribution to the deviance from the ith observation

$$d_i = 2m_i [\{y_i \tilde{\theta}_i - b(\tilde{\theta}_i)\} - \{y_i \hat{\theta}_i - b(\hat{\theta}_i)\}]$$

is called the deviance increment, which measures the local influence of each observation.

Another measure of goodness-of-fit is the Pearson chi-squared statistic

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)/m_i},$$

where $V(\mu)$ is the variance function. The measure is also compared to the chi-squared distribution χ^2_{n-p}. For details on these statistics, see McCullagh and Nelder (1989).

The asymptotic variance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\left\{ \mathrm{E}\left(-\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right) \right\}^{-1} = \phi (X^T W X)^{-1}.$$

The scale parameter ϕ is usually estimated by $\hat{\phi} = D/(n-p)$, where D is the deviance, or $\hat{\phi} = \chi^2/(n-p)$, where χ^2 is the Pearson chi-squared statistic.

The leverage values is defined as the ith diagonal component, A_{ii}, of the hat matrix A on the updating equation (2.15), where $A = X(X^T W X)^{-1} X^T W$, and W is evaluated on the final iteration.

In GLMs, the simple quantity $y_i - \hat{\mu}_i$ might be considered as a residual, but it is inappropriate for the local assessment of goodness-of-fit, because the variance of y_i depends on its mean in GLMs. So the standardized versions of residuals have been proposed on the basis of the goodness-of-fit measure described above. The deviance residual is

$$r_i = \mathrm{sign}(y_i - \hat{\mu}_i) \sqrt{d_i}$$

and the Pearson residual is

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)/m_i}}.$$

To evaluate the departure between each observation and the fitted value, the studentized Pearson residual

$$r_i^* = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)(1 - A_{ii})\hat{\phi}/m_i}}$$

is useful. Other residuals for checking the local goodness-of-fit of a model are introduced in Goto and Tsuchiya (1985) and McCullagh and Nelder (1989).

To investigate influence of the ith observation on the MLE $\hat{\boldsymbol{\beta}}$, it is useful to evaluate the deleted estimate $\hat{\boldsymbol{\beta}}^{(-i)}$ when the ith observation is deleted. Computing the exact $\hat{\boldsymbol{\beta}}^{(-i)}$ is expensive, but the one-step approximation of $\hat{\boldsymbol{\beta}}^{(-i)}$ by the Newton–Raphson method is convenient (Cook and Weisberg, 1982). Let $l^{(-i)}$ be the log-likelihood when the ith observation is deleted. The one-step approximation is written as

$$\hat{\boldsymbol{\beta}}^{(-i)} \approx \hat{\boldsymbol{\beta}} - \left(\left. \frac{\partial^2 l^{(-i)}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right|_{\hat{\boldsymbol{\beta}}} \right)^{-1} \left. \frac{\partial l^{(-i)}}{\partial \boldsymbol{\beta}} \right|_{\hat{\boldsymbol{\beta}}}.$$

If G is the canonical link function (i.e., $\eta_i \equiv G(b'(\theta_i)) = \theta_i$), the Newton–Raphson algorithm is equivalent to the Fisher Scoring algorithm, because all the y_i's in the second derivative of l disappear. Therefore we obtain

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{(-i)} &\approx \hat{\boldsymbol{\beta}} + (X_{(i)}^T W_{(i)} X_{(i)})^{-1} X_{(i)}^T W_{(i)} (\boldsymbol{z}_{(i)} - X_{(i)}^T \hat{\boldsymbol{\beta}}) \\ &= \hat{\boldsymbol{\beta}} - \frac{(X^T W X)^{-1} X^T W (\boldsymbol{z} - X\hat{\boldsymbol{\beta}})}{1 - A_{ii}} \\ &= \hat{\boldsymbol{\beta}} - \frac{(X^T W X)^{-1} X^T \hat{\boldsymbol{s}}}{1 - A_{ii}}, \end{aligned}$$

where the subscript (i) means that the component corresponding to the ith observation is deleted, A_{ii} is the leverage value, and $\hat{\boldsymbol{s}} = (s_1, \ldots, s_n)^T$ with $s_i = m_i(y_i - \hat{\mu}_i)$. Other tools for regression diagnostics are described in McCullagh and Nelder (1989). Especially in logistic regression case, Pregibon (1981) developed many tools for regression diagnostics and sensitivity analysis.

2.3 Maximum Penalized Likelihood Estimation

In this section non(semi)-parametric extension of generalized linear models is discussed, and the penalized version of maximum likelihood approach is described.

2.3.1 Nonparametric Generalized Linear Models

The generalized linear model in Section 2.2 is extended in a nonparametric way. For the stochastic component, the situation of exponential family is supposed on the responses y_i as in the ordinary GLM. On the other hand, for the systematic component, the assumption of linearity on mean structure such as in (2.11) is weaken. Assume that there exists a link function G such that

$$G(\mu_i) = f(t_i) \tag{2.16}$$

for some given explanatory variable t_i, and the unknown function f is to be estimated. In what follows, t_i is supposed to be one-dimensional for simplicity. So this model is called the *nonparametric generalized linear model.* Denote the predictor, the right-hand side of (2.16), by η_i.

Maximizing the ordinary log-likelihood (2.12) yields the estimate of f that interpolates data and shows rapid variation. Instead, similar to the idea of penalized least squares, we consider the penalized version of the log-likelihood. In the log-likelihood (2.12), the nuisance parameter ϕ and the term $c(y_i; \phi)$ is irrelevant to the maximization, so the penalized log-likelihood is defined as

$$\begin{aligned} \Pi(f) &= l(f) - \frac{\lambda}{2} J(f) \\ &= \sum_{i=1}^{n} m_i\{y_i\theta_i - b(\theta_i)\} - \frac{\lambda}{2} J(f). \end{aligned} \tag{2.17}$$

Notice that each of θ_i is linked to $f(t_i)$ by (2.10) and (2.16). Here, as in Section 2.1.1, λ is the smoothing parameter (a positive number, specified in this chapter), and $J(f)$ is a roughness penalty. We find an estimate $\hat{f}$ in a class of smooth functions that maximizes the penalized log-likelihood (2.17). The estimate $\hat{f}$ is called the maximum penalized likelihood estimate (MPLE).

As in the penalized least squares, if we use the roughness penalty $J(f) = \int\{f''(t)\}^2 dt$, the class of $\hat{f}$ becomes cubic smoothing splines with knots at $t_1, \ldots, t_n$. Moreover, $\hat{f}$ tends to the linear estimated function in the ordinary GLM as λ tends to infinity, and $\hat{f}$ shows more rapid variation as λ becomes smaller.

In the special case where each of the responses y_i follows the normal distribution $\mathrm{N}(f(t_i), \sigma^2)$, the penalized log-likelihood (2.17) becomes

$$\Pi(f) = -\frac{1}{2}\sum_{i=1}^{n}\{y_i - f(t_i)\}^2 - \frac{\lambda}{2}J(f),$$

which is equivalent to the penalized sum of squares (2.1). Hence the algorithm described in Section 2.1.1 can be used, and no iterative algorithm is required. The constant factor $\frac{1}{2}$ in (2.17) is multiplied by $\lambda J(f)$ for this reason.

In the other distribution case, the maximum penalized likelihood estimation requires iterative calculation as in the ordinary GLM. The approach of Fisher scoring can be also used in the penalized version. As in Section 2.1.1, assume that f is written as $f(t) = \sum_{k=1}^{q}\xi_k\varphi_k(t)$, and that $J(f) = \boldsymbol{\xi}^T K\boldsymbol{\xi}$ for $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_q)^T$ and some $q \times q$ non-negative definite symmetric matrix K. The updating equation for $\boldsymbol{\xi}$ has the form

$$\boldsymbol{\xi}^{new} = \boldsymbol{\xi}^{old} + \left\{\mathrm{E}\left(-\frac{\partial^2\Pi}{\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^T}\right)\right\}^{-1}\frac{\partial\Pi}{\partial\boldsymbol{\xi}},$$

where both derivatives are evaluated at $\boldsymbol{\xi}^{old}$, and the updating is repeated until $\boldsymbol{\xi}$ converges.

It is easy to show that

$$\frac{\partial\Pi}{\partial\boldsymbol{\xi}} = \frac{\partial l}{\partial\boldsymbol{\xi}} - \lambda K\boldsymbol{\xi}^{old}$$

and

$$\mathrm{E}\left(-\frac{\partial^2\Pi}{\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^T}\right) = \mathrm{E}\left(-\frac{\partial^2 l}{\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^T}\right) + \lambda K.$$

If we denote the $n \times q$ basis matrix $\{\varphi_k(t_i)\}_{i=1,\ldots,n;k=1,\ldots,q}$ by B, then $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T = B\boldsymbol{\xi}$. Some evaluation of derivatives yields

$$\frac{\partial l}{\partial\boldsymbol{\xi}} = \left(\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\xi}}\right)^T\frac{\partial l}{\partial\boldsymbol{\eta}} = B^T W(\boldsymbol{z} - \boldsymbol{\eta}^{old})$$

and

$$\mathrm{E}\left(-\frac{\partial^2 l}{\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^T}\right) = \left(\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\xi}}\right)^T\mathrm{E}\left(-\frac{\partial^2 l}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T}\right)\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\xi}} = B^T WB,$$

where, similarly to the ordinary GLM in Section 2.2.1, $\boldsymbol{z}$ is the working response vector with ith component

$$z_i = (y_i - \mu_i)G'(\mu_i) + \eta_i^{old} \tag{2.18}$$

and W is the working weight matrix with ith component

$$w_i = \{G'(\mu_i)^2 b''(\theta_i)\}^{-1}m_i. \tag{2.19}$$

Hence we obtain

$$\frac{\partial \Pi}{\partial \boldsymbol{\xi}} = B^T W(\boldsymbol{z} - B\boldsymbol{\xi}^{old}) - \lambda K \boldsymbol{\xi}^{old} = B^T W \boldsymbol{z} - (B^T W B + \lambda K)\boldsymbol{\xi}^{old}$$

and

$$\mathrm{E}\left(-\frac{\partial^2 \Pi}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}^T}\right) = B^T W B + \lambda K.$$

Therefore the updating equation for $\boldsymbol{f} = (f(t_1), \ldots, f(t_n))^T$ becomes

$$\boldsymbol{f}^{new} = B\boldsymbol{\xi}^{new} = B(B^T W B + \lambda K)^{-1} B^T W \boldsymbol{z}. \tag{2.20}$$

Notice that, for each iteration, $\boldsymbol{z}$ and W are reevaluated by (2.18) and (2.19), respectively, using μ_i and θ_i that are reevaluated by (2.16) and (2.10) from $\boldsymbol{\xi}^{old}$. The updating equation (2.20) has the same form as the penalized least squares equation (2.3), and can be written as $\boldsymbol{f}^{new} = S_\lambda \boldsymbol{z}$, where S_λ is the smoother matrix.

For further discussion about maximum penalized likelihood estimation in nonparametric GLMs, see also O'Sullivan, Yandell and Raynor (1986) and Green and Silverman (1994).

2.3.2 Semiparametric Generalized Linear Models

Green and Yandell (1985) proposed the semiparametric extension of GLMs. They considered as an assumption on the systematic component

$$G(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} + f(t_i) \tag{2.21}$$

(as in Section 2.1.2, t_i is assumed to be one-dimensional), and called this the *semiparametric generalized linear model.* Both the parameter $\boldsymbol{\beta}$ and the unknown function f are to be estimated. As in the previous subsection, denote the right-hand side of (2.21) by η_i.

Just as in the nonparametric GLM, the penalized log-likelihood

$$\begin{aligned} \Pi(\boldsymbol{\beta}, f) &= l(\boldsymbol{\beta}, f) - \frac{\lambda}{2} J(f) \\ &= \sum_{i=1}^{n} m_i \{y_i \theta_i - b(\theta_i)\} - \frac{\lambda}{2} J(f) \end{aligned}$$

is maximized over $\boldsymbol{\beta}$ and f in a class of smooth functions. Notice that each of θ_i is linked to $\boldsymbol{\beta}$ and $f(t_i)$ by (2.10) and (2.21).

The approach of Fisher scoring can be also applied to the maximum penalized likelihood estimation in the semiparametric GLM. Assume that $f(t) = \sum_{k=1}^{q} \xi_k \varphi_k(t)$ and $J(f) = \boldsymbol{\xi}^T K \boldsymbol{\xi}$ as in the previous subsection. The updating equation for $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ is written as

$$\begin{bmatrix} \boldsymbol{\beta}^{new} \\ \boldsymbol{\xi}^{new} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}^{old} \\ \boldsymbol{\xi}^{old} \end{bmatrix} + \left\{ \mathrm{E} \begin{bmatrix} -\dfrac{\partial^2 \Pi}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} & -\dfrac{\partial^2 \Pi}{\partial \boldsymbol{\beta} \partial \boldsymbol{\xi}^T} \\ -\dfrac{\partial^2 \Pi}{\partial \boldsymbol{\xi} \partial \boldsymbol{\beta}^T} & -\dfrac{\partial^2 \Pi}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}^T} \end{bmatrix} \right\}^{-1} \begin{bmatrix} \dfrac{\partial \Pi}{\partial \boldsymbol{\beta}} \\ \dfrac{\partial \Pi}{\partial \boldsymbol{\xi}} \end{bmatrix},$$

where all the derivatives in the right-hand side are evaluated at $\boldsymbol{\beta}^{old}$ and $\boldsymbol{\xi}^{old}$, and the updating is repeated until both $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ converge.

Let $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T$. The predictor vector is written as $\boldsymbol{\eta} = X\boldsymbol{\beta} + B\boldsymbol{\xi}$, and hence the first derivatives become

$$\frac{\partial \Pi}{\partial \boldsymbol{\beta}} = \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}}\right)^T \frac{\partial l}{\partial \boldsymbol{\eta}} = X^T W(\boldsymbol{z} - \boldsymbol{\eta}^{old})$$

and

$$\begin{aligned} \frac{\partial \Pi}{\partial \boldsymbol{\xi}} &= \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\xi}}\right)^T \frac{\partial l}{\partial \boldsymbol{\eta}} - \lambda K \boldsymbol{\xi}^{old} = B^T W(\boldsymbol{z} - \boldsymbol{\eta}^{old}) - \lambda K \boldsymbol{\xi}^{old} \\ &= B^T W(\boldsymbol{z} - X\boldsymbol{\beta}^{old}) - (B^T W B + \lambda K)\boldsymbol{\xi}^{old}, \end{aligned}$$

where $\boldsymbol{z}$ is the working response vector (2.18), and W is the working weight matrix (2.19). By evaluating the expectations of the second derivatives similarly and performing some matrix calculation, the updating equation becomes

$$\begin{bmatrix} \boldsymbol{\beta}^{new} \\ \boldsymbol{\xi}^{new} \end{bmatrix} = \begin{bmatrix} X^T W X & X^T W B \\ B^T W X & B^T W B + \lambda K \end{bmatrix}^{-1} \begin{bmatrix} X^T \\ B^T \end{bmatrix} W \boldsymbol{z}.$$

If the notation of the smoother matrix $S_\lambda = B(B^T W B + \lambda K)^{-1} B^T W$ is used, the MPLE's $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{f}} = (\hat{f}(t_1), \ldots, \hat{f}(t_n))^T$ can be obtained with the updating equations

$$\hat{\boldsymbol{\beta}}^{new} = \{X^T W(I - S_\lambda)X\}^{-1} X^T W(I - S_\lambda)\boldsymbol{z} \tag{2.22a}$$

and

$$\hat{\boldsymbol{f}}^{new} = S_\lambda(\boldsymbol{z} - X\hat{\boldsymbol{\beta}}^{new}). \tag{2.22b}$$

These representations are just the same as the penalized least squares estimates (2.6) in the semiparametric regression model, and so the algorithm of the semiparametric GLM is the iteration of the semiparametric regression algorithm.

For details about semiparametric GLMs, see also Green (1987) and Green and Silverman (1994).

2.3.3 Generalized Additive Models and Other Models

Hastie and Tibshirani (1986) proposed an extension of GLMs. They considered, for the systematic component, replacing the linear predictor (2.11) by a sum of smooth univariate functions,

$$G(\mu_i) = \alpha + \sum_{j=1}^{p} f_j(t_{ij}), \tag{2.23}$$

and called this the *generalized additive model* (GAM). The additive model described in Section 2.1.3 is a special case of GAM, in which the responses y_i follow the normal distribution.

The generalized additive models can be also estimated using the principle of maximum penalized likelihood estimation. In GAMs, the penalized log-likelihood is written as

$$\Pi(f_1, \ldots, f_p) = l(f_1, \ldots, f_p) - \frac{1}{2}\sum_{j=1}^{p} \lambda_j J(f_j),$$

where $l(f_1, \ldots, f_p)$ is the ordinary log-likelihood. The idea of Fisher scoring can be generalized into GAMs, and Hastie and Tibshirani (1986) called this the *local scoring algorithm.* The algorithm is practically the iteration of the backfitting algorithm, described in Section 2.1.3.

In our notation the local scoring algorithm is represented as follows. Let $\boldsymbol{z}$ be the working response vector (2.18), where η_i is the additive predictor, the right-hand side of (2.23), and W be the working weight matrix (2.19). Each of the equations in the backfitting algorithm (inner loop) is written as

$$\boldsymbol{f}_j = S_{\lambda_j}^{(j)}\left(\boldsymbol{z} - \boldsymbol{\alpha} - \sum_{k \neq j} \boldsymbol{f}_k\right), \quad j = 1, \ldots, p,$$

where $\boldsymbol{\alpha} = (\alpha, \ldots, \alpha)^T$ and $\boldsymbol{f}_j = (f_j(t_{1j}), \ldots, f_j(t_{nj}))^T$. Initially α is set to be $G(\bar{y})$ and $\boldsymbol{f}_j$'s are set to be zero. After getting $\boldsymbol{f}_j$'s and $\boldsymbol{\eta} = \boldsymbol{\alpha} + \sum_{j=1}^{p} \boldsymbol{f}_j$, $\boldsymbol{z}$ and W are reevaluated, and the updating (outer loop) is repeated until $\boldsymbol{\eta}$ converges. For detailed derivation and convergence criterion of the local scoring algorithm, see Hastie and Tibshirani (1986, 1990). Some applications of GAMs are discussed in Hastie and Tibshirani (1987, 1990).

The interaction spline models for non-Gaussian data have been developed by Wang (1994) and Wahba *et al.* (1995). They called these models the smoothing spline analysis of variance (ANOVA) models. Wang (1995) provided the FORTRAN program GRKPACK for fitting the spline smoothing ANOVA models. The code GRKPACK calls RKPACK (Section 2.1.3) as the subroutine.

2.3.4 Inference and Diagnostics in Non(semi)-parametric GLMs and GAMs

In non(semi)-parametric GLMs and GAMs, the overall equivalent degrees of freedom (EDF) for a model, denoted by ν, is defined as the trace of the hat matrix A, evaluated on the final iteration of the Fisher scoring or the local scoring. The leverage values, defined as the diagonal components A_{ii} of the hat matrix A, are also evaluated on the final iteration.

The deviance D and the Pearson chi-squared statistics χ^2 are of the same form as in the ordinary GLM (Section 2.2.4), in which the values of $\hat{\theta}_i$ or $\hat{\mu}_i$ evaluated from the MPLE are used. The measures may be used for analysis of deviance (ANODEV), and compared to the chi-squared distribution with the degrees of freedom for noise, $n-\nu$, but the asymptotic theory on the distribution of D and χ^2 does not seem to have been established. These measures can not be used for selecting the smoothing parameter, because in general they give

smaller values as λ becomes smaller. The scale parameter ϕ is estimated by $\hat{\phi} = D/(n-\nu)$ or $\hat{\phi} = \chi^2/(n-\nu)$.

In the semiparametric GLM in Section 2.3.2, the EDF is written as

$$\nu = \mathrm{tr}S_\lambda + \mathrm{tr}[\{X^TW(I-S_\lambda)X\}^{-1}X^TW(I-S_\lambda)^2X],$$

and the asymptotic variance matrix of $\hat{\boldsymbol{\beta}}$ is represented as

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) = \hat{\phi}\{X^TW(I-S_\lambda)X\}^{-1}X^TW(I-S_\lambda)^2X\{X^TW(I-S_\lambda)X\}^{-1},$$

where W is evaluated on the final iteration (Green and Yandell,1985). Green (1987) provides further discussion about the deviance and the EDF in semiparametric GLMs.

The notion about the residuals and influential observations will be generalized into the non(semi)-parametric case. The one-step approximation of the deleted estimates in nonparametric GLMs is developed in Section 3.2.

2.4 Case Studies

In this section we illustrate some examples in which non(semi)-parametric GLMs are fitted. As representative examples, logistic regression and Poisson regression are taken up. Density estimation for classified data is also attempted. Through this section, cubic B-splines with knots at the values of the nonparametric explanatory variables $t_1, \ldots, t_n$ are used as nonparametric estimated functions.

2.4.1 Logistic Regression

Binary Response Case: Kyphosis in Laminectomy Patients

Hastie and Tibshirani (1990) and Chambers and Hastie (1992) considered fitting a generalized additive model to data on laminectomy surgery. Each of the data is composed of a response that indicates presence or absence of kyphosis after the operation, and three variables: `age`, `number` and `start` (See Appendix). The goal of analysis is to investigate the relationship between the prevalence of kyphosis and these three predictors, and to identify risk factors for kyphosis. The observations No. 15 that has extremely large value of `age` and No. 28 that has extremely large value of `number` are removed from the analysis.

Hastie and Tibshirani (1990, Section 10.2) fitted various form of additive logistic models to the kyphosis data in preliminary analysis, and finally suggested the parametric model

$$\log\frac{p_i}{1-p_i} = \texttt{poly(age,2)} + (\texttt{start} - 12)\cdot I(\texttt{start} > 12),$$

where $p_i = \mathrm{P}(\mathrm{present})$, the probability of presence of kyphosis, `poly(age,2)` is the quadratic polynomial of `age`, and $I(\cdot)$ is the indicator function.

For an illustration of the semiparametric generalized linear model, we fit a semiparametric logistic regression model to the kyphosis data. The boxplots in

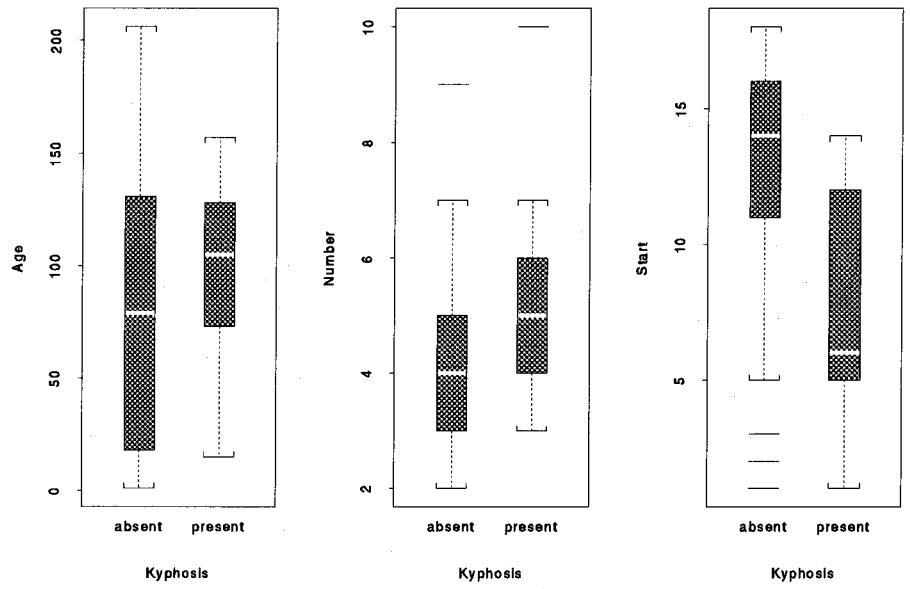


Figure 2.1: Boxplots of **age**, **number** and **start** by presence or absence of kyphosis.

Figure 2.1 and Chambers and Hastie's (1992) consideration with tree models suggest that presence of kyphosis should be typically caused by higher number of level (**number** > 4.5) and lower start level (**start** > 12.5). We fit the semiparametric model of the form

$$\log \frac{p_i}{1-p_i} = f(\mathbf{age}) + \beta_1 I(\mathbf{number} > 4.5) + \beta_2 I(\mathbf{start} > 12.5) + \beta_3 I(\mathbf{number} > 4.5) \cdot I(\mathbf{start} > 12.5).$$

The last term of the right-hand side is incorporated to evaluate the effect of interaction, since both **number** and **start** are related on vertebrae level.

Table 2.1 shows the results of comparing the model in which the interaction term is included, with the model from which the interaction term is excluded. The S-PLUS function **smooth.spline** is used for computation. We select the value of the smoothing parameter, say $\hat{\lambda}$, by minimizing $\mathrm{LCV}_1(\lambda)$ to be described in Section 3.2. (See also Section 3.3.) The existence of the interaction term has little effect on the values of $\hat{\lambda}$ and $\hat{\beta}_k$ $(k = 1, 2, 3)$, and makes the values of the chi-squared statistic χ^2 and the deviance D less decreasing. Hence the model with no interaction term is suggested. It can be seen from Table 2.1 that the term of $I(\mathbf{start} > 12.5)$ associated with $\hat{\beta}_2$ gives stronger effect. In fact, the numbers 1–12 correspond to the thoracic vertebrae, while the numbers 13–17 correspond to the lumbar vertebrae.

Figure 2.2 plots the fitted function $\hat{f}$ concerned with **age** for the model with no interaction term. The values of the parametric term and the Pearson

Table 2.1: The results of analysis to the kyphosis data: the model in which the interaction term is included, and the model from which the interaction term is excluded.

Interaction term	Included		Excluded	
$\hat{\lambda}$	0.066		0.064	
$\hat{\beta}_1$	1.550	(0.555)	1.472	(0.436)
$\hat{\beta}_2$	−2.685	(1.111)	−2.852	(0.615)
$\hat{\beta}_3$	1.550	(0.555)	1.472	(0.436)
ν (EDF)	5.99		5.01	
$n - \nu$	75.01		75.99	
χ^2	62.985		65.387	
D	51.073		51.096	
$\hat{\phi} = \chi^2/(n-\nu)$	0.840		0.860	
$\mathrm{LCV}_1(\hat{\lambda})$	0.7877		0.7567	

Note: The standard errors of $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ are in parentheses.

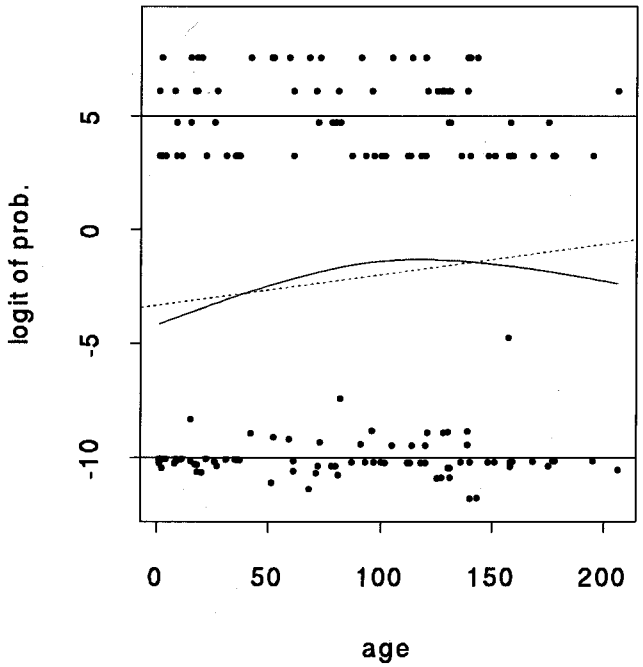


Figure 2.2: Plot of the fitted function $\hat{f}$ to the kyphosis data (in logit scale). The dots on the top side plots the values of the sum of the parametric terms, and the dots on the bottom side plots the Pearson residuals. The broken line shows the ordinary logistic regression fitting.

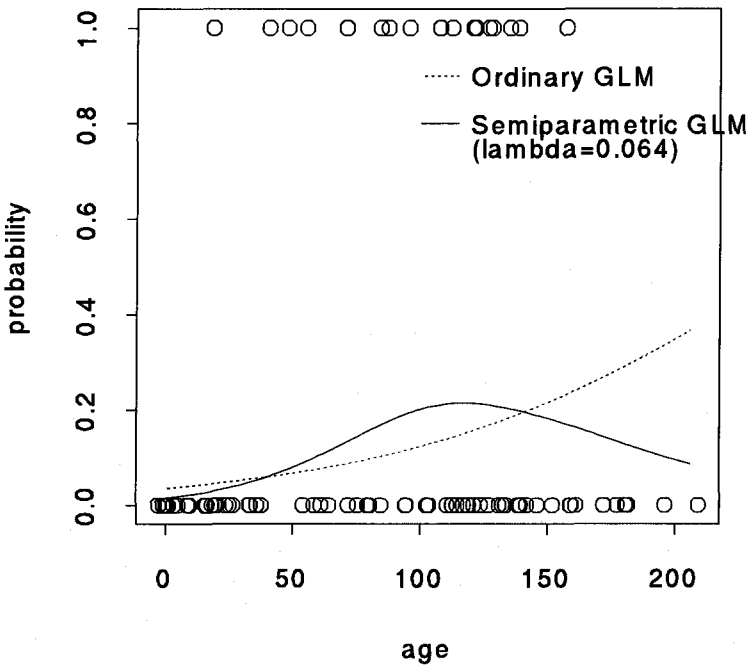


Figure 2.3: Plot of the logit of $\hat{f}$ to the kyphosis data in probability scale. The responses that show presence or absence of kyphosis are also plotted. The broken line shows the ordinary logistic regression fitting.

residuals are also plotted. Figure 2.3 plots the logit of $\hat{f}$ in probability scale with the responses. For comparison, ordinary logistic regression model fitting are superimposed with the broken lines in both figures. The figures show that risk of kyphosis is the highest at about 120 months old.

The interaction between age and other variables cannot be evaluated with the models described in the previous sections. It requires a multidimensional nonparametric models such as an interaction spline model. We would like to consider using such a model in further analysis.

Semiparametric GLMs were also considered by Green and Yandell (1985), where the semiparametric logistic model was fitted to the data on prevalence of bileduct hyperplasia. O'Sullivan, Yandell and Raynor (1986) fitted logistic models using two-dimensional smoothing splines to heart disease data and potato early dying disease data.

Binomial Response Case: Assay of Trypanosomes

Ashford and Walker (1972) illustrated the analysis of quantal response data on Assay of Trypanosome. They considered the mixture of probit model

$$p_i = \theta\Phi(a_1 + b_1 t_i) + (1-\theta)\Phi(a_2 + b_2 t_i),$$

where p_i is the probability of response, the mortality rate, Φ is the standardized cumulative normal integral, t_i is the log dose, and a_1, a_2, b_1, b_2 and θ are parameters to be estimated. Eilers and Marx (1996) fitted a nonparametric logistic regression model to the trypanosome data. They used P-splines proposed by themselves, of degree 3 and penalties of order 2, 3 and 4, without taking the logarithm of the dose, and selected the smoothness of the estimated function by AIC. They suggested a cubic logistic fit as a result.

Denote the number of deaths by y_i and the number of observations by m_i in the ith dose level. We fit the nonparametric logistic regression model

$$y_i \sim \mathrm{B}(m_i, p_i) \quad \text{and} \quad \log\frac{p_i}{1-p_i} = f(t_i).$$

The penalized likelihood becomes

$$\Pi(f) = \sum_{i=1}^{n}\{y_i \log p_i + (m_i - y_i)\log(1-p_i)\} - \frac{\lambda}{2}\int\{f''(t)\}^2 dt.$$

The models of the EDF $\nu = 3$, 4 and 5 are selected, and the results for them are listed with the ordinary GLM fitting in Table 2.2. Moreover, four fitted functions are displayed in Figure 2.4. Underfitting for the ordinary GLM and the nonparametric model of $\nu = 3$ is obvious from the large values of χ^2 and $\hat{\phi} = \chi^2/(n-\nu)$. The model of $\nu = 4$ gives seemingly good fitting, which supports the suggestion of Eilers and Marx (1996). The model of $\nu = 5$ seems to be slightly overfitting since $\chi^2 = 1.695$ and $n - \nu = 3.00$.

Table 2.2: The results of analysis to the trypanosome data. The models of the EDF $\nu = 3$, 4 and 5 are listed with the ordinary GLM fitting.

	Ordinary GLM	Nonparametric GLMs		
$\log_{10}\lambda$	—	−0.98	−1.92	−2.65
ν	2	3.00	4.00	5.00
χ^2	20.039	11.085	3.536	1.695
$\hat{\phi}$	3.340	2.217	0.884	0.566
$\mathrm{LCV}_1(\lambda)$	5.026	4.582	2.668	1.768
$\mathrm{LCV}_2(\lambda)$	3.753	3.120	2.783	3.900

Note: $\mathrm{LCV}_1(\lambda)$ and $\mathrm{LCV}_2(\lambda)$ are the likelihood-based cross-validation scores to be discussed in Section 3.2. See also Section 4.1.1.

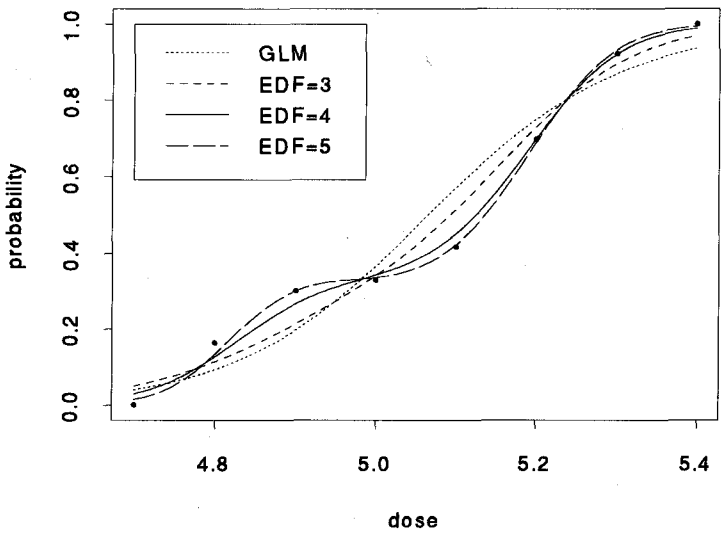


Figure 2.4: Plot of the logit of the fitted functions $\hat{f}$ to the trypanosome data in probability scale, of the EDF 3, 4 and 5, with the fitted function by the ordinary GLM.

2.4.2 Poisson Regression

Poisson regression is effective in situations where the dependent variable is a count within a series of subdivisions of the data. Assume that the count of events y_i in ith category follows the Poisson distribution $\mathrm{Po}(\mu_i)$. The expected count μ_i depends on the size of the ith category, say q_i, such as the person-time, the total time accounted for those who experienced the event and who could have experienced it but did not. In Poisson regression the expected rate μ_i/q_i, rather than the expected count μ_i, is modeled as the function of an independent variable t_i associated with the ith category:

$$\mu_i = q_i e^{\alpha+\beta t_i} \quad \text{or} \quad \log\left(\frac{\mu_i}{q_i}\right) = \alpha + \beta t_i. \tag{2.24}$$

Poisson regression models are expressed in various forms. Let y_{ij} be the number of a two-way contingency table classified by, for example, age categories and exposure categories, and q_{ij} be the person-time. The number y_{ij} can be modeled as

$$y_{ij} \sim \mathrm{Po}(\mu_{ij}) \quad \text{and} \quad \log\left(\frac{\mu_i}{q_i}\right) = \alpha + \beta_i + \gamma_j$$

using the parameters β_i and γ_j that incorporate the effect of the (i,j)th category. See Selvin (1995) for details.

The Poisson regression model also belongs to the class of GLMs. The person-time q_i is included in the model as an additive effect that is not an independent variable, since the model (2.24) is rewritten as

$$\log \mu_i = \log q_i + \alpha + \beta t_i.$$

The term $\log q_i$ is called an offset variable, which is added to the working response for each iteration of the Fisher scoring algorithm.

We consider a nonparametric extension of the Poisson regression model (2.24). Let y_i be the count of events such as deaths, and q_i be the size of the ith category such as the population size or the person-time. The expected rate μ_i/q_i is then modeled as the nonparametric function of an independent variable t_i such as the age:

$$y_i \sim \mathrm{Po}(\mu_i) \quad \text{and} \quad \log\left(\frac{\mu_i}{q_i}\right) = f(t_i).$$

This model can be applied to analysis of a mortality table as described below.

Example: Analysis of a Mortality Table

Green and Silverman (1994) applied the nonparametric GLM to the analysis of a mortality table. Let y_i be the number of death and q_i be the number of individuals in the ith age category. Actuaries estimate *true* death rates p_i from *crude* death rates y_i/q_i by a smoothing technique called graduation to make use of calculating insurance premiums. Green and Silverman (1994) fitted the nonparametric logistic regression model

$$y_i \sim \mathrm{B}(q_i, p_i) \quad \text{and} \quad \log\left(\frac{p_i}{1-p_i}\right) = f(t_i),$$

which ensures that the true death rate p_i is represented as a smooth function of the age t_i.

The approach of Poisson regression is also available for a mortality table. The nonparametric Poisson regression model

$$y_i \sim \mathrm{Po}(\mu_i) \quad \text{and} \quad \log\left(\frac{\mu_i}{q_i}\right) = f(t_i)$$

is fitted to the mortality table, where μ_i is the expected number of deaths. The penalized log-likelihood

$$\Pi(f) = \sum_{i=1}^{n} [y_i\{\log q_i + f(t_i)\} - q_i \exp f(t_i)] - \frac{\lambda}{2}\int \{f''(t)\}^2 dt$$

is maximized to estimate f. We attempt to fit the nonparametric Poisson regression model to the mortality table discussed by Green and Silverman (1994). The estimated death rates $\hat{\mu}_i/q_i = \exp \hat{f}(t_i)$ is plotted with the crude death rates in Figure 2.5. The smoothing parameter λ is selected by minimizing $\mathrm{LCV}_1(\lambda)$. The result is similar to that by the logistic regression in Green and Silverman (1994).

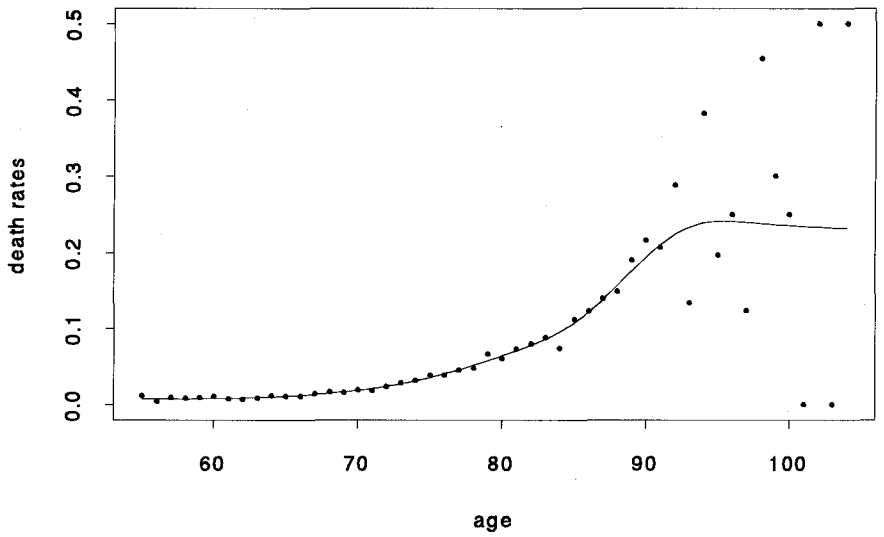


Figure 2.5: The estimated death rates $\exp \hat{f}(t_i)$ (the solid line) and the crude death rates (the dots). to the mortality table in Green and Silverman (1994).

2.4.3 Density Smoothing for Classified Data

The idea of Poisson regression can be applied to density estimation for classified data, or smoothing histograms. Suppose that raw data x_r, $r = 1, \ldots, N$, are extracted from a distribution of the density function $g(x)$. When x_k's are classified into n disjoint classes $C_1, \ldots, C_n$ and the count of data in each class $y_i = \#\{x_r; x_r \in C_i\}$ is observed, the density $g(x)$ is estimated from $y_1, \ldots, y_n$ using nonparametric Poisson regression.

The probability of observing y in C_i is $p_i = \int_{C_i} g(x)dx$ and $\sum_{i=1}^n p_i = 1$. Under the condition that the sum of counts is $\sum_{i=1}^n y_i = N$, the vector of counts $(y_1, \ldots, y_n)$ follows a multinomial distribution with N draws and probability vector $(p_1, \ldots, p_n)$. Equivalently, each of the counts y_i can be also considered to have a Poisson distribution $\mathrm{Po}(\mu_i)$, where the mean parameter is $\mu_i = p_i N$. See Bishop, Fienberg and Holland (1975) for the equivalence between multinomial distribution and Poisson distribution. Let t_i be the ith class mark located at the midpoint of each class. Assume that all the classes have a common width h, that is, t_i's are equally spaced. The probability p_i is then roughly approximated by $p_i \approx hg(t_i)$ and hence $\mu_i \approx Nhg(t_i)$. Therefore a histogram can be smoothed by using the technique of Poisson regression. We fit the nonparametric Poisson regression model

$$y_i = \mathrm{Po}(\mu_i) \quad \text{and} \quad \log \mu_i = f(t_i) \tag{2.25}$$

and estimate f with a cubic smoothing spline. As in the previous subsection,

the penalized log-likelihood

$$\Pi(f) = \sum_{i=1}^{n} \{y_i f(t_i) - \exp f(t_i)\} - \frac{\lambda}{2} \int \{f''(t)\}^2 dt$$

is maximized to obtain $\hat{f}$. The estimated density becomes $\hat{g}(t) = \exp \hat{f}(t)/Nh$.

Eilers and Marx (1996) discussed the density estimation using P-splines as an application of Poisson regression. Efron and Tibshirani (1996) proposed a family of density estimators as a combination of an exponential family and a kernel estimator, and called it the specially designed exponential family. The idea of Poisson regression is also included in their approach. Density estimation for non-classified data in the context of maximum penalized approach is discussed in Good and Gaskins (1971), Silverman (1982, 1986), O'Sullivan (1988) and so on.

Example: Old Faithful Geyser Data

As an example, the density smoothing method with Poisson regression is applied to the data on duration of eruptions of the Old Faithful geyser, also discussed by Silverman (1986) and Eilers and Marx (1996). At first, the domain from 1.5 to 5 is divided into 35 intervals of bin width 0.1 and the histogram is constructed. The nonparametric Poisson regression model (2.25) is fitted using a cubic smoothing spline with knots at class marks of the histogram. The smoothed density function $Nh\hat{g}(t)$ is plotted in Figure 2.6 with the histogram. The smoothing parameter λ that minimizes $\mathrm{LCV}_1(\lambda)$ is selected.

Next we consider changing the origin of the histogram. Figure 2.7 plots the smoothed density function with the histogram when the domain from 1.55 to 5.05 is divided into 35 intervals of bin width 0.1. The two histograms give different impression, while the two smoothed density functions seem to be similar. In general, a histogram gives various impression with the position of the origin as well as the number of classes or bin width. On the other hand, density smoothing has the advantage of little influence from the shape of a histogram, if the division of the histogram is appropriate and the smoothing parameter is selected adaptively.

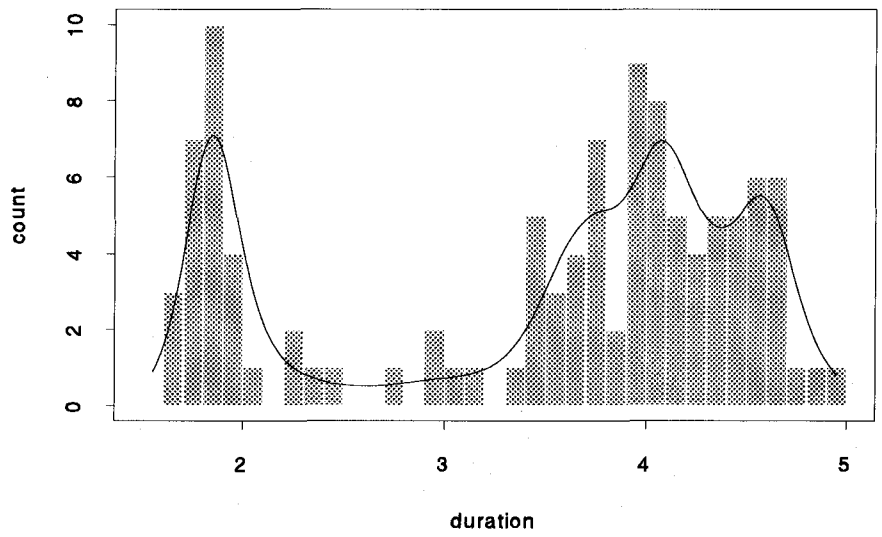


Figure 2.6: The smoothed density and the histogram to the Old Faithful geyser data, in which the domain from 1.5 to 5 is divided into 35 intervals of bin width 0.1.

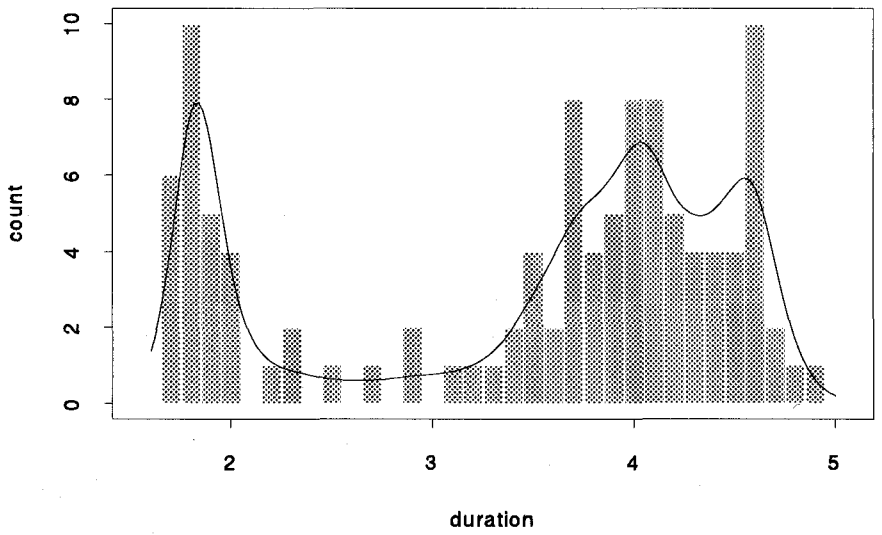


Figure 2.7: The smoothed density and the histogram to the Old Faithful geyser data, in which the domain from 1.55 to 5.05 is divided into 35 intervals of bin width 0.1.

Chapter 3

Selecting the Smoothing Parameter

3.1 Standard Procedures

In this section we discuss some of the standard procedures for selecting the smoothing parameter in maximum penalized likelihood estimation. Through this chapter we take up the case of only one smoothing parameter, that is, the case that only one variable is related to nonparametric component, but similar idea will be adopted to the case of multiple smoothing parameters such as additive models.

3.1.1 Selecting the Smoothing Parameter in Penalized Least Squares Problems

At first we describe some of the common procedures for selecting the smoothing parameter in penalized least squares that correspond to the normal distribution case of maximum penalized likelihood estimation. Here the scores are defined in the weighted form for extension in later section.

In penalized least squares the method of cross-validation is commonly used. The cross-validation is based on the viewpoint of prediction. See Stone (1974) for detail survey and discussion about cross-validation. Denote the predicted value of y_i by $\hat{\eta}_i$, which is obtained as the value η_i $(i = 1, \ldots, n)$ that minimizes the penalized weighted sum of squares

$$\mathcal{S} = \sum_{i=1}^{n} w_i(y_i - \eta_i)^2 + \lambda \boldsymbol{\xi}^T K \boldsymbol{\xi},$$

where η_i stands for $f(t_i)$ in the nonparametric regression model in Section 2.1.1 and for $\boldsymbol{x}_i^T\boldsymbol{\beta} + f(t_i)$ in the semiparametric regression model in Section 2.1.2. The function f is assumed to be written as $f(t_i) = \sum_{k=1}^{q} \xi_k \varphi_k(t_i)$. Let $\hat{\eta}_i^{(-i)}$ be the value of η_i predicted from $n-1$ observations when y_i is deleted. A cross-validation score is then constructed on the basis of the squared-error criterion,

and defined as

$$\mathrm{OCV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} w_i(y_i - \hat{\eta}_i^{(-i)})^2$$

(Wahba and Wold, 1975). The notation OCV stands for ordinary cross-validation to distinguish it from GCV described below. The value of λ that minimizes $\mathrm{OCV}(\lambda)$ is selected.

Calculating $\hat{\eta}_i^{(-i)}$ directly by minimizing the penalized sum of squares constructed from $n-1$ observations except y_i is very expensive, so the method for simple calculation was proposed. Craven and Wahba (1979, Lemma 3.1 and 3.2) proved that the deleted residual $y_i - \hat{\eta}_i^{(-i)}$ is represented as

$$y_i - \hat{\eta}_i^{(-i)} = \frac{y_i - \hat{\eta}_i}{1 - A_{ii}}, \tag{3.1}$$

where η_i is the ordinary predicted value, and A_{ii} is the ith leverage value, the diagonal component of the hat matrix A_λ, i.e., $\hat{\boldsymbol{\eta}} = (\hat{\eta}_1, \ldots, \hat{\eta}_n)^T = A_\lambda \boldsymbol{y}$. The subscript λ is augmented to emphasize the dependence of the hat matrix on λ. The derivation of (3.1) is similar to that of the PRESS criterion, proposed by Allen (1974) in the context of multiple linear regression and ridge regression. See also Green and Silverman (1994, Lemma 3.1 and the following remarks). By using (3.1), the ordinary cross-validation score is written as

$$\mathrm{OCV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} w_i \left(\frac{y_i - \hat{\eta}_i}{1 - A_{ii}}\right)^2. \tag{3.2}$$

A more popular score is the one proposed by Wahba (1977) and Craven and Wahba (1979), called the generalized cross-validation (GCV) score. By replacing A_{ii} in (3.2) with $n^{-1}\mathrm{tr}A_\lambda$, the equivalent degrees of freedom for the model divided by n, the score is written as

$$\mathrm{GCV}(\lambda) = \frac{1}{n}\frac{\sum_{i=1}^{n} w_i(y_i - \hat{\eta}_i)^2}{(1 - n^{-1}\mathrm{tr}A_\lambda)^2},$$

which is minimized to select λ.

In the context of the univariate nonparametric regression (Section 2.1.1), Craven and Wahba (1979) established the optimality of the GCV score in terms of the predicted mean squared error (here we denote in the weighted form)

$$T(\lambda) = \frac{1}{n}\sum_{i=1}^{n} w_i\{\hat{f}(t_i) - f(t_i)\}^2,$$

that is,

$$\frac{ET(\hat{\lambda})}{\min ET(\lambda)} \to 1 \quad \text{as } n \to \infty,$$

where $\hat{\lambda}$ is selected by minimizing $\mathrm{GCV}(\lambda)$. Also, the GCV score has invariance properties for orthogonal transformations (see Wahba, 1990). Golub, Heath and

Wahba (1979) applied the GCV score to ridge regression. Utreras (1981) and Silverman (1985) proposed methods to compute the GCV score using approximations of the eigenvalues of the smoother matrix. Behavior of the GCV score as λ tends to zero is studied by Wahba and Wang (1995).

The method of the cross-validation does not require any knowledge of the error variance. If the error variance (assumed to be common to each observation), say σ^2, is known or estimated using the methods in Section 2.1.4, the unbiased risk (UBR) estimate is valid. Let

$$\hat{T}(\lambda) = \frac{1}{n}\|W^{1/2}(I - A_\lambda)\boldsymbol{y}\|^2 - \frac{\sigma^2}{n}\mathrm{tr}(I - A_\lambda)^2 + \frac{\sigma^2}{n}\mathrm{tr}A_\lambda^2,$$

where $\|\cdot\|$ means the Euclidean norm. Then $E\hat{T}(\lambda) = ET(\lambda)$. Minimizing $\hat{T}(\lambda)$ is equivalent to minimizing

$$\mathrm{UBR}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} w_i(y_i - \hat{\eta}_i)^2 + \frac{2}{n}\sigma^2\mathrm{tr}A_\lambda.$$

The UBR score is an extended form of Mallows' C_p statistic (Mallows, 1973) to nonparametric regression, and is equivalent to the Akaike information criterion (AIC) described in Section 3.1.3 in the normal distribution case.

3.1.2 Cross-validation Scores Based on Squared Error

In general cases of maximum penalized likelihood estimation, the cross-validation method has been also most commonly used for selecting the smoothing parameter. The equation (2.20) for evaluating the MPLE $\hat{f}$ in the nonparametric GLM is viewed as the solution of the penalized least squares problem:

$$\text{Minimize} \quad \mathcal{S} = \sum_{i=1}^{n} w_i(z_i - \eta_i)^2 + \lambda\boldsymbol{\xi}^T K\boldsymbol{\xi} \quad \text{over } f, \tag{3.3}$$

where $\eta_i = f(t_i)$. Similarly, the equation (2.22) in the semiparametric GLM is viewed as the solution of minimizing $\mathcal{S}$ of (3.3) over $\boldsymbol{\beta}$ and f, where $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + f(t_i)$. Here z_i is the working response, and w_i is the working weight, both evaluated on the final iteration of the Fisher scoring algorithm. Therefore, two types of cross-validation scores based on the squared-error criterion might be considered, on the analogue of the discussion in the previous subsection.

Let A_λ be the hat matrix in the equations (2.20) or (2.22), that is, $\hat{\boldsymbol{\eta}} = A_\lambda\boldsymbol{z}$, and A_{ii} be the ith diagonal component of A_λ. Green and Yandell (1985) and O'Sullivan, Yandell and Raynor (1986) used the generalized cross-validation score

$$\mathrm{GCV}(\lambda) = \frac{1}{n}\frac{\sum_{i=1}^{n} w_i(z_i - \hat{\eta}_i)^2}{(1 - n^{-1}\mathrm{tr}A_\lambda)^2}.$$

This score was also suggested by Hastie and Tibshirani (1990, Section 6.9), in the context of generalized additive models (GAM), as an approximation to

the likelihood-based cross-validation score described in the next section. The ordinary cross-validation score

$$\mathrm{OCV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} w_i \left(\frac{z_i - \hat{\eta}_i}{1 - A_{ii}}\right)^2$$

was suggested by Green and Silverman (1994, Section 5.4.3), again as an approximation to the likelihood-based cross-validation score. These scores are evaluated using the values of z_i, w_i and A_{ii} on the final iteration of the Fisher scoring algorithm, and minimized to select λ.

However, it has been known that the GCV score is inappropriate for non-normal distribution case, especially for binary data, in maximum penalized likelihood estimation. Green and Yandell (1985) evaluated the GCV score to select the smoothing parameter in applying semiparametric logistic regression to bioassay data. They pointed out that the GCV score might show bad behavior, that is, it might have no global minimum, and tend to zero as λ tends to zero. Many other authors made a similar suggestion (e.g., Gu, 1992), and our numerical experiments will show the bad behavior of the GCV score (see Chapter 4).

3.1.3 Akaike Information Criterion

The Akaike information criterion (AIC) (Akaike, 1973) has been also often used for selecting the smoothing parameter. The AIC score is constructed from the theory of Kullback–Leibler information, and is defined as $2l(\hat{\boldsymbol{\theta}}) - 2\nu$, the twice of the log-likelihood from which the twice of equivalent degrees of freedom is subtracted. In our notation the AIC score is written, divided by n for the convenience of our comparison, as

$$\begin{aligned}\mathrm{AIC}(\lambda) &= \frac{D}{n} + \frac{2}{n}\phi\nu \\ &= \frac{2}{n}\sum_{i=1}^{n} m_i[\{y_i\tilde{\theta}_i - b(\tilde{\theta}_i)\} - \{y_i\hat{\theta}_i - b(\hat{\theta}_i)\}] + \frac{2}{n}\phi\,\mathrm{tr}A_\lambda\end{aligned}$$

where D is the deviance, $\tilde{\theta}_i$ is the solution of $b'(\tilde{\theta}_i) = y_i$, $\hat{\theta}_i$ is the MPLE of the natural parameter θ_i, and ϕ is the scale parameter. In the case of binomial or Poisson distribution, ϕ is often known and set to be 1. See also Hastie and Tibshirani (1990). Eilers and Marx (1996) also proposed using the AIC score. An asymptotic equivalence of selecting a model by AIC and cross-validation is discussed by Stone (1977).

3.2 Likelihood-based Cross-validation Score

The OCV and GCV scores described in the previous section are based on the squared error criterion, but the performance of these scores, especially the GCV score, is unsatisfactory. We think it is better to construct a cross-validation score based on likelihood in maximum penalized likelihood estimation. The

AIC score is based on likelihood, but the term related to the effective number of parameters is added simply by analogy with the degrees of freedom for a family of parametric models. In this section we construct a likelihood-based cross-validation score, and propose a method for simple calculation of it. We also derive a simpler AIC-like form of the score.

3.2.1 Likelihood-based Cross-validation Score

The likelihood-based cross-validation score, suggested by Green and Silverman (1994), is defined as the sum of the deviance increment (Sections 2.2.4 and 2.3.4) from the model fitted to the delete-one data. Let $d_i^{(-i)}$ be the deviance increment of y_i and $\hat{\theta}_i^{(-i)}$ be the delete-one estimate of the natural parameter θ_i when the ith observation y_i is deleted. The likelihood-based cross-validation (LCV) score is written as

$$\begin{aligned} \mathrm{LCV}(\lambda) &= \frac{1}{n}\sum_{i=1}^{n} d_i^{(-i)} \\ &= \frac{2}{n}\sum_{i=1}^{n} m_i[\{y_i\tilde{\theta}_i - b(\tilde{\theta}_i)\} - \{y_i\hat{\theta}_i^{(-i)} - b(\hat{\theta}_i^{(-i)})\}], \end{aligned} \tag{3.4}$$

where $\tilde{\theta}_i$ is the solution of $b(\theta_i) = y_i$. In the normal distribution case, $\mathrm{LCV}(\lambda)$ coincides with $\mathrm{OCV}(\lambda)$.

However, direct calculation of $\hat{\theta}_i^{(-i)}$ and hence the exact calculation of the LCV score (3.4) is very expensive and not practical, because finding $\hat{\theta}_i^{(-i)}$ requires solving a maximum penalized likelihood equation to the data from which y_i is deleted. So we propose a method for simple calculation of the delete-one estimate $\hat{\theta}_i^{(-i)}$ and the LCV score, also discussed in Sakamoto and Shirahata (1997b).

3.2.2 Simple Calculation of the Delete-one Estimate and the Likelihood-based Cross-validation Score

As described in 3.1.2, the MPLE in a non(semi)-parametric GLM is viewed as the solution of the penalized least squares problem (3.3). By applying the deletion lemma of Craven and Wahba (1979) such as (3.1), an approximation to the predicted values of η_i when y_i is deleted

$$z_i - \hat{\eta}_i^{(-i)} \approx \frac{z_i - \hat{\eta}_i}{1 - A_{ii}}$$

or

$$\hat{\eta}_i^{(-i)} \approx \hat{\eta}_i - \frac{A_{ii}}{1 - A_{ii}}(z_i - \hat{\eta}_i) \tag{3.5}$$

is obtained. Therefore the delete-one estimate $\hat{\theta}_i^{(-i)}$ is obtained from (3.5) using the relationships (2.16) [or (2.21)] and (2.10), that is, $\hat{\eta}_i^{(-i)} = G(b'(\hat{\theta}_i^{(-i)}))$.

If G is the canonical link function, i.e., $\eta_i \equiv G(b'(\theta_i)) = \theta_i$, the expression (3.5) is rewritten explicitly as follows. Since $G'(\mu_i)b''(\theta_i) = 1$, the variance function becomes $V(\mu_i) = G'(\mu_i)^{-1}$, and the working response (2.18) is written as

$$z_i = (y_i - \mu_i)/V(\mu_i) + \eta_i^{old}.$$

Hence, on the final iteration, (3.5) is rewritten without using z_i as

$$\hat{\theta}_i^{(-i)} \approx \hat{\theta}_i - \frac{A_{ii}}{1 - A_{ii}} \frac{y_i - \hat{\mu}_i}{V(\hat{\mu}_i)}, \tag{3.6}$$

where $\hat{\mu}_i = b'(\hat{\theta}_i)$. Therefore the delete-one estimate $\hat{\theta}_i^{(-i)}$ is found by simple calculation using the ordinary estimate $\hat{\theta}_i$ and the leverage value A_{ii}. In the normal distribution case, the relationship (3.6) holds exactly.

An approximation of the LCV score is obtained by substituting the delete-one estimate $\hat{\theta}_i^{(-i)}$ just found into (3.4). If G is the canonical link function, the approximated score is written from (3.6) as

$$\begin{aligned} \mathrm{LCV}_1(\lambda) \;=\; & \frac{2}{n}\sum_{i=1}^{n} m_i \Bigg[\{y_i\tilde{\theta}_i - b(\tilde{\theta}_i)\} \\ & - \left\{ y_i \left(\hat{\theta}_i - \frac{A_{ii}}{1 - A_{ii}} \frac{y_i - \hat{\mu}_i}{V(\hat{\mu}_i)} \right) - b\left(\hat{\theta}_i - \frac{A_{ii}}{1 - A_{ii}} \frac{y_i - \hat{\mu}_i}{V(\hat{\mu}_i)} \right) \right\} \Bigg]. \end{aligned}$$

In the normal distribution case, $\mathrm{LCV}_1(\lambda)$ coincides with $\mathrm{OCV}(\lambda)$.

3.2.3 Equivalence to the One-step Approximation

The expression (3.5) can be associated with the one-step approximation to the delete-one estimate based on the Newton–Raphson method. For simplicity, we describe this in the nonparametric case (Section 2.3.1), but it also holds in the semiparametric case (Section 2.3.2).

Let $\Pi^{(-i)}$ be the penalized log-likelihood when y_i is deleted

$$\Pi^{(-i)}(f) = \sum_{j \neq i} m_j \{y_j\theta_j - b(\theta_j)\} - \frac{\lambda}{2} J(f),$$

where θ_j is linked to $f(t_j)$, and f is assumed to be $f(t) = \sum_{k=1}^{q} \xi_k \varphi_k(t)$. The one-step approximation to $\hat{\boldsymbol{\xi}}^{(-i)}$, the vector of estimates of $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_q)^T$ when y_i is deleted, is written as

$$\hat{\boldsymbol{\xi}}^{(-i)} \approx \hat{\boldsymbol{\xi}} - \left(\frac{\partial^2 \Pi^{(-i)}}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}^{\mathrm{T}}} \bigg|_{\hat{\boldsymbol{\xi}}} \right)^{-1} \frac{\partial \Pi^{(-i)}}{\partial \boldsymbol{\xi}} \bigg|_{\hat{\boldsymbol{\xi}}}$$

in the same way as in Section 2.2.4. If G is the canonical link function, the Newton–Raphson algorithm is equivalent to the Fisher scoring algorithm, and

hence

$$\begin{aligned}
\hat{\boldsymbol{\xi}}^{(-i)} &\approx \hat{\boldsymbol{\xi}} + (B_{(i)}^T W_{(i)} B_{(i)} + \lambda K)^{-1} \{B_{(i)}^T W_{(i)} (\boldsymbol{z}_{(i)} - B_{(i)} \boldsymbol{\xi}) - \lambda K \boldsymbol{\xi}\} \\
&= (B_{(i)}^T W_{(i)} B_{(i)} + \lambda K)^{-1} B_{(i)}^T W_{(i)} \boldsymbol{z}_{(i)} \\
&= (B^T W B - w_i \boldsymbol{b}_i \boldsymbol{b}_i^T + \lambda K)^{-1} (B^T W \boldsymbol{z} - w_i z_i \boldsymbol{b}_i) \\
&= \Bigg\{ (B^T W B + \lambda K)^{-1} \\
&\quad + \frac{w_i (B^T W B + \lambda K)^{-1} \boldsymbol{b}_i \boldsymbol{b}_i^T (B^T W B + \lambda K)^{-1}}{1 - w_i \boldsymbol{b}_i^T (B^T W B + \lambda K)^{-1} \boldsymbol{b}_i} \Bigg\} (B^T W \boldsymbol{z} - w_i z_i \boldsymbol{b}_i),
\end{aligned}$$

where the subscript (i) means that the components related to the ith observation are removed, and $\boldsymbol{b}_i^T$ is the ith row of B.

Noticing that $A_{ii} = w_i \boldsymbol{b}_i^T (B^T W B + \lambda K)^{-1} \boldsymbol{b}_i$ and $\hat{\boldsymbol{\xi}} = (B^T W B + \lambda K)^{-1} B^T W \boldsymbol{z}$, we derive

$$\begin{aligned}
\hat{\boldsymbol{\xi}}^{(-i)} &\approx (B^T W B + \lambda K)^{-1} B^T W \boldsymbol{z} - w_i z_i (B^T W B + \lambda K)^{-1} \boldsymbol{b}_i \\
&\quad + \frac{w_i}{1 - A_{ii}} (B^T W B + \lambda K)^{-1} \boldsymbol{b}_i \boldsymbol{b}_i^T (B^T W B + \lambda K)^{-1} B^T W \boldsymbol{z} \\
&\quad - \frac{A_{ii} w_i z_i}{1 - A_{ii}} (B^T W B + \lambda K)^{-1} \boldsymbol{b}_i \\
&= \hat{\boldsymbol{\xi}} - \frac{w_i (z_i - \boldsymbol{b}_i^T \hat{\boldsymbol{\xi}})}{1 - A_{ii}} (B^T W B + \lambda K)^{-1} \boldsymbol{b}_i.
\end{aligned}$$

Since $\hat{\eta}_i = \boldsymbol{b}_i^T \hat{\boldsymbol{\xi}}$, the one-step approximation to $\hat{\eta}_i^{(-i)}$ becomes

$$\hat{\eta}_i^{(-i)} \approx \hat{\eta}_i - \frac{A_{ii}}{1 - A_{ii}} (z_i - \hat{\eta}_i),$$

which coincides with (3.5).

Similar ideas were proposed by O'Sullivan (1988) in the context of density estimation, and le Cessie and van Houwelingen (1992) on choosing a ridge parameter in logistic ridge regression.

3.2.4 An AIC-like Form of the LCV Score

We derive a still simpler form of the LCV score. We rewrite (3.4) as

$$\begin{aligned}
\mathrm{LCV}(\lambda) &= \frac{2}{n} \sum_{i=1}^{n} m_i [\{y_i \tilde{\theta}_i - b(\tilde{\theta}_i)\} - \{y_i \hat{\theta}_i - b(\hat{\theta}_i)\}] \\
&\quad + \frac{2}{n} \sum_{i=1}^{n} m_i [y_i (\hat{\theta}_i - \hat{\theta}_i^{(-i)}) - \{b(\hat{\theta}_i) - b(\hat{\theta}_i^{(-i)})\}]. \qquad (3.7)
\end{aligned}$$

If we use the first-order approximation

$$\begin{aligned}
b(\hat{\theta}_i) - b(\hat{\theta}_i^{(-i)}) &\approx b'(\hat{\theta}_i)(\hat{\theta}_i - \hat{\theta}_i^{(-i)}) \\
&= \hat{\mu}_i (\hat{\theta}_i - \hat{\theta}_i^{(-i)}),
\end{aligned}$$

then the second term of the right-hand side of (3.7), say D_2, becomes

$$D_2 \approx \frac{2}{n}\sum_{i=1}^{n} m_i(y_i - \hat{\mu}_i)(\hat{\theta}_i - \hat{\theta}_i^{(-i)}).$$

Moreover, if G is the canonical link function, by substituting (3.6), we have

$$D_2 \approx \frac{2}{n}\sum_{i=1}^{n} \frac{A_{ii}}{1 - A_{ii}} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)/m_i}.$$

From the fact that $\mathrm{var}(y_i) = V(\mu_i)\phi/m_i$, a rather rough approximation

$$\frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)/m_i} \approx \phi$$

leads to an AIC-like form of the LCV score

$$\mathrm{LCV}_2(\lambda) = \frac{2}{n}\sum_{i=1}^{n} m_i[\{y_i\tilde{\theta}_i - b(\tilde{\theta}_i)\} - \{y_i\hat{\theta}_i - b(\hat{\theta}_i)\}] + \frac{2}{n}\phi\sum_{i=1}^{n}\frac{A_{ii}}{1 - A_{ii}}.$$

The difference between $\mathrm{AIC}(\lambda)$ and $\mathrm{LCV}_2(\lambda)$ is in the second terms, in which $\mathrm{tr}A_\lambda = \sum_{i=1}^{n} A_{ii}$ for $\mathrm{AIC}(\lambda)$ is replaced with $\sum_{i=1}^{n} A_{ii}/(1 - A_{ii})$ for $\mathrm{LCV}_2(\lambda)$. In the case of binomial or Poisson distribution, ϕ can be set to be 1.

3.3 Comparison with Exact Calculation

In this section we show how the method for simple calculation of the delete-one estimate and the LCV score gives good approximations to the exact calculation method. The LCV scores proposed in the previous section are derived in the case of logistic regression and Poisson regression, and the goodness of approximation is examined by using some of the data introduced in Section 2.4. Diagnostics of influential observation with the delete-one estimate are also illustrated.

3.3.1 Logistic Regression Case

In the case of Logistic regression, the likelihood-based cross-validation score is written as

$$\mathrm{LCV}(\lambda) = \frac{2}{n}\sum_{i=1}^{n} m_i[\{y_i \log y_i + (1 - y_i)\log(1 - y_i)\} - \{y_i\hat{\theta}_i^{(-i)} - \log(1 + \exp\hat{\theta}_i^{(-i)})\}],$$

where we define that $y_i \log y_i = 0$ if $y_i = 0$ and $(1 - y_i)\log(1 - y_i) = 0$ if $y_i = 1$. Since $V(\mu) = \mu(1 - \mu)$, the delete-one estimate $\hat{\theta}_i^{(-i)}$ by simple calculation is

$$\hat{\theta}_i^{(-i)} \approx \hat{\theta}_i - \frac{A_{ii}}{1 - A_{ii}} \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i(1 - \hat{\mu}_i)}$$

from (3.6), and substituting it into the right-hand side of $\mathrm{LCV}(\lambda)$ yields the form of simple calculation. Especially in the case of binary response case, the LCV score by simple calculation becomes

$$\mathrm{LCV}_1(\lambda) = \frac{2}{n}\sum_{i=1}^{n}\left\{-y_i\left(\hat{\theta}_i - \frac{A_{ii}}{1-A_{ii}}\frac{y_i-\hat{\mu}_i}{\hat{\mu}_i(1-\hat{\mu}_i)}\right)\right.$$
$$\left. + \log\left(1+\exp\left(\hat{\theta}_i - \frac{A_{ii}}{1-A_{ii}}\frac{y_i-\hat{\mu}_i}{\hat{\mu}_i(1-\hat{\mu}_i)}\right)\right)\right\}.$$

The AIC-like form of the LCV score becomes

$$\mathrm{LCV}_2(\lambda) = \frac{2}{n}\sum_{i=1}^{n}\{-y_i\hat{\theta}_i + \log(1+\exp\hat{\theta}_i)\} + \frac{2}{n}\phi\sum_{i=1}^{n}\frac{A_{ii}}{1-A_{ii}},$$

where ϕ is usually set to be 1.

The likelihood-based cross-validation scores $\mathrm{LCV}_1(\lambda)$ and $\mathrm{LCV}_2(\lambda)$ for various values of λ are evaluated to the kyphosis data introduced in Section 2.4.1, and are compared with the exact values of $\mathrm{LCV}(\lambda)$ evaluated according to the definition of deleting each one observation. The cross-validation scores $\mathrm{GCV}(\lambda)$ and $\mathrm{OCV}(\lambda)$ based on the squared error criterion are also evaluated for reference. The S-PLUS function `smooth.spline` was used for computation as in Section 2.4.1.

Figure 3.1 plots these cross-validation scores as the function of $\log_{10}\lambda$ between -5 and 2. The scores $\mathrm{LCV}_1(\lambda)$ and $\mathrm{LCV}_2(\lambda)$ give good approximations to the exact score $\mathrm{LCV}(\lambda)$ when $\log_{10}\lambda > -2$. The value $\hat{\lambda} = 0.064$ that minimizes $\mathrm{LCV}_1(\lambda)$ and the value $\hat{\lambda} = 0.1$ that minimizes $\mathrm{LCV}_2(\lambda)$ are near to the value $\hat{\lambda} = 0.08$ that minimizes $\mathrm{LCV}(\lambda)$. Approximations of the two scores become worse when λ becomes small, but it seems to make little trouble because the scores become much greater as λ tends to zero. On the other hand, the plots of $\mathrm{GCV}(\lambda)$ and $\mathrm{OCV}(\lambda)$ show bad behavior. It can be seen that these scores based on the squared error criterion are inadequate for approximations to $\mathrm{LCV}(\lambda)$.

To verify the goodness of approximation by our simple calculation method, the deleted estimates $\hat{\theta}_i^{(-i)}$ and the deleted deviance increments

$$d_i^{(-i)} = \frac{2}{n}\{-y_i\hat{\theta}_i^{(-i)} + \log(1+\exp\hat{\theta}_i^{(-i)})\}$$

obtained by the simple calculation were compared with those obtained by the exact calculation. Table 3.1 lists a part of these values to the kyphosis data in the case of logistic linear regression, and Table 3.2 lists them in the case of nonparametric logistic regression with $\hat{\lambda}$ selected by minimizing $\mathrm{LCV}_1(\lambda)$. The difference between the values $\hat{\theta}_i^{(-i)}$ obtained by both the methods of calculation is small for almost all the observations, although the observations No. 11, 76 and 79 give slightly different values of $\hat{\theta}_i^{(-i)}$. The observations No. 11 and 79 have high deviance increments and the differences of $d_i^{(-i)}$ for these observations between the exact calculation and the simple one become relatively large.

The results as these scores were evaluated to other data are described in Section 4.1.1.

Table 3.1: The values of ordinary estimates $\hat{\theta}_i$, deleted estimates $\hat{\theta}_i^{(-i)}$, deleted deviance increments $d_i^{(-i)}$ and leverages A_{ii} when a logistic linear regression model was fitted to the kyphosis data.

Obs.	$\hat{\theta}_i$	$\hat{\theta}_i^{(-i)}$		$d_i^{(-i)}$		A_{ii}
		Exact	Simple	Exact	Simple	
1	−1.212	−1.138	−1.137	0.00686	0.00687	0.0545
2	−2.998	−2.964	−2.964	0.00124	0.00124	0.0314
3	−0.448	−0.658	−0.654	0.02654	0.02649	0.0746
4	−0.594	−0.414	−0.412	0.01253	0.01255	0.1053
5	−5.102	−5.093	−5.093	0.00015	0.00015	0.0094
6	−5.102	−5.093	−5.093	0.00015	0.00015	0.0094
7	−4.298	−4.285	−4.285	0.00034	0.00034	0.0126
8	−4.620	−4.608	−4.608	0.00024	0.00024	0.0112
9	−3.601	−3.582	−3.582	0.00068	0.00068	0.0182
10	0.170	0.051	0.051	0.01649	0.01649	0.0607
11	−2.473	−3.337	−3.146	0.08326	0.07872	0.0497
12	−3.132	−3.103	−3.102	0.00109	0.00109	0.0272
13	−0.380	−0.217	−0.216	0.01458	0.01460	0.0888
14	−2.150	−2.071	−2.070	0.00293	0.00294	0.0675
16	−2.864	−2.824	−2.823	0.00142	0.00142	0.0366
17	−5.102	−5.093	−5.093	0.00015	0.00015	0.0094
18	−2.527	−2.473	−2.472	0.00200	0.00200	0.0482
19	−1.227	−1.006	−0.998	0.00770	0.00775	0.1503
20	−2.500	−2.445	−2.445	0.00205	0.00205	0.0489
⋮	⋮	⋮		⋮		⋮
71	−1.922	−1.845	−1.843	0.00362	0.00363	0.0645
72	−3.372	−3.335	−3.335	0.00086	0.00086	0.0347
73	−1.455	−1.296	−1.290	0.00597	0.00600	0.1176
74	−0.461	−0.332	−0.331	0.01335	0.01336	0.0737
75	−3.949	−3.934	−3.934	0.00048	0.00048	0.0148
76	0.598	1.260	1.247	0.03728	0.03702	0.1872
77	−4.968	−4.958	−4.958	0.00017	0.00017	0.0098
78	−2.730	−2.682	−2.682	0.00163	0.00163	0.0431
79	−3.011	−3.901	−3.691	0.09682	0.09173	0.0309
80	−3.224	−3.185	−3.185	0.00100	0.00100	0.0363
81	−3.507	−3.486	−3.486	0.00074	0.00074	0.0195
82	−0.058	−0.212	−0.211	0.01986	0.01985	0.0691
83	−4.633	−4.622	−4.622	0.00024	0.00024	0.0111

Table 3.2: The values of ordinary estimates $\hat{\theta}_i$, deleted estimates $\hat{\theta}_i^{(-i)}$, deleted deviance increments $d_i^{(-i)}$ and leverages A_{ii} when a nonparametric logistic regression model was fitted to the kyphosis data, with $\hat{\lambda}$ selected by minimizing $\mathrm{LCV}_1(\lambda)$.

Obs.	$\hat{\theta}_i$	$\hat{\theta}_i^{(-i)}$		$d_i^{(-i)}$		A_{ii}
		Exact	Simple	Exact	Simple	
1	−0.773	−0.618	−0.617	0.01064	0.01066	0.0965
2	−3.335	−3.303	−3.305	0.00089	0.00089	0.0282
3	−0.219	−0.429	−0.428	0.02298	0.02296	0.0853
4	−1.513	−1.307	−1.302	0.00591	0.00594	0.1477
5	−5.874	−5.864	−5.868	0.00007	0.00007	0.0058
6	−5.874	−5.864	−5.868	0.00007	0.00007	0.0058
7	−3.880	−3.859	−3.858	0.00052	0.00052	0.0210
8	−4.604	−4.590	−4.591	0.00025	0.00025	0.0123
9	−3.048	−3.014	−3.012	0.00118	0.00119	0.0332
10	0.390	0.244	0.242	0.01429	0.01430	0.0810
11	−1.913	−2.812	−2.645	0.07088	0.06699	0.0861
12	−3.224	−3.193	−3.194	0.00099	0.00099	0.0279
13	−0.937	−0.762	−0.755	0.00946	0.00951	0.1158
14	−3.021	−2.953	−2.957	0.00126	0.00125	0.0580
16	−3.470	−3.436	−3.439	0.00078	0.00078	0.0293
17	−5.874	−5.864	−5.868	0.00007	0.00007	0.0058
18	−1.995	−1.898	−1.894	0.00345	0.00346	0.0815
19	−2.104	−1.939	−1.938	0.00332	0.00332	0.1286
20	−1.953	−1.852	−1.848	0.00360	0.00361	0.0838
⋮	⋮	⋮		⋮		⋮
71	−2.409	−2.336	−2.337	0.00228	0.00228	0.0620
72	−3.897	−3.868	−3.871	0.00051	0.00051	0.0244
73	−1.863	−1.720	−1.718	0.00407	0.00408	0.1119
74	−0.214	−0.047	−0.047	0.01655	0.01654	0.0849
75	−3.293	−3.263	−3.260	0.00093	0.00093	0.0310
76	−1.248	−0.416	−0.263	0.01251	0.01408	0.4334
77	−5.512	−5.502	−5.505	0.00010	0.00010	0.0069
78	−3.623	−3.585	−3.590	0.00068	0.00067	0.0315
79	−3.323	−4.467	−4.153	0.11057	0.10294	0.0281
80	−3.508	−3.474	−3.476	0.00075	0.00075	0.0297
81	−3.046	−3.013	−3.011	0.00118	0.00119	0.0321
82	−0.116	−0.330	−0.326	0.02152	0.02146	0.0898
83	−4.637	−4.624	−4.625	0.00024	0.00024	0.0120

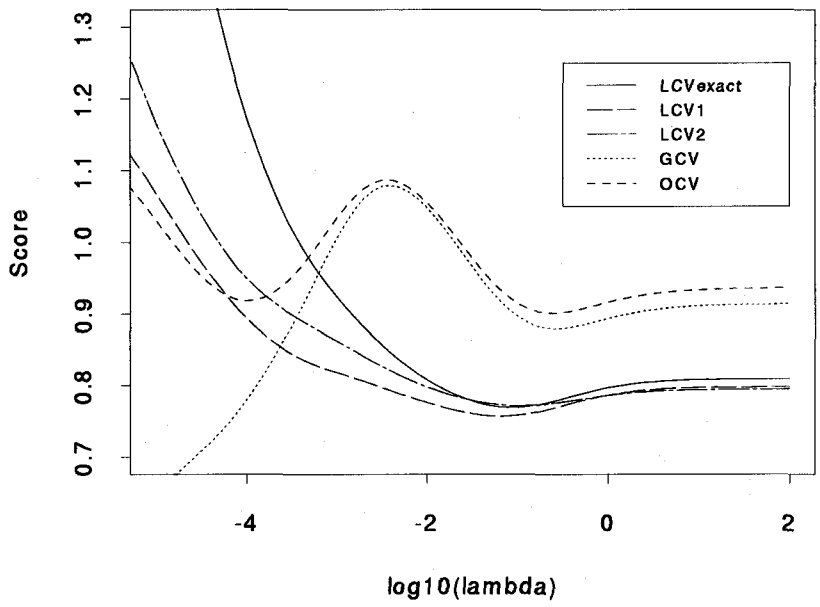


Figure 3.1: Plots of the cross-validation scores to the kyphosis data as the function of $\log_{10}\lambda$. The solid line plots the exact values of $\mathrm{LCV}(\lambda)$.

Diagnosing Influential Observations

The simple calculation of delete-one estimates $\hat{\theta}_i^{(-i)}$ makes it easier to detect influential observations. Figure 3.2 plots the values of $|\hat{\theta}_i - \hat{\theta}_i^{(-i)}|$, the leverage value A_{ii} and the deviance increments d_i for each observation in the nonparametric logistic regression to the kyphosis data. A large value of $|\hat{\theta}_i - \hat{\theta}_i^{(-i)}|$ implies that deleting the ith observation gives strong influence. The observation No. 76 has a high leverage value, and the observations No. 11 and 79 have high deviance increments. Inspection of $|\hat{\theta}_i - \hat{\theta}_i^{(-i)}|$ enables us to identify these three influential observations together.

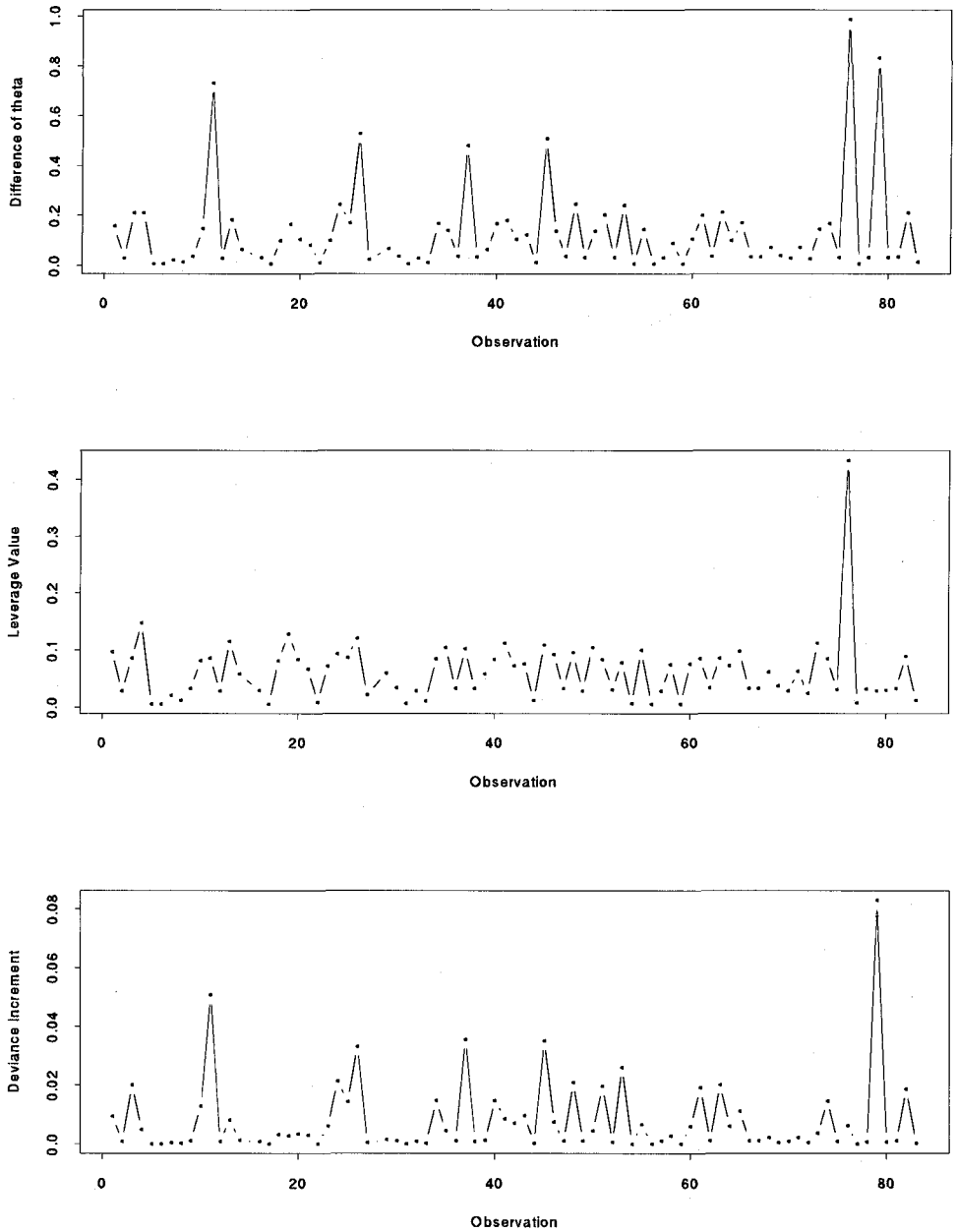


Figure 3.2: The top figure plots $|\hat{\theta}_i - \hat{\theta}_i^{(-i)}|$ versus i, the middle figure plots A_{ii} versus i, and the bottom figure plots d_i versus i in the nonparametric logistic regression to the kyphosis data.

3.3.2 Poisson Regression Case

In the case of Poisson regression and density smoothing for classified data, the likelihood-based cross-validation score is written as

$$\mathrm{LCV}(\lambda) = \frac{2}{n}\sum_{i=1}^{n}\{y_i(\log y_i - 1) - (y_i\hat{\theta}_i^{(-i)} - \exp\hat{\theta}_i^{(-i)})\},$$

where we define that $y_i(\log y_i - 1) = 0$ if $y_i = 0$. Since $V(\mu) = \mu$, the delete-one estimate $\hat{\theta}_i^{(-i)}$ by simple calculation is

$$\hat{\theta}_i^{(-i)} \approx \hat{\theta}_i - \frac{A_{ii}}{1 - A_{ii}}\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}$$

from (3.6). Substituting it into the right-hand side of $\mathrm{LCV}(\lambda)$ yields

$$\mathrm{LCV}_1(\lambda) = \frac{2}{n}\sum_{i=1}^{n}\left[y_i(\log y_i - 1) - \left\{y_i\left(\hat{\theta}_i - \frac{A_{ii}}{1 - A_{ii}}\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right) - \exp\left(\hat{\theta}_i - \frac{A_{ii}}{1 - A_{ii}}\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right)\right\}\right].$$

The AIC-like form of the LCV score becomes

$$\mathrm{LCV}_2(\lambda) = \frac{2}{n}\sum_{i=1}^{n}\{y_i(\log y_i - 1) - (y_i\hat{\theta}_i - \exp\hat{\theta}_i)\} + \frac{2}{n}\phi\sum_{i=1}^{n}\frac{A_{ii}}{1 - A_{ii}},$$

where ϕ is usually set to be 1.

The scores $\mathrm{LCV}_1(\lambda)$ and $\mathrm{LCV}_2(\lambda)$ for various values of λ are evaluated in the situation of density smoothing. As in logistic regression case, the exact values of $\mathrm{LCV}(\lambda)$ and two scores $\mathrm{GCV}(\lambda)$ and $\mathrm{OCV}(\lambda)$ are also evaluated. Figure 3.3 plots these scores when the density smoothing in Section 2.4.3 is applied to the Old Faithful geyser data. The class marks are rescaled in [0,1] and a cubic smoothing spline with knots at the class marks was fitted. The value $\hat{\lambda} = 0.00019$ that minimizes $\mathrm{LCV}_1(\lambda)$ is almost equal to the one that minimizes the exact score $\mathrm{LCV}(\lambda)$, while the value $\lambda = 0.0005$ that minimizes $\mathrm{LCV}_2(\lambda)$ is slightly larger. Moreover $\mathrm{LCV}_2(\lambda)$ becomes greater when λ is small. The behavior of $\mathrm{GCV}(\lambda)$ and $\mathrm{OCV}(\lambda)$ is not so wrong as in the logistic regression case, but our LCV scores approximate the exact score better.

Table 3.3 compares the deleted estimates $\hat{\theta}_i^{(-i)}$ and the deleted deviance increments

$$d_i^{(-i)} = \frac{2}{n}\{y_i(\log y_i - 1) - (y_i\hat{\theta}_i^{(-i)} - \exp\hat{\theta}_i^{(-i)})\}$$

obtained by the simple calculation with those obtained by the exact one, where the value of $\hat{\lambda}$ that minimizes $\mathrm{LCV}_1(\lambda)$ was used. Since there exists no class that has extreme influence, the difference between the values obtained by the two calculation methods is very small for every class.

The results as these scores were evaluated to other data are described in Section 4.1.2.

Table 3.3: The values of ordinary estimates $\hat{\theta}_i$, deleted estimates $\hat{\theta}_i^{(-i)}$, deleted deviance increments $d_i^{(-i)}$ and leverages A_{ii} when density smoothing was applied to the Old Faithful geyser data, with $\hat{\lambda}$ selected by minimizing $\mathrm{LCV}_1(\lambda)$.

Class mark	$\hat{\theta}_i$	$\hat{\theta}_i^{(-i)}$		$d_i^{(-i)}$		A_{ii}
		Exact	Simple	Exact	Simple	
1.55	−0.101	0.779	0.796	0.12453	0.12661	0.473
1.65	0.933	0.837	0.812	0.00540	0.00639	0.401
1.75	1.689	1.739	1.743	0.00251	0.00267	0.417
1.85	1.960	1.716	1.746	0.05103	0.04523	0.444
1.95	1.692	1.635	1.624	0.00402	0.00456	0.395
2.05	1.089	1.275	1.242	0.02369	0.02084	0.318
2.15	0.470	0.805	0.821	0.12777	0.12993	0.260
2.25	0.008	−0.303	−0.287	0.04176	0.04059	0.231
2.35	−0.316	−0.422	−0.417	0.00444	0.00435	0.214
2.45	−0.531	−0.296	−0.276	0.04252	0.04337	0.204
2.55	−0.635	−0.863	−0.857	0.01627	0.01608	0.200
2.65	−0.640	−0.403	−0.384	0.03817	0.03892	0.204
2.75	−0.568	−0.780	−0.776	0.01363	0.01349	0.213
2.85	−0.463	−0.206	−0.175	0.04653	0.04795	0.223
2.95	−0.362	−1.010	−0.917	0.10113	0.09256	0.229
3.05	−0.277	−0.016	0.015	0.05625	0.05803	0.226
3.15	−0.143	−0.185	−0.188	0.00092	0.00095	0.227
3.25	0.098	0.135	0.127	0.00054	0.00048	0.238
3.35	0.466	0.610	0.601	0.01317	0.01273	0.266
3.45	0.911	1.000	0.998	0.00597	0.00587	0.307
3.55	1.301	0.948	0.969	0.09386	0.08974	0.344
3.65	1.531	1.795	1.845	0.10373	0.11557	0.356
3.75	1.625	1.172	1.203	0.19756	0.18730	0.353
3.85	1.658	1.961	1.979	0.14680	0.15233	0.341
3.95	1.789	1.973	1.974	0.04824	0.04849	0.358
4.05	1.933	1.518	1.559	0.18524	0.17052	0.388
4.15	1.875	1.905	1.923	0.00229	0.00308	0.378
4.25	1.676	1.724	1.710	0.00193	0.00150	0.343
4.35	1.547	1.638	1.618	0.00788	0.00665	0.325
4.45	1.597	1.795	1.800	0.05324	0.05395	0.341
4.55	1.705	1.515	1.538	0.03232	0.02922	0.381
4.65	1.589	1.282	1.310	0.07156	0.06606	0.394
4.75	1.134	1.325	1.324	0.02853	0.02843	0.348
4.85	0.483	0.684	0.690	0.01699	0.01737	0.351
4.95	−0.218	−0.439	−0.471	0.00479	0.00545	0.508

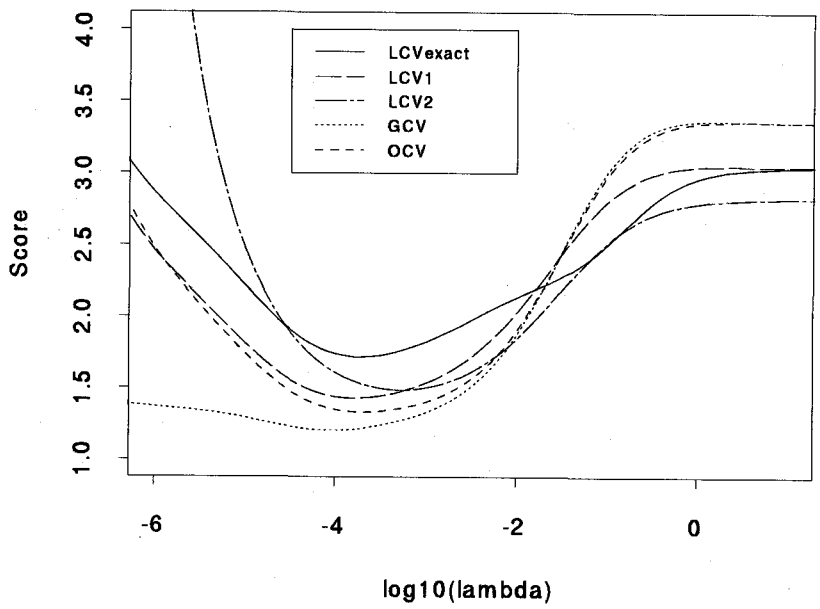


Figure 3.3: Plots of the cross-validation scores to the Old Faithful geyser data as the function of $\log_{10}\lambda$. The solid line plots the exact values of LCV(λ).

3.4 Other Procedures

3.4.1 Gu's Algorithm

Gu (1990) proposed an algorithm to obtain a stable value of $\hat{\lambda}$. His proposal was to select $\hat{\lambda}$ within each cycle of the Fisher scoring algorithm and to estimate a function f for the $\hat{\lambda}$ after stabilized. In Gu (1990) the GCV score was used to select $\hat{\lambda}$, while Gu (1992) suggested that the unbiased risk estimate

$$\mathrm{UBR}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} w_i(z_i - \hat{\eta}_i)^2 + \frac{2}{n}\phi \mathrm{tr}A_\lambda$$

as ϕ is set to be 1 should have better performance.

It is sure that Gu's (1992) method yields a stable $\hat{\lambda}$ and $\hat{f}$, but we think that the method takes much time because searching $\hat{\lambda}$ is replicated, and is inappropriate for large scale algorithms such as the local scoring in generalized additive models. Moreover Gu's algorithm changes the optimization problem for each iteration and so convergence is not guaranteed. To invert the order of iteration to update an estimate and minimization of the score was also discussed by Green and Silverman (1994), where the OCV score was used, but they say that it is more attractive to select $\hat{\lambda}$ on the final iteration.

3.4.2 Generalized Approximate Cross-validation

Xiang and Wahba (1996) also proposed likelihood-based cross-validation scores, but their scores are different from $\mathrm{LCV}_1(\lambda)$ and $\mathrm{LCV}_2(\lambda)$ that we proposed.

At first they defined the delete-one cross-validation score

$$\mathrm{LCV}_{\mathrm{XW}}(\lambda) = \frac{2}{n}\sum_{i=1}^{n}\{-y_i\hat{\theta}_i^{(-i)} + b(\hat{\theta}_i)\},$$

which is multiplied by 2 for comparison. The difference from $\mathrm{LCV}(\lambda)$ defined as (3.4) is, up to constant terms, in the term $b(\hat{\theta}_i)$, which is replaced with $b(\hat{\theta}_i^{(-i)})$ in $\mathrm{LCV}(\lambda)$. Then they derived an approximate cross-validation score

$$\mathrm{ACV}(\lambda) = \frac{2}{n}\sum_{i=1}^{n}\{-y_i\hat{\theta}_i + b(\hat{\theta}_i)\} + \frac{2}{n}\sum_{i=1}^{n}\frac{h_{ii}y_i(y_i - \hat{\mu}_i)}{1 - h_{ii}b''(\hat{\theta}_i)}$$

from $\mathrm{LCV}_{\mathrm{XW}}(\lambda)$, where h_{ii} is the ith diagonal component of the Hessian matrix H of the penalized log-likelihood. In our notation $H = B(B^TWB + \lambda K)^{-1}B^T$ and hence $A_\lambda = HW$, $h_{ii}b''(\hat{\theta}_i) = A_{ii}$ and $h_{ii} = A_{ii}/w_i = A_{ii}/V(\hat{\mu}_i)$. Moreover they derived a generalized approximate cross-validation score

$$\mathrm{GACV}(\lambda) = \frac{2}{n}\sum_{i=1}^{n}\{-y_i\hat{\theta}_i + b(\hat{\theta}_i)\} + \frac{2\mathrm{tr}H}{n}\frac{\sum_{i=1}^{n}y_i(y_i - \hat{\mu}_i)}{n - \mathrm{tr}(W^{1/2}HW^{1/2})}$$

by replacing h_{ii} and $h_{ii}b'(\hat{\theta}_i)$ in $\mathrm{ACV}(\lambda)$ with their averages $n^{-1}\mathrm{tr}H$ and $n^{-1}\mathrm{tr}(W^{1/2}HW^{1/2})$, respectively.

We think that there exists no evidence for the form of $\mathrm{LCV}_{\mathrm{XW}}(\lambda)$ even if it may be an approximation of $\mathrm{LCV}(\lambda)$. What should the form of the score become in the case such as binomial or Poisson distribution? In addition the way of taking the trace in $\mathrm{GACV}(\lambda)$ seems very strange. We cannot understand the reason of "generalization" in the non-normal distribution case, because the weight for each observation in the Fisher scoring algorithm is not homogeneous.

Chapter 4

Comparison of the Scores to Select the Smoothing Parameter

In section 3.2 we proposed the simple calculation method and the AIC-like version of the likelihood-based cross-validation (LCV) score. Goodness of the approximation of the simple calculation to the exact calculation was confirmed in Section 3.3. In this chapter the performance of the simple calculation (LCV_1) and the AIC-like version (LCV_2) is compared with that of other standard scores (GCV, OCV and AIC) to select the smoothing parameter.

4.1 Comparison by Data in Literature

In this section various data sets in literature are examined and the comparison of five scores is attempted. We describe the examination in each case of logistic regression, Poisson regression and density smoothing for classified data. Through the section FORTRAN programs are implemented for computation and cubic smoothing splines are fitted.

4.1.1 Binary Logistic Regression

Nonparametric logistic regression models

$$\mathrm{P}(y_i = 1) = p_i \quad \text{and} \quad \log \frac{p_i}{1 - p_i} = f(t_i), \quad i = 1, \ldots, n,$$

are applied to several data sets, where each zero–one response y_i is observed with a one-dimensional explanatory variable t_i. The data sets applied to are listed in Table 4.1. The data set No. 3 has been also taken up in Sections 2.4.1 and 3.3.1. Each of the data is assumed to have only one explanatory variable for simplicity. The values of the smoothing parameters $\hat{\lambda}$ are chosen by minimizing each of the five scores $\mathrm{GCV}(\lambda)$, $\mathrm{OCV}(\lambda)$, $\mathrm{AIC}(\lambda)$, $\mathrm{LCV}_1(\lambda)$ and $\mathrm{LCV}_2(\lambda)$. The sets of explanatory variables are rescaled in [0,1] to unify the range to search the value of $\log_{10} \hat{\lambda}$ from -7 to 2 at intervals of 0.1.

Table 4.1: The examined data sets: binary logistic regression.

No.	Response	Explanatory var.	n	
1	Tumor prevalence	Age at death	207	Male rats
2			112	Female rats
3	Kyphosis	Age	83	
4	Toxicity	Velban dose	55	
5	Nodal involvement	Age	53	
6		ACP	53	
7	Defects	Purity index	22	Standard process
8			22	Modified process

The column labeled 'n' lists the sample size.

References: 1, 2: Green and Yandell (1986)
3: Hastie and Tibshirani (1990)
4: Brown and Hu (1980)
5, 6: Brown (1980)
7, 8: Cox and Snell (1981)

Table 4.2 shows the values of $\log_{10}\hat{\lambda}$ chosen by minimizing each of the five scores with the values of equivalent degrees of freedom (EDF). When extremely small values of $\hat{\lambda}$ are chosen, other values of $\log_{10}\lambda$ that give local minimums are also listed with EDF in italics. Notice that EDF is at least 2 and that EDF becomes large as λ becomes small. The value of EDF near to 2 implies that a nearly linear function is fitted and the extremely large values of EDF implies that an almost interpolating function is fitted. Figures 4.1–4.4 plot the five scores to the eight data sets as the functions of EDF less than 10. The exact score $\mathrm{LCV}(\lambda)$ are also plotted for reference.

For the data set No. 1 all the five scores become increasing functions of EDF where EDF is greater than 3 and are minimized at the values of EDF less than 3, which suggests that a linear fitted function is appropriate. On the other hand, for the data set No. 2 the four scores except $\mathrm{LCV}_1(\lambda)$ become nearly decreasing functions of EDF where EDF is less than 10 and small values of $\hat{\lambda}$ are chosen. Only the LCV_1 score takes the minimum at the values of EDF between 2 and 3, although the approximation of $\mathrm{LCV}_1(\lambda)$ to $\mathrm{LCV}(\lambda)$ is not so good. We think that the observation No. 106 gives strong influence since it has the extreme value of age.

For the data set No. 3 $\mathrm{LCV}_1(\lambda)$ and $\mathrm{LCV}_2(\lambda)$ take the minimums at the values of EDF between 3 and 4, which suggests apparent nonlinearity, and $\mathrm{OCV}(\lambda)$ takes the minimum at the value of EDF between 2 and 3. The scores $\mathrm{GCV}(\lambda)$ and $\mathrm{AIC}(\lambda)$ also have local minimums at the values of EDF less than 4, but these scores become smaller as λ tends to zero. For the data set No. 4 only the LCV_1 score provides appropriate fitting with the EDF 3.22. The other scores also have local minimums but become smaller as λ tends to zero.

The data sets No. 5 and 6 have common responses but are composed of

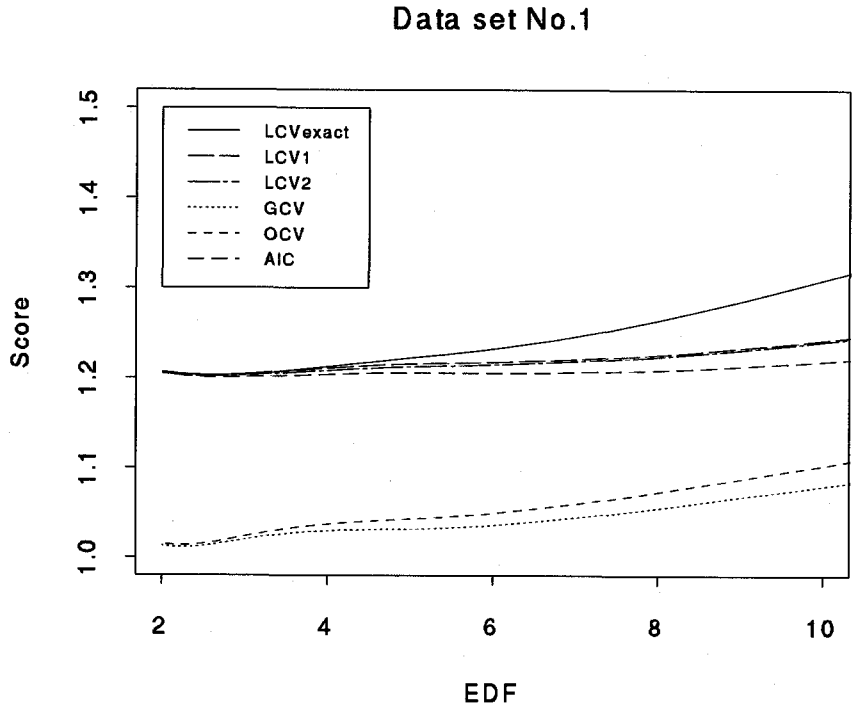
Table 4.2: The values of $\log_{10}\hat{\lambda}$ and EDF (in parentheses) to the data sets in Table 4.1. The values that give a local minimum are in italics.

No.	GCV	OCV	AIC	LCV_1	LCV_2
1	−0.3 (2.26)	−0.2 (2.21)	−0.9 (2.73)	−0.8 (2.63)	−0.8 (2.63)
2	≤ -7 *−2.2 (3.80)*	−4.8 (10.16)	≤ -7	−1.1 (2.60)	−4.7 (9.76)
3	≤ -7 *−1.5 (2.84)*	−1.5 (2.84)	≤ -7 *−2.1 (3.46)*	−1.8 (3.12)	−2.1 (3.46)
4	≤ -7 *−3.9 (5.56)*	−6.3 (11.80) *−3.2 (4.28)*	≤ -7 *−3.2 (4.28)*	−2.4 (3.22)	−6.4 (11.90) *−3.0 (3.99)*
5	−0.4 (2.16)	−0.5 (2.20)	−2.8 (4.82)	−2.8 (4.82)	−0.2 (2.11)
6	−1.3 (2.54)	−1.2 (2.47)	−2.0 (3.17)	−1.8 (2.98)	−1.2 (2.47)
7	−0.1 (2.03)	−2.5 (3.29)	≥ 2 (2.00)	−0.9 (2.17)	≥ 2 (2.00)
8	−1.2 (2.38)	−1.1 (2.32)	≤ -7 *−1.8 (2.82)*	−1.3 (2.44)	−1.5 (2.58)

different explanatory variables. In both cases there exist the minimums of the five scores. For the data set No. 5 the AIC and LCV_1 scores select almost the same values of $\hat{\lambda}$ but the minimum of the scores are hardly different from the scores when EDF=2, and the other scores seem to increase monotonously with EDF. The ground of nonlinearity on age is thought to be little. On the other hand, for the data No. 6 the minimum of each score is about 0.1 lower than the score when EDF=2, and hence weak nonlinearity on ACP can be admitted, although the curves of the scores are very different.

For the data sets No. 7 the algorithm does not converge when $\log_{10}\lambda < -5.4$ (EDF > 6.7). The scores of GCV(λ) and AIC(λ) trace strange curves while the curves of the other scores go up as EDF becomes large. For the data set No. 8 AIC(λ) becomes small as λ tends to zero while the other scores take the minimums at the values of EDF less than 3.

Overall, the GCV score has local minimums at appropriate values of λ and EDF but often becomes smaller as λ tends to zero. The OCV score shows similar behavior to the GCV score but the behavior as λ tends to zero is not so bad as the GCV score. The AIC score also has local minimums but traces the curve with smaller variation than the other scores. The LCV_1 score seems to show the most stable behavior. The score becomes large when λ tends to zero for many data sets, and when it takes the minimum all the values of EDF are between 2 and 5. Moreover the LCV_1 score approximates the exact LCV score well. The LCV_2 score is a crude approximation of the LCV score but the behavior of the LCV_2 curve seems to be next best to the LCV_1 score.



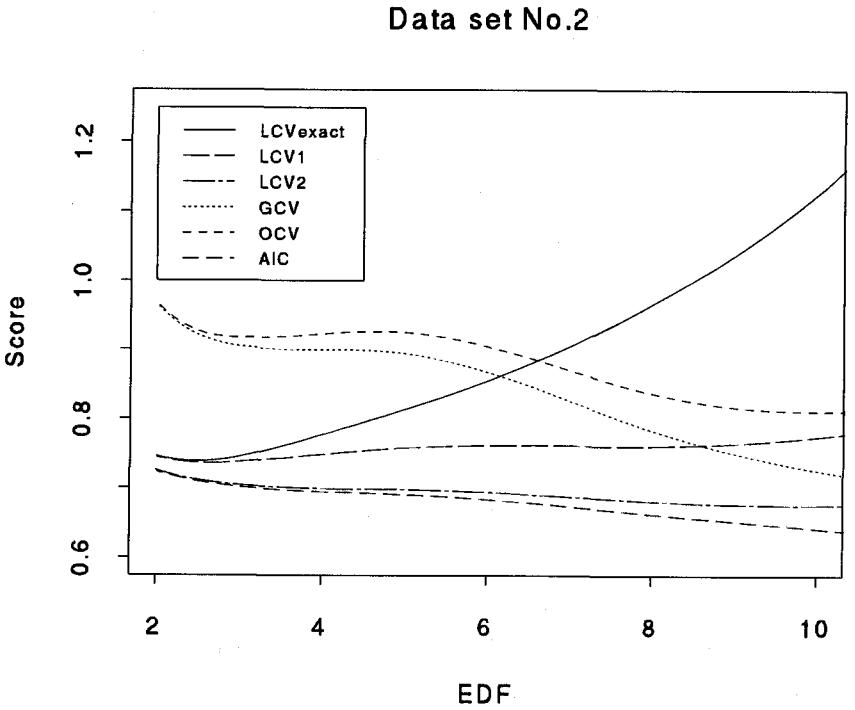


Figure 4.1: Plots of six scores to the data sets No. 1 and 2.

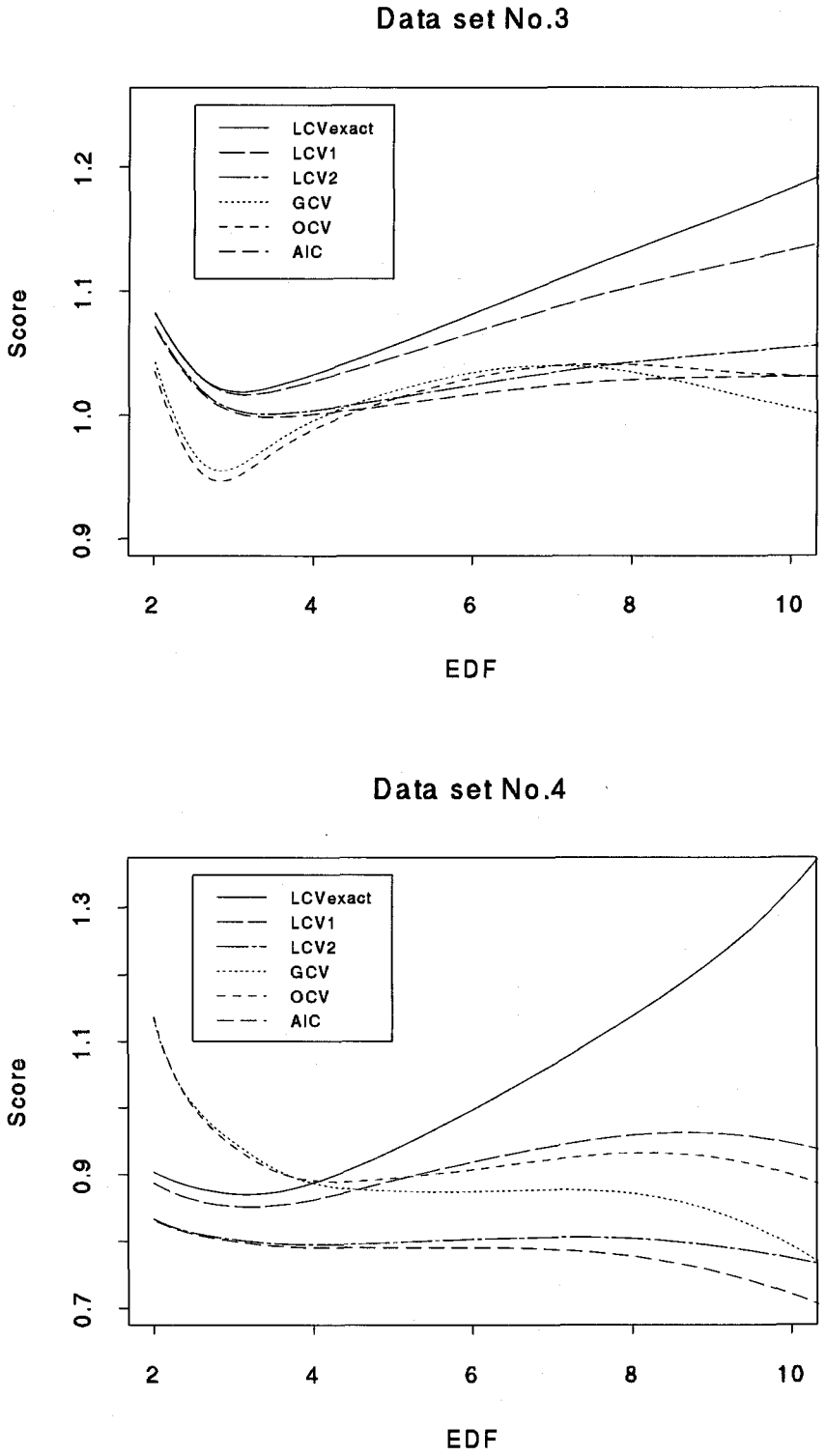


Figure 4.2: Plots of six scores to the data sets No. 3 and 4.

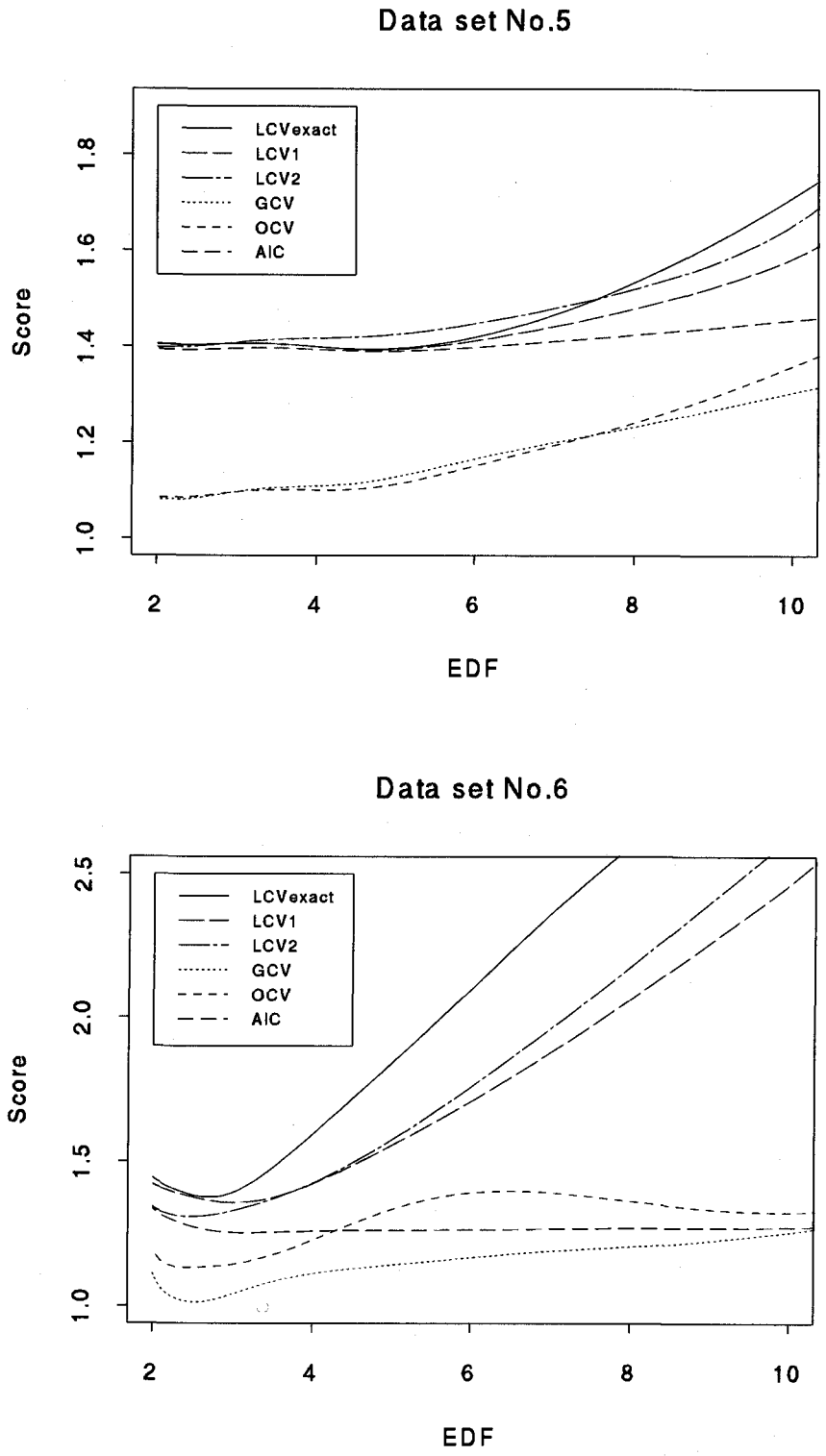
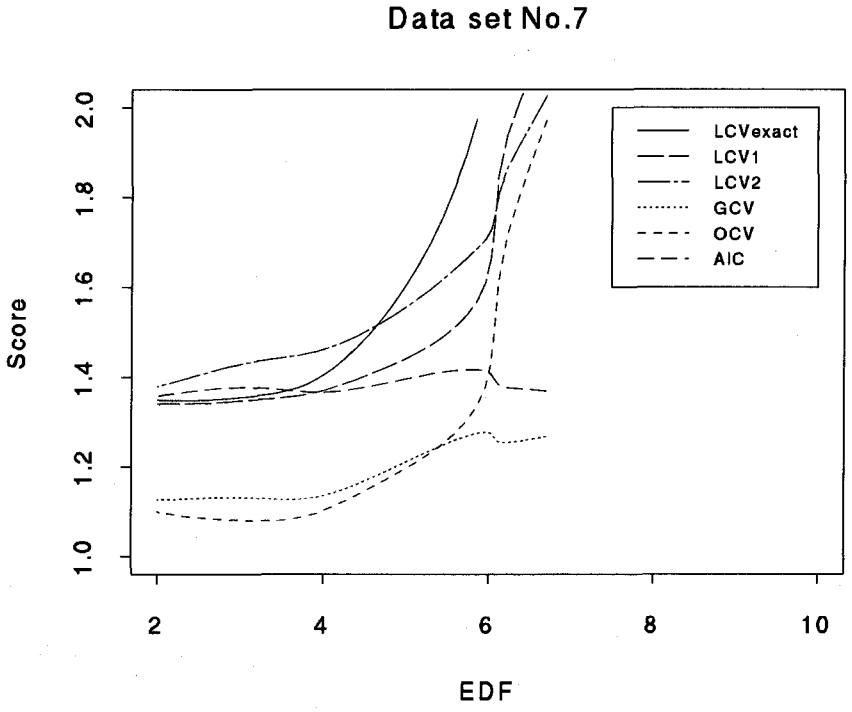


Figure 4.3: Plots of six scores to the data sets No. 5 and 6.



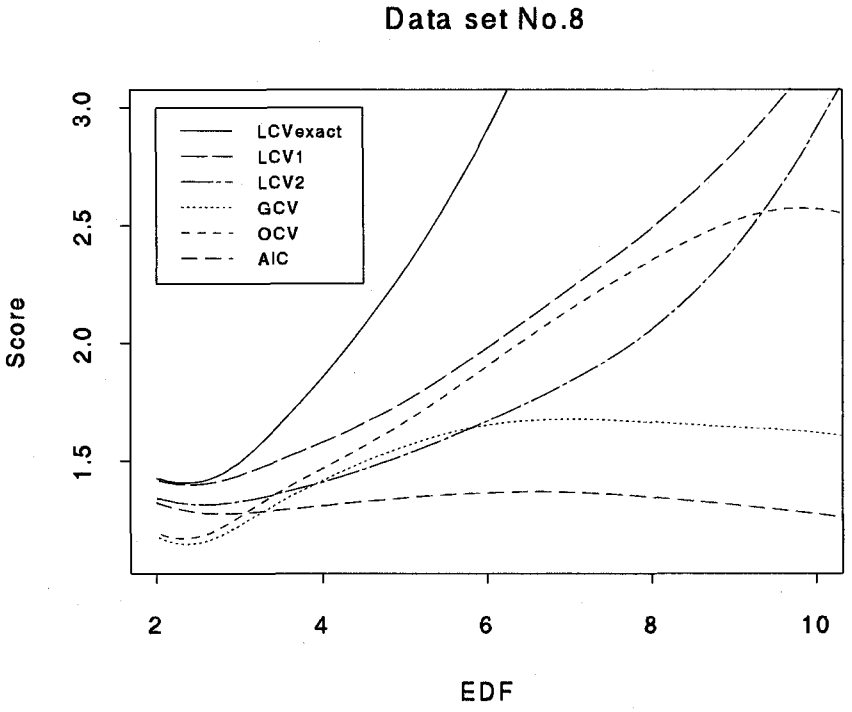


Figure 4.4: Plots of six scores to the data sets No. 7 and 8.

4.1.2 Binomial Logistic Regression

Nonparametric logistic regression models

$$y_i \sim \mathrm{B}(m_i, p_i) \quad \text{and} \quad \log \frac{p_i}{1-p_i} = f(t_i), \quad i = 1, \ldots, n,$$

are applied to the data sets listed in Table 4.3, where each response y_i is observed with a one-dimensional explanatory variable t_i. The data set No. 9 has a relatively larger number of age categories, while the data sets No. 10 and 11 have smaller numbers of age categories. The data sets No. 12–16 from dose response experiments are a part of the ones that were referred and analized by Kawai (1997) and have small numbers of dose levels. The data set No. 9 is the same as the one taken up in Section 2.4.2 and here we attempt fitting a logistic regression model. The data set No. 12 has been also taken up in Section 2.4.1.

The values of $\hat{\lambda}$ are chosen by minimizing each of the five scores GCV(λ), OCV(λ), AIC(λ), $\mathrm{LCV}_1(\lambda)$ and $\mathrm{LCV}_2(\lambda)$ in the same way as in the previous subsection. Table 4.4 shows the values of $\log_{10}\hat{\lambda}$ with the values of EDF. Figures 4.5–4.9 plot the five scores with the exact score LCV(λ) as the functions of EDF.

For the data set No. 9 quite a large value of EDF is chosen by the AIC score while the other scores select the values of EDF between 7 and 10. For the data set No. 10 all the scores take minimums when EDFs are near 3 and the minimum scores are about half of the scores when EDFs are 2, which suggests strong nonlinearity. On the other hand, for the data set No. 11 all the scores take the minimums when EDFs are 2 and suggest linear fitting. It is interesting that these two data sets separated only by presence or absence of breathlessness provide different shapes of fitted functions.

For the data sets No. 12–14 the results are similar. The score GCV(λ) finds local minimums but no global minimums. The scores OCV(λ) and $\mathrm{LCV}_1(\lambda)$ select appropriate values of EDF between about 4 and 5. The curves of AIC(λ) seem to have smaller variation than the other curves. The score $\mathrm{LCV}_2(\lambda)$ becomes extremely large when EDFs are large and hence relatively small EDFs are selected. For the data set No. 12 taken up in Section 2.4.1 the value of EDF about 5 can be suggested.

The data set No. 15 provides quite strange behavior of the scores as shown on the top of Figure 4.9. The data set contains an unusual number of responses at the dose level 0.11 and so almost interpolating fitted functions are chosen by all the scores except $\mathrm{LCV}_2(\lambda)$. The curves after the dose level 0.11 is removed are moderate as shown in the bottom of Figure 4.9. For the data set No. 16 all the scores except GCV(λ) provide linear fitted functions.

The overall results are similar to the binary logistic regression case. In addition, the OCV and LCV_1 scores take very close values and show similarly good behavior. The LCV_2 score is close to the AIC score when EDF is small but becomes quite larger when EDF is large, and hence it tends to give a small EDF especially when n is small.

Table 4.3: The examined data sets: binomial logistic regression.

No.	Response	Explan. var.	n	N	
9	Crude death rates	Age category	50	364,440	Mortality Table
10	Subjects respond-	Age category	9	15,855	Breathlessness: No
11	ing to wheeze		9	2,427	Breathlessness: Yes
12	Number of deaths	Log dose	8	426	
13			7	350	
14			6	292	
15			8	80	
16			9	54	

The column labeled 'n' lists the numbers of categories or dose levels, and the column labeled 'N' lists the total numbers of zero–one observations.

References:
- 9: Green and Silverman (1994)
- 10, 11: Cox and Snell (1981)
- 12: Ashford and Walker (1972)
- 13: Thompson (1947)
- 14: Finney (1971)
- 15: Bliss (1938)
- 16: Reed and Muench (1938)

Table 4.4: The values of $\log_{10} \hat{\lambda}$ and EDF (in parentheses) to the data sets in Table 4.3. The values that give a local minimum are in italics.

No.	GCV	OCV	AIC	LCV_1	LCV_2
9	−1.2 (7.88)	−1.2 (7.88)	−6.4 (46.78)	−1.3 (8.27)	−1.8 (10.48)
10	0.3 (3.30)	0.3 (3.30)	0.3 (3.30)	0.3 (3.30)	0.7 (2.88)
11	$\geq$ 2 (2.00)	$\geq$ 2 (2.00)	$\geq$ 2 (2.00)	$\geq$ 2 (2.00)	$\geq$ 2 (2.00)
12	≤ -7 *−2.4 (4.63)*	−2.7 (5.08)	−2.5 (4.78)	−2.8 (5.24)	−1.7 (3.74)
13	≤ -7 *−3.1 (4.10)*	−3.0 (4.01)	−4.9 (5.36)	−3.2 (4.19)	−1.9 (3.03)
14	≤ -7 *−1.1 (2.86)*	−2.4 (4.02)	−1.4 (3.10)	−2.3 (3.93)	−0.3 (2.32)
15	≤ -7 *$\geq$ 2 (2.00)*	−6.4 (7.49) *−0.5 (2.16)*	−3.2 (4.69)	−6.7 (7.55) *−0.2 (2.04)*	$\geq$ 2 (2.00)
16	≤ -7 *$\geq$ 2 (2.00)*	$\geq$ 2 (2.00)	$\geq$ 2 (2.00)	$\geq$ 2 (2.00)	$\geq$ 2 (2.00)

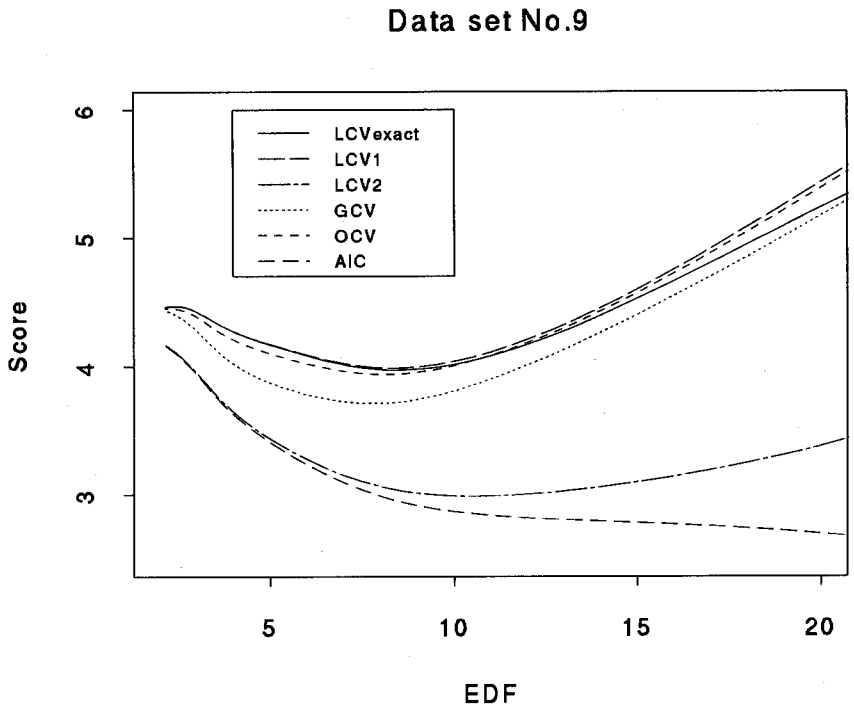
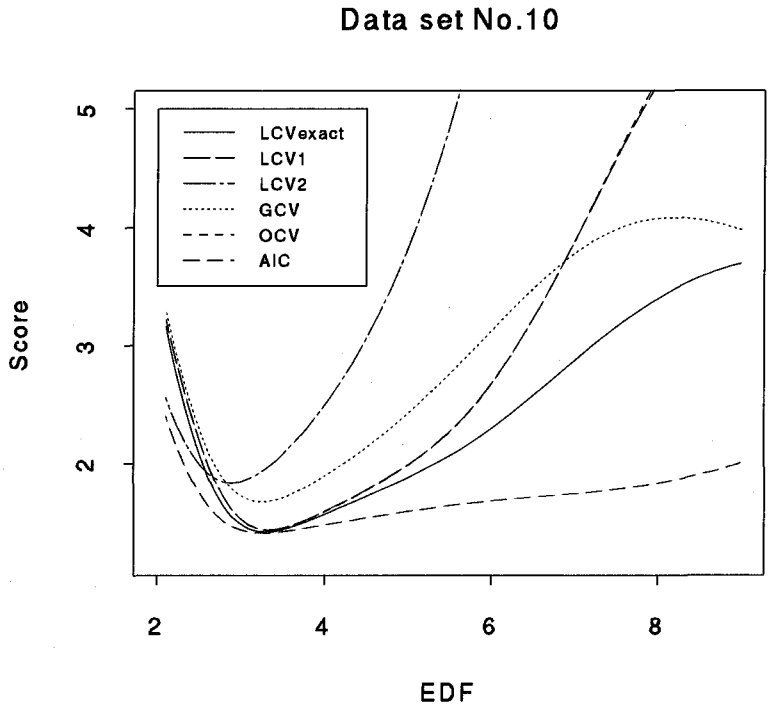


Figure 4.5: Plots of six scores to the data set No. 9.



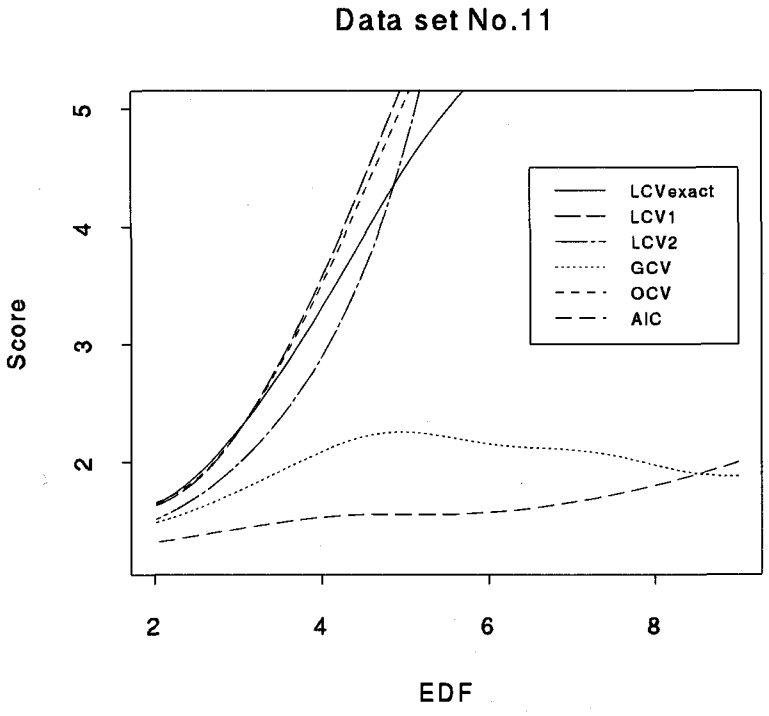


Figure 4.6: Plots of six scores to the data sets No. 10 and 11.

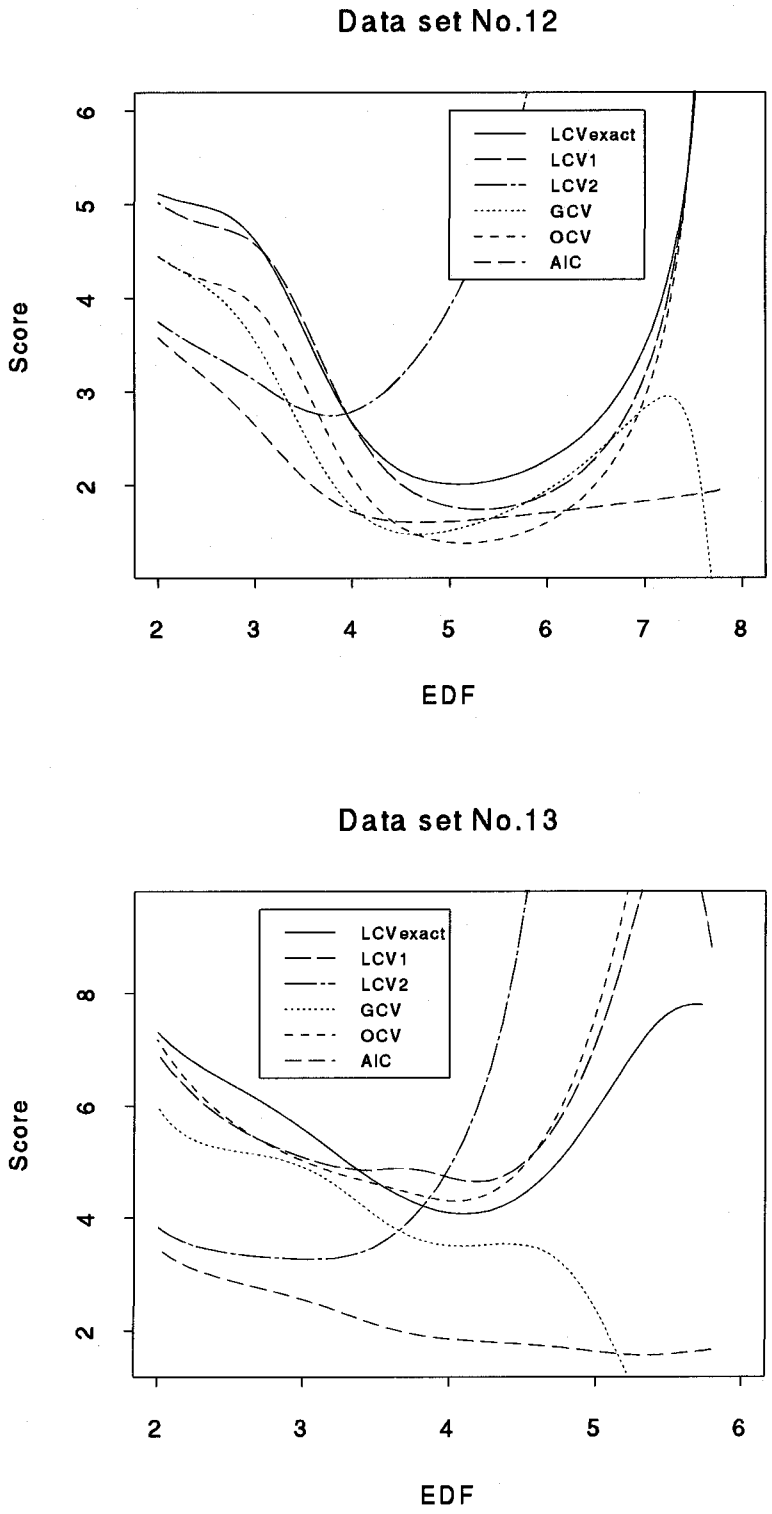
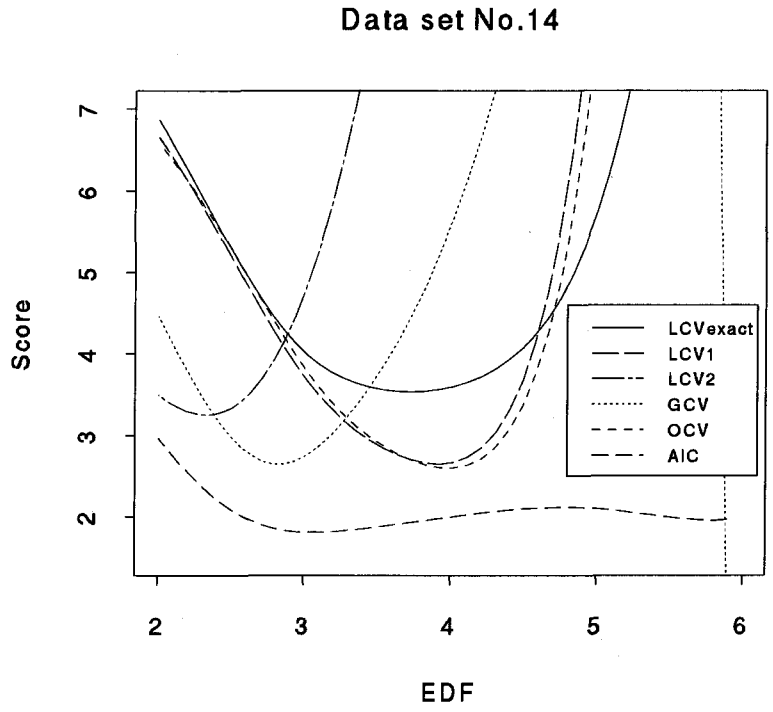


Figure 4.7: Plots of six scores to the data sets No. 12 and 13.



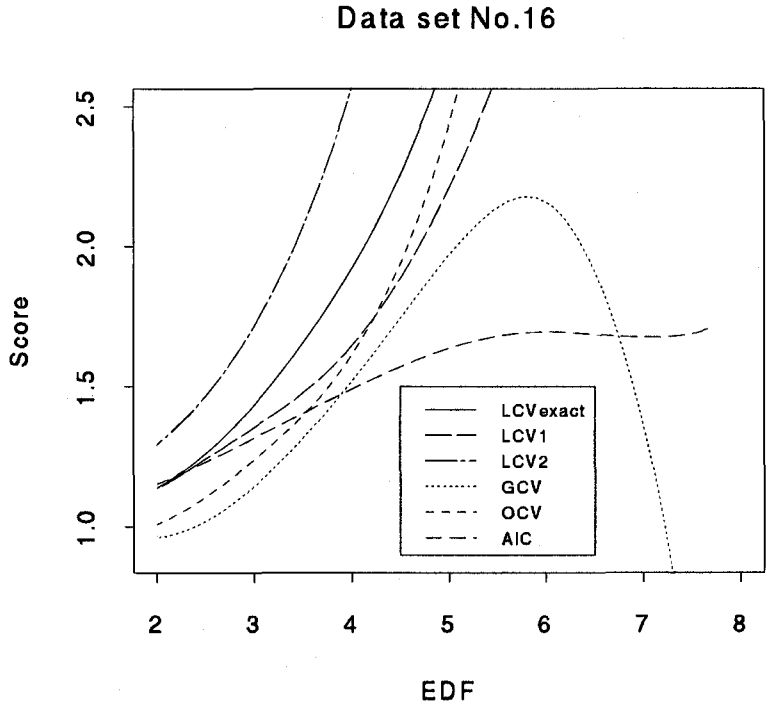
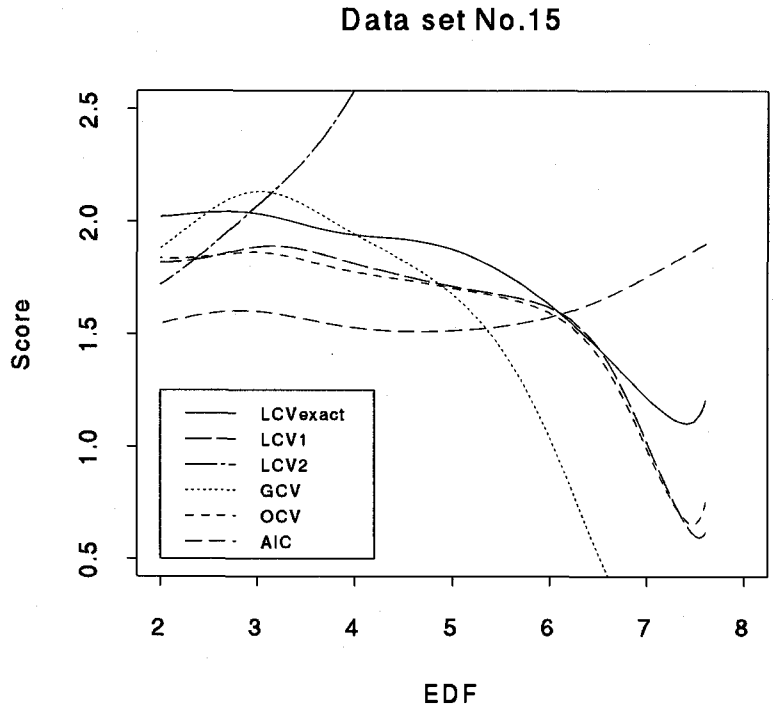


Figure 4.8: Plots of six scores to the data sets No. 14 and 16.



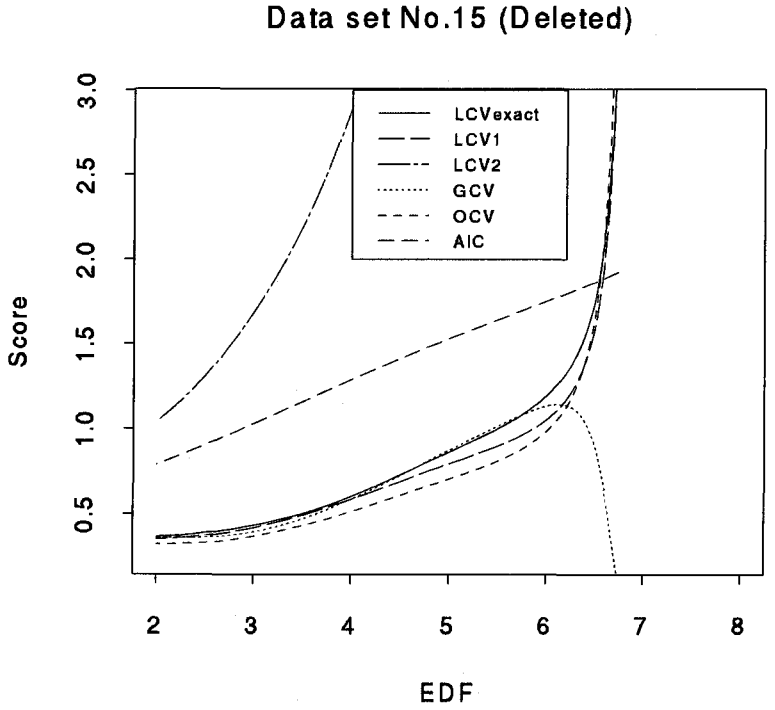


Figure 4.9: Plots of six scores to the data set No. 15.

4.1.3 Poisson Regression

Nonparametric Poisson regression models

$$y_i \sim \mathrm{Po}(\mu_i) \quad \text{and} \quad \log \frac{\mu_i}{q_i} = f(t_i), \quad i = 1, \ldots, n,$$

are applied to the data sets listed in Table 4.5, where each response y_i is observed with an offset variable q_i and a one-dimensional explanatory variable t_i. The data set No. 17 has no offset variable and hence all q_i's are set to be 1. The data set No. 18 has been also taken up in Section 2.4.2 and is the same as the data set No. 9. The values of $\log_{10} \hat{\lambda}$ chosen by minimizing each of the five scores GCV(λ), OCV(λ), AIC(λ), $\mathrm{LCV}_1(\lambda)$ and $\mathrm{LCV}_2(\lambda)$ are shown in Table 4.6 with the values of EDF. Figures 4.10 and 4.11 plot the five scores with the exact score LCV(λ) as the functions of EDF.

For the data set No. 17 all the scores trace similar curves and select almost equal values of $\hat{\lambda}$, whereas for the data set No. 18 extremely small $\hat{\lambda}$ is chosen by the AIC score. The data set No. 18 provides the result very similar to the data set No. 9 where a logistic regression model is fitted. The data sets No. 19 and 20 have relatively small numbers of age categories. For both data sets the AIC and LCV_2 scores trace different curves from the other scores and select slightly smaller $\hat{\lambda}$, while the scores OCV($\hat{\lambda}$) and $\mathrm{LCV}_1(\lambda)$ take very similar values and approximate the exact LCV score well.

Table 4.5: The examined data sets: Poisson regression.

No.	Response	Explan. var.	Offset variable	n	
17	Number of accidents	Year	—	112	
18	Number of deaths	Age category	Category size	50	
19			Person-years	12	Male
20				12	Female

The column labeled 'n' lists the numbers of categories.

References: 17: Jarrett (1979)
18: Green and Silverman (1994)
19, 20: Selvin (1994)

Table 4.6: The values of $\log_{10} \hat{\lambda}$ and EDF (in parentheses) to the data sets in Table 4.5. The values that give a local minimum are in italics.

No.	GCV	OCV	AIC	LCV_1	LCV_2
17	−2.0 (6.23)	−2.0 (6.23)	−2.3 (7.19)	−2.3 (7.19)	−2.1 (6.53)
18	−1.1 (7.66)	−1.1 (7.66)	−5.9 (44.94)	−1.2 (8.03)	−1.7 (10.18)
19	−0.9 (3.90)	−0.8 (3.76)	−0.2 (3.02)	−0.8 (3.76)	0.2 (2.63)
20	−1.9 (5.22)	−2.1 (5.60)	−1.7 (4.87)	−2.1 (5.60)	−0.7 (3.43)

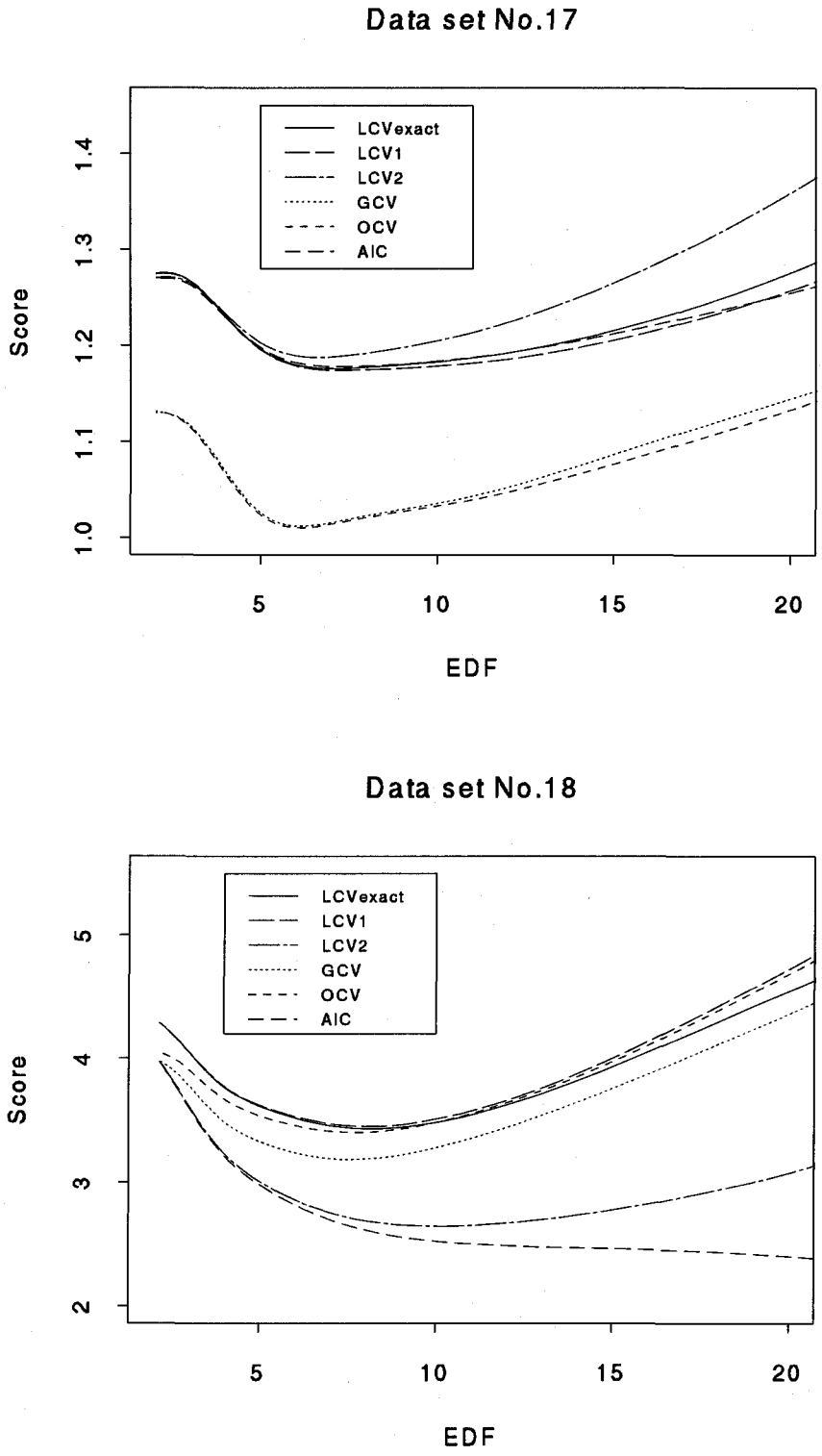


Figure 4.10: Plots of six scores to the data sets No. 17 and 18.

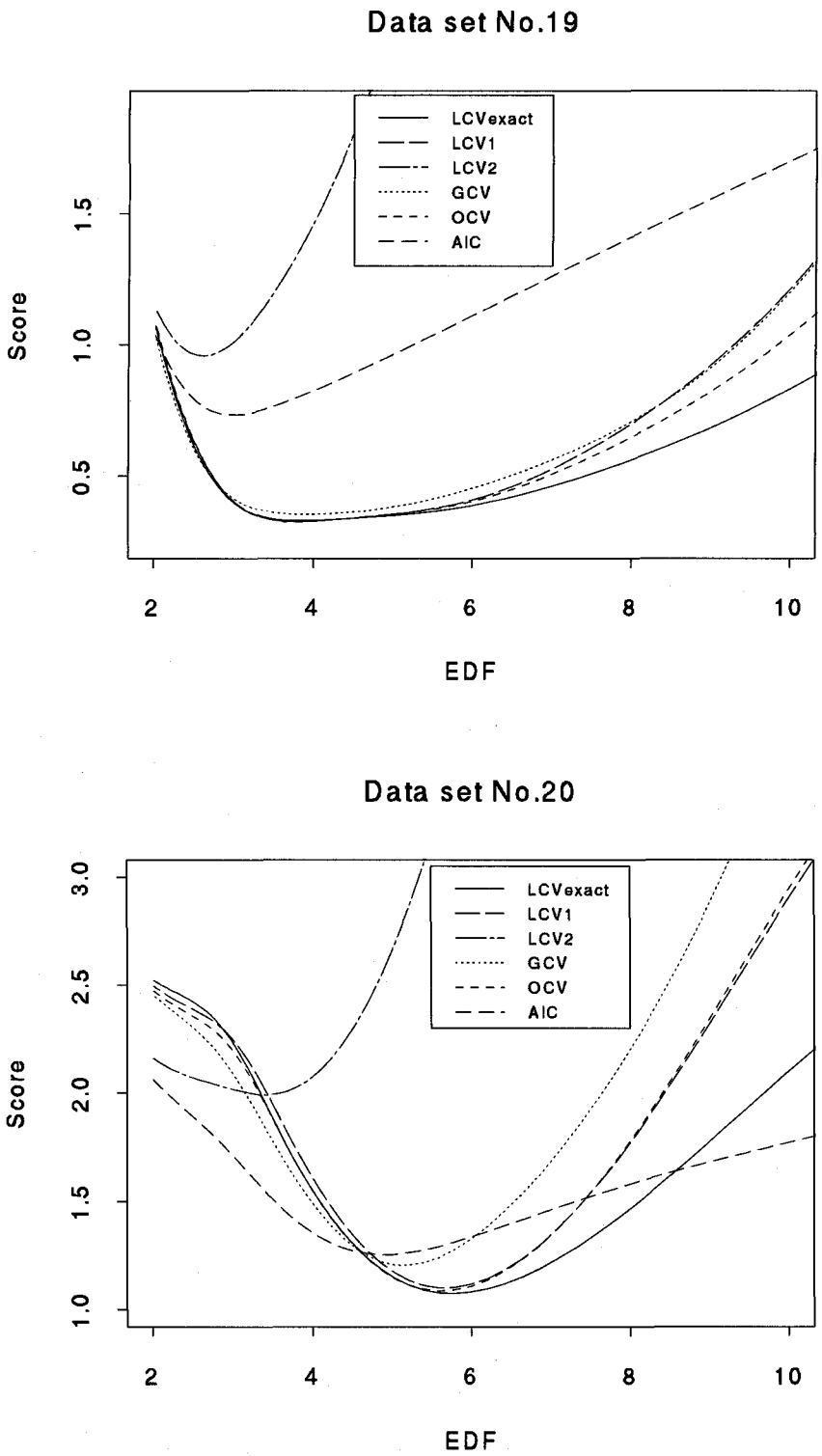


Figure 4.11: Plots of six scores to the data sets No. 19 and 20.

4.1.4 Density Smoothing

The method of density smoothing described in Section 2.4.3 is applied to several data sets listed in Table 4.7. The data sets No. 23 and 24 has been also taken up in Sections 2.4.3 and 3.3.2. The values of $\log_{10}\hat{\lambda}$ chosen by minimizing each of the five scores GCV(λ), OCV(λ), AIC(λ), $\mathrm{LCV}_1(\lambda)$ and $\mathrm{LCV}_2(\lambda)$ are shown in Table 4.8 with the values of EDF. Figures 4.12–4.14 plot the five scores with the exact score LCV(λ) as the functions of EDF.

For the data sets No. 23, 24 and 25 all the scores trace similar shapes of curves and almost the same values of $\hat{\lambda}$ are chosen. These data sets have relatively large number of observations. For the data set No. 21 OCV(λ) and $\mathrm{LCV}_1(\lambda)$ select the EDF about 10 while AIC(λ) and $\mathrm{LCV}_2(\lambda)$ select small values of EDF. For the data set No. 22 all the scores except $\mathrm{LCV}_2(\lambda)$ have two local minimums. The scores OCV(λ) and $\mathrm{LCV}_1(\lambda)$ are the lowest where EDFs are about 3, while AIC(λ) is the lowest where EDF=13.18 and GCV(λ) becomes smaller as EDF becomes larger. For the data set No. 26 all the scores suggest the linear fitted function, which implies shifted exponential distribution that has the density of the shape $p(x) = \exp\{-\beta(x-\alpha)\}$.

As a whole, the GCV score has local minimums but often can not select $\hat{\lambda}$. The OCV score takes close values to the GCV score when EDF is small while to the LCV_1 score when EDF is large. The values of $\hat{\lambda}$ selected by the OCV and LCV_1 scores are very similar. The AIC score seems to go up more slowly than the other scores as λ becomes small, while the LCV_2 score seems to go up steeply and select slightly large $\hat{\lambda}$. We think that the LCV_1 score or the OCV score is useful for determining the optimal degree of smoothing.

Table 4.7: The examined data sets: density smoothing.

No.	References	Observations of raw data	n	N
21	Simonoff (1996)	Time intervals	55	109
22	Efron and Tibshirani (1996)	Pain scores	40	67
23	Silverman (1986)	Duration of eruptions	35	107
24			35	107
25	Simonoff (1996)	Salary	28	147
26	Silverman (1986)	Length of treatment spells	20	86

The column labeled 'n' lists the numbers of classes, and the column labeled 'N' lists the numbers of observations.

Note: 23: The domain $[1.5, 5]$ is divided into 35 intervals.
24: The domain $[1.55, 5.05]$ is divided into 35 intervals.

Table 4.8: The values of $\log_{10} \hat{\lambda}$ and EDF (in parentheses) to the data sets in Table 4.7. The values that give a local minimum are in italics.

No.	GCV	OCV	AIC	LCV_1	LCV_2
21	≤ -7 *−3.8 (10.81)*	−3.8 (10.81)	−2.3 (5.50)	−3.6 (9.89)	−1.6 (4.08)
22	≤ -7 *−0.9 (2.93)*	−0.8 (2.83)	−4.6 (13.18) *−1.1 (3.16)*	−1.0 (3.04)	−1.0 (3.04)
23	−4.0 (11.89)	−3.7 (10.54)	−3.9 (11.42)	−3.8 (10.97)	−3.3 (8.98)
24	−3.1 (8.60)	−3.7 (11.05)	−3.8 (11.52)	−3.9 (12.00)	−3.1 (8.60)
25	−1.0 (3.69)	−0.9 (3.53)	−2.0 (5.61)	−1.0 (3.67)	−1.2 (3.98)
26	≤ -7 *≥ 2 (2.00)*	≥ 2 (2.00)	≥ 2 (2.00)	≥ 2 (2.00)	≥ 2 (2.00)

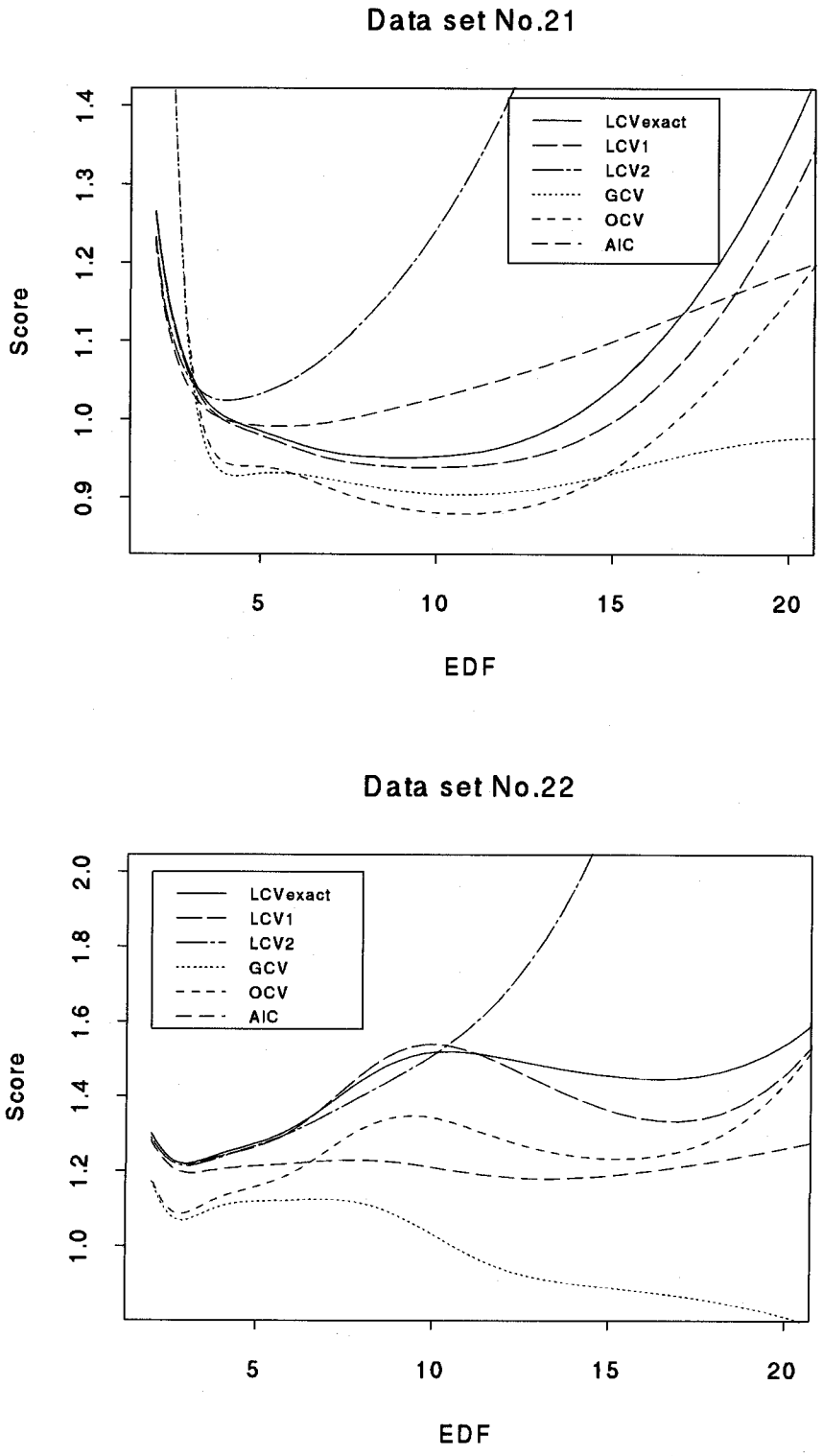
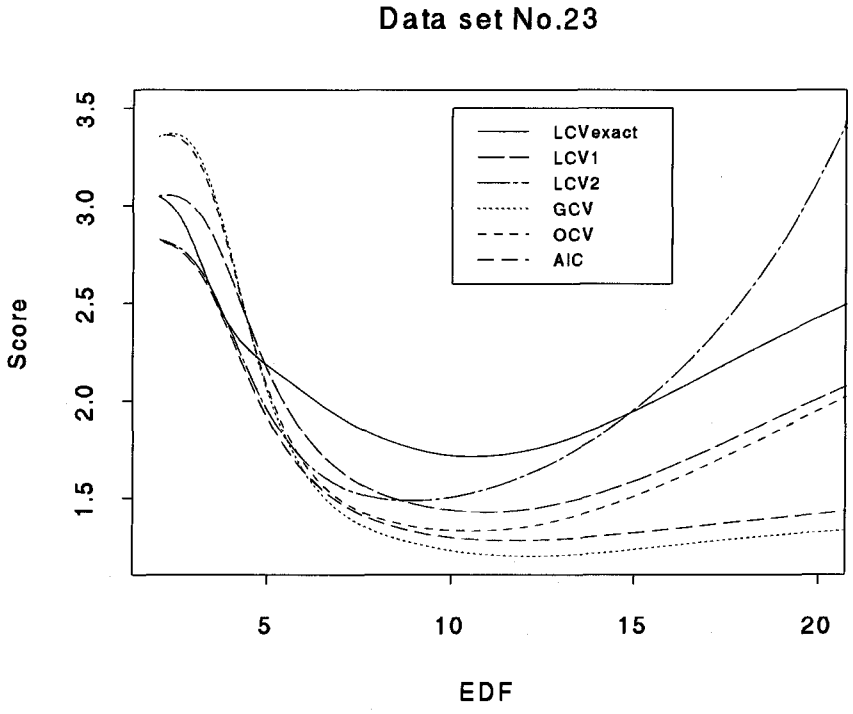


Figure 4.12: Plots of six scores to the data sets No. 21 and 22.



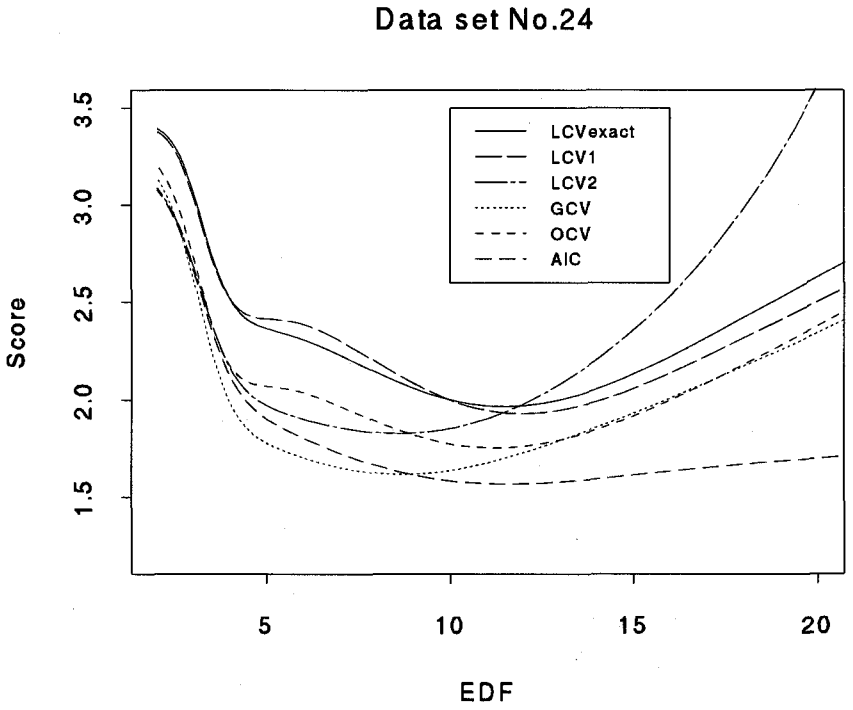


Figure 4.13: Plots of six scores to the data sets No. 23 and 24.

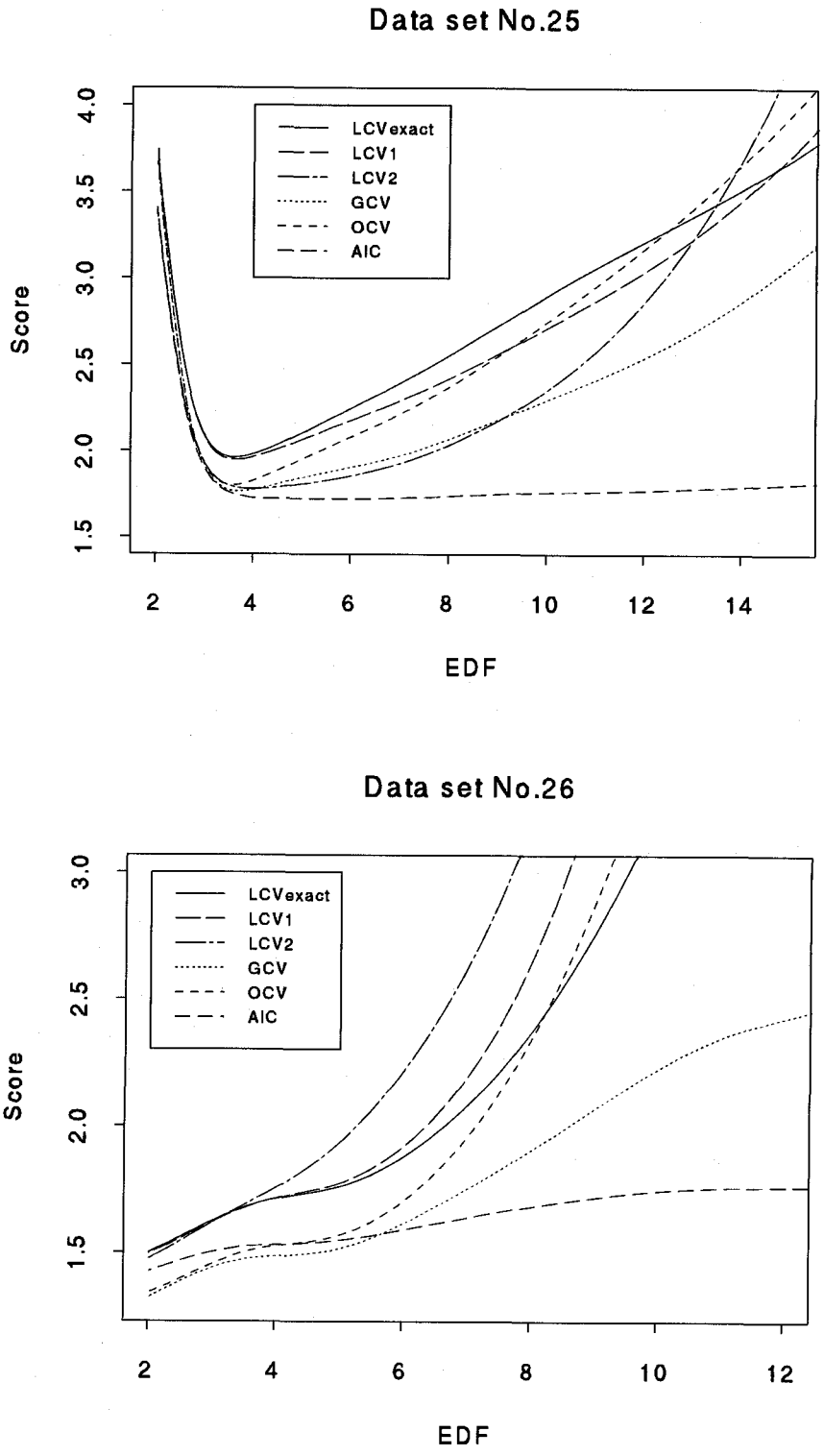


Figure 4.14: Plots of six scores to the data sets No. 25 and 26.

4.2 Simulation Studies

In this section some simulation is performed to assess characteristics of the LCV_1 and the LCV_2 scores and to show usefulness of these scores. The performance of the LCV_1 and the LCV_2 scores are compared with that of the GCV, OCV and the AIC scores and whether the LCV_1 and the LCV_2 scores improve overall goodness-of-fit of estimates is examined. Moreover the effect of factors such as sample size and smoothness of a true function on the bias and the dispersion of estimates is investigated.

4.2.1 Logistic Regression Case

In this section the simulation is performed in the context of binary logistic regression case. Binary responses y_i are produced according to

$$\mathrm{P}(y_i = 1) = p_i \quad \text{and} \quad \log \frac{p_i}{1 - p_i} = f(t_i), \quad i = 1, \ldots, n,$$

for some true function f and explanatory variables t_i.

We use the fitted functions to the data sets examined in Section 4.1.1 as the true functions f to perform more practical simulation. Out of the eight data sets, No. 2 and 4 are excluded because extremely small values of $\hat{\lambda}$ are chosen by the four scores except $\mathrm{LCV}_1(\lambda)$, and the data set No. 7 is excluded because the iteration of the algorithm often does not converge. The data sets No. 1 and 5 provide similar results and hence we use No. 1 to which all the scores select the values of EDF less than 3. Therefore the data sets No. 1, 3, 6 and 8 are used. For each data set three fitted functions with the EDFs 2, 3 and 6 are selected as the true functions. These are plotted in Figures 4.15 and 4.16. The fitted function with EDF=2 is a straight line, that is, the fitted function of the ordinary logistic regression model, the one with EDF=3 is a relatively smooth function and the one with EDF=6 is a relatively rough function. The values of $\hat{\lambda}$ and $\log_{10} \hat{\lambda}$ corresponding to EDF=3 and 6 are listed in Table 4.9. The data set No. 3 is used because we want to examine how the low probability affects the performance of the scores.

The sample size n is taken as 25, 50, 100 and 200 corresponding to the sample size of the data sets in Section 4.1.1. To unify the range to search λ, the domain of explanatory variables of each original data set is rescaled so

Table 4.9: The values of $\hat{\lambda}$ and $\log_{10} \hat{\lambda}$ (in parentheses) corresponding to EDF=3 and 6 for the fitted functions to the data sets used in the simulation.

No.	EDF=3		EDF=6	
1	0.072	(−1.143)	0.00183	(−2.738)
3	0.021	(−1.678)	0.00034	(−3.469)
6	0.0015	(−2.824)	0.00017	(−3.770)
8	0.0097	(−2.013)	0.000107	(−3.971)

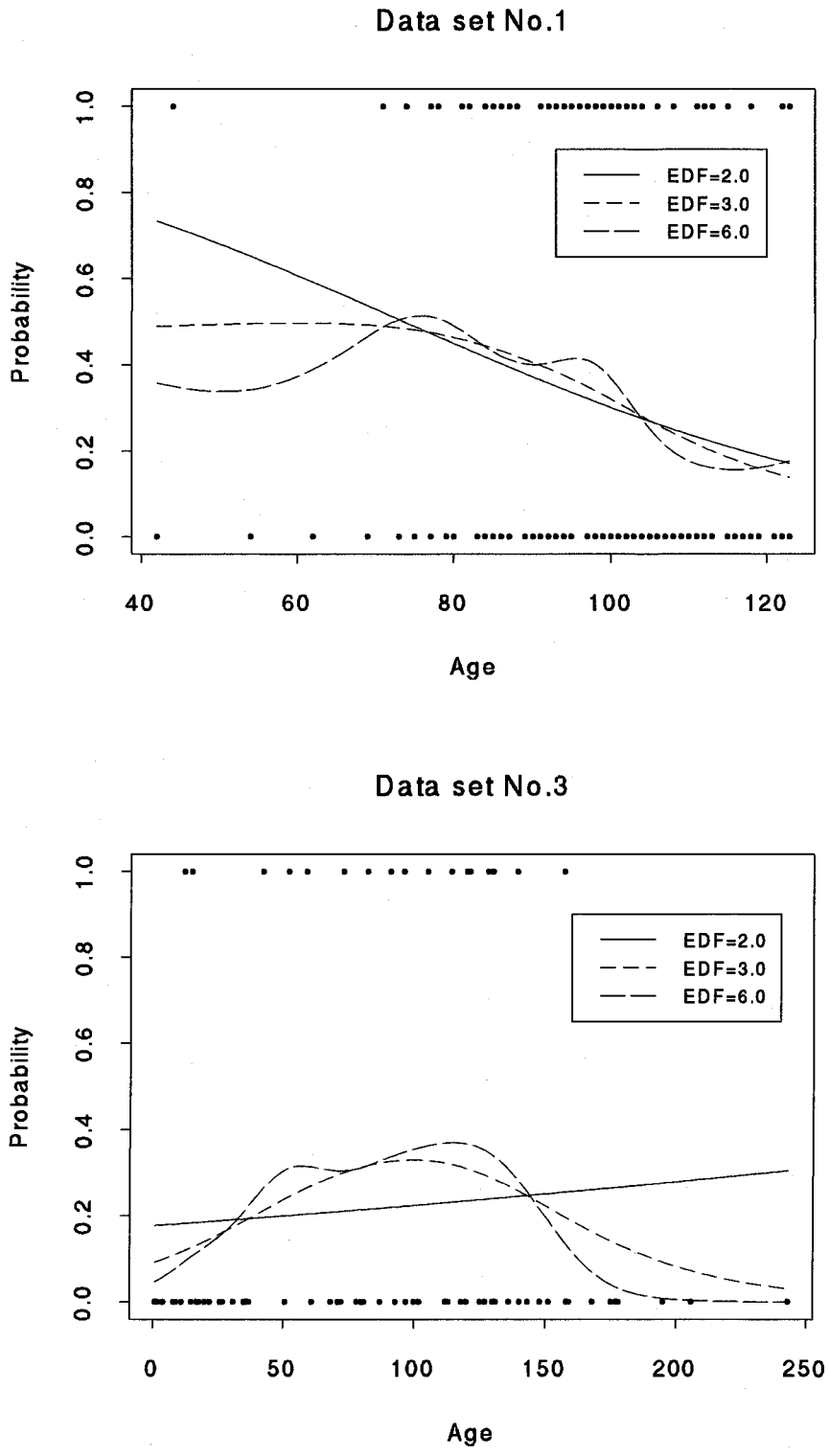


Figure 4.15: Fitted functions to the data set No. 1 and 3.

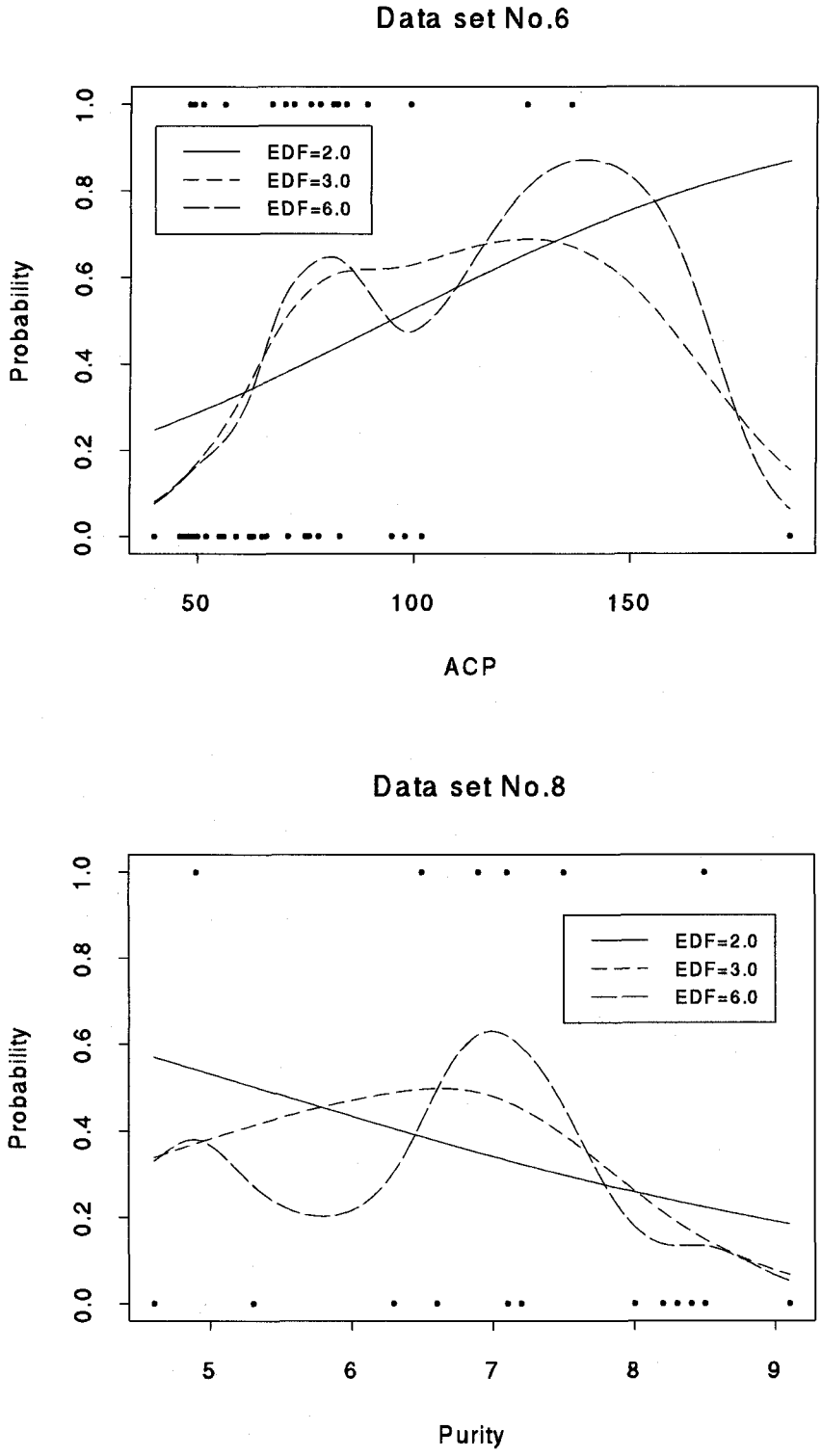


Figure 4.16: Fitted functions to the data set No. 6 and 8.

that the lower and upper bounds coincide with 0 and 1 respectively, and $\{t_i\}$ are set to be equally spaced on [0,1] for simplicity. The nonparametric logistic regression model is fitted to each of 200 data sets produced as above. The values of $\hat{\lambda}$ are chosen by minimizing any of the five scores GCV(λ), OCV(λ), AIC(λ), $\mathrm{LCV}_1(\lambda)$ and $\mathrm{LCV}_2(\lambda)$, and the natural cubic B-splines $\hat{f}$ with knots at $t_1, \ldots, t_n$ are estimated from the $\hat{\lambda}$'s just chosen.

The binary responses are made by transforming uniform pseudo-random numbers produced by using the IMSL FORTRAN subroutine DRNUN. Some other IMSL FORTRAN subroutines are also used in computation. The simulation is performed with the program compiled by Microsoft FORTRAN Visual Workbench ver. 1.00 on the PC with Pentium processor of 150MHz and the RAM of 32MB in which Windows 95 has been installed.

The values of $\log_{10}\lambda$ are searched at first at intervals of 0.5 from -6 to 2, and still more searched at intervals $0.5^6 \approx 0.016$ on the neighborhood of $\log_{10}\lambda$ just chosen. The distribution of $\hat{\lambda}$ chosen by minimizing each of the five scores is shown in Figures 4.17–4.24, which are plotted by the S-PLUS command `boxplot`. White marks in the box indicate the median and both bounds of the box indicate the first and the third quartiles. Whiskers are drawn to the nearest value not beyond 1.5 times of the inter-quartile range from the quartiles, and points beyond are drawn individually. The numbers of times that $\hat{\lambda} \geq 10^2$ and $\hat{\lambda} \leq 10^{-6}$ are counted and are listed in Tables 4.10, 4.12, 4.14 and 4.16. Notice that $\hat{\lambda} \geq 10^2$ implies that a nearly linear function is fitted and that $\hat{\lambda} \leq 10^{-6}$ implies that an almost interpolating function is fitted.

The GCV score tends to choose quite small $\hat{\lambda}$ when sample size is small and have many times that $\hat{\lambda} \leq 10^{-6}$. The AIC score also does so but seems not to be so bad as the GCV score. The GCV score provides distribution of $\hat{\lambda}$ similar to the OCV score when sample size is large, although these scores choose relatively larger values of $\hat{\lambda}$ even when the true function is rough. The AIC score provides distribution of $\hat{\lambda}$ similar to the LCV_1 and the LCV_2 scores when sample size is large. The LCV_1 and the LCV_2 scores provide similar distribution of $\hat{\lambda}$, although the LCV_1 score provides distribution of slightly wider range and the LCV_2 score selects slightly larger $\hat{\lambda}$. These scores tend to select $\hat{\lambda}$ larger than or close to 10^2 when the true function is linear, and select $\hat{\lambda}$ closer to the one with which the true function is fitted to the original data when the true function is nonlinear. Moreover these scores hardly choose extremely small $\hat{\lambda}$ which leads to an interpolating function. The LCV_2 score never selects $\hat{\lambda}$ less than 10^{-6}. There are many times that $\hat{\lambda} \geq 10^2$ even if the true function is rough when the fitted functions to the data set No. 1 are used. We think it is because the true function has smaller variation even when EDF is large.

To investigate overall goodness-of-fit of estimated functions selected with each of the five scores, the averaged mean squared error

$$\mathrm{AMSE}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\{\hat{f}(t_i) - f(t_i)\}^2$$

is evaluated. The logarithms of the estimates $\widehat{\mathrm{AMSE}}(\hat{f})$ are taken since the distribution of $\widehat{\mathrm{AMSE}}(\hat{f})$ is skewed to the left side. Tables 4.11, 4.13, 4.15 and

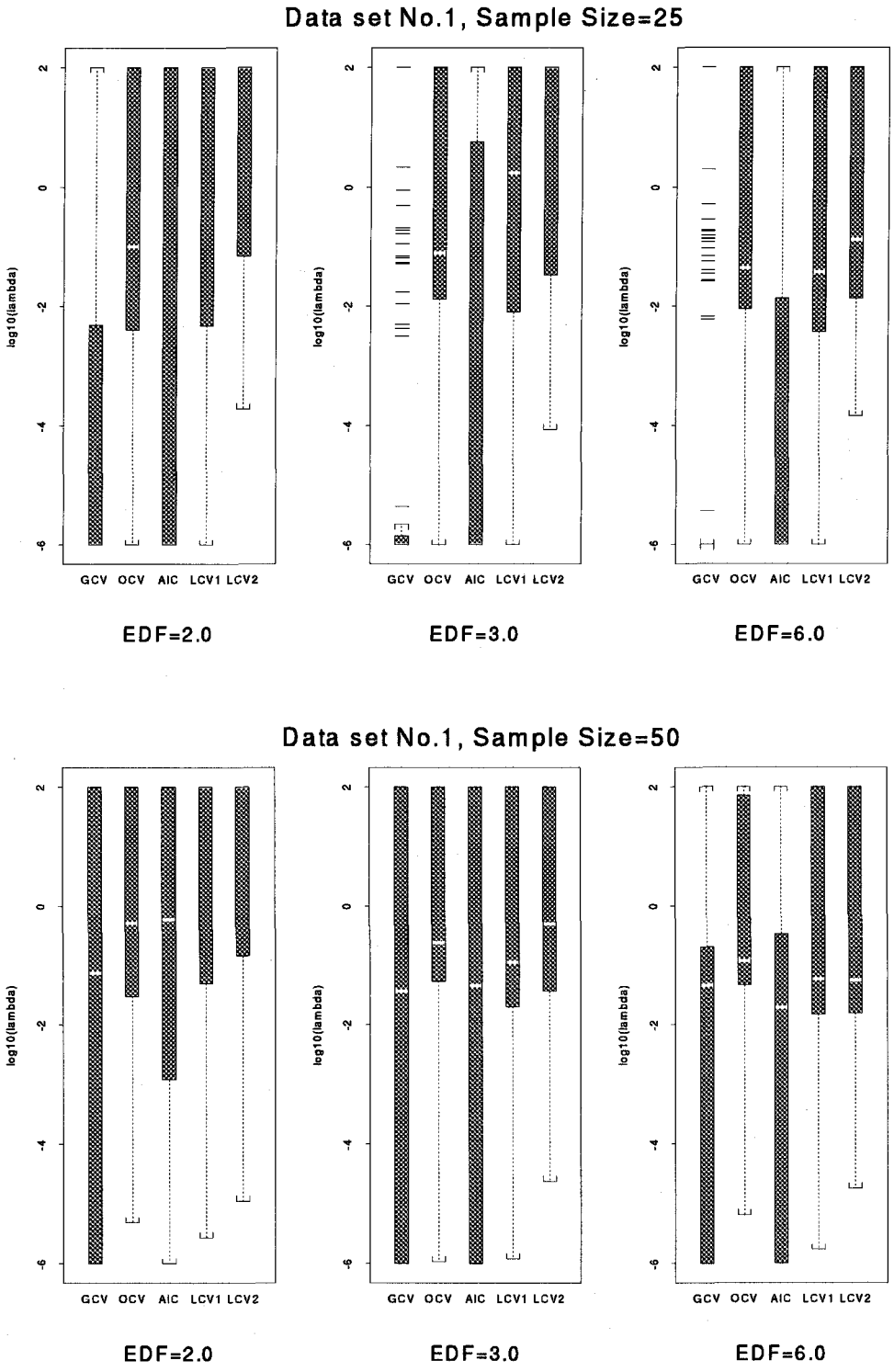


Figure 4.17: Boxplots of distribution of $\hat{\lambda}$ when the fitted function to the data set No. 1 is used.

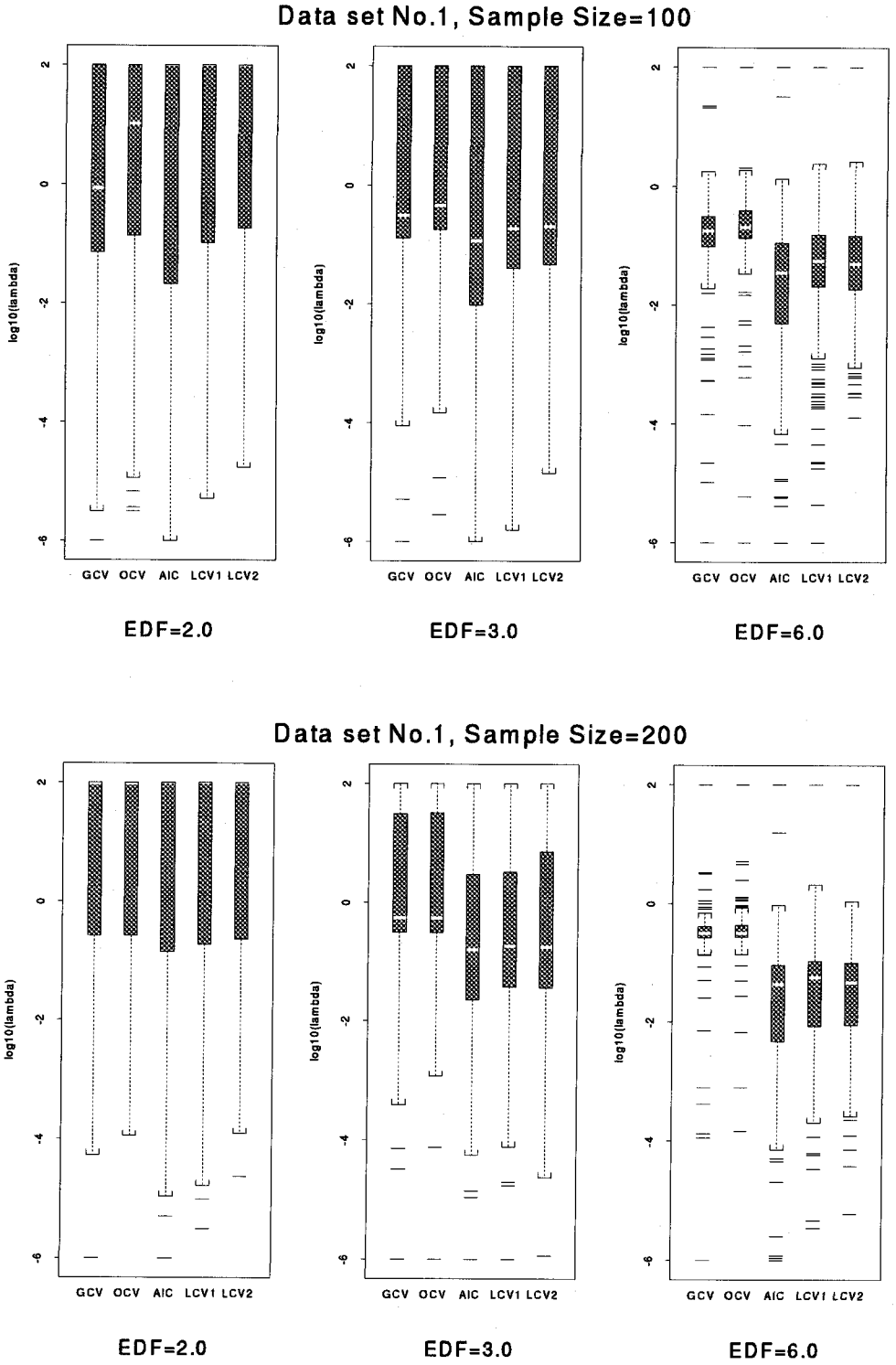


Figure 4.18: Boxplots of distribution of $\hat{\lambda}$ when the fitted function to the data set No. 1 is used (continued).

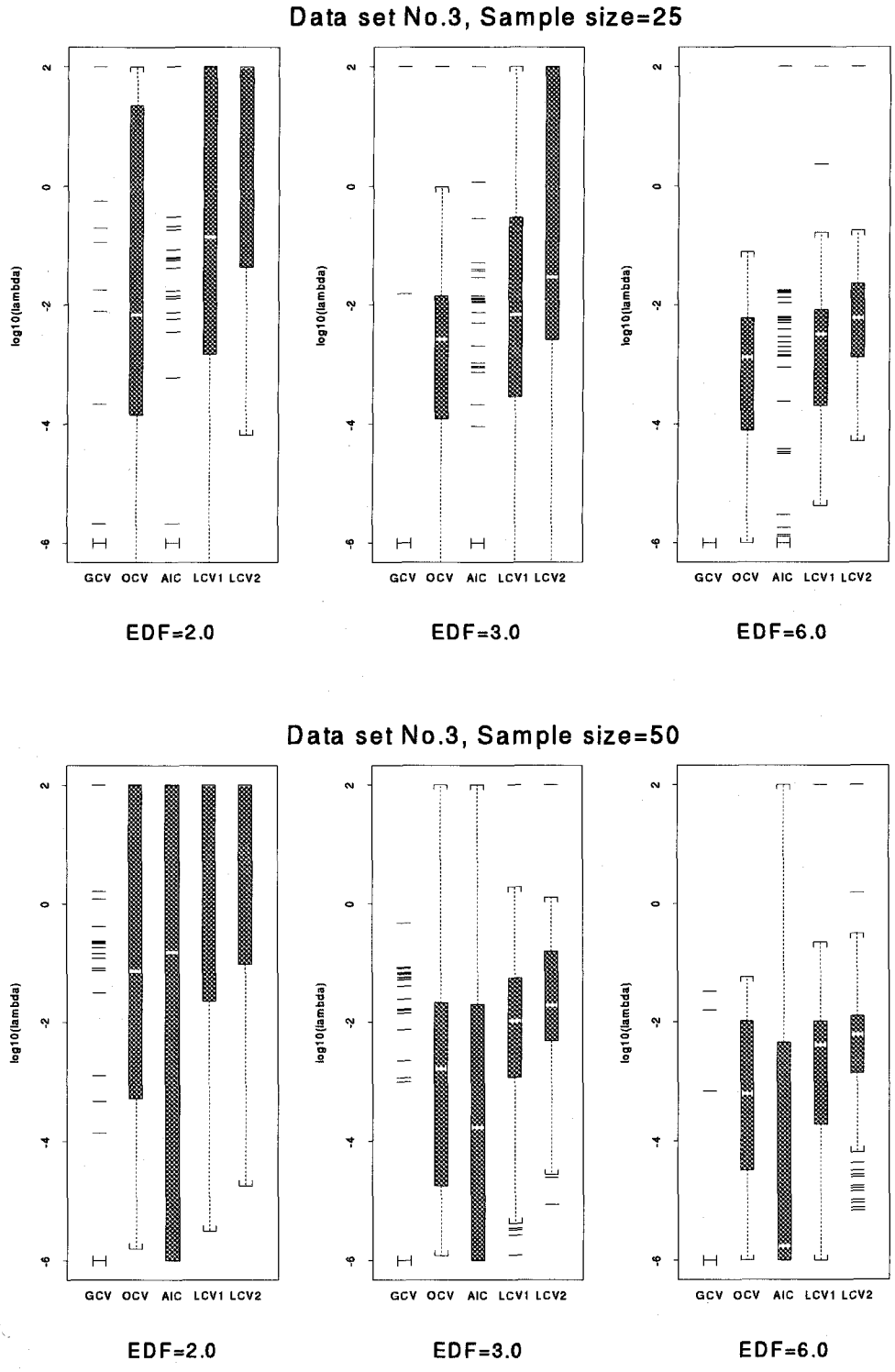


Figure 4.19: Boxplots of distribution of $\hat{\lambda}$ when the fitted function to the data set No. 3 is used.

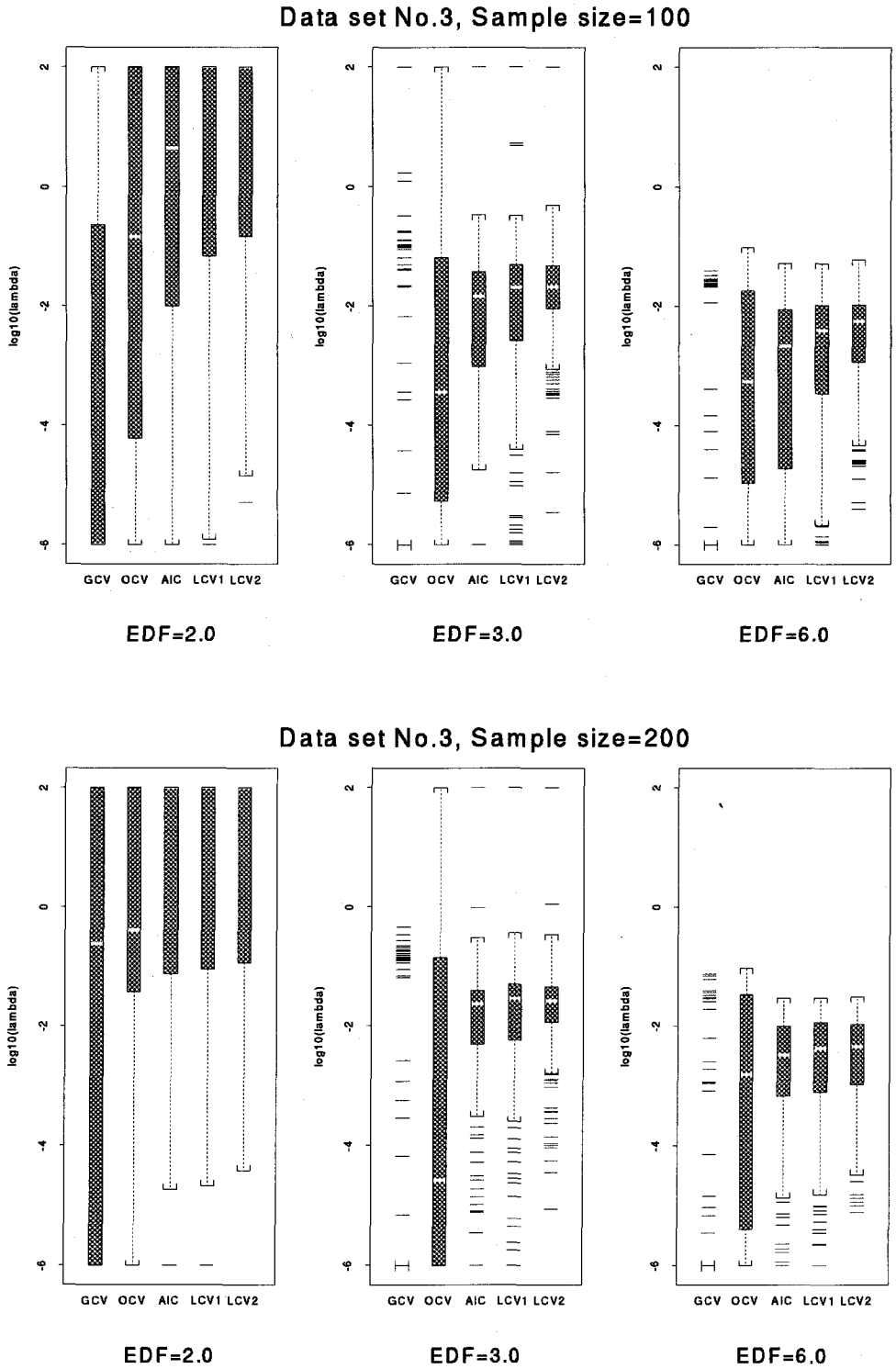


Figure 4.20: Boxplots of distribution of $\hat{\lambda}$ when the fitted function to the data set No. 3 is used (continued).

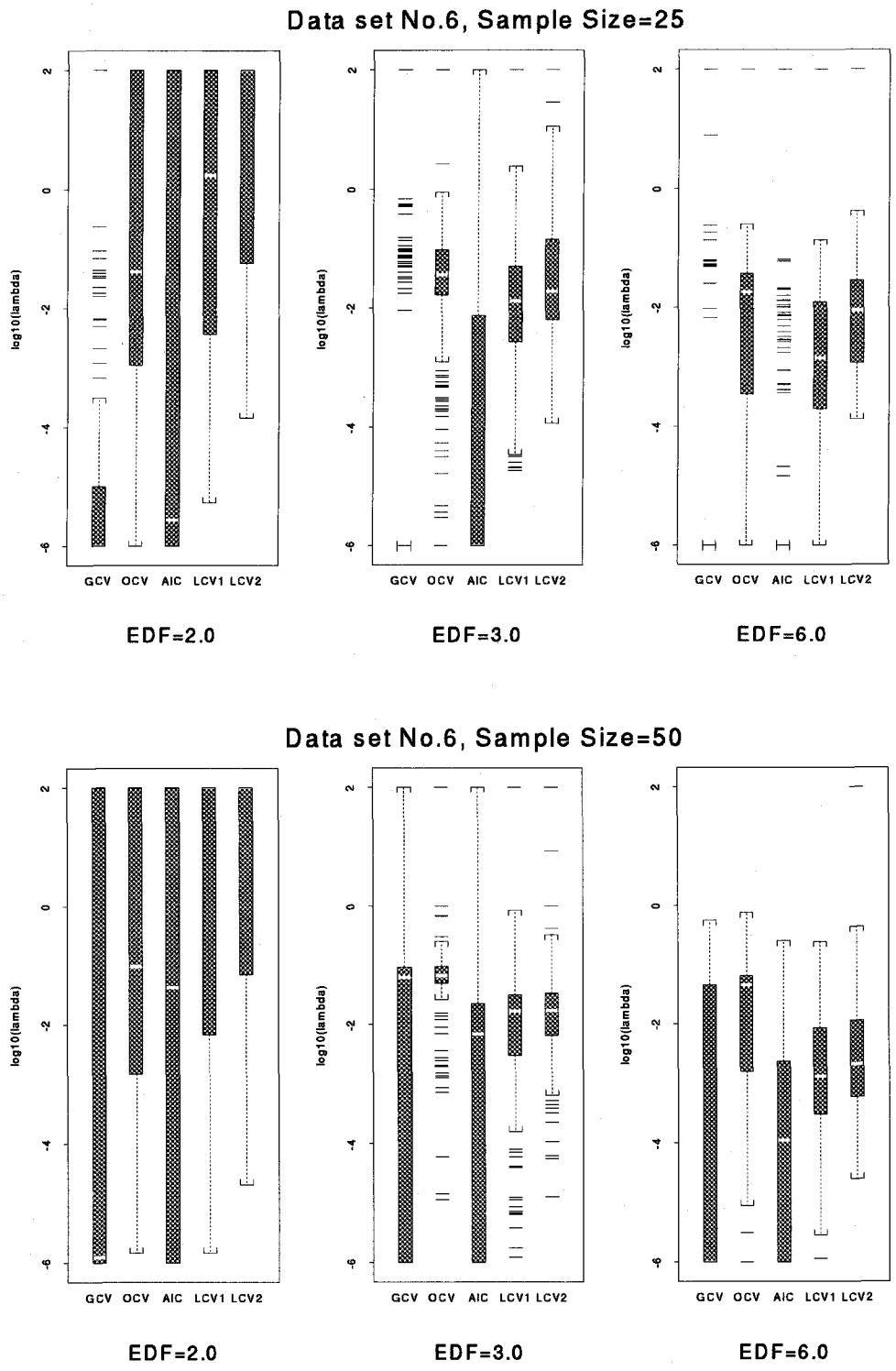


Figure 4.21: Boxplots of distribution of $\hat{\lambda}$ when the fitted function to the data set No. 6 is used.

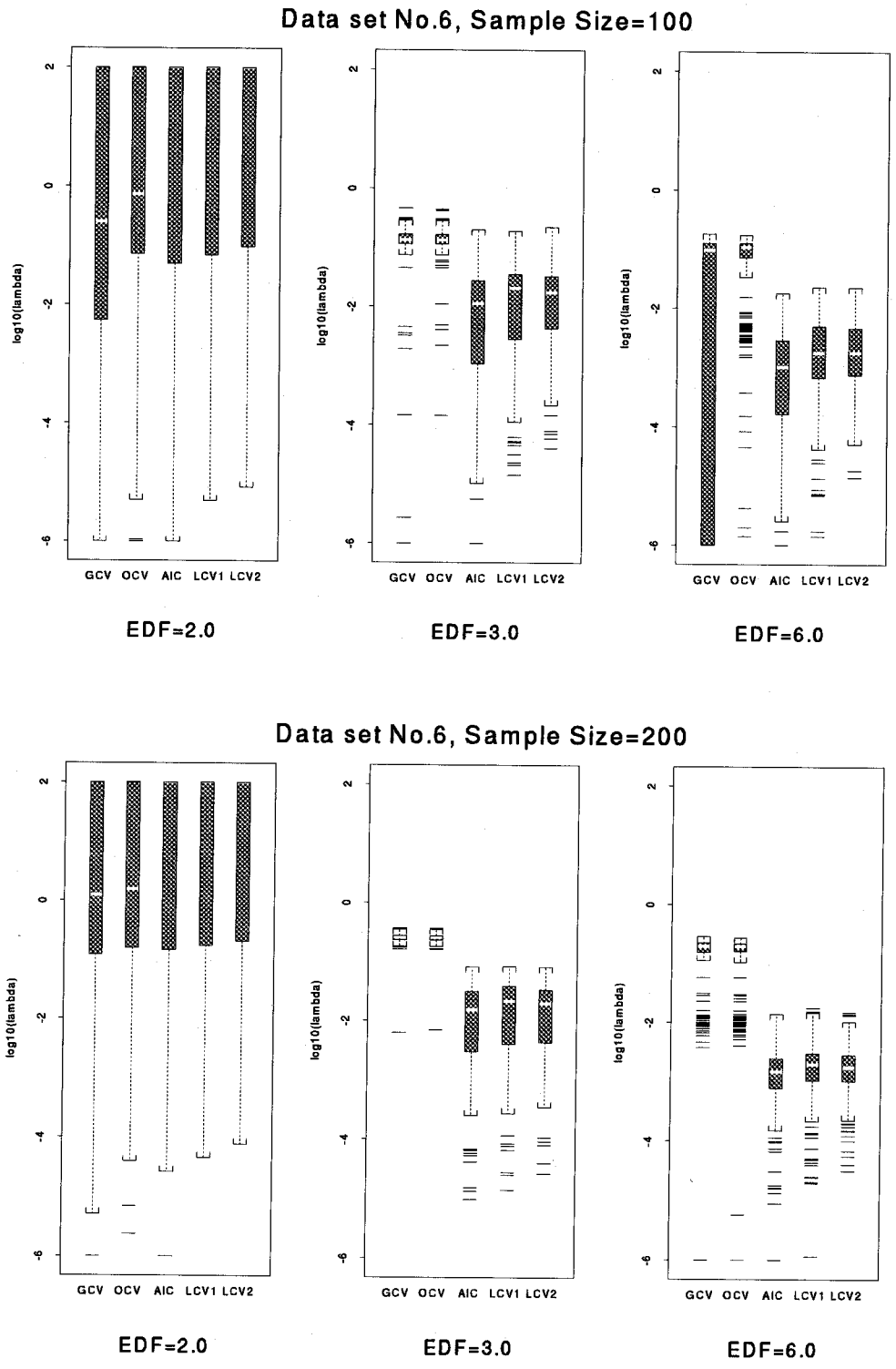


Figure 4.22: Boxplots of distribution of $\hat{\lambda}$ when the fitted function to the data set No. 6 is used (continued).

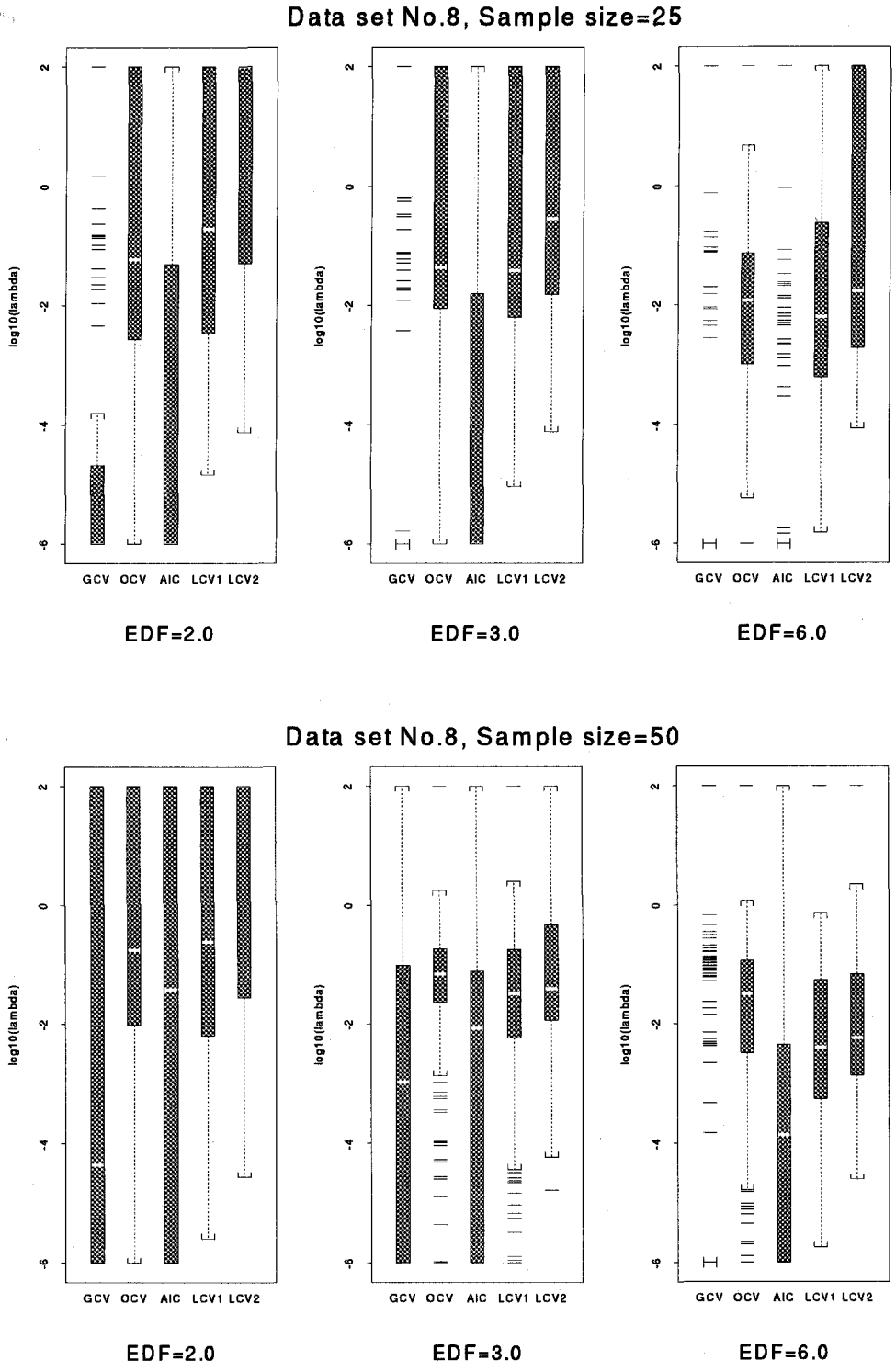


Figure 4.23: Boxplots of distribution of $\hat{\lambda}$ when the fitted function to the data set No. 8 is used.

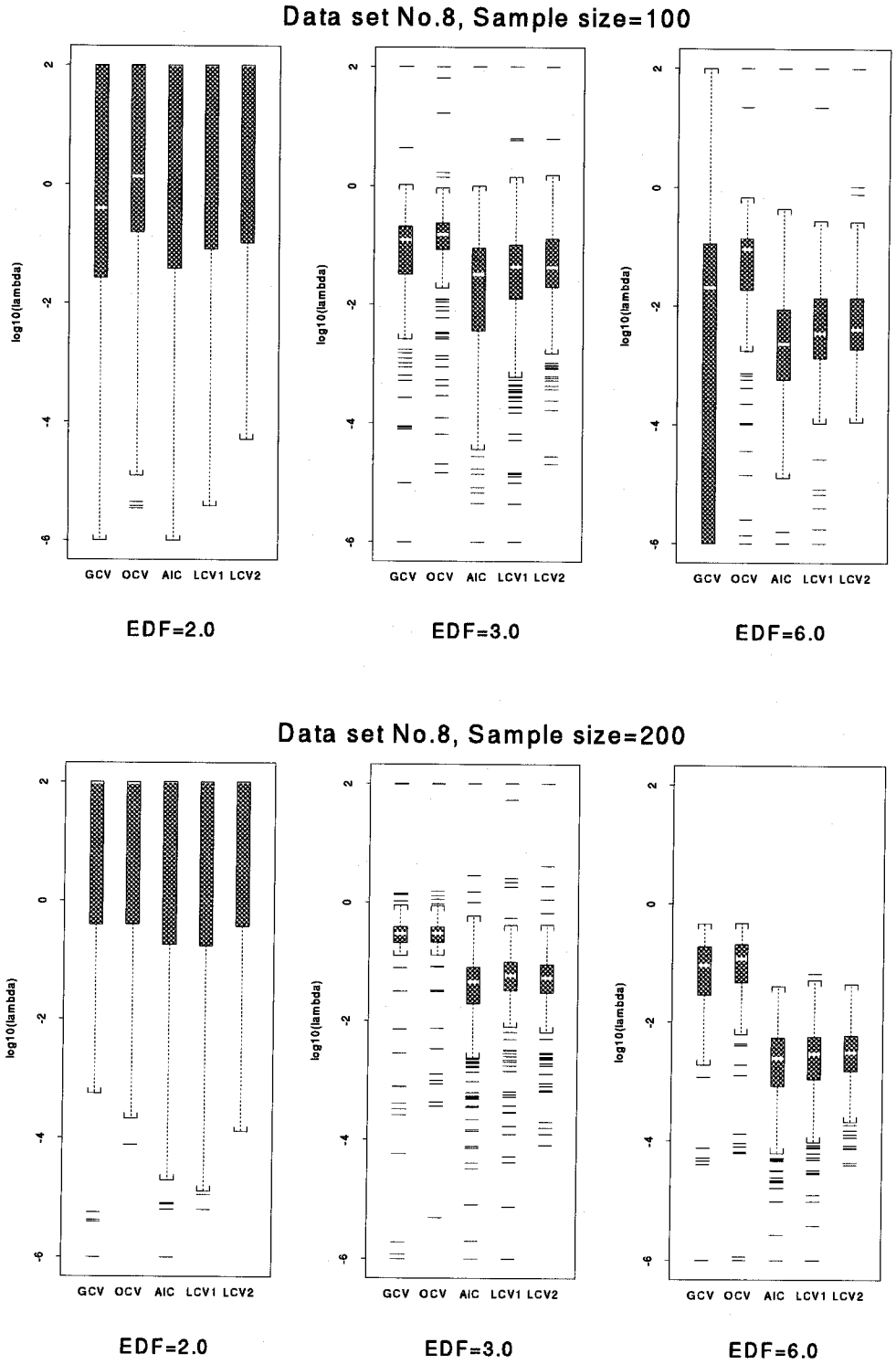


Figure 4.24: Boxplots of distribution of $\hat{\lambda}$ when the fitted function to the data set No. 8 is used (continued).

Table 4.10: The numbers of the times that $\hat{\lambda} \geq 10^2$ and $\hat{\lambda} \leq 10^{-6}$ out of 200 replications when the fitted functions to the data set No. 1 are used.

Design		$\hat{\lambda} \geq 10^2$					$\hat{\lambda} \leq 10^{-6}$				
EDF	n	GCV	OCV	AIC	LCV_1	LCV_2	GCV	OCV	AIC	LCV_1	LCV_2
2.0	25	41	84	60	110	129	131	5	108	2	0
	50	69	92	96	109	129	81	0	44	0	0
	100	88	96	103	117	126	20	0	15	0	0
	200	86	88	121	123	127	2	0	1	0	0
3.0	25	29	80	50	97	123	149	7	119	2	0
	50	52	80	64	83	96	76	0	55	0	0
	100	62	71	65	73	78	21	0	14	0	0
	200	39	39	45	45	47	1	1	1	1	0
6.0	25	26	53	30	70	91	153	8	132	2	0
	50	37	50	45	58	63	78	0	56	0	0
	100	22	26	27	28	30	24	1	11	1	0
	200	5	5	4	5	5	3	0	1	0	0

Table 4.11: The values of the mean and the standard error (in parentheses) of $\log \widehat{\mathrm{AMSE}}(\hat{f})$ when the fitted functions to the data set No. 1 are used.

EDF	n	GCV	OCV	AIC
2.0	25	3.035 (0.212)	−0.410 (0.147)	2.229 (0.220)
	50	0.249 (0.209)	−1.430 (0.114)	−0.570 (0.178)
	100	−2.106 (0.143)	−2.512 (0.105)	−1.966 (0.136)
	200	−3.226 (0.101)	−3.262 (0.096)	−3.065 (0.110)
3.0	25	3.172 (0.186)	−0.490 (0.129)	2.381 (0.198)
	50	0.271 (0.179)	−1.360 (0.085)	−0.106 (0.155)
	100	−1.649 (0.119)	−2.077 (0.071)	−1.543 (0.100)
	200	−2.710 (0.064)	−2.729 (0.061)	−2.465 (0.073)
6.0	25	3.493 (0.201)	−0.130 (0.122)	2.960 (0.202)
	50	0.490 (0.173)	−1.181 (0.070)	0.182 (0.154)
	100	−1.203 (0.117)	−1.712 (0.059)	−1.274 (0.089)
	200	−2.080 (0.047)	−2.142 (0.036)	−2.036 (0.055)

EDF	n	LCV_1	LCV_2
2.0	25	−0.529 (0.142)	−0.936 (0.110)
	50	−1.437 (0.112)	−1.656 (0.095)
	100	−2.391 (0.109)	−2.512 (0.107)
	200	−3.121 (0.106)	−3.185 (0.101)
3.0	25	−0.453 (0.116)	−0.800 (0.083)
	50	−1.199 (0.091)	−1.376 (0.080)
	100	−1.809 (0.076)	−1.922 (0.067)
	200	−2.546 (0.073)	−2.573 (0.071)
6.0	25	−0.062 (0.108)	−0.420 (0.078)
	50	−0.920 (0.080)	−1.095 (0.062)
	100	−1.539 (0.066)	−1.632 (0.057)
	200	−2.095 (0.050)	−2.126 (0.047)

Table 4.12: The numbers of the times that $\hat{\lambda} \geq 10^2$ and $\hat{\lambda} \leq 10^{-6}$ out of 200 replications when the fitted functions to the data set No. 3 are used.

Design		$\hat{\lambda} \geq 10^2$					$\hat{\lambda} \leq 10^{-6}$				
EDF	n	GCV	OCV	AIC	LCV_1	LCV_2	GCV	OCV	AIC	LCV_1	LCV_2
2.0	25	9	50	26	95	132	182	5	157	0	0
	50	23	68	91	110	129	161	0	66	0	0
	100	31	58	98	106	115	122	11	19	3	0
	200	56	68	115	114	117	63	8	2	1	0
3.0	25	2	14	9	48	79	196	7	169	0	0
	50	0	8	17	28	42	182	0	95	0	0
	100	1	3	15	17	18	174	17	28	2	0
	200	0	1	3	3	3	167	69	7	6	0
6.0	25	0	0	3	18	31	200	13	165	0	0
	50	0	0	2	2	8	197	1	100	1	0
	100	0	0	0	0	0	183	12	41	7	0
	200	0	0	0	0	0	176	28	7	2	0

Table 4.13: The values of the mean and the standard error (in parentheses) of $\log \widehat{\mathrm{AMSE}}(\hat{f})$ when the fitted functions to the data set No. 3 are used.

EDF	n	GCV	OCV	AIC
2.0	25	4.839 (0.163)	0.414 (0.175)	4.069 (0.222)
	50	2.344 (0.169)	−0.771 (0.144)	0.100 (0.205)
	100	0.412 (0.154)	−1.264 (0.141)	−1.640 (0.135)
	200	−1.896 (0.162)	−2.735 (0.120)	−2.889 (0.104)
3.0	25	6.186 (0.109)	1.200 (0.148)	5.493 (0.172)
	50	4.395 (0.143)	0.921 (0.145)	2.241 (0.213)
	100	2.823 (0.132)	0.437 (0.148)	−0.447 (0.141)
	200	1.482 (0.120)	0.249 (0.132)	−1.514 (0.098)
6.0	25	6.933 (0.064)	1.858 (0.147)	6.193 (0.156)
	50	5.931 (0.079)	1.637 (0.137)	3.501 (0.209)
	100	4.927 (0.101)	1.639 (0.140)	1.542 (0.172)
	200	4.261 (0.113)	1.564 (0.147)	0.190 (0.132)

EDF	n	LCV_1	LCV_2
2.0	25	−0.068 (0.157)	−0.643 (0.135)
	50	−1.168 (0.126)	−1.572 (0.104)
	100	−1.944 (0.112)	−2.163 (0.094)
	200	−2.921 (0.103)	−2.973 (0.096)
3.0	25	0.902 (0.130)	0.420 (0.111)
	50	0.198 (0.123)	−0.313 (0.094)
	100	−0.834 (0.108)	−1.203 (0.083)
	200	−1.514 (0.095)	−1.781 (0.073)
6.0	25	1.689 (0.108)	1.401 (0.087)
	50	1.427 (0.130)	0.956 (0.097)
	100	0.893 (0.128)	0.310 (0.096)
	200	0.057 (0.116)	-0.159 (0.097)

Table 4.14: The numbers of the times that $\hat{\lambda} \geq 10^2$ and $\hat{\lambda} \leq 10^{-6}$ out of 200 replications when the fitted functions to the data set No. 6 are used.

Design		$\hat{\lambda} \geq 10^2$					$\hat{\lambda} \leq 10^{-6}$				
EDF	n	GCV	OCV	AIC	LCV_1	LCV_2	GCV	OCV	AIC	LCV_1	LCV_2
2.0	25	26	71	59	97	129	102	3	84	0	0
	50	52	78	84	109	133	96	0	55	0	0
	100	74	89	111	113	119	39	1	7	0	0
	200	82	81	113	115	117	10	0	4	0	0
3.0	25	16	27	15	30	38	152	8	142	0	0
	50	7	7	7	8	8	69	0	54	0	0
	100	0	0	0	0	0	12	0	17	0	0
	200	0	0	0	0	0	0	0	0	0	0
6.0	25	4	10	4	11	16	182	15	164	1	0
	50	0	0	0	1	0	135	1	86	0	0
	100	0	0	0	0	0	59	0	28	0	0
	200	0	0	0	0	0	10	1	4	0	0

Table 4.15: The values of the mean and the standard error (in parentheses) of $\log \widehat{\text{AMSE}}(\hat{f})$ when the fitted functions to the data set No. 6 are used.

EDF	n	GCV	OCV	AIC
2.0	25	3.204 (0.181)	−0.233 (0.140)	2.233 (0.204)
	50	1.085 (0.212)	−1.158 (0.139)	0.043 (0.194)
	100	−1.471 (0.157)	−2.110 (0.105)	−1.950 (0.112)
	200	−2.862 (0.105)	−3.038 (0.082)	−2.878 (0.093)
3.0	25	3.211 (0.157)	−0.140 (0.110)	2.909 (0.172)
	50	0.620 (0.158)	−0.989 (0.052)	0.263 (0.151)
	100	−0.964 (0.077)	−1.266 (0.036)	−1.134 (0.089)
	200	−1.403 (0.024)	−1.408 (0.024)	−2.076 (0.055)
6.0	25	4.067 (0.108)	0.625 (0.113)	3.685 (0.137)
	50	2.380 (0.139)	−0.182 (0.062)	1.467 (0.157)
	100	0.428 (0.111)	−0.494 (0.047)	−0.416 (0.097)
	200	−0.503 (0.052)	−0.608 (0.039)	−1.580 (0.060)

EDF	n	LCV_1	LCV_2
2.0	25	−0.364 (0.130)	−0.850 (0.103)
	50	−1.210 (0.133)	−1.633 (0.111)
	100	−2.132 (0.095)	−2.213 (0.092)
	200	−2.989 (0.084)	−3.007 (0.083)
3.0	25	−0.176 (0.083)	−0.376 (0.061)
	50	−0.861 (0.077)	−1.028 (0.064)
	100	−1.440 (0.063)	−1.514 (0.059)
	200	−2.127 (0.052)	−2.153 (0.051)
6.0	25	0.440 (0.081)	0.110 (0.051)
	50	−0.074 (0.081)	−0.333 (0.057)
	100	−0.851 (0.060)	−0.961 (0.052)
	200	−1.658 (0.048)	−1.700 (0.044)

Table 4.16: The numbers of the times that $\hat{\lambda} \geq 10^2$ and $\hat{\lambda} \leq 10^{-6}$ out of 200 replications when the fitted functions to the data set No. 8 are used.

Design		$\hat{\lambda} \geq 10^2$					$\hat{\lambda} \leq 10^{-6}$				
EDF	n	GCV	OCV	AIC	LCV_1	LCV_2	GCV	OCV	AIC	LCV_1	LCV_2
2.0	25	34	72	42	93	125	138	1	121	0	0
	50	52	76	73	94	108	96	1	58	0	0
	100	81	93	102	110	115	26	0	10	0	0
	200	96	95	115	116	123	2	0	1	0	0
3.0	25	24	60	35	74	95	159	5	136	0	0
	50	18	29	26	38	43	97	1	68	1	0
	100	15	16	17	18	18	26	0	12	1	0
	200	2	2	3	3	4	2	0	2	1	0
6.0	25	6	30	13	47	62	181	13	162	0	0
	50	8	23	18	31	33	151	1	92	0	0
	100	5	5	3	5	5	74	3	21	2	0
	200	0	0	0	0	0	26	1	5	2	0

Table 4.17: The values of the mean and the standard error (in parentheses) of $\log \widehat{\mathrm{AMSE}}(\hat{f})$ when the fitted functions to the data set No. 8 are used.

EDF	n	GCV	OCV	AIC
2.0	25	3.102 (0.204)	−0.568 (0.126)	2.511 (0.206)
	50	0.810 (0.201)	−1.303 (0.116)	0.041 (0.175)
	100	−1.856 (0.145)	−2.367 (0.099)	−2.024 (0.120)
	200	−3.273 (0.099)	−3.341 (0.090)	−3.020 (0.100)
3.0	25	3.847 (0.185)	−0.212 (0.126)	3.221 (0.202)
	50	1.188 (0.195)	−1.015 (0.089)	0.572 (0.175)
	100	−1.220 (0.119)	−1.746 (0.062)	−1.316 (0.106)
	200	−1.965 (0.056)	−2.019 (0.048)	−2.248 (0.075)
6.0	25	4.685 (0.139)	0.646 (0.133)	4.258 (0.164)
	50	2.725 (0.150)	−0.094 (0.079)	1.639 (0.155)
	100	0.389 (0.129)	−0.769 (0.046)	−0.647 (0.086)
	200	−0.746 (0.075)	−1.075 (0.033)	−1.423 (0.068)

EDF	n	LCV_1	LCV_2
2.0	25	−0.623 (0.121)	−0.946 (0.106)
	50	−1.179 (0.118)	−1.440 (0.105)
	100	−2.251 (0.098)	−2.335 (0.093)
	200	−3.054 (0.098)	−3.189 (0.091)
3.0	25	−0.140 (0.101)	−0.407 (0.078)
	50	−0.744 (0.097)	−1.007 (0.077)
	100	−1.641 (0.082)	−1.758 (0.073)
	200	−2.323 (0.069)	−2.381 (0.065)
6.0	25	0.585 (0.100)	0.228 (0.076)
	50	0.063 (0.085)	−0.223 (0.060)
	100	−0.936 (0.062)	−1.064 (0.055)
	200	−1.503 (0.060)	−1.581 (0.055)

4.17 list the values of the mean and the standard error of $\log \widehat{\mathrm{AMSE}}(\hat{f})$. The GCV and AIC scores give larger values of the mean of $\log \widehat{\mathrm{AMSE}}(\hat{f})$ when sample size is small because extremely small $\hat{\lambda}$'s are chosen. The mean to the GCV and especially the AIC scores become quickly smaller as sample size becomes larger. The OCV score gives the smallest values of the mean of $\log \widehat{\mathrm{AMSE}}(\hat{f})$ when the data set No. 1 is used, and when the linear true functions to the data set No. 6 and 8 are used and sample size is large. But it gives larger values than the AIC, the LCV_1 and the LCV_2 scores do when the nonlinear true functions to the data sets No. 3, 6 and 8 are used. In many cases the LCV_2 score gives the smallest values of the mean of $\log \widehat{\mathrm{AMSE}}(\hat{f})$ and the LCV_1 score gives the second smallest values to the LCV_2 score. We think that the LCV_2 score provides better results in overall goodness-of-fit because the way of constructing it likes the AIC score.

The criterion $\mathrm{AMSE}(\hat{f})$ contains the effect of both bias and dispersion to be controlled in the process of smoothing. We attempt to divide the effect of $\mathrm{AMSE}(\hat{f})$ into two criteria: the averaged squared bias

$$\mathrm{ASB}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\{\mathrm{E}\hat{f}(t_i) - f(t_i)\}^2$$

and the averaged variance

$$\mathrm{AV}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\{\hat{f}(t_i) - \mathrm{E}\hat{f}(t_i)\}^2.$$

Tables 4.18, 4.20, 4.22 and 4.24 list the ratios of the estimates $\widehat{\mathrm{ASB}}(\hat{f})$ and $\widehat{\mathrm{AV}}(\hat{f})$ with the value

$$100 \times \frac{\widehat{\mathrm{ASB}}(\hat{f})}{\widehat{\mathrm{ASB}}(\hat{f}) + \widehat{\mathrm{AV}}(\hat{f})}.$$

Relatively higher ratios of $\widehat{\mathrm{ASB}}(\hat{f})$ to the GCV and the AIC scores when the sample size is small are due to the poor fitting. The OCV score gives much higher ratio of $\widehat{\mathrm{ASB}}(\hat{f})$ than the other three scores when the true function is nonlinear and especially when the sample size is large, which implies that the OCV score has a tendency to select $\hat{\lambda}$ so as to reduce the variance of $\hat{f}$. On the other hand the LCV_1 and the LCV_2 scores always give much lower ratios of $\widehat{\mathrm{ASB}}(\hat{f})$, which implies that these scores have a tendency to select $\hat{\lambda}$ so as to reduce the bias of $\hat{f}$.

In addition, the values $\log \widehat{\mathrm{ASB}}(\hat{f})$ and $\log \widehat{\mathrm{AV}}(\hat{f})$ are evaluated four times repeatedly for each 50 estimates $\hat{f}$ to perform the two-way analysis of variance with the EDF of three levels and the sample size of four levels as factors. Tables 4.19, 4.21, 4.23 and 4.25 show the proportions and the p-values for each factor in the analysis of variance to $\log \widehat{\mathrm{ASB}}(\hat{f})$ and $\log \widehat{\mathrm{AV}}(\hat{f})$. It can be seen that both $\log \widehat{\mathrm{ASB}}(\hat{f})$ and $\log \widehat{\mathrm{AV}}(\hat{f})$ for the GCV and the AIC scores are affected strongly by the sample size, which indicates that the fitting is improved quickly as the sample size becomes large. On the other hand $\log \widehat{\mathrm{ASB}}(\hat{f})$ for the OCV

Table 4.18: The values (in percentage) of $100 \times \widehat{\mathrm{ASB}}(\hat{f})/(\widehat{\mathrm{ASB}}(\hat{f}) + \widehat{\mathrm{AV}}(\hat{f}))$ when the fitted functions to the data set No. 1 are used.

EDF	n	GCV	OCV	AIC	LCV_1	LCV_2
2.0	25	24.18	4.63	17.96	4.39	4.76
	50	13.52	4.74	7.89	3.77	2.93
	100	4.43	3.25	3.68	3.27	3.29
	200	2.43	2.07	2.58	2.52	2.43
3.0	25	23.02	2.11	16.62	2.11	1.93
	50	11.71	3.11	8.07	2.90	2.90
	100	3.32	9.77	2.39	2.63	3.65
	200	16.27	18.65	3.49	4.15	4.21
6.0	25	26.56	4.48	22.46	3.76	4.82
	50	14.74	9.53	12.28	5.61	7.55
	100	3.42	17.96	4.10	5.46	7.65
	200	32.84	51.43	7.99	10.90	11.81

Table 4.19: The result of analysis of variance: proportions (in percentage) and p-values (in parentheses), when the fitted functions to the data set No. 1 are used.

$\log \widehat{\mathrm{ASB}}(\hat{f})$ (Averaged squared bias)

Factor	GCV		OCV		AIC	
ν (EDF)	1.90	(.0002)	10.48	(.0000)	1.04	(.0029)
n (Sample size)	90.34	(.0000)	53.86	(.0000)	95.35	(.0000)
$\nu \times n$	4.66	(.0000)	21.41	(.0000)	0.90	(.0908)
Factor	LCV_1		LCV_2			
ν	4.70	(.0171)	21.47	(.0000)		
n	69.59	(.0000)	44.28	(.0000)		
$\nu \times n$	7.18	(.0536)	11.65	(.0152)		

$\log \widehat{\mathrm{AV}}(\hat{f})$ (Averaged variance)

Factor	GCV		OCV		AIC	
ν (EDF)	0.37	(.0483)	0.55	(.3834)	0.20	(.2006)
n (Sample size)	97.06	(.0000)	88.23	(.0000)	97.38	(.0000)
$\nu \times n$	0.57	(.1501)	1.20	(.6362)	0.31	(.5289)
Factor	LCV_1		LCV_2			
ν	0.16	(.8411)	0.44	(.4856)		
n	82.13	(.0000)	87.33	(.0000)		
$\nu \times n$	0.77	(.9459)	1.45	(.5695)		

Table 4.20: The values (in percentage) of $100 \times \widehat{\mathrm{ASB}}(\hat{f})/(\widehat{\mathrm{ASB}}(\hat{f}) + \widehat{\mathrm{AV}}(\hat{f}))$ when the fitted functions to the data set No. 3 are used.

EDF	n	GCV	OCV	AIC	LCV_1	LCV_2
2.0	25	28.46	7.53	25.55	10.57	10.33
	50	17.68	5.62	8.72	4.54	4.20
	100	11.82	6.03	4.59	4.18	3.47
	200	5.14	2.77	2.24	2.12	2.43
3.0	25	51.25	13.07	45.39	14.21	10.68
	50	47.27	19.02	26.95	13.09	7.25
	100	39.54	15.35	8.24	5.88	3.34
	200	25.64	14.57	3.41	3.41	3.15
6.0	25	67.93	8.08	58.36	7.75	6.48
	50	72.72	13.13	37.46	8.84	2.16
	100	72.28	17.39	16.69	7.45	1.37
	200	69.88	19.08	7.27	5.08	3.31

Table 4.21: The result of analysis of variance: proportions (in percentage) and p-values (in parentheses), when the fitted functions to the data set No. 3 are used.

$\log \widehat{\mathrm{ASB}}(\hat{f})$ (Averaged squared bias)

Factor	GCV		OCV		AIC	
ν (EDF)	45.99	(.0000)	57.31	(.0000)	18.21	(.0000)
n (Sample size)	44.54	(.0000)	22.00	(.0000)	76.59	(.0000)
$\nu \times n$	8.75	(.0000)	11.41	(.0000)	2.30	(.0011)
Factor	LCV_1		LCV_2			
ν	30.70	(.0000)	28.28	(.0000)		
n	44.88	(.0000)	53.16	(.0000)		
$\nu \times n$	6.49	(.0689)	3.80	(.1922)		

$\log \widehat{\mathrm{AV}}(\hat{f})$ (Averaged variance)

Factor	GCV		OCV		AIC	
ν (EDF)	19.31	(.0000)	46.38	(.0000)	18.36	(.0000)
n (Sample size)	70.43	(.0000)	35.43	(.0000)	72.58	(.0000)
$\nu \times n$	9.01	(.0000)	6.68	(.0081)	6.25	(.0000)
Factor	LCV_1		LCV_2			
ν	47.17	(.0000)	51.62	(.0000)		
n	33.05	(.0000)	35.25	(.0000)		
$\nu \times n$	6.66	(.0164)	5.00	(.0059)		

Table 4.22: The values (in percentage) of $100 \times \widehat{\mathrm{ASB}}(\hat{f})/(\widehat{\mathrm{ASB}}(\hat{f}) + \widehat{\mathrm{AV}}(\hat{f}))$ when the fitted functions to the data set No. 6 are used.

EDF	n	GCV	OCV	AIC	LCV_1	LCV_2
2.0	25	25.20	5.83	18.11	6.80	5.23
	50	17.38	6.34	10.57	5.52	4.00
	100	6.33	2.43	3.31	2.13	1.93
	200	3.60	3.47	3.38	3.13	3.10
3.0	25	25.85	1.56	23.60	1.82	6.10
	50	9.31	28.27	9.02	2.29	4.76
	100	5.45	62.94	1.71	3.49	4.55
	200	81.99	81.92	4.64	6.65	6.42
6.0	25	35.08	3.72	32.93	4.59	11.04
	50	23.62	8.81	18.86	6.81	5.54
	100	5.20	46.52	5.44	5.74	7.45
	200	37.47	65.12	5.07	9.12	9.72

Table 4.23: The result of analysis of variance: proportions (in percentage) and p-values (in parentheses), when the fitted functions to the data set No. 6 are used.

$\log \widehat{\mathrm{ASB}}(\hat{f})$ (Averaged squared bias)

Factor	GCV		OCV		AIC	
ν (EDF)	4.17	(.0000)	32.27	(.0000)	3.80	(.0000)
n (Sample size)	82.98	(.0000)	14.01	(.0000)	93.51	(.0000)
$\nu \times n$	8.61	(.0000)	40.88	(.0000)	0.75	(.0524)
Factor	LCV_1		LCV_2			
ν	15.01	(.0000)	26.75	(.0000)		
n	52.56	(.0000)	49.93	(.0000)		
$\nu \times n$	20.61	(.0000)	8.12	(.0126)		

$\log \widehat{\mathrm{AV}}(\hat{f})$ (Averaged variance)

Factor	GCV		OCV		AIC	
ν (EDF)	4.32	(.0000)	5.75	(.0000)	1.92	(.0000)
n (Sample size)	90.46	(.0000)	84.45	(.0000)	95.60	(.0000)
$\nu \times n$	2.87	(.0000)	2.35	(.1081)	0.85	(.0141)
Factor	LCV_1		LCV_2			
ν	3.90	(.0002)	6.12	(.0000)		
n	86.21	(.0000)	83.06	(.0000)		
$\nu \times n$	3.69	(.0069)	3.26	(.0348)		

Table 4.24: The values (in percentage) of $100 \times \widehat{\mathrm{ASB}}(\hat{f})/(\widehat{\mathrm{ASB}}(\hat{f}) + \widehat{\mathrm{AV}}(\hat{f}))$ when the fitted functions to the data set No. 8 are used.

EDF	n	GCV	OCV	AIC	LCV_1	LCV_2
2.0	25	22.93	3.90	17.05	5.60	4.31
	50	16.66	5.27	10.00	5.61	4.86
	100	7.19	5.98	5.58	7.18	6.35
	200	2.27	2.46	2.48	2.37	2.22
3.0	25	32.95	3.04	27.01	5.78	4.90
	50	20.92	4.26	12.10	4.52	4.22
	100	3.97	20.52	4.39	3.63	4.96
	200	29.18	41.48	3.45	6.17	7.04
6.0	25	39.55	6.46	35.99	8.75	7.88
	50	27.65	7.78	18.84	8.75	9.49
	100	10.69	28.67	7.08	10.52	13.96
	200	11.10	50.32	8.07	9.91	11.97

Table 4.25: The result of analysis of variance: proportions (in percentage) and p-values (in parentheses), when the fitted functions to the data set No. 8 are used.

$\log \widehat{\mathrm{ASB}}(\hat{f})$ (Averaged squared bias)

Factor	GCV		OCV		AIC	
ν (EDF)	8.17	(.0000)	37.33	(.0000)	4.36	(.0000)
n (Sample size)	85.56	(.0000)	41.70	(.0000)	91.74	(.0000)
$\nu \times n$	3.42	(.0000)	8.85	(.0020)	0.72	(.2611)
Factor	LCV_1		LCV_2			
ν	23.89	(.0000)	37.11	(.0000)		
n	55.90	(.0000)	38.53	(.0000)		
$\nu \times n$	1.87	(.7190)	4.93	(.1991)		

$\log \widehat{\mathrm{AV}}(\hat{f})$ (Averaged variance)

Factor	GCV		OCV		AIC	
ν (EDF)	4.54	(.0000)	5.77	(.0000)	2.02	(.0001)
n (Sample size)	91.48	(.0000)	84.61	(.0000)	94.91	(.0000)
$\nu \times n$	1.89	(.0004)	2.93	(.0325)	0.31	(.6783)
Factor	LCV_1		LCV_2			
ν	9.25	(.0000)	11.78	(.0000)		
n	80.53	(.0000)	78.20	(.0000)		
$\nu \times n$	0.28	(.9837)	0.79	(.7941)		

score is affected by both of the EDF of the true function and the sample size. The values $\log \widehat{\mathrm{ASB}}(\hat{f})$ for the LCV_1 and the LCV_2 scores are also affected by both factors but the sample size gives stronger effect. The values $\log \widehat{\mathrm{AV}}(\hat{f})$ for all the five scores are affected mainly by the sample size. The proportions when the data set No. 3 is used is very different from that when the other data sets are used. If the data set No. 3 is not taken account of, it can be said that for the LCV_1 and LCV_2 scores the smoothness of the true function gives stronger effect on $\log \widehat{\mathrm{ASB}}(\hat{f})$ than $\log \widehat{\mathrm{AV}}(\hat{f})$.

To summarize, the GCV and the AIC scores often choose extremely small λ when sample size is small, although the fitting is improved as the sample size becomes larger. In some cases the OCV score shows good performance and it has a tendency to reduce the variance of estimates. The LCV_1 and the LCV_2 scores select λ more adaptively, that is, these scores have stronger possibility of estimating linear functions when the true function is linear, smooth functions when the true function is smooth and rough but not interpolating functions when the true function is rough. These scores have a tendency to reduce the bias of estimates.

4.2.2 Density Smoothing Case

Next, the simulation is performed in the context of density smoothing based on Poisson regression. Raw data x_r, $r = 1, \ldots, N$, are observed at one of the disjoint classes C_i, $i = 1, \ldots, n$, according to $\mathrm{P}(x_r \in C_i) = p_i = \int_{C_i} g(x)dx$ for some density $g(x)$, and $y_i = \#\{x_r; x_r \in C_i\}$ is counted for each class C_i.

Here we report the case that $g(x)$ is the density of gamma distribution. The gamma distribution $\mathrm{Ga}(\alpha, \beta)$ has the density

$$\tilde{g}(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right),$$

where $\Gamma(\alpha)$ is the gamma function. The values of the shape parameter α are selected as 0.5, 1 and 2. In the case of $\alpha = 1$ the gamma distribution coincides with the exponential distribution and $\log \tilde{g}(x)$ becomes a linear function, while in other cases $\log \tilde{g}(x)$ is nonlinear. The values of the scale parameter β are determined so that a random variable that follows $\mathrm{Ga}(\alpha, \beta)$ has the variance 1/25 and that more than 99% of the data extracted from $\mathrm{Ga}(\alpha, \beta)$ are observed within (0,1), and hence $\beta = \sqrt{2}/5$ if $\alpha = 0.5$, $\beta = 1/5$ if $\alpha = 1$ and $\beta = \sqrt{2}/10$ if $\alpha = 2$. For simplicity suppose that all data are observed on the interval [0,1), which is divided into n classes $C_i = [\frac{i-1}{n}, \frac{i}{n})$, $i = 1, \ldots, n$. Therefore the density that produces random numbers is truncated at $x \geq 1$ and becomes $g(x) = \tilde{g}(x)/\int_0^1 \tilde{g}(x)dx$.

The numbers n of classes are taken as 10, 20, 30, 50 and 100 and the sample size N of raw data is taken as 2, 3, 5 and 10 times of n. Two hundred data sets composed of the counts y_i, $i = 1, \ldots, n$, are produced as above and density smoothing is applied to each data sets. The values $\hat{\lambda}$ are chosen by minimizing any of the $\mathrm{GCV}(\lambda)$, $\mathrm{OCV}(\lambda)$, $\mathrm{AIC}(\lambda)$, $\mathrm{LCV}_1(\lambda)$ and $\mathrm{LCV}_2(\lambda)$ as in the previous subsection, and density functions $\hat{g}(t) = n \exp \hat{f}(t)/N$ are estimated, where $\hat{f}$ is

the natural cubic B-spline estimate with knots at class marks $t_i = (2i-1)/2n$, $i = 1, \ldots, n$.

Some boxplots of the distribution of $\hat{\lambda}$ chosen by minimizing each of the five scores are shown in Figures 4.25–4.27. As a whole $\hat{\lambda}$'s are selected better as number of classes and sample size become larger. The GCV score has a tendency to choose extremely small $\hat{\lambda}$ when number of classes or sample size is small. The other four scores do not seem to have the trouble such as the GCV score has. The OCV score provides distribution of $\hat{\lambda}$ of a little wider range than the AIC and the LCV_1 scores, and tends to choose smaller $\hat{\lambda}$ even if $\alpha = 1$ where the log of the true density function is linear. The AIC score provides distribution of $\hat{\lambda}$ of narrower range than the OCV and the LCV_1 scores. The LCV_1 score seems to select better than the OCV score but not so good as the AIC score. The LCV_2 score provides distribution of $\hat{\lambda}$ of the narrowest range of all the five scores, although it tends to select larger $\hat{\lambda}$ than the other scores especially when the number of classes is small.

To investigate overall goodness-of-fit of estimated density, the Kullback–Leibler distance

$$\mathrm{KL}(\hat{g}) = \int g(x) \log \frac{g(x)}{\hat{g}(x)} dx$$

between the true density $g(x)$ and the estimated density $\hat{g}(x)$ is evaluated. We estimate it by

$$\widehat{\mathrm{KL}}(\hat{g}) = \frac{1}{n} \sum_{i=1}^{n} g(t_i) \log \frac{g(t_i)}{\hat{g}(t_i)}.$$

The logarithms of the estimates $\widehat{\mathrm{KL}}(\hat{g})$ are taken because the distribution of it is skewed to the left side. Tables 4.26–4.28 list the values of the mean and the standard error of $\log \widehat{\mathrm{KL}}(\hat{f})$. The GCV score gives larger values of the mean of $\log \widehat{\mathrm{KL}}(\hat{f})$ but they become quickly smaller as number of classes and sample size become larger. In some cases the OCV score gives smaller values than the other scores in the case of $\alpha = 0.5$ and 2 and small number of classes, but the differences are not significant. The AIC score gives the largest values when $\alpha = 0.5$ and sample size is large. The LCV_1 score provides good overall fitting when both number of classes and sample size are small but not better than the LCV_2 score. The LCV_2 score provides the best performance in terms of Kullback–Leibler distance especially when sample size is small.

The effect of the Kullback–Leibler distance are divided into the effect of bias and dispersion, and two criteria $\widehat{\mathrm{ASB}}(\hat{f})$ and $\widehat{\mathrm{AV}}(\hat{f})$ are evaluated from 200 data sets as in the previous subsection. The three-way analysis of variance is applied to $\log \widehat{\mathrm{ASB}}(\hat{f})$ and $\log \widehat{\mathrm{AV}}(\hat{f})$ with shape of true density α of three levels, number of classes n of five levels and sample size N of four levels as factors. Table 4.29 shows the proportions and the p-values for each factor. The values $\log \widehat{\mathrm{ASB}}(\hat{f})$ and $\log \widehat{\mathrm{AV}}(\hat{f})$ for the GCV and the AIC scores have strong dependence on the number of classes and the sample size. The shape of true density affects on $\log \widehat{\mathrm{ASB}}(\hat{f})$ for the OCV, the LCV_1 and especially the LCV_2 scores rather than $\log \widehat{\mathrm{AV}}(\hat{f})$.

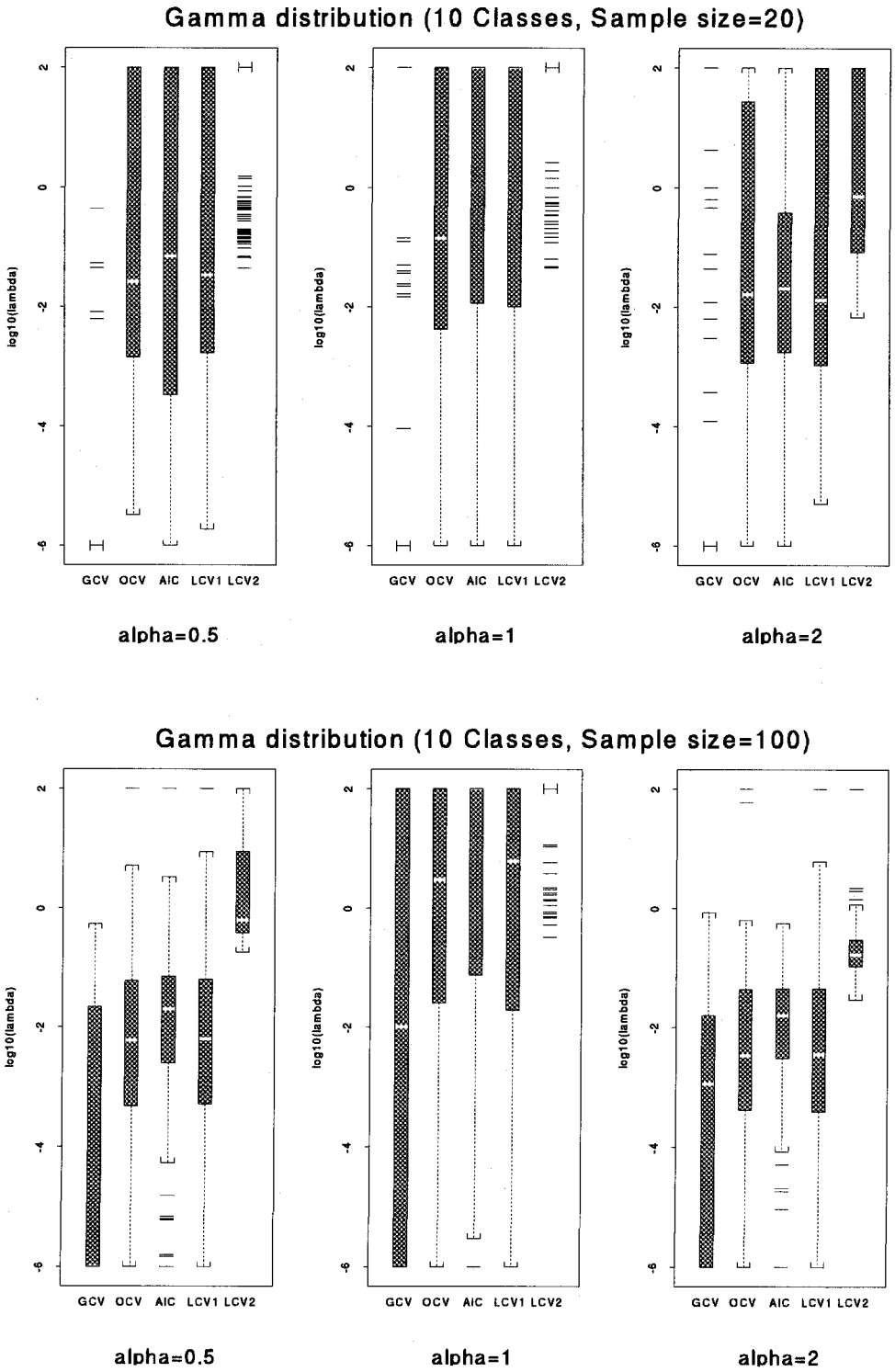
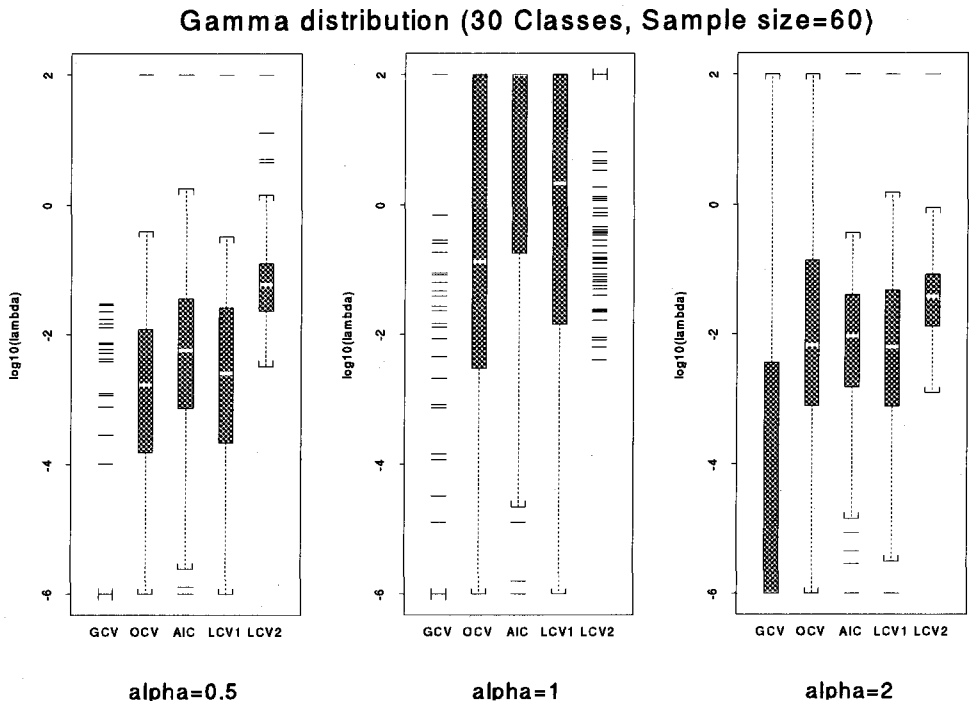


Figure 4.25: Boxplots of distribution of $\hat{\lambda}$ in the simulation of density smoothing.



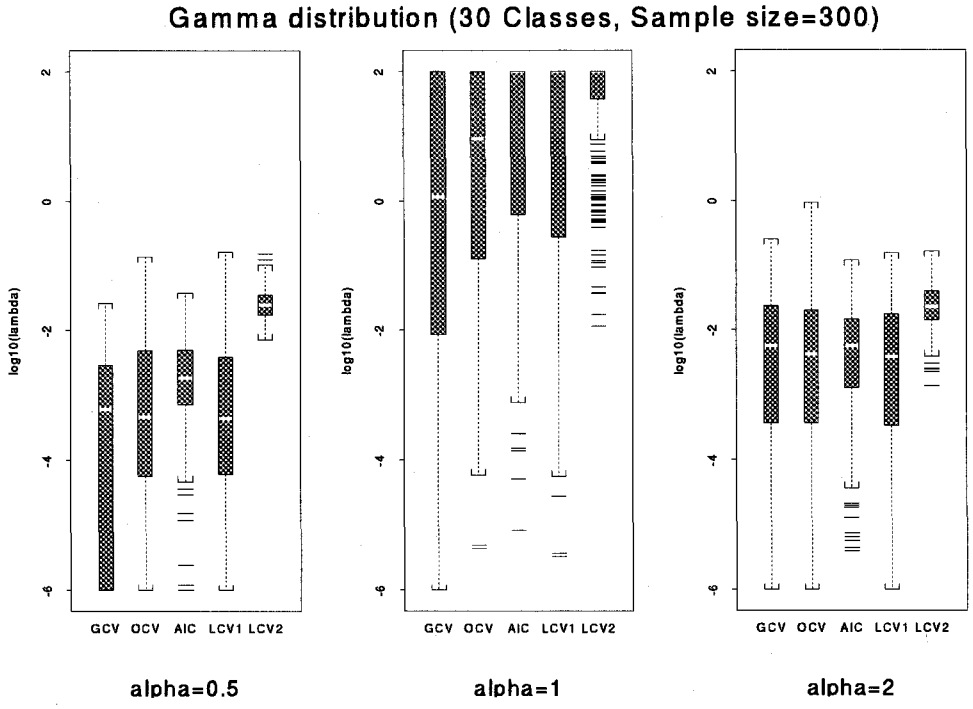


Figure 4.26: Boxplots of distribution of $\hat{\lambda}$ in the simulation of density smoothing (continued).

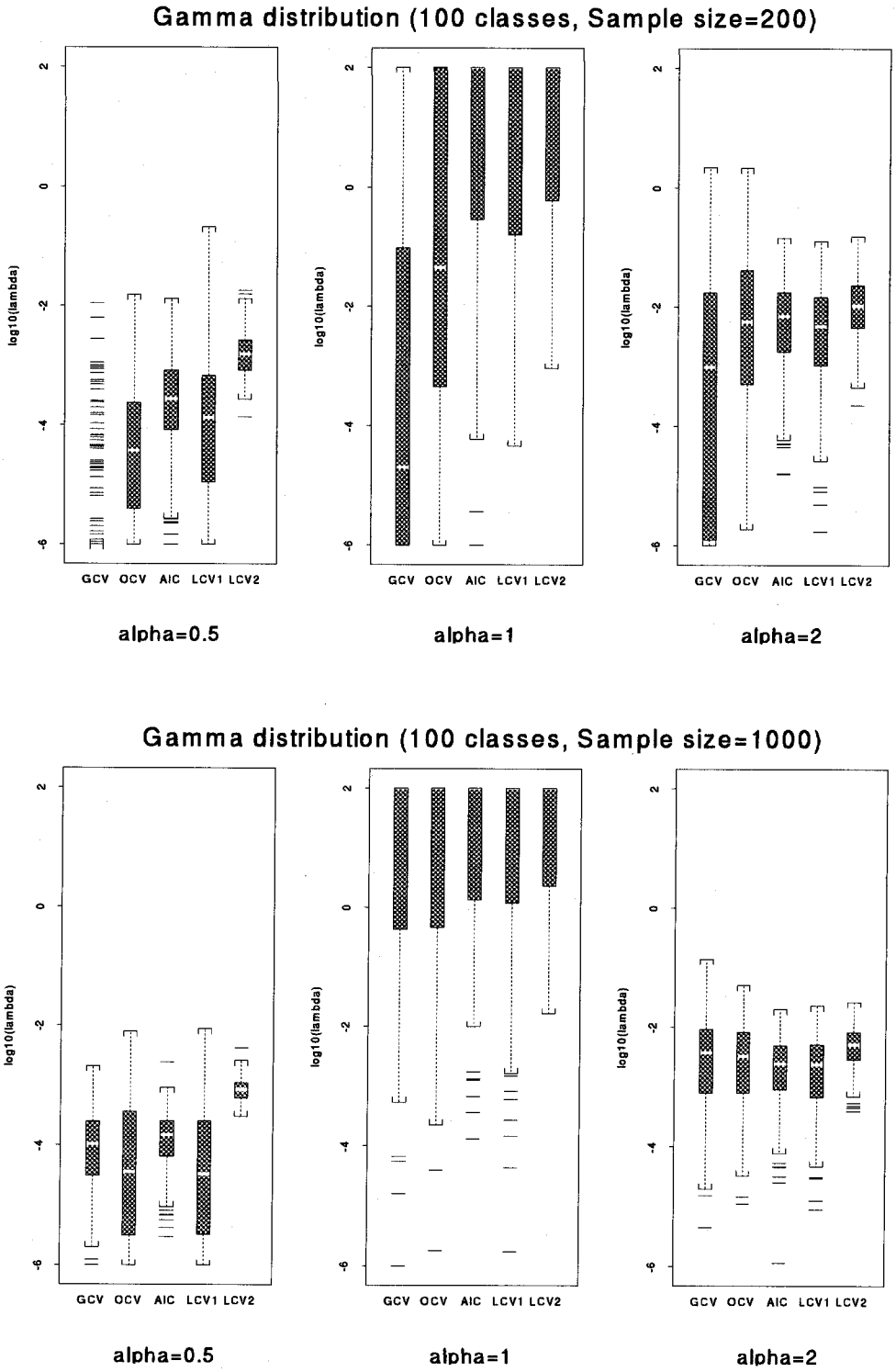


Figure 4.27: Boxplots of distribution of $\hat{\lambda}$ in the simulation of density smoothing (continued).

Table 4.26: The values of the mean and the standard error (in parentheses) of $\log \widehat{\mathrm{KL}}(\hat{f})$ when $\alpha = 0.5$.

N	n	GCV	OCV	AIC
10	20	−0.497 (0.056)	−2.727 (0.071)	−2.237 (0.089)
	30	−1.051 (0.071)	−2.876 (0.071)	−2.714 (0.080)
	50	−1.663 (0.076)	−3.227 (0.059)	−3.026 (0.079)
	100	−2.975 (0.085)	−3.744 (0.053)	−3.865 (0.060)
20	40	−0.742 (0.056)	−2.649 (0.059)	−2.540 (0.066)
	60	−1.454 (0.074)	−3.004 (0.056)	−3.007 (0.061)
	100	−2.270 (0.080)	−3.374 (0.055)	−3.502 (0.059)
	200	−3.245 (0.077)	−3.852 (0.039)	−4.001 (0.046)
30	60	−1.194 (0.058)	−2.794 (0.054)	−2.825 (0.057)
	90	−1.778 (0.066)	−3.095 (0.048)	−3.154 (0.056)
	150	−2.676 (0.078)	−3.408 (0.044)	−3.554 (0.046)
	300	−3.767 (0.071)	−3.979 (0.041)	−4.222 (0.041)
50	100	−1.681 (0.059)	−2.879 (0.044)	−3.013 (0.051)
	150	−2.254 (0.067)	−3.160 (0.047)	−3.278 (0.047)
	250	−3.199 (0.060)	−3.587 (0.037)	−3.787 (0.036)
	500	−4.200 (0.046)	−4.102 (0.034)	−4.333 (0.034)
100	200	−2.368 (0.045)	−2.990 (0.044)	−3.288 (0.036)
	300	−2.889 (0.046)	−3.349 (0.037)	−3.608 (0.031)
	500	−3.667 (0.045)	−3.785 (0.036)	−4.031 (0.026)
	1000	−4.462 (0.030)	−4.297 (0.029)	−4.569 (0.021)
N	n	LCV_1	LCV_2	
10	20	−2.721 (0.070)	−2.759 (0.044)	
	30	−2.945 (0.070)	−2.930 (0.036)	
	50	−3.206 (0.058)	−3.112 (0.034)	
	100	−3.733 (0.054)	−3.453 (0.031)	
20	40	−2.638 (0.054)	−2.795 (0.037)	
	60	−2.964 (0.056)	−3.011 (0.035)	
	100	−3.406 (0.054)	−3.387 (0.031)	
	200	−3.863 (0.039)	−3.766 (0.027)	
30	60	−2.748 (0.054)	−2.954 (0.033)	
	90	−3.091 (0.049)	−3.250 (0.031)	
	150	−3.420 (0.046)	−3.528 (0.028)	
	300	−3.998 (0.042)	−3.960 (0.025)	
50	100	−2.916 (0.045)	−3.211 (0.027)	
	150	−3.168 (0.049)	−3.396 (0.028)	
	250	−3.616 (0.038)	−3.781 (0.024)	
	500	−4.145 (0.035)	−4.212 (0.024)	
100	200	−3.110 (0.040)	−3.456 (0.022)	
	300	−3.412 (0.037)	−3.672 (0.022)	
	500	−3.813 (0.036)	−3.985 (0.021)	
	1000	−4.323 (0.030)	−4.454 (0.017)	

Table 4.27: The values of the mean and the standard error (in parentheses) of $\log \widehat{\mathrm{KL}}(\hat{f})$ when $\alpha = 1$.

N	n	GCV	OCV	AIC
10	20	−0.724 (0.084)	−3.540 (0.131)	−3.449 (0.142)
	30	−1.293 (0.092)	−4.106 (0.148)	−4.010 (0.148)
	50	−2.409 (0.152)	−4.488 (0.147)	−4.745 (0.166)
	100	−4.061 (0.164)	−5.390 (0.134)	−5.337 (0.151)
20	40	−1.206 (0.099)	−4.373 (0.148)	−4.533 (0.181)
	60	−2.151 (0.144)	−4.663 (0.147)	−4.983 (0.172)
	100	−3.423 (0.181)	−5.363 (0.137)	−5.605 (0.146)
	200	−5.012 (0.183)	−5.910 (0.150)	−6.157 (0.166)
30	60	−1.641 (0.120)	−4.612 (0.157)	−5.134 (0.167)
	90	−3.009 (0.180)	−5.114 (0.161)	−5.679 (0.179)
	150	−4.527 (0.201)	−5.786 (0.156)	−6.058 (0.163)
	300	−5.951 (0.167)	−6.484 (0.145)	−6.709 (0.144)
50	100	−2.543 (0.142)	−4.910 (0.184)	−5.832 (0.187)
	150	−4.200 (0.196)	−5.545 (0.159)	−6.190 (0.161)
	250	−5.438 (0.168)	−6.006 (0.141)	−6.615 (0.149)
	500	−6.885 (0.147)	−7.122 (0.139)	−7.200 (0.142)
100	200	−3.950 (0.181)	−5.471 (0.175)	−6.348 (0.156)
	300	−5.060 (0.179)	−6.122 (0.156)	−6.742 (0.164)
	500	−6.599 (0.171)	−6.911 (0.158)	−7.361 (0.160)
	1000	−7.764 (0.147)	−7.805 (0.146)	−7.891 (0.146)
N	n	LCV_1	LCV_2	
10	20	−3.612 (0.133)	−4.534 (0.132)	
	30	−4.146 (0.148)	−5.065 (0.132)	
	50	−4.469 (0.153)	−5.636 (0.158)	
	100	−5.337 (0.134)	−6.356 (0.128)	
20	40	−4.703 (0.171)	−5.434 (0.150)	
	60	−4.796 (0.153)	−5.768 (0.154)	
	100	−5.367 (0.142)	−6.130 (0.127)	
	200	−5.877 (0.155)	−6.815 (0.164)	
30	60	−4.843 (0.157)	−5.686 (0.151)	
	90	−5.327 (0.169)	−6.189 (0.160)	
	150	−5.832 (0.162)	−6.524 (0.157)	
	300	−6.574 (0.150)	−7.192 (0.145)	
50	100	−5.485 (0.187)	−6.330 (0.172)	
	150	−6.009 (0.174)	−6.559 (0.161)	
	250	−6.431 (0.152)	−6.927 (0.146)	
	500	−7.189 (0.156)	−7.642 (0.135)	
100	200	−6.260 (0.153)	−6.564 (0.146)	
	300	−6.676 (0.168)	−7.036 (0.155)	
	500	−7.264 (0.164)	−7.668 (0.163)	
	1000	−7.823 (0.149)	−8.098 (0.141)	

Table 4.28: The values of the mean and the standard error (in parentheses) of $\log \widehat{\mathrm{KL}}(\hat{f})$ when $\alpha = 2$.

N	n	GCV	OCV	AIC
10	20	−0.813 (0.065)	−2.588 (0.060)	−2.439 (0.080)
	30	−1.427 (0.081)	−2.822 (0.067)	−2.778 (0.078)
	50	−2.221 (0.082)	−3.203 (0.059)	−3.196 (0.069)
	100	−3.352 (0.080)	−3.776 (0.053)	−3.876 (0.059)
20	40	−1.127 (0.076)	−2.855 (0.055)	−2.793 (0.069)
	60	−1.885 (0.085)	−3.115 (0.053)	−3.153 (0.060)
	100	−2.543 (0.079)	−3.469 (0.048)	−3.543 (0.049)
	200	−3.734 (0.064)	−4.000 (0.043)	−4.153 (0.036)
30	60	−1.720 (0.083)	−2.918 (0.052)	−3.128 (0.059)
	90	−2.418 (0.087)	−3.319 (0.047)	−3.395 (0.051)
	150	−3.153 (0.077)	−3.746 (0.043)	−3.833 (0.043)
	300	−4.068 (0.058)	−4.187 (0.042)	−4.316 (0.039)
50	100	−2.388 (0.076)	−3.217 (0.044)	−3.427 (0.043)
	150	−3.066 (0.073)	−3.555 (0.048)	−3.761 (0.038)
	250	−3.740 (0.061)	−3.994 (0.038)	−4.128 (0.033)
	500	−4.510 (0.036)	−4.554 (0.030)	−4.619 (0.031)
100	200	−3.186 (0.063)	−3.665 (0.041)	−3.921 (0.034)
	300	−3.797 (0.050)	−4.006 (0.035)	−4.204 (0.029)
	500	−4.357 (0.042)	−4.456 (0.032)	−4.558 (0.029)
	1000	−4.941 (0.026)	−4.956 (0.026)	−4.999 (0.027)
N	n	LCV_1	LCV_2	
10	20	−2.588 (0.058)	−2.832 (0.041)	
	30	−2.791 (0.068)	−3.083 (0.044)	
	50	−3.178 (0.059)	−3.324 (0.044)	
	100	−3.742 (0.053)	−3.779 (0.038)	
20	40	−2.849 (0.056)	−3.176 (0.041)	
	60	−3.164 (0.056)	−3.440 (0.038)	
	100	−3.489 (0.049)	−3.700 (0.035)	
	200	−4.001 (0.044)	−4.111 (0.032)	
30	60	−3.018 (0.057)	−3.356 (0.037)	
	90	−3.352 (0.050)	−3.648 (0.035)	
	150	−3.808 (0.040)	−3.937 (0.030)	
	300	−4.200 (0.042)	−4.371 (0.027)	
50	100	−3.369 (0.045)	−3.605 (0.031)	
	150	−3.671 (0.045)	−3.843 (0.033)	
	250	−4.042 (0.037)	−4.175 (0.026)	
	500	−4.555 (0.035)	−4.679 (0.022)	
100	200	−3.882 (0.036)	−4.014 (0.027)	
	300	−4.167 (0.030)	−4.266 (0.026)	
	500	−4.530 (0.031)	−4.619 (0.024)	
	1000	−4.981 (0.026)	−5.038 (0.022)	

To summarize, in the case of density smoothing the AIC score selects $\hat{\lambda}$ in a stable way especially when sample size and number of classes are large. The LCV_1 score is effective if sample size and number of classes are small. The LCV_2 score provides the best performance in terms of the Kullback–Leibler distance, although it selects slightly larger $\hat{\lambda}$.

Table 4.29: The result of analysis of variance: proportions (in percentage) and p-values (in parentheses) in the simulation of density smoothing.

$\log \widehat{\mathrm{ASB}}(\hat{f})$ (Averaged squared bias)

Factor	GCV		OCV		AIC	
α (Shape)	6.39	(.0000)	34.17	(.0000)	19.50	(.0000)
N (Classes)	47.30	(.0000)	22.46	(.0000)	40.66	(.0000)
n (Sample size)	40.55	(.0000)	31.52	(.0000)	31.77	(.0000)
$\alpha \times N$	0.81	(.3015)	4.02	(.0005)	4.30	(.0001)
$\alpha \times n$	0.82	(.1610)	4.18	(.0001)	1.10	(.0612)
$N \times n$	2.23	(.0368)	1.46	(.2583)	0.82	(.5697)
Factor	LCV_1		LCV_2			
α	32.87	(.0000)	62.18	(.0000)		
N	30.12	(.0000)	13.82	(.0000)		
n	27.88	(.0000)	11.82	(.0000)		
$\alpha \times N$	4.75	(.0001)	6.16	(.0001)		
$\alpha \times n$	1.57	(.0192)	2.03	(.0242)		
$N \times n$	0.84	(.5981)	1.30	(.5032)		

$\log \widehat{\mathrm{AV}}(\hat{f})$ (Averaged variance)

Factor	GCV		OCV		AIC	
α (Shape)	6.68	(.0000)	19.91	(.0000)	11.43	(.0000)
N (Classes)	49.27	(.0000)	32.41	(.0000)	47.30	(.0000)
n (Sample size)	37.76	(.0000)	36.51	(.0000)	34.64	(.0000)
$\alpha \times N$	0.46	(.5662)	5.43	(.0001)	1.97	(.0183)
$\alpha \times n$	0.42	(.4228)	1.41	(.0658)	0.58	(.3571)
$N \times n$	3.81	(.0006)	1.91	(.1641)	2.08	(.0597)
Factor	LCV_1		LCV_2			
α	20.30	(.0000)	33.64	(.0000)		
N	37.38	(.0000)	23.74	(.0000)		
n	29.00	(.0000)	38.65	(.0000)		
$\alpha \times N$	6.58	(.0006)	2.12	(.0002)		
$\alpha \times n$	1.04	(.3680)	0.39	(.2176)		
$N \times n$	2.08	(.3698)	0.43	(.6243)		

Chapter 5

Conclusions and Further Developments

5.1 Conclusions

We have summarized the subject on maximum penalized likelihood estimation in non(semi)-parametric regression problems in Chapter 2. The maximum penalized likelihood estimation is a natural extension of the penalized least squares and the maximum likelihood method, and is useful as a method of estimation in non(semi)-parametric generalized linear models. The algorithm is based on Fisher scoring and is easily constructed as the iteration of the penalized least squares algorithm by using some packages with matrix manipulation. The maximum penalized likelihood estimation is also incorporated into large scale models such as generalized additive models. The equivalent degrees of freedom (EDF) for a model defined as the trace of the hat matrix indicates the number of effective parameters. The deviance and the chi-squared statistic evaluates goodness-of-fit of a model. The non(semi)-parametric generalized linear model contains the ordinary linear model at the limit as $\lambda \to \infty$ and so it is appropriate for detecting nonlinear relationships in logistic regression, Poisson regression and so on. The method of Poisson regression is also applied to density smoothing for classified data.

We have proposed the method for simple calculation of the delete-one estimate and the likelihood-based cross-validation score (LCV_1) in Chapter 3. The method is an analogy of the deletion lemma of Craven and Wahba (1979) and coincides with the one-step approximation based on the Newton–Raphson method in the case of canonical link. The AIC-like form of the likelihood-based cross-validation score (LCV_2) has been also derived. The LCV_1 and the LCV_2 scores have been compared with the one by exact calculation. They provide good approximations if the smoothing parameter is not extremely small. The simple calculation of the delete-one estimate also enables us to diagnose influential observations.

We have compared the LCV_1 and LCV_2 scores with other standard scores: the GCV, the OCV and the AIC scores, through examination of data sets in literature and simulation studies in Chapter 4. The GCV score often chooses ex-

tremely small value of the smoothing parameter and sometimes fails to choose. We think the "generalization" in the context of non-normal regression is inadequate because the weight for each observation is not homogeneous. The AIC score provides good fitting when sample size is large and in the situation of Poisson regression, but it sometimes fails in selecting the smoothing parameter when sample size is small and in the situation of logistic regression. This problem might be improved if the term of the degrees of freedom in the AIC score were modified. The OCV score diminishes the defect that GCV and AIC have to some extent and has the effect of reducing the variance of estimates.

However, in many cases the LCV_1 and the LCV_2 scores can select the smoothing parameter more adaptively. These scores take quite larger values as the smoothing parameter becomes smaller and hence they have little risk of choosing extremely small smoothing parameter. The LCV_1 score provides distribution of the smoothing parameter of a little wider range, while the LCV_2 score selects slightly larger smoothing parameter. They have the effect of reducing the bias of estimates and provide better performance in the sense of overall goodness-of-fit. We consider that these scores are useful especially in the case of small sample size and in the case of binary logistic regression, and that they are available when other scores can not select an appropriate value of the smoothing parameter.

5.2 Further Developments

It is necessary to give further examination in various cases on selecting the smoothing parameter, especially on the likelihood-based cross-validation scores. In this thesis the simple calculation of the likelihood-based cross-validation score is described only in the case of one smoothing parameter. We must consider the adaptation of the simple calculation to more complicated models such as generalized additive models where several smoothing parameters determine the smoothness of a surface. The LCV_1 score does not require any knowledge of the scale parameter and will be possible to apply to the models that contain over(under)-dispersion. In addition the method of diagnosing influential observations by the simple calculation of the delete-one estimate must be developed further.

All of the procedures for selecting the smoothing parameter that have been taken up such as the cross-validation and the AIC are based on the viewpoint of prediction. However few procedures that reflect the structure of explanatory variables have been developed. The penalized approach is familiar in the context of the ridge regression, where various procedures for selecting the ridge parameter have been developed. We want to consider procedures that make good use of the structure of a model such as the semiparametric or the additive models.

The problem on the bias of the estimate of the parameter $\boldsymbol{\beta}$ in semiparametric regression models discussed in Section 2.1.2 will be possible also in non-normal distribution case. It is necessary to investigate how the relationship between explanatory variables affects the estimation of $\boldsymbol{\beta}$ and the selection of

the smoothing parameter, and to develop a technique to reduce the bias of $\hat{\boldsymbol{\beta}}$.

In this thesis the maximum penalized likelihood estimation has been discussed in the framework of generalized linear models. We would like to take the notion of the penalized approach into quasi-likelihood in generalized estimation equations (GEE) and to consider the method for selecting the smoothing parameter such as cross-validation. Furthermore we want to study applying non(semi)-parametric approaches to proportional hazard models, random effect models and so on.

Appendices

A.1 Splines

In this section we briefly summarize splines that have been used as the fitted functions. Ichida and Yoshimoto (1979) and Green and Silverman (1994) are referred for the description here. Further description about splines are also seen in de Boor (1978), Eubank (1988) and so on.

A.1.1 Natural Splines and Smoothing Splines

Suppose that we have real numbers $t_1, \ldots, t_n$ $(n \geq 3)$ on some interval $[a, b]$ satisfying $a < t_1 < t_2 < \cdots < t_n < b$. A function f defined on $[a, b]$ is a *spline of degree* $m-1$ if f is a polynomial of degree at most $m-1$ on each of the intervals (a, t_1), (t_1, t_2), $\ldots$, (t_n, b), and if f itself and its derivatives up to order $m-2$ are continuous at each point t_i, and hence on the whole interval $[a, b]$. The points t_i are called *knots*. Cubic splines, the case of $m = 4$, are most commonly used. Furthermore the spline of degree $2m'-1$ is said to be a *natural spline* if f is a polynomial of degree at most $m'-1$ on each of the two extreme intervals (a, t_1) and (t_n, b).

When constructing a roughness penalty $J(f)$ of a function f, it is desirable that $J(f)$ is not affected by adding a constant or a lower-degree function to f. For example, $J(f) = \int_a^b \{f''(t)\}^2 dt$, the integral of the squared second derivative of f, is not affected by adding a constant or a linear function. Let $\mathcal{W}_2[a, b]$ be the space of functions that are differentiable and have an absolutely continuous first derivative on $[a, b]$. All twice-differentiable functions such that their second derivatives are square integrable are included in $\mathcal{W}_2[a, b]$. It can be proved that the function $\tilde{f}$ that minimizes $J(f) = \int_a^b \{f''(t)\}^2 dt$ over all functions f in $\mathcal{W}_2[a, b]$ that interpolate n points (t_i, f_i), $i = 1, \ldots, n$, is the natural cubic spline with knots at $t_1, \ldots, t_n$. Uniqueness of $\tilde{f}$ as such is also guaranteed. See Green and Silverman (1994), Theorem 2.3.

For the natural cubic splines f, the roughness penalty $J(f) = \int_a^b \{f''(t)\}^2 dt$ is represented as the quadratic form $J(f) = \boldsymbol{f}^T K \boldsymbol{f}$ of $\boldsymbol{f} = (f(t_1), \ldots, f(t_n))^T$ for some positive definite matrix $K = QR^{-1}Q^T$. Here, letting $h_i = t_i - t_{i-1}$ for $i = 1, \ldots, n$, Q is the $n \times (n-2)$ band matrix with entries q_{ij} for $i = 1, \ldots, n$ and $j = 2, \ldots, n-1$ given by

$$q_{i-1,i} = h_{i-1}^{-1}, \quad q_{ii} = -h_{i-1}^{-1} - h_i^{-1} \quad \text{and } q_{i+1,i} = h_i^{-1}$$

for $j = 2, \ldots, n-1$, and $q_{ij} = 0$ for $|i - j| \geq 2$, and R is the $(n-2) \times (n-2)$ band matrix with entries r_{ij} for $i, j = 2, \ldots, n-1$ given by

$$r_{ii} = \frac{h_{i-1} + h_i}{3} \quad \text{for } i = 2, \ldots, n-1,$$

$$r_{i,j+1} = r_{i+1,i} = \frac{h_i}{6} \quad \text{for } i = 2, \ldots, n-2$$

and $r_{ij} = 0$ for $|i - j| \geq 2$.

Similarly, if we denote by $W_2^{m'}[a,b]$ the space of functions that are $(m'-1)$ times differentiable and have an absolutely continuous $(m'-1)$-th derivative, the function $\tilde{f}$ that minimizes $J(f) = \int_a^b \{f^{(m')}(t)\}^2 dt$, the integral of the squared m'th derivative of f, over all functions f in $W_2^{m'}[a,b]$ that interpolate (t_i, f_i), $i = 1, \ldots, n$ uniquely exists and it is the natural spline of degree $2m'-1$ with knots at $t_1, \ldots, t_n$. In addition, for the natural spline of degree $2m'-1$, $J(f)$ is represented as the quadratic form $J(f) = \boldsymbol{f}^T K \boldsymbol{f}$ of $\boldsymbol{f} = (f(t_1), \ldots, f(t_n))^T$ for some positive definite matrix K constructed from $t_1, \ldots, t_n$. Notice that the value $f(t)$ of the natural spline for any t can be computed from the values $f(t_1), \ldots, f(t_n)$ at n knots.

Now suppose that we have n observations $y_1, \ldots, y_n$ at some design points $t_1, \ldots, t_n$, respectively. We consider the nonparametric regression problem

$$y_i = f(t_i) + \epsilon_i, \quad i = 1, \ldots, n.$$

As described in Section 2.1.1, the penalized sum of squares

$$\mathcal{S}(f) = \sum_{i=1}^{n} \{y_i - f(t_i)\}^2 + \lambda J(f)$$

is minimized over f. If we define $J(f) = \int_a^b \{f^{(m')}(t)\}^2 dt$, it can be proved that the function $\hat{f}$ that minimizes $\mathcal{S}(f)$ over $f \in W_2^{m'}[a,b]$ is the natural cubic spline of degree $2m'-1$ with knots at $t_1, \ldots, t_n$, called the *smoothing spline*. The values $\hat{f}(t_i)$, $i = 1, \ldots, n$, are given by

$$\hat{\boldsymbol{f}} = (\hat{f}(t_1), \ldots, \hat{f}(t_n))^T = (I + \lambda K)^{-1} \boldsymbol{y}, \tag{A.1}$$

which is the special case of (2.3). See Green and Silverman (1994), Theorem 2.4.

It is expensive to compute the inverse of $I + \lambda K$ in (A.1). Reinsch's (1967) algorithm makes it possible to compute $\hat{\boldsymbol{f}}$ in $O(n)$ algebraic operations. In the case of cubic smoothing splines, the equation for the $(n-2)$-vector $\boldsymbol{\gamma} = (f''(t_2), \ldots, f''(t_{n-1}))^T$ of the second derivative of f is given by

$$(R + \lambda Q^T Q)\boldsymbol{\gamma} = Q^T \boldsymbol{y} \tag{A.2}$$

and the estimates $\hat{f}(t_i)$ are obtained by

$$\boldsymbol{f} = \boldsymbol{y} - \lambda Q \boldsymbol{\gamma}.$$

The matrix $R + \lambda Q^T Q$ in (A.2) is banded with bandwidth 5 and so (A.2) is solved by the Cholesky decomposition of $R + \lambda Q^T Q$ and forward and back substitution.

A.1.2 B-splines

The use of a spline function sometimes causes the problem that the linear system to determine its coefficients become ill-conditioned. To avoid this problem, it is convenient to represent a spline function as a linear combination of some basis

functions each of which has a local support. B-splines satisfy this condition and have good computational property.

Consider the sequence of real numbers $\cdots < s_{-2} < s_{-1} < s_0 < s_1 < \cdots < s_k < s_{k+1} < \cdots$. For a given natural number m, let

$$M_m(s;t) = (t-s)_+^{m-1} = \begin{cases} (t-s)^{m-1} & \text{if } s < t \\ 0 & \text{if } s \geq t \end{cases}$$

and

$$M_m(t; s_k, \ldots, s_{k+l}) = \frac{M_m(t; s_{k+1}, \ldots, s_{k+l}) - M_m(t; s_k, \ldots, s_{k+l-1})}{s_{k+l} - s_k}$$

for $l = 2, 3, \ldots, m$. The B-spline of degree $m-1$ with knots at $s_k, \ldots, s_{k+m}$ is defined as

$$B_{mk}(t) = M_m(t; s_k, s_{k+1}, \ldots, s_{k+m}).$$

The B-spline $B_{mk}(t)$ satisfies the conditions of splines of degree $m-1$ and has the support on $[s_k, s_{k+m}]$, that is, $B_{mk} = 0$ if $t \in (-\infty, s_k] \cup [s_{k+m}, \infty)$ and $B_{mk}(t) > 0$ if $t \in (s_k, s_{k+m})$.

The B-splines can be the basis of spline functions. Let $a < s_0 < s_1 < s_2 < \cdots < s_q < b$. If $f(t)$ is the spline function of degree $m-1$ on $[a, b]$ with knots at $s_1, \ldots, s_q$, $f(t)$ is represented as the linear combination of B-splines

$$f(t) = \sum_{k=-m+1}^{q} \xi_k B_{mk}(t)$$

for some constants $\xi_{-m+1}, \ldots, \xi_q$.

To construct $B_{mk}(t)$ in practice, $2m$ additional knots $s_{-m+1} < s_{-m+2} < \cdots < s_{-1} < s_0 (\leq a)$ and $(b \leq) s_{q+1} < s_{q+2} < \cdots < s_{q+m}$ are introduced. Then the value of the B-spline $B_{mk}(t)$ is easily computed by recursive formulas

$$B_{1k}(t) = \begin{cases} (s_{k+1} - s_k)^{-1} & \text{if } s_k \leq t < s_{k+1} \\ 0 & \text{otherwise} \end{cases} \tag{A.3}$$

and

$$B_{rk}(t) = \frac{(t - s_k) B_{r-1,k}(t) + (s_{k+r} - t) B_{r-1,k+1}(t)}{s_{k+r} - s_k} \tag{A.4}$$

for $r = 2, 3, \ldots, m$. This is called de Boor–Cox's algorithm. See Ichida and Yoshimoto (1979), Section 3.3. For example, if the value of the cubic B-spline ($m = 4$) at some $t \in [a, b]$ is wanted, find k such that $s_k \leq t < s_{k+1}$ at first, and compute $B_{1k}(t)$ from (A.3). Then from (A.4) obtain $B_{2k}(t)$ and $B_{2,k+1}(t)$; next $B_{3k}(t)$, $B_{3,k+1}$ and $B_{3,k+2}(t)$; and at last $B_{4k}(t), \ldots, B_{4,k+3}(t)$. The B-splines not listed here are all zero in $[s_k, s_{k+1})$. In such a way the basis matrix using B-splines $B = \{B_{mk}(t_i)\}$ can be easily constructed for any sequence of design points $t_1, \ldots, t_n$. Assumptions on the uniqueness and the order of $t_1, \ldots, t_n$ are not necessary.

Correction to satisfy the condition of natural splines is also possible. For example, the natural cubic spline $f(t)$ with knots at $s_1, \ldots, s_q$ is represented as the linear combination of q B-splines $f(t) = \sum_{k=1}^{q} \xi_k \tilde{B}_{4,k+2}(t)$, where $\tilde{B}_{4,k+2}(t)$ for $k = 1, 2, q-1$ and q are adjusted from $B_{4,k+2}(t)$ so that they are linear on $[a, s_1)$ and $(s_q, b]$ and twice differentiable at $t = s_1$ and s_q. Notice that $\tilde{B}_{4,k+2}(t)$ is the same as $B_{4,k+2}(t)$ for $k = 3, \ldots, q-2$. Especially in constructing a smoothing spline, the number of B-spline basis q is set equal to n and knots are taken at $t_1, \ldots, t_n$.

The roughness penalty $J(f) = \int_a^b \{f^{(m')}(t)\}^2 dt$ for the B-spline can be also computed. The case of natural cubic splines ($m' = 2$) is described. If a natural cubic spline $f(t)$ is written as $f(t) = \sum_{k=1}^{q} \xi_k \varphi_k(t)$ where $\varphi_k(t) = \tilde{B}_{4,k+2}(t)$, the penalty is represented as $J(f) = \boldsymbol{\xi}^T K \boldsymbol{\xi}$, where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_q)^T$ and K is the $q \times q$ symmetric band matrix with (k, l)th component

$$K_{kl} = \int_a^b \varphi_k''(t)\varphi_l''(t)dt.$$

Notice that $K_{kl} = 0$ if $|j - k| \geq 3$ since $\tilde{B}_{4,k+2}(t)$ has the support on $[s_k, s_{k+4}]$.

At first compute the second derivatives of B-spline basis at the knots by

$$\begin{aligned}
\varphi_{k-1}''(s_k) &= \frac{6}{(s_{k+1} - s_{k-3})(s_{k+1} - s_{k-2})(s_{k+1} - s_{k-1})}, \\
\varphi_k''(s_k) &= -\frac{6}{(s_{k+2} - s_{k-2})(s_{k+1} - s_{k-1})}\left(\frac{1}{s_{k+1} - s_{k-2}} + \frac{1}{s_{k+2} - s_{k-1}}\right), \\
\varphi_{k+1}''(s_k) &= \frac{6}{(s_{k+3} - s_{k-1})(s_{k+2} - s_{k-1})(s_{k+1} - s_{k-1})}
\end{aligned}$$

for $k = 2, \ldots, q-1$, and put $\varphi_k''(s_j) = 0$ for the other combinations of (j, k). On each interval (s_j, s_{j+1}), the second derivatives $\varphi_k''(t)$ for $k = j, \ldots, j+3$ are linear functions

$$\varphi_k''(t) = \frac{\varphi_k''(s_{j+1}) - \varphi_k''(s_j)}{s_{j+1} - s_j}(t - s_j) + \varphi_k''(s_j)$$

and hence the partitioned integral $\int_{s_j}^{s_{j+1}} \varphi_k''(t)\varphi_l''(t)dt$ is exactly computed as the integral of a quadratic function. Therefore the penalty matrix K can be obtained.

If the B-splines are used as the basis functions, the smoother matrix $S_\lambda = B(B^T W B + \lambda K)^{-1} B^T W$ is easily constructed. Since the basis matrix B contains many zero entries and K is banded, $B^T W B + \lambda K$ is also banded. If cubic B-splines are used, $B^T W B + \lambda K$ has the bandwidth 7. Therefore, once $B^T W B$ and $B^T W \boldsymbol{y}$ are computed, the Cholesky decomposition of $B^T W B + \lambda K$ and forward and back substitution make it possible to compute the estimate of the coefficient vector $\hat{\boldsymbol{\xi}} = (B^T W B + \lambda K)^{-1} B^T W \boldsymbol{y}$ in $O(q)$ operations. Each leverage value $A_{ii} = w_i \boldsymbol{b}_i^T (B^T W B + \lambda K)^{-1} \boldsymbol{b}_i$ is also computed in $O(q)$ operations.

A.2 Data

Data sets taken up in Section 2.4

Kyphosis in Laminectomy Patients (Hastie and Tibshirani, 1990; Chambers and Hastie, 1992)
The table below shows the data on 83 patients (almost children) undergoing laminectomy surgery. Each of the data is composed of a response that indicates presence or absence of kyphosis after the operation, and three variables: age of patients (`age`), the number of vertebrae levels involved in the surgery (`number`), and the starting vertebrae level of the surgery (`start`).

Obs.	Kyphosis	Age	Number	Start	Obs.	Kyphosis	Age	Number	Start
1	absent	71	3	5	43	present	73	5	1
2	absent	158	3	14	44	absent	35	3	13
3	present	128	4	5	45	absent	143	9	3
4	absent	2	5	1	46	absent	61	4	1
5	absent	1	4	15	47	absent	97	3	16
6	absent	1	2	16	48	present	139	3	10
7	absent	61	2	17	49	absent	136	4	15
8	absent	37	3	16	50	absent	131	5	13
9	absent	113	2	16	51	present	121	3	3
10	present	59	6	12	52	absent	177	2	14
11	present	82	5	14	53	absent	68	5	10
12	absent	148	3	16	54	absent	9	2	17
13	absent	18	5	2	55	present	139	10	6
14	absent	1	4	12	56	absent	2	2	17
15	absent	243	8	8	57	absent	140	4	15
16	absent	168	3	18	58	absent	72	5	15
17	absent	1	3	16	59	absent	2	3	13
18	absent	78	6	15	60	present	120	5	8
19	absent	175	5	13	61	absent	51	7	9
20	absent	80	5	16	62	absent	102	3	13
21	absent	27	4	9	63	present	130	4	1
22	absent	22	2	16	64	present	114	7	8
23	present	105	6	5	65	absent	81	4	1
24	present	96	3	12	66	absent	118	3	16
25	absent	131	2	3	67	absent	118	4	16
26	present	15	7	2	68	absent	17	4	10
27	absent	9	5	13	69	absent	195	2	17
28	present	12	14	2	70	absent	159	4	13
29	absent	8	3	6	71	absent	18	4	11
30	absent	100	3	14	72	absent	15	5	16
31	absent	4	3	16	73	absent	158	5	14
32	absent	151	2	16	74	absent	127	4	12
33	absent	31	3	16	75	absent	87	4	16
34	absent	125	2	11	76	absent	206	4	10
35	absent	130	5	13	77	absent	11	3	15
36	absent	112	3	16	78	absent	178	4	15
37	absent	140	5	11	79	present	157	3	13
38	absent	93	3	16	80	absent	26	7	13
39	absent	1	3	9	81	absent	120	2	13
40	present	52	5	6	82	present	42	7	6
41	absent	20	6	9	83	absent	36	4	13
42	present	91	5	12					

Assay of Trypanosome (Ashford and Walker, 1972)
The table below shows data on the numbers of trypanosome organisms killed at different doses of a certain poison.

Dose	Observations	Killed
4.7	55	0
4.8	49	8
4.9	60	18
5.0	55	18
5.1	53	22
5.2	53	37
5.3	51	47
5.4	50	50

Mortality Table (Green and Silverman, 1994)
The table below gives, for a particular population of retired American white females, the age structure of the population and the annualized number of deaths in each age group. The column `size` gives the population size, and the column `death` gives the number of death.

age	size	death	age	size	death	age	size	death
55	84	1	72	20116	480	89	510	97
56	418	2	73	18876	537	90	430	93
57	1066	10	74	17461	566	91	362	75
58	2483	21	75	15012	581	92	291	84
59	3721	35	76	11871	464	93	232	31
60	5460	62	77	10002	461	94	196	75
61	6231	50	78	8949	433	95	147	29
62	8061	55	79	7751	515	96	100	25
63	9487	88	80	6140	374	97	161	20
64	10770	132	81	4718	348	98	11	5
65	24267	267	82	3791	304	99	10	3
66	26791	300	83	2806	249	100	8	2
67	29174	432	84	2240	167	101	5	0
68	28476	491	85	1715	192	102	4	2
69	25840	422	86	1388	171	103	2	0
70	23916	475	87	898	126	104	2	1
71	21412	413	88	578	86			

Duration of Eruption of Old Faithful Geyser (Silverman, 1986)
The data set is the duration (in minutes) of 107 eruptions of Old Faithful geyser in Yellow National Park, USA.

4.37	3.87	4.00	4.03	3.50	4.08	2.25	4.70	1.73	4.93
1.73	4.62	3.43	4.25	1.68	3.92	3.68	3.10	4.03	1.77
4.08	1.75	3.20	1.85	4.62	1.97	4.50	3.92	4.35	2.33
3.83	1.88	4.60	1.80	4.73	1.77	4.57	1.85	3.52	4.00
3.70	3.72	4.25	3.58	3.80	3.77	3.75	2.50	4.50	4.10
3.70	3.80	3.43	4.00	2.27	4.40	4.05	4.25	3.33	2.00
4.33	2.93	4.58	1.90	3.58	3.73	3.73	1.82	4.63	3.50
4.00	3.67	1.67	4.60	1.67	4.00	1.80	4.42	1.90	4.63
2.93	3.50	1.97	4.28	1.83	4.13	1.83	4.65	4.20	3.93
4.33	1.83	4.53	2.03	4.18	4.43	4.07	4.13	3.95	4.10
2.72	4.58	1.90	4.50	1.95	4.83	4.12			

Data sets examined in Chapter 4

1, 2: Tumor Prevalence (Green and Yandell, 1986)
The table below show the data from the toxicology experiment in which 207 male and 112 female rats were exposed to six dose levels of flame retardant, and presence or absence of bileduct hyperplasia at death of each rat was reported as the zero-one response (presence is indicated by 1). Here only one explanatory variable **Age** at death (in month) is employed, although the original data contain the dose levels and other two explanatory variables.

Male rats

Obs.	Dose	Age	Resp.	Obs.	Dose	Age	Resp.	Obs.	Dose	Age	Resp.	Obs.	Dose	Age	Resp.
1	0	75	0	53	1	112	0	105	2	97	1	157	4	105	0
2	0	100	0	54	1	113	0	106	2	99	1	158	4	108	0
3	0	101	0	55	1	117	0	107	2	106	1	159	4	108	0
4	0	102	0	56	1	118	0	108	2	106	1	160	4	112	0
5	0	104	0	57	1	118	0	109	2	108	1	161	4	112	0
6	0	106	0	58	1	122	0	110	2	111	1	162	4	112	0
7	0	107	0	59	1	122	0	111	2	113	1	163	4	77	1
8	0	108	0	60	1	123	0	112	3	69	0	164	4	85	1
9	0	108	0	61	1	123	0	113	3	79	0	165	4	93	1
10	0	109	0	62	1	123	0	114	3	80	0	166	4	93	1
11	0	111	0	63	1	123	0	115	3	85	0	167	4	95	1
12	0	111	0	64	1	77	1	116	3	90	0	168	4	96	1
13	0	111	0	65	1	81	1	117	3	90	0	169	4	96	1
14	0	111	0	66	1	94	1	118	3	91	0	170	4	98	1
15	0	112	0	67	1	99	1	119	3	94	0	171	4	98	1
16	0	113	0	68	1	100	1	120	3	97	0	172	4	98	1
17	0	116	0	69	1	102	1	121	3	97	0	173	4	98	1
18	0	117	0	70	1	108	1	122	3	100	0	174	4	102	1
19	0	119	0	71	1	112	1	123	3	103	0	175	4	103	1
20	0	121	0	72	1	115	1	124	3	107	0	176	4	104	1
21	0	122	0	73	2	85	0	125	3	107	0	177	5	42	0
22	0	122	0	74	2	90	0	126	3	110	0	178	5	73	0
23	0	123	0	75	2	93	0	127	3	112	0	179	5	77	0
24	0	123	0	76	2	94	0	128	3	117	0	180	5	84	0
25	0	123	0	77	2	95	0	129	3	118	0	181	5	84	0
26	0	71	1	78	2	99	0	130	3	78	1	182	5	85	0
27	0	71	1	79	2	104	0	131	3	86	1	183	5	86	0
28	0	84	1	80	2	105	0	132	3	87	1	184	5	87	0
29	0	103	1	81	2	105	0	133	3	88	1	185	5	90	0
30	0	118	1	82	2	105	0	134	3	91	1	186	5	91	0
31	0	122	1	83	2	106	0	135	3	92	1	187	5	92	0
32	0	123	1	84	2	106	0	136	3	97	1	188	5	92	0
33	0	123	1	85	2	107	0	137	3	98	1	189	5	94	0
34	1	54	0	86	2	107	0	138	3	101	1	190	5	94	0
35	1	87	0	87	2	108	0	139	3	102	1	191	5	94	0
36	1	89	0	88	2	109	0	140	3	106	1	192	5	94	0
37	1	101	0	89	2	110	0	141	3	111	1	193	5	95	0
38	1	102	0	90	2	111	0	142	3	112	1	194	5	97	0
39	1	102	0	91	2	112	0	143	4	62	0	195	5	97	0
40	1	103	0	92	2	112	0	144	4	83	0	196	5	101	0
41	1	105	0	93	2	112	0	145	4	85	0	197	5	101	0
42	1	105	0	94	2	112	0	146	4	87	0	198	5	103	0
43	1	107	0	95	2	113	0	147	4	92	0	199	5	44	1
44	1	107	0	96	2	113	0	148	4	94	0	200	5	74	1
45	1	107	0	97	2	113	0	149	4	95	0	201	5	81	1
46	1	108	0	98	2	113	0	150	4	98	0	202	5	82	1
47	1	108	0	99	2	115	0	151	4	99	0	203	5	87	1
48	1	108	0	100	2	116	0	152	4	100	0	204	5	94	1
49	1	110	0	101	2	116	0	153	4	100	0	205	5	96	1
50	1	110	0	102	2	78	1	154	4	102	0	206	5	99	1
51	1	111	0	103	2	87	1	155	4	103	0	207	5	101	1
52	1	112	0	104	2	94	1	156	4	104	0				

Female rats

Obs.	Dose	Age	Resp.	Obs.	Dose	Age	Resp.	Obs.	Dose	Age	Resp.	Obs.	Dose	Age	Resp.
1	0	50	0	29	1	106	0	57	2	123	0	85	4	123	0
2	0	59	0	30	1	108	0	58	2	123	0	86	4	123	0
3	0	77	0	31	1	108	0	59	2	123	0	87	4	124	0
4	0	77	0	32	1	113	0	60	2	124	0	88	4	124	0
5	0	84	0	33	1	114	0	61	2	124	0	89	4	81	1
6	0	87	0	34	1	116	0	62	2	124	0	90	4	81	1
7	0	106	0	35	1	116	0	63	3	112	0	91	4	102	1
8	0	106	0	36	1	123	0	64	3	114	0	92	4	105	1
9	0	111	0	37	1	123	0	65	3	120	0	93	5	36	0
10	0	111	0	38	1	124	0	66	3	122	0	94	5	41	0
11	0	111	0	39	1	124	0	67	3	123	0	95	5	86	0
12	0	112	0	40	1	124	0	68	3	123	0	96	5	94	0
13	0	114	0	41	1	124	0	69	3	123	0	97	5	103	0
14	0	114	0	42	2	95	0	70	3	123	0	98	5	103	0
15	0	119	0	43	2	96	0	71	3	124	0	99	5	105	0
16	0	120	0	44	2	100	0	72	3	124	0	100	5	106	0
17	0	123	0	45	2	103	0	73	3	96	1	101	5	108	0
18	0	124	0	46	2	103	0	74	4	87	0	102	5	109	0
19	0	98	1	47	2	104	0	75	4	92	0	103	5	113	0
20	0	108	1	48	2	109	0	76	4	94	0	104	5	116	0
21	1	78	0	49	2	110	0	77	4	109	0	105	5	117	0
22	1	82	0	50	2	113	0	78	4	109	0	106	5	34	1
23	1	83	0	51	2	119	0	79	4	110	0	107	5	74	1
24	1	98	0	52	2	119	0	80	4	112	0	108	5	98	1
25	1	99	0	53	2	120	0	81	4	121	0	109	5	106	1
26	1	101	0	54	2	122	0	82	4	123	0	110	5	112	1
27	1	103	0	55	2	123	0	83	4	123	0	111	5	115	1
28	1	105	0	56	2	123	0	84	4	123	0	112	5	118	1

3: Kyphosis in Laminectomy Patients See p. 109.

4: Dose Levels for the Treatment of Testicular Cancer (Brown and Hu, 1980)

The table below is on 55 courses of treatment received by 14 patients with testicular cancer. Each patient received one to six courses of a combination of three days. Here the dose of velban used for each patient is given in mg/day per body weight (in kg) with the occurrence of serious gastrointestinal toxicity.

Patient	Course	Dose	Toxicity	Patient	Course	Dose	Toxicity	Patient	Course	Dose	Toxicity
A	1	201	Yes	F	1	184	No	J	1	158	No
	2	207	Yes		2	181	No		2	159	No
B	1	199	Yes		3	183	No		3	152	No
C	1	199	Yes		4	143	No		4	154	No
	2	111	No		5	89	No		5	171	No
	3	218	Yes	G	1	132	No	L	1	185	No
	4	123	No		2	147	No		2	186	Yes
	5	124	No		3	177	Yes		3	156	No
	6	122	No		4	188	No		4	156	No
D	1	202	Yes		5	182	No		5	156	No
	2	204	Yes		6	143	No	M	1	188	Yes
	3	214	Yes	H	1	189	Yes		2	194	No
	4	144	Yes		2	123	No		3	203	No
E	1	218	Yes		3	129	No	N	1	217	Yes
	2	191	No		4	125	No	O	1	181	No
	3	180	No		5	126	No		2	127	No
	4	183	No	I	1	220	Yes		3	156	No
	5	179	Yes		2	208	Yes		4	156	Yes
	6	173	No								

5, 6: Nodal Involvement for Prostate Cancer Patients (Brown, 1980)
The table below gives the data on 53 prostate cancer patients receiving surgery. Here for each case presence (1) or absence (0) of nodal involvement is given with two quantitative variables to predict whether or not the lymph nodes were affected: age at diagnosis and level of serum acid phosphatase (ACP) ($\times 100$).

Case	Age	ACP	Nodes	Case	Age	ACP	Nodes	Case	Age	ACP	Nodes
1	66	48	0	19	52	83	0	37	59	63	0
2	68	56	0	20	56	98	0	38	61	102	0
3	66	50	0	21	67	52	0	39	53	76	0
4	56	52	0	22	63	75	0	40	67	95	0
5	58	50	0	23	59	99	1	41	53	66	0
6	60	49	0	24	64	187	0	42	65	84	1
7	65	46	0	25	61	136	1	43	50	81	1
8	60	62	0	26	56	82	1	44	60	76	1
9	50	56	1	27	64	40	0	45	45	70	1
10	49	55	0	28	61	50	0	46	56	78	1
11	61	62	0	29	64	50	0	47	46	70	1
12	58	71	0	30	63	40	0	48	67	67	1
13	51	65	0	31	52	55	0	49	63	82	1
14	67	67	1	32	66	59	0	50	57	67	1
15	67	47	0	33	58	48	1	51	51	72	1
16	51	49	0	34	57	51	1	52	64	89	1
17	56	50	0	35	65	49	1	53	68	126	1
18	60	78	0	36	65	48	0				

7, 8: Effect of Process and Purity Index on Fault Occurrence (Cox and Snell, 1981)
Batches of raw material were selected and each batch was divided into two equal sections: for each batch, one of the sections was processed by the standard method and the other by a slightly modified process. Before processing, a purity index was measured for the whole batch of material. For the product from each section of material it was recorded whether the minor faults did (F) or did not occur (NF). Results for 22 batches are given in the table below.

Purity index	Standard process	Modified process	Purity index	Standard process	Modified process
7.2	NF	NF	6.5	NF	F
6.3	F	NF	4.9	F	F
8.5	F	NF	5.3	F	NF
7.1	NF	F	7.1	NF	F
8.2	F	NF	8.4	F	NF
4.6	F	NF	8.5	NF	F
8.5	NF	NF	6.6	F	NF
6.9	F	F	9.1	NF	NF
8.0	NF	NF	7.1	F	NF
8.0	F	NF	7.5	NF	F
9.1	NF	NF	8.3	NF	NF

9, 18: Mortality Table See p. 110.

10, 11: Incidence of Pneumoconiosis among Coalminers (Cox and Snell, 1981)
The table below is on 18,282 coalminers who were known to be smokers but who showed no radiological abnormality were grouped by age and classified according to whether or not they showed signs of two symptoms, breathlessness and wheeze.

Breathlessness		Yes		No		Total
Wheeze		Yes	No	Yes	No	
Age group	20–24	9	7	95	1,841	1,952
(years)	25–29	23	9	105	1,654	1,791
	30–34	54	19	177	1,863	2,113
	35–39	121	48	257	2,357	2,783
	40–44	169	54	273	1,778	2,274
	45–49	269	88	324	1,712	2,393
	50–54	404	117	245	1,324	2,090
	55–59	406	152	225	967	1,750
	60–64	372	106	132	526	1,136
Total		1,827	600	1,833	14,022	18,282

12: Assay of Trypanosome See p. 110.

13: Quantal Response Data (Thompson, 1947; Kawai, 1997)
The data is on 350 male mice that were divided into 7 groups of each 50 mice and 7 dose levels of toxin were injected. The number of death is observed for each dose levels.

Dose	Subjects	Responses
1.0625	50	6
1.125	50	7
1.25	50	33
1.5	50	39
2	50	45
3	50	50
5	50	50

14: Quantal Response Data (Finney, 1971; Kawai, 1997)
The data set below counts the numbers of death out of 50 macrosiphoniella sanbornis on a day after 6 levels of the concentration (mg/l) of deguelin were sprayed on them.

Dose	Subjects	Responses
2.6358941	49	16
3	48	18
3.4794154	48	34
3.7894873	49	47
4.0525184	50	47
4.2490096	48	48

15: Quantal Response Data (Bliss, 1938; Kawai, 1997)
Eight levels of the concentration of gelsemicine hydrochloride were prescribed to 80 rabbits and the numbers of death were observed.

Dose	Subjects	Responses
0.06	10	0
0.07	10	1
0.08	10	3
0.09	10	6
0.1	10	8
0.11	10	5
0.12	10	9
0.13	10	10

16: Quantal Response Data (Reed and Muench, 1938)
The number of death observed in each group composed of 6 subjects for 9 levels of stimulus was observed.

Dose	Subjects	Responses
1	6	0
2	6	0
4	6	1
8	6	0
16	6	2
32	6	4
64	6	4
128	6	6
256	6	5

17: Numbers of Coal-mining Disasters (Jarrett, 1979)
Jarrett (1979) corrected the data on time intervals between 191 coal mine explosions involving more than ten men killed in Great Britain between 1851 and 1962. The table below rewrites it into the count of disasters in each year.

Year	Count	Year	Count	Year	Count	Year	Count	Year	Count
1851	4	1874	4	1897	0	1919	0	1941	3
1852	5	1875	4	1898	0	1920	0	1942	3
1853	4	1876	1	1899	1	1921	0	1943	0
1854	1	1877	5	1900	0	1922	2	1944	0
1855	0	1878	5	1901	1	1923	1	1945	0
1856	4	1879	3	1902	1	1924	0	1946	1
1857	3	1880	4	1903	0	1925	0	1947	4
1858	4	1881	2	1904	0	1926	0	1948	0
1859	0	1882	5	1905	3	1927	1	1949	0
1860	6	1883	2	1906	1	1928	1	1950	0
1861	3	1884	2	1907	0	1929	0	1951	1
1862	3	1885	3	1908	3	1930	2	1952	0
1863	4	1886	4	1909	2	1931	3	1953	0
1864	0	1887	2	1910	2	1932	3	1954	0
1865	2	1888	1	1911	0	1933	1	1955	0
1866	6	1889	3	1912	1	1934	1	1956	0
1867	3	1890	2	1913	1	1935	2	1957	1
1868	3	1891	2	1914	1	1936	1	1958	0
1869	5	1892	1	1915	0	1937	1	1959	0
1870	4	1893	1	1916	1	1938	1	1960	1
1871	5	1894	1	1917	0	1939	1	1961	0
1872	3	1895	1	1918	1	1940	2	1962	1
1873	1	1896	3						

19, 20: Hodgkin's Disease Mortality Data (Selvin, 1994)
The table below lists the numbers of death by Hodgkin's disease observed in males and females for residents of California in 1989.

	Male		Female	
Age	Person-Years	Death	Person-Years	Death
30–34	1,299,868	55	1,300,402	37
35–39	1,240,595	49	1,217,896	29
40–44	1,045,453	38	1,045,801	23
45–49	795,776	26	810,260	12
50–54	645,991	19	665,612	7
55–59	599,729	17	633,646	12
60–64	568,109	22	650,686	9
65–69	506,475	21	600,455	19
70–74	368,751	18	474,609	13
75–79	252,581	11	376,781	14
80–84	140,053	10	255,412	5
85+	81,850	4	313,603	3
Total	7,545,231	290	8,345,163	183

21: Time Intervals between Mine Explosions (Simonoff, 1996)
The table below gives counts that correspond to a discretization into 5 cells of 109 time intervals between explosions in mines from December 8, 1875 to May 29, 1951, which is probably a part of the data set No.17.

Cell	Days	Count	Cell	Days	Count	Cell	Days	Count
1	0– 30	18	20	571– 600	1	38	1111–1140	0
2	31– 60	14	21	601– 630	0	39	1141–1170	0
3	61– 90	9	22	631– 660	0	40	1171–1200	0
4	91–120	8	23	661– 690	1	41	1201–1230	1
5	121–150	6	24	691– 720	0	42	1231–1260	0
6	151–180	4	25	721– 750	0	43	1261–1290	0
7	181–210	6	26	751– 780	1	44	1291–1320	1
8	211–240	7	27	781– 810	0	45	1321–1350	0
9	241–270	1	28	811– 840	0	46	1351–1380	1
10	271–300	6	29	841– 870	0	47	1381–1410	0
11	301–330	6	30	871– 900	0	48	1411–1440	0
12	331–360	5	31	901– 930	1	49	1441–1470	0
13	361–390	5	32	931– 960	0	50	1471–1500	0
14	391–420	0	33	961– 990	0	51	1501–1530	0
15	421–450	0	34	991–1020	0	52	1531–1560	0
16	451–480	2	35	1021–1050	0	53	1561–1590	0
17	481–510	1	36	1051–1080	0	54	1591–1620	1
18	511–540	1	37	1081–1110	0	55	1621–1650	1
19	541–570	1						

22: Pain Scores Data (Efron and Tibshirani, 1996)
The table below shows the discretized version of pain scores for 67 women, each obtained by averaging the results from a questionnaire administered after an operation. The score 0 means no pain and the score 4 means worst pain. The domain [0,4] is partitioned into 40 cells of length 0.1.

Cell	Scores	Count	Cell	Scores	Count	Cell	Scores	Count
1	0.0– 0.1	3	15	1.4– 1.5	3	28	2.7– 2.8	0
2	0.1– 0.2	7	16	1.5– 1.6	5	29	2.8– 2.9	1
3	0.2– 0.3	6	17	1.6– 1.7	0	30	2.9– 3.0	1
4	0.3– 0.4	1	18	1.7– 1.8	1	31	3.0– 3.1	1
5	0.4– 0.5	2	19	1.8– 1.9	0	32	3.1– 3.2	0
6	0.5– 0.6	3	20	1.9– 2.0	0	33	3.2– 3.3	0
7	0.6– 0.7	3	21	2.0– 2.1	2	34	3.3– 3.4	0
8	0.7– 0.8	1	22	2.1– 2.2	2	35	3.4– 3.5	0
9	0.8– 0.9	7	23	2.2– 2.3	0	36	3.5– 3.6	0
10	0.9– 1.0	5	24	2.3– 2.4	0	37	3.6– 3.7	0
11	1.0– 1.1	4	25	2.4– 2.5	0	38	3.7– 3.8	0
12	1.1– 1.2	4	26	2.5– 2.6	1	39	3.8– 3.9	0
13	1.2– 1.3	1	27	2.6– 2.7	0	40	3.9– 4.0	0
14	1.3– 1.4	3						

23, 24: Duration of Eruption of Old Faithful Geyser See p. 110.

25: Monthly Salary Data (Simonoff, 1996)
The table below gives a 28-cell discretizations of data representing the month spray salary (in dollars) of 147 nonsupervisory female employees holding the Bachelors (but no higher) degree who were practicing mathematics or statistics in 1981.

Cell	Salary	Count	Cell	Salary	Count	Cell	Salary	Count
1	951–1050	5	11	1951–2050	6	20	2851–2950	5
2	1051–1150	1	12	2051–2150	9	21	2951–3050	4
3	1151–1250	0	13	2151–2250	5	22	3051–3150	2
4	1251–1350	5	14	2251–2350	12	23	3151–3250	1
5	1351–1450	2	15	2351–2450	7	24	3251–3350	2
6	1451–1550	10	16	2451–2550	3	25	3351–3450	0
7	1551–1650	5	17	2551–2650	10	26	3451–3550	1
8	1651–1750	10	18	2651–2750	4	27	3551–3650	1
9	1751–1850	10	19	2751–2850	6	28	3651–3750	1
10	1851–1950	20						

26: Length of Treatment Spells (Silverman, 1986)
The data set comprises the lengths of 86 spells of psychiatric treatment undergone by patients used as controls in a study of suicide risks. In applying the density smoothing to the data set the domain [0,800] is divided into 20 intervals of length 40.

1	1	1	5	7	8	8	13	14	14
17	18	21	21	22	25	27	27	30	30
31	31	32	34	35	36	37	38	39	39
40	49	49	54	56	56	62	63	65	65
67	75	76	79	82	83	84	84	84	90
91	92	93	93	103	103	111	112	119	122
123	126	129	134	144	147	153	163	167	175
228	231	235	242	256	256	257	311	314	322
369	415	573	609	640	737				

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory* (Ed: B. N. Petrov and F. Czáki), 267–281. Akademiai Kiadó, Budapest.

Anderson, J. A. and Blair, V. (1982). Penalized maximum likelihood estimation in logistic regression and discrimination. *Biometrika*, **69**, 123–136.

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, **16**, 125–127.

Ashford, J. R. and Walker, P. J. (1972). Quantal response analysis for a mixture of populations. *Biometrics*, **28**, 981–988.

Bian, Q., Sakamoto, W. and Shirahata, S. (1997). Likelihood ratio test in the two phase linear regression problem. *To appear in Jpn. J. Appl. Statist.*

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* MIT Press, Cambridge, Mass.

Bliss, C. I. (1938). The determination of the dosage-mortality curves from small numbers. *Quart. J. Pharm. Pharmacol.*, **11**, 192–216.

de Boor, C. (1978). *A Practical Guide to Splines.* Springer-Verlag, New York.

Braiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Amer. Statist. Assoc.*, **80**, 580–619.

Brown, B. W. (1980). Prediction analysis for binary data. In *Biostatistics Casebook* (Ed: R. G. Miller, B. Efron, B. W. Brown and L. E. Moses), 3–18. John Wiley and Sons, New York.

Brown, B. W. and Hu, M. S. J. (1980). Setting dose levels for the treatment of testicular cancer. In *Biostatistics Casebook* (Ed: R. G. Miller, B. Efron, B. W. Brown and L. E. Moses), 123–152. John Wiley and Sons, New York.

Buckley, M. J., Eagleson, G. K. and Silverman, B. W. (1988). The estimation of residual variance in nonparametric regression. *Biometrika*, **75**, 189–199.

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.*, **17**, 453–555.

le Cessie, S. and van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Appl. Statist.*, **41**, 191–201.

Chambers, J. M. and Hastie, T. J. (1992). *Statistical Models in S.* Wadsworth & Brooks/Cole, Pacific Grove, California.

Chen, H. and Shiau, J. H. (1991). A two-stage spline smoothing method for partially linear models. *J. Statist. Plann. Inference*, **27**, 187–201.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* Chapman and Hall, London.

Cox, D. R. and Snell, E. J. (1981). *Applied Statistics: Principles and Examples.* Chapman and Hall, London.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Num. Math.*, **31**, 377–403.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and Penalties (with discussion). *Statist. Sci.*, **11**, 89–121.

Efron, B. and Tibshirani, R. (1996). Using specially designed exponential families for density estimation. *Ann. Statist.*, **6**, 2431–2461.

Eubank, R. L. (1984). The hat matrix for smoothing splines. *Statist. Probab. Letters*, **2**, 9–14.

Eubank, R. L. (1985). Diagnostics for smoothing splines. *J. Roy. Statist. Soc. Ser. B*, **47**, 332–341.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* Marcel Dekker, New York.

Eubank, R. L. and Gunst, R. F. (1986). Diagnostics for penalized least-squares estimators. *Statist. Probab. Letters*, **4**, 265–272.

Finney, D. J. (1971). *Probit Analysis (3rd edition).* Cambridge University Press.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.*, **19**, 1–141.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.*, **76**, 817–823.

Gasser, T., Sroka, L. and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika*, **73**, 625–633.

Golub, G. H., Heath, M. and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, **21**, 215–223.

Good, I. J. and Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika*, **58**, 255–277.

Goto, M. and Tsuchiya, Y. (1985). *Ouyou Toukei Jissen Kyohon* (Japanese Translation of Cox and Snell, 1981). MPC, Tokyo.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *Intl. Statist. Lev.*, **55**, 245–259.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall, London.

Green, P. J. and Yandell, B. S. (1985). Semiparametric generalized linear models. In *Generalized Linear Models, Lecture Notes in Statistics*, **32**, 44–55. Springer, Berlin.

Gu, C. (1989). RKPACK and its apprications: fitting smoothing spline models. *Proceedings of the Statistical Computing Section, Amer. Statist. Assoc.*, 42–51.

Gu, C. (1990). Adaptive spline smoothing in non-Gaussian regresion models. *J. Amer. Statist. Assoc.*, **85**, 801–807.

Gu, C. (1992). Cross-validating non-Gaussian data. *J. Comput. Graph. Statist.*, **1**, 169–179.

Gu, C., Bates, D. M., Chen, Z. and Wahba, G. (1989). The computation of GCV function through Householder tridiagonalization with application to the fitting of interaction spline models. *SIAM J. Matrix Anal.*, **10**, 457–480.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statist. Sci.*, **1**, 297–318.

Hastie, T. and Tibshirani, R. (1987). Generalized additive models: some applications. *J. Amer. Statist. Assoc.*, **82**, 371–386.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Ichida, K. and Yoshimoto, F. (1979). *Spline Kansuu-to Sono Ouyou* (in Japanese). Kyoiku Shuppan, Tokyo.

Jarrett, R. G. (1979). A note on the intervals between coal-mining disasters. *Biometrika*, **66**, 191–193.

Kawai, N. (1997). Shitsumu outou-gata shiken-ni-okeru data-kaiseki katei (in Japanese). Master's Thesis, Graduate School of Engineering Science, Osaka University.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models (2nd edition).* Chapman and Hall, London.

Nagai, T., Goto, M., Tsuchiya, Y., Uragari, Y., Watanabe, H. and Eto, T. (1994). *Keisanki Toukeigaku Nyumon* (Japanese Translation of Yang and Robinson, 1986). MPC, Tokyo.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A*, **135**, 370–384.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimates. *SIAM J. Sci. Stat. Comput.*, **9**, 363–379.

O'Sullivan, F., Yandell, B. S., and Raynor, W. J. (1988). Automatic smoothing of regression functions in generalized linear models. *J. Amer. Statist. Assoc.*, **81**, 96–103.

Pregibon, D. (1981). Logistic regression diagnostics. *Ann. Statist.*, **9**, 705–724.

Reed, L. J. and Muench, H. (1938). A simple method of estimating fifty per cent endpoints. *American Journal of Hygiene.*, **27**, 493–497.

Reinsch, C. H. (1967). Smoothing by spline functions. *Numer. Math.*, **10**, 177–183.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.*, **12**, 1215–1230.

Rice, J. (1986). Convergence rates for partially splined models. *Statist. Probab. Letters*, **4**, 203–208.

Sakamoto, W. and Shirahata, S. (1997a). Spline smoothing on semiparametric regression models (in Japanese). *Bull. Comput. Statist. Jpn.*, **9**, 13–35.

Sakamoto, W. and Shirahata, S. (1997b). Simple calculation of likelihood-based cross-validation score in maximum penalized likelihood estimation of regression functions. *To appear in J. Jpn. Soc. Comput. Statist.*

Selvin, S. (1995). *Practical Biostatistical Methods.* Duxbury Press, North Scituate, Mass.

Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.*, **10**, 795–810.

Silverman, B. W. (1984). A fast and efficient cross-validation method for smoothing parameter choice in spline regression. *J. Amer. Statist. Assoc.*, **79**, 584–589.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *J. Roy. Statist. Soc. Ser. B*, **47**, 1–52.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Simonoff, J. S. (1986). *Smoothing Methods in Statistics.* Springer, New York.

Speckman, P. (1988). Kernel smoothing in partial linear models. *J. Roy. Statist. Soc. Ser. B*, **50**, 413–436.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *J. Roy. Statist. Soc. Ser. B*, **36**, 111–147.

Stone, M. (1977). Asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Statist. Soc. Ser. B*, **39**, 44–47.

Thompson, W. R. (1947). Use of moving averages and interpolation to estimate median-effective dose: I. Fundamental formulas, estimation of error, and relation to other methods. *Bacteriological Review*, **11**, 115–145.

Utreras, D. F. (1981). Optimal smoothing of noisy data using spline functions. *SIAM J. Sci. Stat. Comput.*, **2**, 349–362.

Wahba, G. (1977). A survey of some smoothing problems and the method of generalized cross-validation for solving them. In *Application of Statistics* (Ed: P. R. Krishnaiah), 507–523. North-Holland, Amsterdam.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B*, **45**, 133–150.

Wahba, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.

Wahba, G. and Wang, Y. (1995). Behavior near zero of the distribution of GCV smoothing parameter estimates. *Statist. Probab. Letters*, **25**, 105–111.

Wahba, G., Wang, Y., Gu. C. Klein, R. and Klein, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *Ann. Statist.*, **23**, 1865–1895.

Wahba, G. and Wold, S. (1975). A completely automatic french curve: fitting spline functions by cross validation. *Commun. Statist.*, **4**, 1–17.

Wang, Y. (1994). Smoothing spline analysis of variance of data from exponential families. Technical Report No. 928, Dept. of Statistics, University of Wisconsin-Madison.

Wang, Y. (1995). GRKPACK: fitting smoothing spline ANOVA models for exponential families. Technical Report No. 942, Dept. of Statistics, University of Wisconsin-Madison.

Xiang. D. and Wahba, G. (1996). A generalized approximate cross validation for smoothing solines with non-Gaussian data. *Statistica Sinica*, **6**, 675–692.

Yang, M. C. K. and Robinson, D. H. (1986). *Understanding and Learning Statistics by Computer.* World Scientific Publishing.