

# On model selection in additive regression

**Author(s):**
Gosoniu, Nicoleta-Francisca

**Publication Date:**
2008

**Permanent Link:**

https://doi.org/10.3929/ethz-a-005561290 →

Diss. ETH No. 17637

# On Model Selection in Additive Regression

A dissertation submitted to

ETH ZURICH

for the degree of

Doctor of Mathematics

presented by

NICOLETA FRANCISCA GOŞONIU
MSc., Dipl. Math. University of Bucharest, Romania
born June 2, 1977
citizen of Romania

accepted on the recommendation of

Prof. Dr. Hans Rudolf Künsch, examiner
Prof. Dr. Peter Bühlmann, co–examiner
Prof. Dr. Theo Gasser, co–examiner
Prof. Dr. Burkhardt Seifert, co–examiner

2008

# Abstract

Nonparametric additive models were introduced in order to reduce the complexity of a general regression model and as a possible way to deal with the curse of dimensionality (Stone, 1985, Hastie and Tibshirani, 1990).

Several approaches have been proposed in the literature to estimate the additive components of such models: backfitting (Breiman and Friedman, 1985), marginal integration (Linton and Nielsen, 1995), smooth backfitting (Mammen et al., 1999) or penalized regression splines (Wood, 2000, 2001).

In the present thesis, we focus on estimation by smooth backfitting, motivated by the method's good theoretical and practical properties. In particular, we investigate properties of the local linear smooth backfitting estimates (SBE) for additive models (Mammen et al. 1999, Nielsen and Sperlich, 2005) in $d$ dimensions. To gain insight into the structure of the additive estimators, we choose an equidistant design.

Under a new appropriate identifiability condition (Mammen and Park, 2005), the SBE estimates of the additive components are shown to be the centered 1-dimensional local linear estimates.

We prove the averaged squared error (ASE) optimality of a smoothing parameter selection criterion, $AIC_T$ (Studer et al., 2005) related to the Akaike Information Criterion (AIC), Akaike (1973). The $AIC_T$ turns out to be equivalent to the PLS (Penalized Least Squares) criterion, introduced by Mammen and Park (2005).

In a nonparametric setup, model selection requires simultaneous selection of a model and of a smoothing parameter. The assumptions of Mammen and Park (2005) are therefore generalized by extending the admissible range of the smoothing parameter to $(0, \infty]^d$.

The standard assumption that the second derivative of the true regression function is different from zero is not sufficient to get the optimal smoothing parameter of order $n^{-1/5}$. It is therefore proven that it is necessary to also impose the additional assumption that the second derivative of the true regression function is not a constant. As a result, the optimal smoothing parameters in each dimension when estimating the whole regression function are different from the corresponding 1-dimensional smoothing parameters when separately estimating each additive component.

Simulation studies are provided to address the probability of over- and underfitting in additive models with model selection. In our nonparametric setup, for $d = 1$, the probability of overfitting by 1 is approximately 0.3, exceeding the linear case by roughly a factor two. For $d = 2$, the correct model is selected with a probability of about 0.5.

Also, the thesis presents the extension to nonparametric regression of a type of $R^2$ coefficient, as introduced by Rousson and Goşoniu (2007) in the context of linear regression problems. The nonparametric coefficient is again evaluated by means of simulation experiments.

**Keywords:** additive models, SBE estimates, optimal bandwidth, model selection, AIC-type criteria, R-squared.

# Zusammenfassung

Nichtparametrische additive Modelle wurden eingeführt, um die Komplexität eines allgemeinen Regressionsmodelles zu reduzieren, und bieten eine Möglichkeit, mit dem sogenannten Fluch der Dimensionalität umzugehen (Stone, 1985, Hastie und Tibshirani, 1990).

Mehrere Verfahren zur Schätzung der additiven Modellkomponenten sind in der Fachliteratur verfügbar: Backfitting (Breiman und Friedman, 1985), Marginal Integration (Linton und Nielsen, 1995), Smooth Backfitting (Mammen et al., 1999) und Penalized Regression Splines (Wood, 2000, 2001).

Die vorliegende Arbeit untersucht die Schätzung durch Smooth Backfitting, angeregt durch die guten theoretischen und praktischen Eigenschaften der Methode. Insbesondere studieren wir Eigenschaften lokal-linearer Smooth Backfitting Schäzer (SBE) für $d$-dimensionale, additive Modelle (Mammen et al. 1999, Nielsen und Sperlich, 2005). Um die Struktur der additiven Schätzer herauszustellen, verwenden wir ein äquidistantes Design.

Unter einer neuen, geeigneten Identifizierbarkeits-Bedingung (Mammen und Park, 2005) wird gezeigt, dass die SBE-Schätzwerte der additiven Komponenten gerade die zentrierten, eindimensionalen, lokal-linearen Schätzungen sind.

Für $\mathrm{AIC}_T$ (Studer et al., 2005), ein zum Akaike Informationskriterium (AIC; Akaike, 1973) verwandtes Auswahlkriterium für Glättungsparameter, zeigen wir Optimalität im Sinne des mittleren quadratischen Fehlers (Averaged Square Error, ASE). Das $\mathrm{AIC}_T$-Kriterium erweist sich als äquivalent zu dem von Mammen und Park (2005) eingeführten Kriterium PLS (Penalized Least Squares).

In einem nichtparametrischen Regime erfordert Modellselektion die gleichzeitige Wahl eines Modells und eines Glättungsparameters. Wir verallgemeinern daher die Annahmen von Mammen und Park (2005) derart, dass der Wertebereich des Glättungsparameters auf $(0, \infty]^d$ erweitert wird.

Um optimale Werte des Glättungsparameters von der Ordnung $n^{-1/5}$ zu garantieren, erweist sich die herkömmliche Voraussetzung, dass die wahre Regressionsfunktion eine nicht-verschwindende zweite Ableitung besitzt, als nicht hinreichend. Als zusätzliche Annahme fordern wir daher, dass die genannte Ableitung nicht konstant ist. Die für die Schätzung der gesamten Regressionsfunktion optimalen Werte der Glättungsparameter in jeder Dimension unterscheiden sich daher von den optimalen Werten für die entsprechenden eindimensionalen Glättungsparameter, die sich bei separater Schätzung der additiven Komponenten ergeben.

Die Wahrscheinlichkeit für Over- und Underfitting bei der Modellwahl in additiven Modellen wird mit Hilfe von Simulationsexperimenten studiert. In unserem nichtparametrischen Szenario ist die Wahrscheinlichkeit des Overfittings um 1, für $d = 1$, etwa 0.3, und damit circa zweimal so hoch wie im linearen Fall. Für $d = 2$ wird das korrekte Modell mit einer Wahrscheinlichkeit nahe 0.5 gewählt.

Desweiteren stellt die Doktorarbeit die Erweiterung eines Koeffizienten vom Typ $R^2$ auf die nichtparametrische Regression vor, der von Rousson und Goşoniu (2007) im Rahmen der linearen Regressionsprobleme eingeführt wurde. Dieser nichtparametrische Koeffizient wird ebenfalls experimentell durch Simulationsexperimente untersucht.

**Stichwörter:** additive Modelle, SBE Schätzungen, optimaler Glättungsparameter, Modellauswahl, AIC-Typ Kriterien, R-Quadrat.