Journal of the Korean
Statistical Society
Vol. 24, No. 2, 1995

Detecting Influential Observations on the Smoothing Parameter in Nonparametric Regression †

Choongrak Kim[1] and Jong-Woo Jeon[2]

ABSTRACT

We present formula for detecting influential observations on the smoothing parameter in smoothing spline. Further, we express them as functions of basic building blocks such as residuals and leverage, and compare it with the local influence approach by Thomas(1991). An example based on a real data set is given.

KEYWORDS: Cross-validation, Influential observation, Leverage, Masking effect, Smoothing parameter, Studentized residual.

1. INTRODUCTION

One of the important issues in smoothing spline diagnostics is finding influential observations on the smoothing parameter, however, quite a few attention was paid to the nonparametric regression diagnostics. Since the estimator of the smoothing parameter is quite sensitive to outliers it is important to detect influential observations. Eubank (1985) and Silverman (1985) studied basic

†This research was supported by the Basic Science Research Institute Program, Ministry of Education, 1994, Project No. BSRI-94-1418.

[1]Department of Statistics, Pusan National University, Pusan, 609-735, Korea.
[2]Department of Statistics, Seoul National University, Seoul, 151-742. Korea.

building blocks such as residuals and leverage in smoothing spline. Hastie and Tibshirani (1990) defined a version of Cook's distance for a single case in the generalized additive model. Thomas (1991) derived local influence of observations on the cross-validated smoothing parameter in smoothing spline.

Consider a nonparametric regression model

$$y_j = \mu(t_j) + \varepsilon_j, \quad j = 1, \ldots, n, \tag{1.1}$$

where μ belongs to the m-th order Sobolev space $W_2^m[a,b]$ of functions f $(a \le t_1 < \cdots < t_n \le b)$ and the errors ε_j are uncorrelated, with mean zero and variance σ^2. One of many possible estimators of μ in (1.1) (see Eubank (1988)) is the minimizer over $f \in W_2^m$ of

$$\frac{1}{n}\sum_{j=1}^{n}\{y_j - f(t_j)\}^2 + \lambda \int_a^b \{f^{(m)}(t)\}^2\, dt, \quad \lambda > 0. \tag{1.2}$$

If $n \ge m$, the minimizer $\hat{\mu}_\lambda$ is a natural polynomial spline of order $2m$ with knots at the t_j that is known as a smoothing spline. Discussions of smoothing splines and their statistical applications may be found in Wegman and Wright (1983), Silverman (1985), Eubank (1988), and Wahba (1990). The parameter λ is called a smoothing parameter, and the choice of λ is usually done by minimizing cross-validation or generalized cross-validation by Craven and Wahba (1979). Throughout this paper we will use the cross-validation criterion for estimation of λ, and it is given by

$$CV(\lambda) = \frac{1}{n}\sum_{j=1}^{n}\{y_j - \hat{\mu}_{j(j)}\}^2, \tag{1.3}$$

where $\hat{\mu}_{j(j)}$ indicates the fit at t_j, computed by leaving out the j-th data point. $CV(\lambda)$ is computed for a number of values of λ over a suitable range and then the minimizing $\hat{\lambda}$ is selected.

In this paper we present formula for detecting influential observations on smoothing parameter in smoothing spline. Deletion formula in the linear model and the analogous results in smoothing spline are derived in Section 2. An

example based on the German hyperinflation data (Eubank 1985, 1988) is given in Section 3.

2. INFLUENCE ON THE SMOOTHING PARAMETER

For a given λ in (1.1), the fitted values $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)'$ are computed by $\hat{\boldsymbol{\mu}} = \boldsymbol{H}\boldsymbol{y}$ where $\boldsymbol{H} = \boldsymbol{H}_\lambda$ is the hat matrix. See Eubank (1988) for details. Let h_{ij} be the ij-th component of $\boldsymbol{H}$, and $\boldsymbol{H}_K$ be a $k \times k$ submatrix of $\boldsymbol{H}$ with elements in $K = \{i_1, \ldots, i_k\}$. Also, let $\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$ and $\boldsymbol{r}_K = (r_{i_1}, \ldots, r_{i_k})'$.

One or few observations can be influential on the estimate of smoothing parameter λ. The most naive method is evaluating $\hat{\lambda} - \hat{\lambda}_{(i)}$, where $\hat{\lambda}_{(i)}$ is the minimizer of CV or GCV with the i-th case omitted. $\hat{\lambda}$ can be found by a grid search, and determining $\hat{\lambda}_{(i)}$ requires this process for each of n cases. Further, $\hat{\lambda}$ can be sensitive to groups of observations acting together rather than single outlying points. If we continue this process to a set $K = \{i_1, \ldots, i_k\}$, we have to compute $\sum_{j=1}^{k} \binom{n}{j}$ times. This approach is clearly computationally infeasible. To overcome with this difficulty, Thomas (1991) suggested diagnostics based on simultaneous perturbation of all obervations rather than deleting single cases. He examined the effect of two types of data perturbations (case and response, response only) on the GCV estimate $\hat{\lambda}$ and derived diagnostics by applying the local influence method of Cook (1986). By perturbing the response $\boldsymbol{y}$ only, Thomas (1991) derived the direction of maximum slope;

$$t_{\max}(\boldsymbol{y}) \propto (c\boldsymbol{I} - \boldsymbol{H}_{\hat{\lambda}})(\boldsymbol{I} - \boldsymbol{H}_{\hat{\lambda}})^2\boldsymbol{y}, \tag{2.1}$$

where $c = \mathrm{tr}\{\boldsymbol{H}_{\hat{\lambda}}(\boldsymbol{I} - \boldsymbol{H}_{\hat{\lambda}})\} / \mathrm{tr}(\boldsymbol{I} - \boldsymbol{H}_{\hat{\lambda}})$ and $\hat{\lambda}$ is the GCV estimate.

Here we suggest a computationally feasible method of evaluating $\hat{\lambda}_{(i)}$ using the CV criterion, and this method can be easily extended to $\hat{\lambda}_{(K)}$, the minimizer of CV with k observations in $K = \{i_1, \ldots, i_k\}$ omitted. Let $CV_{(i)}(\lambda)$ be the cross-validation with the i-th case omitted, then we can define it as

$$CV_{(i)}(\lambda) = \frac{1}{n-1} \sum_{j \neq i} \{y_j - \hat{\mu}_{j(i,j)}\}^2, \tag{2.2}$$

where $\hat{\mu}_{j(i,j)}$ indicates the fit at t_j, computed by leaving out i-th and j-th data point. Equation (2.2) would be very useful if we can write it as functions of residuals and leverage, however, it is not easy. One alternative way is using the deletion formula in the linear model. So, consider a linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.3}$$

where $\boldsymbol{y}$ is an n-vector of response, $\boldsymbol{X}$ is an $n \times q$ full column rank matrix of known covariates. $\boldsymbol{\beta}$ is q-vector of unknown coefficients, and $\boldsymbol{\varepsilon}$ is an n-vector of independent variables with mean zero and unknown variance σ^2. We use y_i and $\boldsymbol{x}_i$ to denote the i-th row of $\boldsymbol{y}$ and $\boldsymbol{X}$, respectively, and use the subscript (i) to indicate the deletion of the i-th observation, thus, for example, $\boldsymbol{X}_{(i)}$ denotes the matrix $\boldsymbol{X}$ with the i-th row deleted. For a set $K = \{i_1, \ldots, i_k\}$ of size k, we define similarly. After fitting the model by the method of least squares, we have $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, $\hat{\boldsymbol{y}} = \boldsymbol{P}\boldsymbol{y}$, where $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is hat matrix, and residual vector $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$. Let $s^2 = \boldsymbol{e}'\boldsymbol{e}/(n-q)$ be the unbiased estimator of σ^2, and $p_{ij} = \boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_j'$ be the ij-th component of $\boldsymbol{P}$. Under model (2.3),

$$\begin{aligned}
\hat{y}_{j(i)} &= \boldsymbol{x}_j\hat{\boldsymbol{\beta}}_{(i)} \\
&= \boldsymbol{x}_j\left(\hat{\boldsymbol{\beta}} - \frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i'e_i}{1-p_{ii}}\right) \\
&= \hat{y}_j - \frac{p_{ji}e_i}{1-p_{ii}}
\end{aligned} \tag{2.4}$$

and

$$\begin{aligned}
\hat{y}_{j(K)} &= \boldsymbol{x}_j\hat{\boldsymbol{\beta}}_{(K)} \\
&= \boldsymbol{x}_j\left[\hat{\boldsymbol{\beta}} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_K'(\boldsymbol{I}-\boldsymbol{P}_K)^{-1}\boldsymbol{e}_K\right] \\
&= \hat{y}_j - \boldsymbol{x}_j(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_K'(\boldsymbol{I}-\boldsymbol{P}_K)^{-1}\boldsymbol{e}_K \\
&= \hat{y}_j - \boldsymbol{p}_{j,K}(\boldsymbol{I}-\boldsymbol{P}_K)^{-1}\boldsymbol{e}_K,
\end{aligned} \tag{2.5}$$

where $\boldsymbol{p}_{j,K} = (p_{j,i_1}, \ldots, p_{j,i_k})$, $\boldsymbol{p}_K = \boldsymbol{x}_K(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_K'$, $\boldsymbol{x}_K = (\boldsymbol{x}_{i_1}', \ldots, \boldsymbol{x}_{i_k}')'$, and $\boldsymbol{e}_K = (e_{i_1}, \ldots, e_{i_k})'$. The rationale is that both $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{\mu}}$ are linear combinations of hat matrix ($\boldsymbol{P}$ and $\boldsymbol{H}$, respectively) and $\boldsymbol{y}$. The only difference is that $\boldsymbol{P}$ is idempotent, while $\boldsymbol{H}$ is not. By using the result in (2.5), we have

$$\begin{aligned} CV_{(i)}(\lambda) &= \frac{1}{n-1}\sum_{j\neq i}\{y_j - \hat{\mu}_j + \hat{\mu}_j - \hat{\mu}_{j(i,j)}\}^2 \\ &= \frac{1}{n-1}\sum_{j\neq i}\left\{r_j + \boldsymbol{h}_{j,N}(\boldsymbol{I} - \boldsymbol{H}_N)^{-1}\boldsymbol{r}_N\right\}^2, \end{aligned} \tag{2.6}$$

where $N = \{i, j\}$ and $\boldsymbol{h}_{j,N} = (h_{ji}, h_{jj})$. Similarly, for a group of observations in $K = \{i_1, \ldots, i_k\}$, we have, by (2.5),

$$\begin{aligned} CV_{(K)}(\lambda) &= \frac{1}{n-k}\sum_{j\notin K}\{y_j - \hat{\mu}_{j(j\cup K)}\}^2 \\ &= \frac{1}{n-k}\sum_{j\notin K}\{r_j + \boldsymbol{h}_{j,K\cup j}\,(\boldsymbol{I} - \boldsymbol{H}_{K\cup j})^{-1}\boldsymbol{r}_{K\cup j}\}^2, \end{aligned} \tag{2.7}$$

where $K \cup j = \{i_1, \ldots, i_k, j\}$.

Hence, to determine $\hat{\lambda}_{(i)}$, $i = 1, \ldots, n$ we don't need to compute $CV_{(i)}(\lambda)$ in (2.2); we are only to compute $\boldsymbol{r}$ and $\boldsymbol{H}$ which are avaliable from $\hat{\lambda}$ by (2.6). The computational amount required in this process is just $1/n$ compared to the naive grid search method, and for groups of size k it is $1/\binom{n}{k}$ of the naive method. The real computation time depends on data, but that of this process is approximately $1/n$ compared to the naive method based on our limited experience.

To be more specific, the computational amount via the naive method requires n times of computing $\hat{\boldsymbol{\mu}}_{(i)}$, $i = 1, \ldots, n$ for fixed λ. Therefore, if the number of grid is m, the total number of computation becomes mn. However, if we use (2.6), we are only to compute $\hat{\boldsymbol{\mu}}$ for fixed λ. Hence, the total number of computation required is m.

The accuracy of computing $\hat{\lambda}$ totally depends on the number of grid. There are other methods such as the golden section method and the bisection method,

however, the grid search is the safest even though it can be slower than the golden section or bisection method.

3. EXAMPLE

As an illustrative example we consider the German hyperinflation data shown in Figure 1. This data consists of values for the logarithm of the money supply as a function of the logarithm for the premium on a forward contract for foreign exchange during the German hyperinflation.

Table 1 contains h_{ii}, $r_i^* = r_i/s\sqrt{1-h_{ii}}$, where λ is chosen as minimizing CV, i.e., $\hat{\lambda} = 1.28 \times 10^{-4}$. We find influential observations and groups of observations on the CV estimate $\hat{\lambda}$ using the results in Section 2, and four largest influential groups of observations on $\hat{\lambda}$ for $k = 1, 2, 3$ are summarized in Table 2. As shown in this Table, for $k = 1$, observations 19 and 30 are very influential. For $k = 2$, sets (29,30), (29,31), (30,31) are quite influential, and swamping phenomenon by (29,30) is clear for $k = 3$. Conclusively, 19, (29,30), (30,31) are influential on $\hat{\lambda}$. Figure 1 shows the spline fit to the German hyperinflation data with 19, (29,30), (30,31), and (29,30,31) deleted, respectively. The local influence approach by Thomas (1991) gives very different results. It shows that cases 24 and 27 are influential. (See Figure 2). They are neither singly influential nor jointly influential. Note that $\hat{\lambda}_{(24)} = 1.02 \times 10^{-4}$, $\hat{\lambda}_{(27)} = 0.80 \times 10^{-4}$, $\hat{\lambda}_{(24,27)} = 0.73 \times 10^{-4}$ while $\hat{\lambda} = 1.28 \times 10^{-4}$. This difference may be due to perturbing response only, however, not clear.

After detecting influential observations, we can interpret them in several ways. First, by removing influential observations the fit could be smoother if $\hat{\lambda}_{(K)}$ is large. Second, if influential observations show sequential pattern, they might suggest that we better use variable smoothing parameter instead of constant smoothing parameter.

The choice of k is quite subjective so far because some appropriate cutoff value is not suggested yet. Local influence approach suggested by Thomas(1991) suffers from the same problem. For example, it is very subjective to judge the case 23 is influential or not.

Table 1. Leverage and Residual in the German Hyperinflation Data.

case	y_i	t_i	h_{ii}	r_i^*
1	6.5605	−1.8202	.52400	0.09966
2	6.5474	−1.7958	.48273	−0.11183
3	6.5802	−1.1087	.55732	−0.06486
4	6.5927	−.9927	.42523	0.11303
5	6.5019	−.6832	.33930	−0.84619
6	6.5896	−.6539	.32425	0.29326
7	6.5414	−.3960	.17308	0.50160
8	6.4580	−.3930	.17106	−0.42615
9	6.5381	−.3653	.15680	0.67090
10	6.4977	−.3271	.14882	0.52652
11	6.4129	−.3093	.14890	−0.26075
12	6.4225	−.1863	.16425	1.21156
13	6.2669	−.1839	.16457	−0.51531
14	6.0839	−.0429	.18958	−0.86404
15	6.1841	−.0837	.17842	−0.20013
16	6.0578	.0000	.21054	−0.71485
17	6.0774	.0999	.30138	0.40923
18	5.9321	.3343	.67361	−0.03976
19	5.7858	1.1845	.83081	0.58333
20	5.5203	1.6369	.47226	1.00237
21	5.2718	1.7630	.37636	−1.02903
22	5.2421	1.9243	.51376	−0.42553
23	5.4116	2.4336	.34464	1.33872
24	5.1504	2.4774	.28146	−1.97715
25	5.4239	2.5908	.24098	1.18494
26	5.3290	2.6053	.24613	0.05886
27	5.1921	2.7955	.33968	−1.19926
28	5.4269	2.9565	.37513	2.92158
29	4.9010	3.1122	.52483	−2.35193
30	4.7712	3.6169	.70216	−0.15324
31	4.7589	3.9176	.89925	0.45570

Table 2. Four Largest Influential Groups of Observations on the CV Estimate $\hat{\lambda}$ for $k = 1, 2, 3$ in the German Hyperinflation Data. ($\hat{\lambda} = 1.28 \times 10^{-4}$)

k	sets	$\hat{\lambda}_{(K)} \times 10^4$	$\hat{\lambda}_{(K)}/\hat{\lambda}$
1	19	24.28	18.97
	30	16.53	12.91
	29	7.47	5.84
	31	5.66	4.42
2	29,30	209.10	163.36
	29,31	144.63	112.99
	30,31	52.90	41.33
	19,29	21.68	16.94
3	6,29,30	245.60	191.88
	7,29,30	243.20	190.00
	20,29,30	240.50	187.89
	9,29,30	239.80	187.34

Figure 1. Spline Fit to the German Hyperinflation Data: Original Data (Solid), Cases Deleted (Broken) (a) Case 19 Deleted (b) Cases 29,30 Deleted (c) Cases 30,31 Deleted (d) Cases 29,30,31 Deleted.

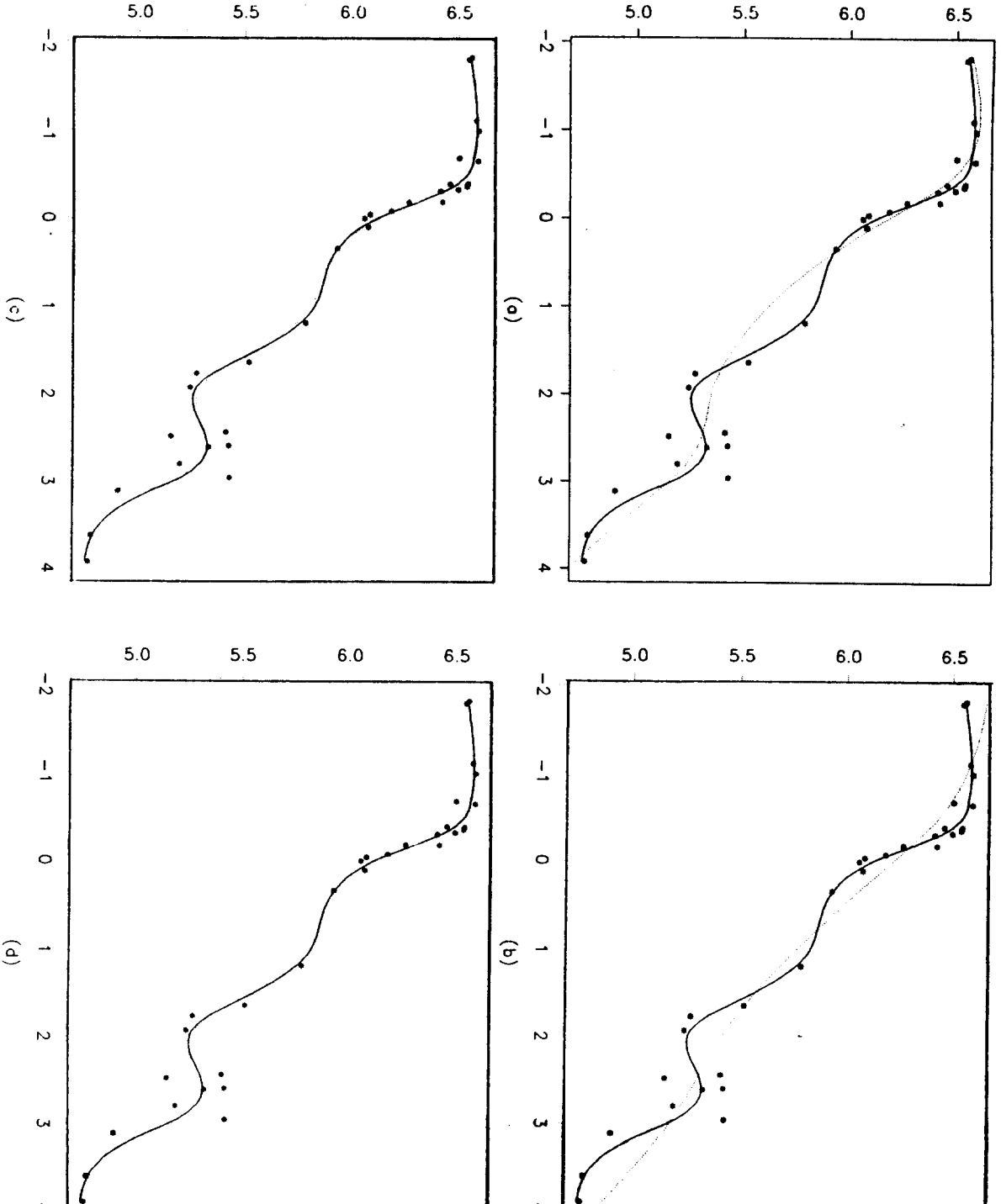
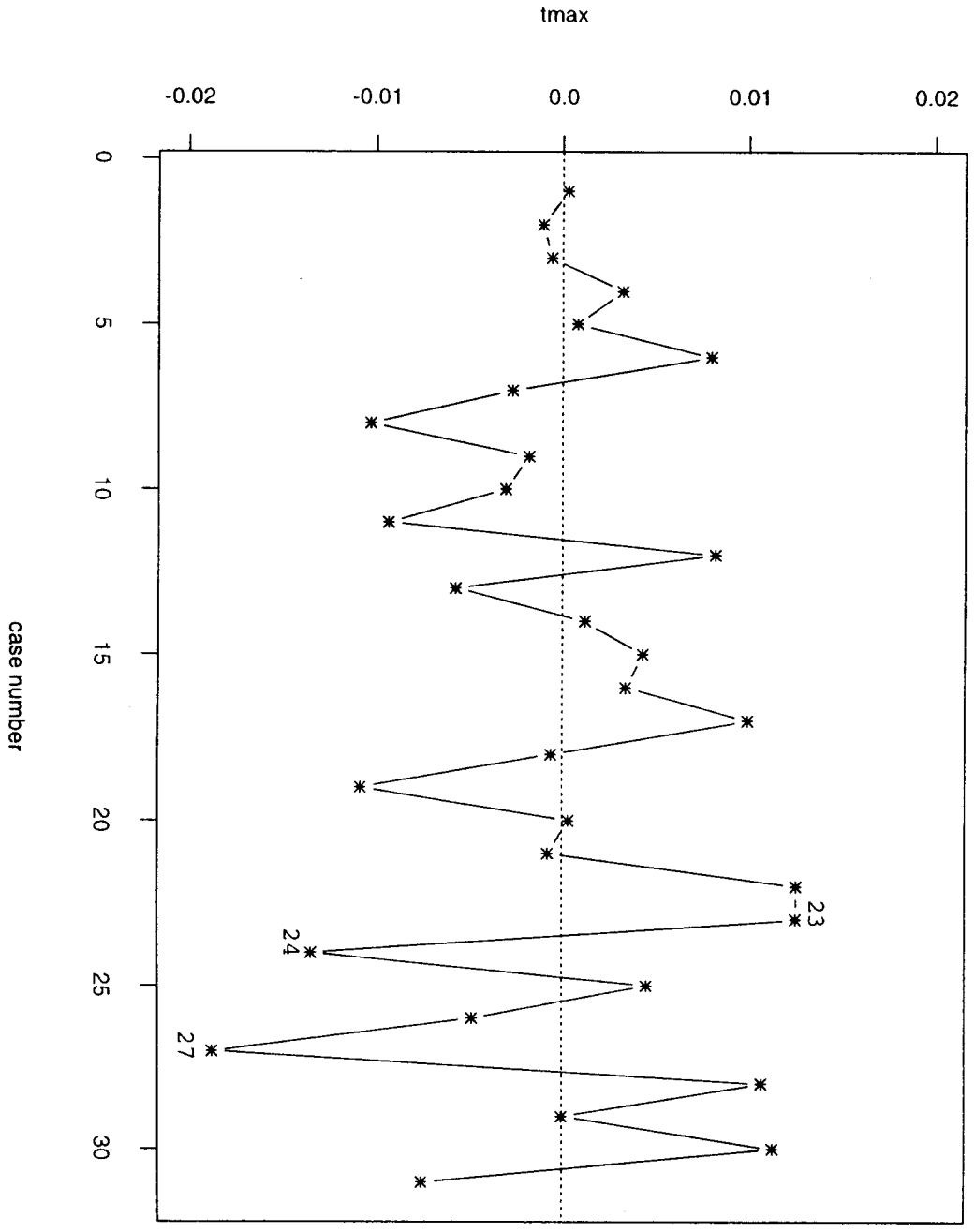


Figure 2. Local Influence Approach in the German Hyperinflation Data.



REFERENCES

(1) Cook, R.D. (1986). Assessment of Local Influence (with Discussion). *Journal of the Royal Statistical Society*, Ser. **B**. **48**, 133–169.

(2) Craven, P. and Wahba, G. (1979). Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation. *Numerische Mathematik*, **31**, 377–403.

(3) Eubank, R.L. (1985). Diagnostics for Smoothing Splines. *Journal of the Royal Statistical Society*, Ser. **B**. **47**, 332–341.

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

(5) Frenkel, J.A. (1977). The Forward Exchange Rate, Expectations, and the Demand for Money: the German Hyperinflation. *American Economic Review*, **67**, 653–670.

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

(7) Silverman, B.W. (1985). Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting (with Discussion). *Journal of the Royal Statistical Society*, Ser. **B**. **47**, 1–52.

(8) Thomas, W. (1991). Influence Diagnostics for the Cross-Validated Smoothing Parameter in Spline Smoothing. *Journal of the American Statistical Association*, **86**, 693–698.

(9) Wahba, G. (1985). A Comparison of GCV and GML for Choosing the Smoothing Parameter in the Generalized Spline Smoothing Problem. *The Annals of Statistics*, **13**, 1378–1402.

(10) Wahba, G. (1990). *Spline Models in Statistics*. SIAM, Philadelphia.

(11) Wegman, E.J. and Wright, I.W. (1983). Splines in Statistics. *Journal of the American Statistical Association*, **78**, 351–365.