How to Select Polynomial Models with an Accurate Derivative

Piet M. T. Broersen

***Abstract*—Derivatives of estimated static relations are often used for linearization in control and in extended Kalman filtering. However, the structure of selected models may only be an approximation to the true relationship, which can cause problems in taking derivatives. Polynomial models, estimated from noisy observations, may give accurate descriptions of the data while at the same time their derivatives may be poor approximations of the true derivative. The explanation of the strong degradation of the derivative of selected models is straightforward: estimating polynomial models of increasing order from a set of data gives not only a description of the true underlying process, but also of the accidental realization of the additive noise. The higher order polynomial models will crinkle around the true process; therefore, they will mostly have an irregular derivative. Models with a better derivative can be selected by using a higher penalty factor in the selection criterion.**

***Index Terms*—Control nonlinearity, error criterion, function approximation, model selection, penalty factor, static model, stochastic modeling.**

I. Introduction

POLYNOMIAL and straight-line regressions are historically treated as separate subjects in regression [1]. In polynomial regression analysis, noisy observations of the dependent variable $\boldsymbol{y}$ are presented as a nonlinear function of the independent variable $\boldsymbol{x}$. A polynomial in $\boldsymbol{x}$, with an additional constant, is used to model the unknown relationship. The model parameters have to be estimated from the data, but the highest order for the polynomials also has to be selected. Suppose that an accurate derivative of that estimated polynomial is the main goal, as it is when the model is used for variations around an operating control point. The question is how many terms have to be included in the polynomial to obtain the best accuracy for the derivative. More specifically, the problem addressed in this paper is whether the maximum likelihood properties of polynomial models of a *known* best order are also applicable to polynomials with a *selected* order. The influence of biased estimation is also considered by taking true nonlinear relations with an infinite polynomial expansion or Taylor series, like exponential functions.

Linearization of an approximated static polynomial relationship by using the local derivative at an *arbitrary* operating point inside the range of the independent variable is studied. This differs from the situation where the derivative at a *particular* value of $\boldsymbol{x}$ is wanted [2]. Data are collected by adjusting the value of the independent variable $\boldsymbol{x}$ and measuring its noisy dependent variable $\boldsymbol{y}$. The principle of randomization requires that the data should be obtained in an arbitrary sequence of $\boldsymbol{x}$ values, so that they are not a sequential time series. Hence, reducing the noise by applying dynamic filtering is impossible. Another approach, differencing the data and looking for a model for differenced data is highly inefficient because it causes a deterioration of the signal-to-noise ratio (SNR) and it introduces correlation in the error signal of the differences. Of course, it would be advantageous to collect observations of the derivative, or directly of the finite differences, if a model for the derivative is wanted. However, this paper is limited to the practical question of selecting, from given data, a model that has an accurate derivative over the whole range.

The intended use of the model should determine the order selection criterion and the search procedure for the model [2]. Mallows's criterion C_p [3] is often used for selection in practice, if the predictive or descriptive ability of the model to new data is important. The penalty factor for additional regressors has an important influence on the resulting models in a selection procedure. An asymptotical evaluation for the penalty factor in C_p has been given in autoregressive modeling [4]. It turns out that the usual factor 2 reduces the probability of underfit almost to zero, at the cost of a high probability of overfit, resulting often in too many regressors being selected to appear in the final model. An approach to improve this behavior is to replace the penalty factor of 2 in the C_p formula with something larger. Optimal values for the penalty factor have been derived in autoregression [5]. These values can also be used for selection in linear regression, yielding a penalty factor of about 3 for the selection of models for the data, with a good balance between the errors due to underfit and overfit.

This paper shows that the quality loss due to selection can be moderate for the model of the data, but at the same time very much greater for the derivative of that model. The inaccuracy is due to overfit; the inclusion of too many terms in the selected model is shown to be the cause of this degradation. A practical solution is: taking a higher penalty factor for additional regressors in the order selection criterion. Recommendations are given for that factor in order to obtain a better accuracy of the derivative.

II. Notation

Suppose that we have data consisting of N noisy observations $\boldsymbol{y}$, that can be described as

$$\boldsymbol{y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon}. \tag{1}$$

Manuscript received April 11, 1997; revised March 22, 2000. The original version of this paper was presented at the IMTC/96 Conference, Brussels, Belgium, June 4–6, 1996.

The author is with the Department of Applied Physics, Delft University of Technology, 2600 GA Delft, The Netherlands.

Publisher Item Identifier S 0018-9456(00)07022-4.

The $N \times 1$ matrix $\boldsymbol{y}$ has elements $y(1), \cdots, y(N)$ and the $N \times 1$ vector $\boldsymbol{\varepsilon}$ is additive noise consisting of independent normally distributed random variables with zero mean and variance $\boldsymbol{\sigma^2}$. A data model with p parameters is given by an $N \times 1$ vector:

$$\hat{\boldsymbol{y}}_{\boldsymbol{p}} = \boldsymbol{X}_{\boldsymbol{p}} \hat{\boldsymbol{b}}_{\boldsymbol{p}}. \tag{2}$$

K regressor variables $\boldsymbol{x_i}, \boldsymbol{i} = \boldsymbol{1}, \cdots, \boldsymbol{K}$, are available, which together define the $N \times K$ matrix $\boldsymbol{X_K}$. In this paper, the regressors are a constant and $K-1$ polynomials $\boldsymbol{x}, \boldsymbol{x^2}, \boldsymbol{x^{K-1}}$, where N values for $\boldsymbol{x}$ are taken equidistant on the interval [0,1] with increments $1/(N-1)$. So the mean is treated as an ordinary regressor. A hierarchical model of order p includes the constant and the first $p-1$ polynomials. Subset models separate the K regressors into two arbitrary parts $[\boldsymbol{X_p}|\boldsymbol{X_r}]$. For ease of notation, the regressors can always be rearranged such that the leading subset $\boldsymbol{X_p}$ contains the p selected regressors; r is the number of regressors that are not in the subset, so $p+r=K$. This is only a simplification in notation without practical consequences. Both for hierarchical and for subset models, the parameters $\hat{\boldsymbol{b}}_{\boldsymbol{p}}$ are estimated as

$$\hat{\boldsymbol{b}}_{\boldsymbol{p}} = [\boldsymbol{X_p^T X_p}]^{-1} \boldsymbol{X_p^T y}. \tag{3}$$

The Residual Sum of Squares RSS_p is a monotonously decreasing function of p and is given by

$$\mathbf{RSS}_{\boldsymbol{p}} = (\boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{p}} \hat{\boldsymbol{b}}_{\boldsymbol{p}})^{\boldsymbol{T}} (\boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{p}} \hat{\boldsymbol{b}}_{\boldsymbol{p}}). \tag{4}$$

An objective measure for the distance between an estimated model (2) and the true noise-free response $\boldsymbol{\eta}$ in (1) can be defined as

$$\boldsymbol{J_p(y)} = (\boldsymbol{X_p} \hat{\boldsymbol{b}}_{\boldsymbol{p}} - \boldsymbol{\eta})^{\boldsymbol{T}} (\boldsymbol{X_p} \hat{\boldsymbol{b}}_{\boldsymbol{p}} - \boldsymbol{\eta}) / \boldsymbol{\sigma^2}. \tag{5}$$

This measure is an excellent indication of the quality of selected models when used to predict response values for new values of x in the given interval [0,1] that is used, but it is only useful in simulations because η has to be known exactly. For use in practice, the selection criterion C_p is a transformation of RSS_p [3], defined as

$$\boldsymbol{C_p} = \mathbf{RSS}_{\boldsymbol{p}} / \boldsymbol{s^2} - \boldsymbol{N} + \boldsymbol{2p}, \tag{6}$$

where s^2 is given by $\mathrm{RSS}_K/(N-K)$, which is the estimate for the residual variance, as obtained from the complete model with all K regressors included. The criterion C_p is popular because it has the same mathematical expectation as $J_p(y)$ [3]. C_p and $J_p(y)$ contain contributions of variance due to estimation and of bias due to the omission of regressors. Their expectations are equal, asymptotically, if the true process η is exactly described by (a subset of) the K candidate regressors. This common expectation equals the number of regressors p for models including at least all regressors with nonzero true parameter values in η [2].

In agreement with (5), an objective quality criterion for the derivative $\boldsymbol{y'}$ of $\boldsymbol{y}$ with respect to $\boldsymbol{x}$ is defined as

$$\boldsymbol{J_p(y')} = (\hat{\boldsymbol{y}}'_{\boldsymbol{p}} - \boldsymbol{\eta'})^{\boldsymbol{T}} (\hat{\boldsymbol{y}}'_{\boldsymbol{p}} - \boldsymbol{\eta'}) / \boldsymbol{\sigma^2}. \tag{7}$$

This is scaled with the same σ^2 as the data. The formula for the derivative of hierarchical models is given by

$$\hat{\boldsymbol{y}}'_{\boldsymbol{p}} = \hat{\boldsymbol{b}}_{\boldsymbol{1}} + \boldsymbol{2} \hat{\boldsymbol{b}}_{\boldsymbol{2}} \boldsymbol{x} + \cdots \boldsymbol{p} \hat{\boldsymbol{b}}_{\boldsymbol{p}} \boldsymbol{x^{p-1}}. \tag{8}$$

For $\boldsymbol{x}$ in the interval [0,1], this means that the energy in the derivative is generally greater than the energy in the data, because the parameters are multiplied by the order in (8). The SNR, used to indicate the noise level, is defined as

$$\mathbf{SNR} = \boldsymbol{\eta^T \eta} / \boldsymbol{\sigma^2}. \tag{9}$$

The total energy in N observations of the true process η is scaled with the variance of the additive noise.

III. Theoretical Background

It is well known that a lower bound exists for the accuracy of unbiased estimated parameters: the Cramér-Rao bound. Asymptotically, that lower bound can be obtained by the maximum likelihood estimator. For normally distributed variables, the least squares algorithm is the maximum likelihood estimator. Under rather mild conditions, the maximum likelihood estimator of a function of a stochastic variable is equal to the function of the maximum likelihood estimate of the stochastic variable itself [6]. This property is used in the accuracy of spectral estimators and of nonlinear transformations of a stochastic variable [7]. It means that the derivative of an unbiased estimated polynomial data model is at the same time the best estimator for the derivative of the data, if the order is known and the model is estimated with the maximum likelihood principle. The main condition for a strict application of this rule is that the estimate must be unbiased. This means that the model type and order must be known in advance. In this paper, the influence of bias on the accuracy is investigated and also the influence of order selection for unknown nonlinear relations.

Polynomials can be considered as basic functions for smoothing over the whole interval, and order selection is necessary to determine the best order of the polynomial. The use of splines or other local smoothing techniques with piecewise polynomial fitting is advised if the polynomial fit over the whole range remains unsatisfactory for high degrees [1]. Splines give a type of local or piecewise smoothing, with individually fitted low-order polynomials for every piece between two knots. Generally no statistics are involved in the choice of the number and the location of the knots and the order of the polynomial [1]. Hence, a statistical evaluation of the quality of the derivative over the whole range of the independent variable is not possible. The quality of the fit and its derivative would be strongly dependent on the location of the knots, and the statistical quality would vary over the interval. This paper deals with one polynomial model for the whole interval.

Taking differences of data as a basis for order selection of an accurate derivative has been investigated. An obvious method to find a model for the derivative is to approximate finite differences, obtained from the data, with a polynomial model. These differences are for equidistant values of $\boldsymbol{x}$ given by

$$\boldsymbol{\Delta y(i)} = \frac{\boldsymbol{y(i) - y(i-1)}}{\boldsymbol{\Delta x}}, \boldsymbol{\Delta x} = \frac{\boldsymbol{1}}{\boldsymbol{N-1}}. \tag{10}$$

The model is a polynomial approximation of $\Delta y(i)$ as a function of x. The variance of $\Delta y(i)$ is $2(N-1)^2\sigma^2$ instead of the value σ^2 that is present in $y(i)$. If the total energy in the derivative is of the same order of magnitude as the energy in the data, the SNR for the differences is very much worse than for the data. Moreover, the errors in $\Delta y(i)$ and in $\Delta y(i+1)$ have -0.5 as correlation coefficient because both contain the same noisy data point $y(i)$. Hence, the weighted least squares estimator is the optimal estimator for parameters for $\Delta y(i)$. The best weighted estimator for the parameters of differences is found theoretically by using the inverse of the covariance matrix of the errors in the differences. This, however, is exactly the same as using the original data for estimation, if the errors in the original data are uncorrelated. Therefore, this approach agrees with the maximum likelihood estimation of the parameters for the data model and taking the derivative afterwards. The theoretical analysis of the data differences has been supported by simulations.

IV. Degradation of Derivative

The degradation of the derivative of a polynomial model is illustrated in Table I and in Fig. 1 with an example where the true process is given by

$$\eta = e^{-x} = \sum_{i=0}^{\infty} \frac{(-x)^i}{i!}. \tag{11}$$

A special property of this process is that the derivative equals $-e^{-x}$, so the energy in the data is equal to the energy in the derivative. This facilitates the interpretation of the results, because the magnitudes of the errors in $J_p(y)$ and $J_p(y')$ are directly comparable, as the same variance σ^2 has been used in both definitions (5) and (7). The true process in (11) has order infinity and it cannot be described exactly by K candidate regressors, so bias will be present. Fig. 1 shows that hierarchical models of order 2 (with an additional estimated constant) and also of order 7 remain close to the true process but that the derivative becomes wild for the higher model order. The derivative of the first-order model, a constant, gives a poor approximation due to underfit. The second- and third-order models are the best approximations for the data model and also for the derivative, as is also seen in Table I from $J_p(y)$ and $J_p(y')$, respectively. Higher order models with overfit give a moderate increase of $J_p(y)$ in Table I and at the same time a very irregular approximation for the derivative, as seen in $J_p(y')$. This shows that a good fitting combination of parameters for the data can give an irregular derivative with (8). In the simulations, the accuracy of both $J_p(y)$ and $J_p(y')$ is optimal for order 3 + 1, where +1 denotes the estimation of the mean with the regressor a constant. *For fixed orders, the best polynomial model for the derivative is found as the derivative of the best fixed-order data model.* This conclusion of unbiased maximum likelihood theory [6] remains valid in simulations for this fixed-order model with bias. Hence, the highest polynomial order in the best model for the derivative is one lower than for the data. This is quite remarkable because the true shape of the data and of the derivative is the same in (11).

In Table I, the best accuracy for $J_p(y')$ is about 60 times worse than the smallest $J_p(y)$. Moreover, $J_p(y')$ increases very

TABLE I
Average of Criteria in 100 Simulation Runs as a Function of the Polynomial Model Order. $\eta = \exp(-x)$, SNR = 100000, N = 30, Hierarchical Models. The Constant Regressor for the Mean is Denoted as +1 in the Column for the Model Order p

p	C_p	$J_p(y)$	$J_p(y')$
0	112110	100000	100000
+1	9016	8094	100000
1+1	159.6	141.1	8082
2+1	5.1	3.93	239
3+1	4.5	3.91	234
4+1	5.3	4.9	586
5+1	6.2	6.0	1285
6+1	7.1	6.9	2207
7+1	8.2	7.8	3902
8+1	9.0	8.9	7111

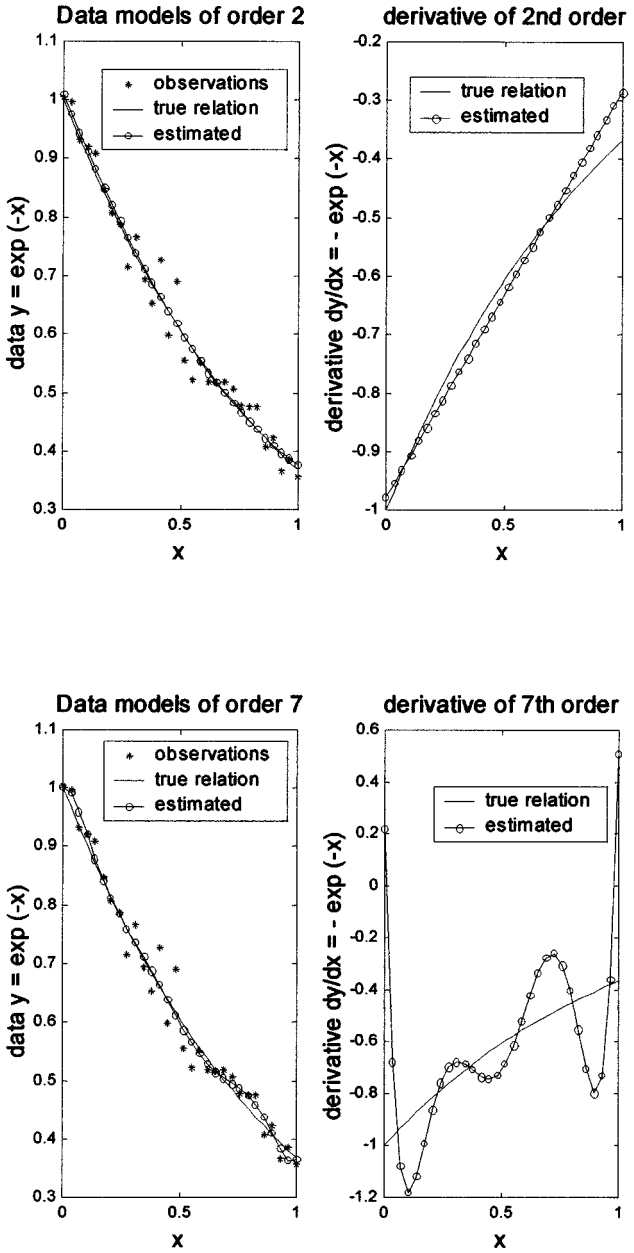

Fig. 1. Data model and its derivative for hierarchical models of order 2 and 7 for a single realization. $y = \exp(-x)$, SNR = 10000, N = 30.

strongly for each additional order. Taking higher orders in overfitted models gives a moderate increase of $J_p(y)$ but at the same time a severe degradation in the quality $J_p(y')$ for the derivative. The polynomials have not been made orthogonal in the example of Table I, but this has no influence at all on the characteristics in the behavior for increasing hierarchical model orders [1].

Of course, much better results for the derivative would have been obtained by fitting a nonlinear exponential to the data and taking the derivative of that estimated exponential model. But that would demand that *a priori* knowledge is available about the true structure of η. In contrast, the problem addressed in this paper is the enormous degradation in the quality if the true order or structure has to be selected from a number of candidates. The degradation depends on the covariance structure of the candidate regressors. Table I shows the results that have been obtained by estimating hierarchical models of increasing fixed orders. Additional problems occur if models with selected orders are studied. The conclusion to be drawn from this example is that the accuracy of the derivative is much more sensitive for too high model orders than the accuracy of the data model.

V. The Penalty in Selection Criteria

The factor 2 in C_p of (6) has been derived by giving equal weights to bias and to variance in a selection criterion [3]. Bias becomes more important in underfitting models with too few regressors included; variance is the problem of overfit and C_p selects a compromise. Looking to the accuracy of data models and their derivatives, $J_p(y)$ and $J_p(y')$ in Table I, it is clear that the inaccuracy due to bias at the lowest model orders is of a comparable magnitude, but the degradation by variance at high orders is much greater for derivatives, up to 700 times. In other words, overfit is more dangerous for the quality of derivatives than for data models. A proper order selection criterion for good derivatives must reduce the probability of selecting overfitting models as much as possible. The overfit problem has been studied for autoregressive processes [4], [5]. It can be diminished by using a higher penalty for additional regressors [5]. A new selection criterion with a penalty factor α is defined as

$$\boldsymbol{C_p(\alpha) = \mathrm{RSS}_p/s^2 - N + \alpha p.} \tag{12}$$

This is a generalization of (6); the usual C_p becomes now equal to $C_p(2)$. The model with the minimum of $C_p(\alpha)$ is selected. That model can also be characterized by critical values for RSS_p: it follows from (12) that arbitrary groups of m regressors that will be included in the selected model must give reductions $\mathrm{RSS}_{p\text{-}m}$ - RSS_p that are greater than α ms^2.

The theory for order selection in hierarchical models has been developed for autoregressive modeling [4], [5]. It has been applied to the choice of an optimal value for α in regression [8]. Unfortunately, theoretical arguments cannot provide a single optimal value for α for all possible situations. In an example, the specific value of α that selects the model with the smallest $J_p(y)$ would be the best for describing the data, but that value depends on the SNR. If the noise level is extremely high, no regressor will be significant and penalty $\alpha = \infty$ will be the best to make sure that no regressor is selected. On the other hand, if there is no noise, all regressors with nonzero estimates for the parameters are significant and they are selected by taking a penalty α close to zero. For less extreme values of the SNR, a certain balance between underfit and overfit yielded the value $\alpha = 3$ as a good compromise for autoregressive models [5]. The discussions about α applied only to data models. The prime interest here is the accuracy of the derivative. From Fig. 1 and Table I we know that the derivative is much more sensitive to overfit than the data model. Hence, the balance between overfit and underfit should be different for the data and for the derivative; the optimal penalty should be higher for the derivative than for the data.

VI. Simulation Results

Simulations have been carried out to find good values for the penalty factor $\boldsymbol{\alpha}$, both for data models and for accurate derivatives, in a variety of examples like polynomials, cosine data, triangles, exponentials, and Gaussian bell shapes. The examples given in Table II are with seven different true processes $\boldsymbol{\eta}$:

A:	$\eta = \exp(-x)$,	SNR = 1 650 000
B:	$\eta = \sin(1.5\pi x)$,	SNR = 250
C:	$\eta = \cos(0.5\pi x)$,	SNR = 100 000
D:	$\eta =$ triangle with basis x =[0,1], top at $x = 0.75$,	SNR = 200
E:	$\eta = \exp(-x^2) - 1$,	SNR = 13 000
F:	$\eta = \sqrt{x}, x > 0, \eta_0 = 100$,	SNR = 3000
G:	$\eta = x + x^2 + x^3$,	SNR = 90 000

The value for the SNR is chosen such that the best polynomial model order is 3 or 4, thus leaving the possibility of overfit and of underfit. Example G is different from the others because it is the only example with a true polynomial model of finite order, that can be estimated without bias. Nevertheless, the performance of G is comparable to the other examples. Therefore, the bias in selected models turns out to be not very important for the conclusions.

Table II gives results of simulations for the average accuracy $J_p(y)$ of *selected* data models for different values of the penalty α. The most accurate selected models are found in practice for α = 3 or α = 4 in $C_p(\alpha)$ of (12). This agrees reasonably with the preference for α = 3 that has been derived for autoregressive order selection [5].

The average accuracy of the derivative of *selected* models, $J_p(y')$, is given in Table III for various values of the penalty α in simulations. This clearly shows that the penalty factors 3 and 4 are too low for an accurate derivative. The possibility of overfit that belongs to those lower values of α creates a problem for all SNR's. Penalty factors 15, 50, and 100 are generally too high. The increase of $J_p(y')$ for those values of α is always caused by underfit: not every important polynomial order is selected with those penalties. The distance to underfit depends on the SNR; hence, optimal values for the penalty may also depend on the SNR. A penalty factor between 5 and 8 will be a compromise with always a reasonable accuracy. Based on these and numerous other simulation results, a good value for the penalty factor α is 6 if the derivative of the selected model is important.

The use of consistent criteria has been investigated thoroughly in time series, from different points of view [9]. The penalty $\log(N)$ is not a constant in consistent criteria, but

TABLE II
AVERAGE IN 1000 SIMULATION RUNS OF THE QUALITY $J_p(y)$ OF SELECTED HIERARCHICAL MODELS AS A FUNCTION OF THE PENALTY FACTOR α FOR DIFFERENT TRUE PROCESSES η AS DESCRIBED IN THE TEXT, N = 50

	True process η						
α	A	B	C	D	E	F	G
1.8	6.36	7.07	6.51	7.77	6.10	8.05	6.27
2	6.09	6.94	6.24	7.68	5.89	7.95	5.97
3	5.15	***6.43***	5.29	***7.43***	5.03	***7.67***	4.98
4	***4.84***	6.44	***5.06***	7.55	***4.82***	7.86	***4.53***
5	4.90	6.59	5.11	7.68	4.83	8.21	4.59
6	5.12	6.85	5.33	8.29	5.07	8.61	4.68
7	5.37	7.13	5.73	8.99	5.52	9.16	4.90
8	5.68	7.61	6.21	9.81	5.90	9.83	5.20
10	6.86	8.98	7.39	11.92	6.80	11.32	6.34
15	10.70	12.48	11.02	19.60	10.58	14.51	9.87
50	18.95	28.18	18.74	54.00	19.78	20.89	21.30
100	19.03	200.58	18.82	65.69	59.32	20.92	21.43

TABLE III
AVERAGE IN 1000 SIMULATION RUNS OF THE QUALITY OF THE DERIVATIVE, $J_p(y')$, OF SELECTED HIERARCHICAL MODELS AS A FUNCTION OF THE PENALTY FACTOR α FOR DIFFERENT TRUE PROCESSES η AS DESCRIBED IN THE TEXT, N = 50

	True process η						
α	A	B	C	D	E	F	G
1.8	2503	2801	2581	3859	1970	2634	2371
2	2147	2462	2257	3493	1621	2408	2088
3	1003	1607	1112	2351	793	1579	1076
4	693	1240	614	1890	445	1238	610
5	517	1073	501	1537	379	1061	482
6	***464***	1073	***471***	***1411***	***363***	***985***	***407***
7	495	***1067***	513	1416	433	1022	425
8	554	1116	573	1443	520	1062	442
10	740	1247	749	1524	656	1159	584
15	1307	1698	1337	1915	1261	1384	1212
50	2693	2973	2588	3338	2569	1867	2998
100	2707	5147	2601	3353	4891	1870	3040

depends on the number of observations N. Example A of Tables II and III has been studied for N between 20 and 1 000 000. For all N, the best order for the data model was found for $\alpha = 4$ and the best derivative for $\alpha = 6$, for the given SNR. So the best penalty does not depend on N, and consistent criteria provide no solution for the problem of derivatives. These simulations showed that the balance between overfit and underfit is important, requiring a constant value for the penalty factor α.

The best data model was selected in all examples of Table II with 3 or 4 as values for α. This is for lower values of α than the optimal values 6 or 7 found in Table III for the derivative. Therefore, the argument from likelihood theory that "the best model for the derivative is the derivative of the best model for the data" is not true for polynomial models with selected orders.

VII. CONCLUSION

The quality loss, due to *selection* of a model order in comparison with the true or best model order, is moderate for the response itself but very much greater for the derivative of the estimated polynomial model. Too high polynomial orders give a huge degradation of the accuracy of the derivative, but high orders are easily selected. A higher penalty factor α in an order selection criterion gives models with better derivatives. The value 6 is a good value for α in simulations with a range of different examples and SNR's, and is a compromise between overfit and underfit especially for derivatives. If the order has to be selected, the best model for the derivative is generally not found as the derivative of the best model for the data.

REFERENCES

[1] G. A. F. Seber, *Linear Regression Analysis*. New York: Wiley, 1977.
[2] R. R. Hocking, "The analysis and selection of variables in linear regression," *Biometrics*, vol. 32, pp. 1–9, 1976.
[3] C. L. Mallows, "Some comments on C_p," *Technometrics*, vol. 15, pp. 661–675, 1973.
[4] R. Shibata, "Selection of the order of an autoregressive model by Akaike's information criterion," *Biometrika*, vol. 63, pp. 117–126, 1976.
[5] P. M. T. Broersen and H. E. Wensink, "On the penalty factor for autoregressive order selection in finite samples," *IEEE Trans. Signal Processing*, vol. 44, pp. 748–752, 1996.
[6] S. Zacks, *The Theory of Statistical Inference*. New York: Wiley, 1971.
[7] P. Stoica and R. Moses, *Introduction to Spectral Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1997.
[8] R. Shibata, "Approximate efficiency of a selection procedure for the number of regression variables," *Biometrika*, vol. 71, pp. 43–49, 1984.
[9] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 41, pp. 465–471, 1978.

Piet M. T. Broersen was born in Zijdewind, The Netherlands, in 1944. He received the M.Sc. degree in applied physics in 1968 and the Ph.D degree in 1976, both from the Delft University of Technology, Delft, The Netherlands.

He is currently in the Department of Applied Physics, Delft University. His research interests are the selection of order and type of time series models and the application to spectral analysis, model building. and feature extraction.