

ACTA

UNIVERSITATIS OULUENSIS

Ilmari Juutilainen

MODELLING OF
CONDITIONAL VARIANCE
AND UNCERTAINTY USING
INDUSTRIAL PROCESS DATA

FACULTY OF TECHNOLOGY,
DEPARTMENT OF ELECTRICAL AND INFORMATION ENGINEERING,
UNIVERSITY OF OULU

C
TECHNICA



ACTA UNIVERSITATIS OULUENSIS
C Technica 258

ILMARI JUUTILAINEN

**MODELLING OF CONDITIONAL
VARIANCE AND UNCERTAINTY
USING INDUSTRIAL PROCESS DATA**

Academic dissertation to be presented, with the assent of
the Faculty of Technology of the University of Oulu, for
public defence in Auditorium TA105, Linnanmaa,
on November 24th, 2006, at 12 noon

OULUN YLIOPISTO, OULU 2006

Copyright © 2006
Acta Univ. Oul. C 258, 2006

Supervised by
Professor Juha Röning

Reviewed by
Professor Lasse Holmström
Professor Olli Simula

ISBN 951-42-8261-2 (Paperback)
ISBN 951-42-8262-0 (PDF) <http://herkules.oulu.fi/isbn9514282620/>
ISSN 0355-3213 (Printed)
ISSN 1796-2226 (Online) <http://herkules.oulu.fi/issn03553213/>

Cover design
Raimo Ahonen

OULU UNIVERSITY PRESS
OULU 2006

Juutilainen, Ilmari, Modelling of conditional variance and uncertainty using industrial process data

Faculty of Technology, University of Oulu, P.O.Box 4000, FI-90014 University of Oulu, Finland,
Department of Electrical and Information Engineering, University of Oulu, P.O.Box 4500, FI-90014 University of Oulu, Finland

Acta Univ. Oul. C 258, 2006

Oulu, Finland

Abstract

This thesis presents methods for modelling conditional variance and uncertainty of prediction at a query point on the basis of industrial process data. The introductory part of the thesis provides an extensive background of the examined methods and a summary of the results. The results are presented in detail in the original papers.

The application presented in the thesis is modelling of the mean and variance of the mechanical properties of steel plates. Both the mean and variance of the mechanical properties depend on many process variables. A method for predicting the probability of rejection in a qualification test is presented and implemented in a tool developed for the planning of strength margins. The developed tool has been successfully utilised in the planning of mechanical properties in a steel plate mill.

The methods for modelling the dependence of conditional variance on input variables are reviewed and their suitability for large industrial data sets are examined. In a comparative study, neural network modelling of the mean and dispersion narrowly performed the best.

A method is presented for evaluating the uncertainty of regression-type prediction at a query point on the basis of predicted conditional variance, model variance and the effect of uncertainty about explanatory variables at early process stages. A method for measuring the uncertainty of prediction on the basis of the density of the data around the query point is proposed. The proposed distance measure is utilised in comparing the generalisation ability of models. The generalisation properties of the most important regression learning methods are studied and the results indicate that local methods and quadratic regression have a poor interpolation capability compared with multi-layer perceptron and Gaussian kernel support vector regression.

The possibility of adaptively modelling a time-varying conditional variance function is disclosed. Two methods for adaptive modelling of the variance function are proposed. The background of the developed adaptive variance modelling methods is presented.

Keywords: joint modelling of mean and dispersion, model uncertainty, process data, tensile properties, time-varying parameter, variance estimation, variance function, variance heterogeneity

Acknowledgements

First of all I want to thank my supervisor prof. Juha Rönning, who organised the opportunity and the projects which made the research possible. Without his friendly support and wise advices the thesis could not have been finished. I am deeply indebted to my reviewers, prof. Lasse Holmström who pointed me out several places to improve the thesis, and prof. Olli Simula. I greatly thank prof. Timo Koski who has kindly promised to be my opponent in the public defence.

I heartily thank my friend and colleague Dr. Perttu Laurinen, whose daily talking and jokes makes the research work enjoyable and whose help solved countless of problems that I encountered in the research.

I am grateful to Ruukki Corporation and its employees whose needs and interest motivated the research and who provided the data and opportunity to real applications. Especially I thank quality manager Lassi Myllykoski who excellently led the interaction between research and application and whose discussions gave me understanding of applied industrial statistics.

I honestly thank my colleagues whose pleasant company at meetings, lunch and coffee breaks et al. gatherings creates a wonderful working atmosphere.

I thank the Graduate School on Metallurgy and Metals Technology, Ruukki corporation and the National Technology Agency who mainly funded the research. I thank Technological Foundation who supported the research by a personal scholarship.

I am deeply thankful to my darling wife Anna-Riikka, whose loving presence ensures my happiness. I sincerely appreciate my family and friends whose fellowship greatly contribute to the fact that I have a very enjoyable life. I thank God for allowing me to live happily here among nice friends, beautiful nature and interesting tasks.

Oulu, October 2006

Ilmari Juutilainen

List of original publications

- I Juutilainen I, Röning J, Myllykoski L (2003): *Modelling the Strength of Steel Plates Using Regression Analysis and Neural Networks*. In: Proceedings of International Conference on Computational Intelligence for Modelling, Control and Automation (CIMCA' 2003), Vienna, Austria, 681–691.
- II Juutilainen I, Röning J (2004): *Heteroscedastic Linear Models for Analysing Process Data*. WSEAS Transactions on Mathematics 2: 179–187.
- III Juutilainen I, Röning J (2004): *Modelling The Probability of Rejection in a Qualification Test Based on Process Data*. In: Proceedings of 16th Symposium of IASC (COMPSTAT 2004), Prague, Czech Republic, 1271–1278.
- IV Juutilainen I, Röning J (2006): *Planning of Strength Margins Using Joint Modelling of Mean and Dispersion*. Materials and Manufacturing Processes 21: 367 – 373.
- V Juutilainen I, Röning J (2005): *A Comparison of Methods for Joint Modelling of Mean and Dispersion*. In: Proceedings of International Symposium on Applied Stochastic Models and Data Analysis (ASMDA 2005), Brest, France, 1499–1506.
- VI Juutilainen I, Röning J (forthcoming) *A Method for Measuring Distance from a Training Data Set*. Communications in Statistics (accepted for publication).
- VII Juutilainen I, Röning J, Laurinen P (2005): *A Study on the Differences in the Interpolation Capabilities of Models*. In: Proceedings of IEEE Mid-Summer Workshop on Soft Computing in Industrial Applications (SMCia/05), Espoo, Finland, 202–207.
- VIII Juutilainen I, Röning J (2006): *Adaptive Modelling of Conditional Variance Function*. In: Proceedings of 17th Symposium of IASC (COMPSTAT 2006), Rome, Italy, 1517–1524.

The author has originated the main ideas and performed a majority of the experiments in all the publications. Mr. Myllykoski wrote a section concerning the production of steel plates in original paper I, but otherwise the author has completely written all the publications. Prof. Röning has commented all the manuscripts and proposed several improvements, and Mr. Laurinen has commented and originated some ideas in original paper VII.

Symbols and abbreviations

$\ x\ $	<i>L2-norm of the vector x</i>
i.i.d.	<i>identically independently distributed</i>
$a = o(x)$	<i>a is order of x</i>
$I(x)$	<i>indicator function: returns 1 if x is true and otherwise returns 0</i>
$f(\beta, x_i)$	<i>the regression function for the mean</i>
$g(\tau, x_i, \mu_i)$	<i>the variance function</i>
$K(\cdot)$	<i>kernel function</i>
$l(\cdot)$	<i>log-likelihood function</i>
$L(\cdot)$	<i>loss function</i>
$L_\lambda(y)$	<i>Box-Cox transformation function</i>
$V(\mu)$	<i>the variance function in GLM</i>
$N(\mu, \sigma^2)$	<i>Gaussian distributed with expectation μ and variance σ^2</i>
$\text{Gamma}(\mu, \phi)$	<i>Gamma distributed with expectation μ and dispersion ϕ</i>
$F(s)$	<i>cumulative distribution function</i>
$p(s)$	<i>probability density function</i>
$\Phi(s)$	<i>standard normal cumulative distribution</i>
n	<i>number of training data observations</i>
P_t	<i>the update matrix of adaptive models at time t</i>
p	<i>number of input variables</i>
q	<i>number of parameters in variance model</i>
T	<i>training data set</i>
V	<i>validation data set</i>
w_i	<i>prior estimation weight for observation i</i>
x_i	<i>the i th observation of the explanatory variable vector</i>
X	<i>matrix of all input variable observations</i>
x_{ij}	<i>the i th observation of the j th input variable</i>
Y	<i>vector of all response variable observations</i>
y_i	<i>the i th response variable observation</i>

z_i	<i>the i th input vector of variance model</i>
β	<i>the parameter vector related to the mean function</i>
β_t	<i>the mean parameter vector at time t</i>
$\hat{\varepsilon}_i$	<i>the residual of the i th observation</i>
η_t	<i>learning rate</i>
μ_i	<i>the conditional mean of the i th observation</i>
σ_i^2	<i>the conditional variance of the i th observation</i>
τ	<i>the parameter vector related to the variance function</i>
τ_t	<i>the variance parameter vector at the time moment t</i>
ϕ	<i>dispersion parameter in GLM</i>
A5	<i>elongation in the tensile test</i>
EQL	<i>extended quasi-likelihood</i>
GLM	<i>generalised linear model</i>
IWLS	<i>iteratively weighted least squares</i>
MCMC	<i>Monte-Carlo Markov chain</i>
ML	<i>maximum likelihood</i>
ReH	<i>yield strength</i>
REML	<i>restricted maximum likelihood</i>
Rm	<i>tensile strength</i>

Contents

Abstract

Acknowledgements

List of original publications

Symbols and abbreviations

Contents

1	Introduction	13
1.1	Background	13
1.2	Motivation	14
1.3	Contribution	15
1.4	Summary of original papers	16
2	Variance Modelling	17
2.1	Reasons for modelling conditional variance	17
2.2	Background - Quality improvement experiments	18
2.2.1	The Taguchi method	18
2.2.2	The dual response surface approach	19
2.2.3	Dual response surface optimisation	20
2.3	Response variables in variance function estimation	20
2.3.1	Squared residual	21
2.3.2	Logarithm of residual	22
2.3.3	Absolute residual	22
2.3.4	Quadratic forms	22
2.3.5	Results for steel plate data	23
2.4	Estimation methods of variance function	23
2.4.1	Maximum likelihood	24
2.4.2	Restricted maximum likelihood	24
2.4.3	Pseudo-likelihood	25
2.4.4	Extended quasi-likelihood	26
2.4.5	Robust estimation	26
2.4.6	Results for steel plate data	27
2.5	Models for joint modelling of mean and variance	28
2.5.1	Heteroscedastic regression	28
2.5.2	Generalised linear models in dispersion modelling	29
2.5.3	Local methods in variance function estimation	29
2.5.4	Mean and dispersion additive models	30
2.5.5	Reproducing kernels	31
2.5.6	Bayesian methods	32
2.5.7	Neural networks	32
2.5.8	Results for steel plate data	33
2.6	Modelling of conditional distribution	33

2.6.1	Conditional density estimation	34
2.6.2	Quantile regression	35
2.6.3	Modelling the parameters of distribution	35
3	Adaptive modelling of a variance function	36
3.1	Recursive estimation	36
3.1.1	Forgetting factors	37
3.1.2	Rolling regression	37
3.2	On-line learning	38
3.2.1	On-line quasi-Newton algorithm	39
3.2.2	On-line neural networks	39
3.2.3	Reproducing adaptive kernels	39
3.2.4	Real-time lazy learning	40
3.3	Time-varying parameter regression	40
3.3.1	Change point models	40
3.3.2	Stochastic coefficient models	41
3.3.3	Functional coefficient models	41
3.4	Models with time-varying variance	42
3.4.1	Conditional autoregressive heteroscedasticity	42
3.4.2	Stochastic volatility models	42
3.5	Adaptive modelling of conditional variance	43
3.5.1	Variance function estimation with recursive smoothing	44
3.5.2	Moving window modelling of conditional variance	44
3.5.3	Adaptive on-line quasi-Newton algorithm	44
3.5.4	Results in steel plate data	45
3.6	Simultaneous adaptive modelling of mean and dispersion	45
4	Industrial process data in predictive modelling	47
4.1	Dispersion modelling using process data	47
4.2	Model selection	48
4.3	Utilisation of industrial prediction models	49
4.4	Model maintenance	50
4.5	Measuring distance between data point and training data set	51
4.6	Interpolation ability of models	52
5	Application to the mechanical properties of steel plates	54
5.1	Modelling the probability of rejection in a qualification test	54
5.2	Steel plate data set	56
5.3	Planning the strength margins of steel plates	57
5.4	Modelling the mean and variance of elongation	58
6	Discussion	61
6.1	Variance function estimation	62
6.2	Process data-based modelling	62
6.3	Adaptive modelling	63
6.4	Uncertainty of prediction	64
7	Summary	65
	References	66
	Original communications	

1 Introduction

1.1 Background

The purpose of regression modelling is to find out how the conditional distribution of the response depends on the explanatory variables. In the usual approach, the conditional distribution is derived from the distributional assumption and the predicted mean. A more general approach is to also model the dependence of variance on the explanatory variables and to approximate the conditional distribution using both the estimated variance and the estimated mean. Further, the most general approach is to model the whole conditional distribution function as a function of the explanatory variables. The last approach is the most difficult, whereas the first ones are special cases where some simplifying assumptions about the form of the conditional distribution functions are made.

A large number of statistical methods have been developed for predicting the conditional mean. It is said that the response depends on the explanatory variables via a regression function. The true values of the regression function are not observed, because the response observations involve a stochastic component. The learning method defines how the regression function is presented and how its parameters are estimated.

Learning methods are often divided into parametric and non-parametric methods. In the parametric methods, the number of model parameters is relatively small, so assumptions about the model structure restrict the possible regression functions. Non-parametric methods employ so many model parameters that the regression function is practically not at all restricted by the assumptions. The model parameters are estimated by minimising a loss function in a training data set. The most usual loss function is the sum of the squared errors. A penalty term is often included in the loss function to penalise complex models and force the models to be smoother. Too complex, overfitted models do not generalise well. Too simple models can not fully capture the interesting relationship revealed by the data.

Statistical prediction models are commonly utilised for industrial purposes, for example in process control, process planning and product planning. In an industrial production process, millions of measurements of the process can be made daily. The term 'process data' refers to data that are automatically measured in an industrial production process. Process data are commonly stored and analysed for the purpose of finding some useful information. Process data are employed to fit statistical models for predicting the con-

ditional distribution of interesting variables. The developed models can be implemented into the production line.

The first method for modelling the dependence of conditional variance on input variables was proposed by Bartlett & Kendall (1946). During the last decades, joint modelling of the mean and dispersion has received more attention and modelling methodology has been rapidly developed. In addition to growing computing power, understanding the relationship between quality and variance has boosted this development. Most commonly, joint modelling of the mean and dispersion has been used to analyse industrial quality improvement experiments. A simple method for modelling conditional variance is heteroscedastic linear regression, but tens of other parametric and non-parametric learning methods have been proposed.

1.2 Motivation

Information about dispersion is essential in prediction: The predicted mean without any other information about distribution does not give much information about the expected realisations of the response. The usual approach of assuming constant variance or some relationship between the mean and variance is sometimes inadequate: The distribution of the response can depend on the explanatory variables in a way which can not be described meaningfully by modelling only the mean and estimating a single dispersion parameter. A model for the dependence of variance on the explanatory parameters may be needed.

When variance depends on the process settings, the model describes the interesting conditional distribution more accurately when the variance heteroscedasticity is also modelled. The more accurate model gives additional efficiency to the industrial utilisation of prediction models. A variance model can work as a helpful tool in actions aiming to decrease variance. Despite that, production line implementations of models for conditional variance have not been reported earlier.

In process data, the numbers of both observations and variables are often large. The modelling methods must be able to handle several input variables and large data sets. Process data mostly consist of measurements made using the normal process settings. For process development, it is equally important to predict well also outside the normal process settings. The ability to generalise at the boundaries of the training data set is therefore an important property, but methods for measuring it have not been proposed. The interpolation abilities of different learning methods have not been studied or compared.

When predictive regression models are utilised in industry, predictions are often queried from the boundaries of the data region. For planning purposes, the benefits of the model may be mainly in the information that the model gives from the rarely observed data region - the behaviour of the process with familiar settings is already known. Unfortunately, the accuracy of a model is not guaranteed at the boundaries: Information about the reliability of prediction is needed. Conditional variance and uncertainty of prediction are closely related topics, and the application may determine which of them is more useful.

A successful industrial plant must develop its processes, which means that industrial processes are changing. Changes in the process may change the relationships between

the variables: The regression function varies over time. One of the major aims of quality improvement is to decrease variance. Thus, variance can be expected to vary even more than the mean. Adaptive models can adapt to time-varying changes in the regression function. Adaptive statistical modelling of the mean is a well-known topic, but adaptive modelling of a conditional variance function has not been studied before. In industrial utilisation it is, however, important that the models can be kept up-to-date.

Although many methods have been proposed for joint modelling of the mean and dispersion, reviews and comparative studies about the different methods are missing. Also, reported applications to large data sets are almost completely missing. Joint modelling of the mean and dispersion has been used in industry to analyse the results of designed experiments, but applications of modelling variance on the basis of process data have been rarely reported.

1.3 Contribution

Process data-based prediction and joint modelling of the mean and dispersion are both well-known topics that have not previously been applied together. In this thesis the two approaches are combined. The research studies the question "*How is joint modelling of the mean and dispersion efficiently applied to industrial process data?*". Modelling of conditional variance is successfully applied to steel industry data. This thesis concentrates on regression problems where the mean and variance of a quantitative response variable depend on several input variables.

The main application is development of a tool for planning working allowances for the strength of steel plates. The application utilises joint modelling of the mean and variance to predict the probability of rejection in tensile testing. The proposed approach is generally applicable for predicting the probability of rejection in qualification tests and adjusting working allowances on the basis of the predicted rejection probabilities.

The existing methods for variance modelling are reviewed and their suitability to large data sets are considered. A comparative study of the prediction accuracy of the most important methods is conducted.

Some novel methodology related to dispersion modelling is presented. A novel neural network approach for variance modelling is presented. A procedure for model selection in joint modelling of the mean and dispersion is presented. A method is proposed for estimating uncertainty of prediction at early process stages when there is uncertainty about some explanatory variables.

Statistical prediction models allowing time-varying variance are reviewed. Two novel algorithms are developed for adaptive modelling of conditional variance. An approach for adaptive joint modelling of the mean and dispersion is proposed.

A method for measuring the distance of a single observation from a training data set is proposed. It is proposed that the distance measure can be utilised in examining the interpolation abilities of models. A comparative study about the interpolating capabilities of learning methods is conducted.

1.4 Summary of original papers

This thesis consists of eight publications. The publications discuss methods for modelling conditional variance and measuring the uncertainty of prediction with applications to steel plate data.

In original paper I, the strength of steel is modelled using linear regression and neural networks. In original paper II, heteroscedastic linear models are applied to steel plate data. A method for model selection in joint modelling of the mean and variance is also proposed. It is shown that the production method has a clear effect on the variance of strength. In original paper IV an implementation of the planning model for optimising the working allowances of steel plates is described. The planning model utilises the developed heteroscedastic models to predict the risk of rejection in a qualification test. It is shown that the model of the variance of strength improves the ability of the planning model to gain economical benefits. In original paper III, the approach developed for predicting the probability of rejection in a qualification test is presented in a more general framework.

Original paper V is a comparative study on the suitability of different methods for joint modelling of the mean and dispersion with large data sets. The proposed novel neural network modelling of the mean and dispersion performed the best in the study. A short review of the methods for joint modelling of the mean and dispersion is also given.

Several results related to evaluating the uncertainty of prediction are presented in the original papers. A method for assessing the uncertainty of prediction at early process stages, where complete information about input variables is not available, is suggested in original paper III and successfully utilised with steel plate data in original paper IV. Original paper VI presents a novel measure of the distance between the training data set and a single observation. The distance measure reflects the uncertainty of prediction when the observation is predicted on the basis of the data set. The distance measure is utilised in original paper VII, where the differences between the interpolation capabilities of models are compared. The main results of the study are that there are differences between learning methods but, model complexity does not have a clear effect on the interpolation capability.

Original paper VIII discusses adaptive modelling of conditional variance. Two methods are proposed for adaptive modelling of the conditional variance function. An approach for adaptive joint modelling of the mean and dispersion is proposed.

2 Variance Modelling

Many methods for statistical modelling of the dependence of variance on input variables have been developed and applied in various fields. Also, one textbook about variance modelling has been written (Carroll & Ruppert 1988). This chapter reviews methods for modelling conditional variance. Industrial quality improvement experiments have been a major field of application of variance modelling: Dispersion modelling has been employed to analyse experiments designed for finding the process settings that minimise variance under given conditions. The development of dispersion modelling has been connected to quality improvement experiments, and therefore the history of quality improvement methods is presented in detail.

2.1 Reasons for modelling conditional variance

Variance modelling has been justified by several reasons. The following motivations for variance function estimation are the most common:

1. Information about variance is needed in the construction of confidence intervals and in predicting conditional distribution. Prediction assuming constant variance produces misleading results under variance heterogeneity.
2. The efficiency of mean model estimation can be improved using weighted estimation with weights proportional to the inverses of error variances. In practice, the variances are not known and considerable attention has been paid to obtaining efficient estimates of the mean with an unknown variance function.
3. In many applications variance is not a nuisance parameter, but a parameter of interest. Especially in quality engineering, the purpose of data analysis can be to obtain information about conditional variance. Small variance means good quality, and therefore variance reduction is a desired goal in industry.
4. A model for variance is utilised in optimisation problems.

If there are enough replications from each of the design points, the variance at each design point can be directly measured. In that case, modelling of variance is simple, and all the usual modelling methods can be used without needing to care about the mean.

Process data usually do not contain many replications, and thus direct measurements of variance are not available. In practice, the mean is not known, so the mean and variance are modelled jointly. A single observation does not give any information about variance, and more observations are needed to model variance than the mean.

In many types of models, like generalised linear models, variance component models and models with noise factors, the model structure implicitly defines a non-constant variance structure. A model for a transformed response also leads to non-constant variance in the original response scale. Often the need for variance modelling can be avoided with an appropriate selection of the model type. In regression analysis, variance stabilising transformations of the response, for example logarithm or Box-Cox transformations, have been widely used.

2.2 Background - Quality improvement experiments

Quality improvement on the basis of statistical analysis of designed experiments has been an important application since the beginning of the development of industrial statistics. The basic modelling concept was to first estimate response surfaces for the mean using regression techniques. The process variables were then optimised on the basis of the estimated response surface. Montgomery (1999) has written a review paper about using experimental design for process and product development.

The importance of variance reduction as the aim of quality improvement experiments was also understood early. Bartlett & Kendall (1946) proposed using the logarithm of the sample variance as the response in variance modelling. Later, variance modelling became a known and occasionally discussed topic (Park 1966, Rutemiller & Bowers 1968, Harvey 1976, Jobson & Fuller 1980, Carroll 1982). Variance modelling became a popular topic at the end of the 1980s, when robust design was introduced to quality improvement experiments in industry. Japanese engineer Taguchi introduced parameter design, later called robust design, to reduce variation in products or processes. Taguchi's ideas became popular and invoked much discussion among statisticians and quality engineers (Nair 1992). The importance of variance in quality engineering was recognised, which led to the rapid development of statistical methods for dispersion modelling. At the same time, Taguchi's ideas have made statistical methods better known among engineers. (Nair & Pregibon 1988.)

2.2.1 The Taguchi method

Taguchi's philosophy aims at minimising the total cost of products. Total cost consists of production costs and the costs for the customer. Quality can often be measured as variation in the properties of the product. In robust design, the process variables are classified into two categories: control factors, x , which can be easily controlled with minimal cost, and noise factors, z , which are difficult or expensive to control. Variation in the noise factors during the process causes variation in the outputs of the process. Robinson *et al.* (2004)

has written a review about robust design.

The basic idea in robust design is to find optimal settings for the control factors, so that the process is insensitive to variation in the noise factors and at the same time the process results in the desired outputs (Nair 1992). Taguchi recommended experiments with factorial, orthogonal designs to obtain data for statistical analysis. Taguchi defined a performance criterion called signal-to-noise ratio for measuring the quality of the process (Taguchi 1986). For example, when the process has a fixed target value, the signal-to-noise ratio is proportional to the logarithm of the proportion between the squared mean and the variance

$$\text{SNR} = \log \left(\frac{(E y)^2}{\text{var}(y)} \right). \quad (1)$$

At each of the design points, replications are used to measure the signal-to-noise ratio. Then, standard analysis of the variance techniques are applied to the signal-to-noise ratio with the purpose of identifying robust settings for control factors. (Taguchi 1987.) The statistical efficiency of Taguchi's analysis methods and the complexity of his experimental design have later been criticised (Nair 1992).

Other methods for minimising variance on the basis of factorial experiments were published. For example, Box (1988) proposed to modify Taguchi's analysis with a power transform of the response to make the mean and variance independent. Another proposal was to base optimisation directly on the sample means and sample variances (Shoemaker *et al.* 1991).

2.2.2 The dual response surface approach

The term 'response surface' is broadly used to mean the surface of predictions of a regression-type model. The term 'response surface methodology' has been used to mean the analysis of factorial experiments using the second order polynomial

$$E y_i = x_i^T \beta + x_i^T \Lambda x_i \quad (2)$$

response surface (Myers 1999). Response surface methodology had been intensively used in industry. The aim is usually to optimise the control factors due to some objective.

In many applications there are two or more responses of interest. In the dual response approach (Myers & Carter 1973), a response surface is fitted for two variables. In the basic optimisation of a dual response system, the first response is optimised under constraints on the other response. Vining & Myers (1990) proposed the dual response surface approach for finding the optimal settings for control factors in the presence of noise factors. They used the first response surface for the mean and the other for variance. In their approach the minimisation of variance is based on the setting

$$y_i = x_i^T \beta + z_i^T \tau + x_i^T \Lambda z_i + \varepsilon_i \quad (3)$$

where β and τ are parameter vectors, Λ is a parameter matrix for the interactions, x is fixed and z is random. The response surface of variance

$$\text{var}(y_i) = (\tau^T + x_i^T \Lambda) \text{cov}(z_i) (\tau^T + x_i^T \Lambda)^T + \sigma^2 \quad (4)$$

is a function of control factors. Here $\text{var}(\varepsilon_i) = \sigma^2 \quad \forall i$, which means variance is constant if there are no interactions between control and noise factors. The optimal settings for control factors are found using direct constrained optimisation for the dual response surface of the mean and variance (Myers *et al.* 1992).

Both Taguchi (1986) and Vining & Myers (1990) assume that noise variables cause variance heterogeneity. However, variance heterogeneity often can not be explained by uncontrollable noise factors, but variance also depends on the control factors. In that case, a model for conditional variance has to be fitted separately.

Several regression approaches based on modelling of conditional variance have been proposed for identifying the optimal values of the process variables. A linear variance model fitted to the squared residuals of the mean model (Hamada & Nelder 1997, Chan & Mak 2001) or the logarithm of squared residuals (Chan & Mak 1995) has been used to analyse quality improvement experiments. Also, double generalised linear models have been suggested for analysing experiments with the purpose of decreasing variation (Nelder & Lee 1991).

2.2.3 Dual response surface optimisation

Industrial applications of dispersion modelling have often been related to optimisation of the process settings, so the optimisation methodology related to dispersion models has been regularly discussed (Kim & Lin 2006). Several methods have been proposed for optimising a dual response surface

$$\begin{aligned}\mu_i &= \mu(\beta, x_i) \\ \sigma_i^2 &= \sigma^2(\tau, x_i)\end{aligned}\tag{5}$$

(Fan 2000, Tang & Xu 2002, K ksoy & Dognaksoy 2003). The papers do not discuss how the response surface of the standard deviation or variance is estimated.

There are three common optimising problems: maximising the response, minimising the response and minimising the difference between the response and the target (Vining & Myers 1990). Tang & Xu (2002) present a high-level general formulation that includes some of the existing optimisation methods as special cases. Their paper also includes a review of existing techniques for dual response surface optimisation. Mak & Nebebe (2003) presented a methodology for optimising a general loss function using the mean, variance and the conditional distribution function of standardised residuals. Carlyle *et al.* (2000) gives a good review of statistical optimisation techniques related to quality engineering.

2.3 Response variables in variance function estimation

Several alternative response variables can be used in the modelling of dispersion. Usually, a model for the mean is fitted and a transformation of the obtained residuals is used as the response in variance function estimation.

2.3.1 Squared residual

Because $E(y_i - \mu_i)^2 = \sigma_i^2$, the squared residual $\widehat{\varepsilon}_i^2 = (y_i - \widehat{\mu}_i)^2$ is a natural response for variance modelling. Especially in the case of a normally distributed response, squared residuals are very attractive because of the result

$$\varepsilon \sim N(0, \sigma^2) \Rightarrow \varepsilon^2 \sim \text{Gamma}(\sigma^2, 2). \quad (6)$$

The notation $y \sim \text{Gamma}(\mu, s)$ means that y is Gamma distributed with expectation μ and variance $s\mu^2$.

Normal distribution of an additive error term is often a reasonable assumption. Even when the response itself is not normally distributed, some monotonic transformation of the response can satisfy the normality assumption. The transformed response is used in modelling, and as the result the predicted distribution in the original scale approximates the distribution of the original response.

Rigby & Stasinopoulos (2000) proposed a method for joint modelling of the mean and dispersion where the response is transformed using the Box-Cox transformation

$$\begin{aligned} L_\lambda(y) &= \frac{y^\lambda - 1}{\lambda}, \quad \text{when } \lambda > 0 \\ L_\lambda(y) &= \log y, \quad \text{when } \lambda = 0. \end{aligned} \quad (7)$$

The authors proposed to first estimate preliminary models for the mean and variance, and then to find the transformation parameter λ that minimises deviance in the original scale

$$D(\lambda) = \sum_{i=1}^n \frac{[L_\lambda(y_i) - \widehat{\mu}_i]^2}{\widehat{\sigma}_i^2} + \sum_{i=1}^n \log(2\pi\widehat{\sigma}_i^2) - 2(\lambda - 1) \sum_{i=1}^n \log y_i. \quad (8)$$

Deviance is defined to be twice the difference between the maximum achievable log-likelihood and the log-likelihood at the maximum likelihood estimates of the parameters. The deviance in the original scale includes the extra contribution to the likelihood from the Jacobian of the Box-Cox transformation: $(\lambda - 1) \sum_{i=1}^n \log y_i$. When testing a transformation L_λ for a linear mean model, the same link function for the mean should be used: $EL_\lambda(y_i) = L_\lambda(x^T \beta)$. Finally, the models for the mean and variance are specified using the optimally transformed response.

When the model is unbiased, it holds that

$$E\widehat{\varepsilon}_i^2 = \sigma_i^2 + \text{var}(\widehat{\mu}_i) - 2\text{cov}(y_i, \widehat{\mu}_i). \quad (9)$$

Fitting a model to the mean biases the variance function estimation, because the measured and predicted responses are correlated whenever the observation is used to fit the mean model. It has been proposed that the bias could be corrected by using a corrected response that takes into account the difference $\text{var}(\widehat{\mu}_i) - 2\text{cov}(y_i, \widehat{\mu}_i)$ (Carroll & Ruppert 1988). When the fit can be expressed with a smoother matrix $\widehat{Y} = SY$, the corrected response

$$\frac{\widehat{\varepsilon}_i^2}{1 - \Delta_i} \quad (10)$$

leads to an unbiased variance function estimation scheme (Ruppert *et al.* 1997). Here, $\Delta = (\Delta_1, \Delta_2, \dots, \Delta_n) = \text{diag}(2S - SS^T)$. The correction can be interpreted as using squared leave-one-out residuals to eliminate the bias of mean model fitting (Cawley *et al.* 2004).

2.3.2 Logarithm of residual

Heteroscedasticity of errors is usually modelled using a framework

$$\varepsilon_i = \sigma_i u_i, u_i \sim \text{i.i.d.}, E u_i = 0, \text{var}(u_i) = 1. \quad (11)$$

In this framework the variance of the logarithm of the squared error term is constant

$$\text{var}(\log \varepsilon_i^2) = \text{var}(\log \sigma_i^2 + \log u_i^2) = \text{var}(\log u_i^2) \forall i. \quad (12)$$

Thus, the dispersion parameters involved in σ_i^2 can be estimated using the standard least squares method, using the logarithms of the squared residuals as the response (Harvey 1976). In the prediction it should be noticed that $E \log \varepsilon_i^2 \neq \log \sigma_i^2$, for example

$$\varepsilon \sim N(0, \sigma^2) \Rightarrow E \log \varepsilon^2 \approx -1.27 + \log \sigma^2 \quad (13)$$

(Harvey 1976). This method is quite robust, but not very efficient, can be biased and the near-zero residuals are problematic (Carroll & Ruppert 1988).

2.3.3 Absolute residual

If it seems reasonable to assume zero median errors, $\text{median}(\varepsilon_i) = 0$, rather than zero mean errors $E \varepsilon_i = 0$, the absolute residuals $|\widehat{\varepsilon}_i|$ become a natural response variable (Welsh *et al.* 1994). Under the assumption Eq. (11) it holds that $E|\varepsilon_i| = c_u \sigma_i$ where the constant $c_u \leq 1$ depends on the distribution of u_i 's (Dadidian & Carroll 1987). For example, under normal distribution it holds that

$$\varepsilon \sim N(0, \sigma^2) \Rightarrow E|\varepsilon| = \sqrt{2\sigma^2/\pi}. \quad (14)$$

So the variances of the absolute error terms are proportional to their squared means,

$$\text{var}(|\varepsilon_i|) = E \varepsilon_i^2 - (E|\varepsilon_i|)^2 = (1 - c_u^2) \sigma_i^2 = (1/c_u^2 - 1) (E|\varepsilon_i|)^2, \quad (15)$$

and the model can be fitted by IWLS with weights proportional to $1/\widehat{\sigma}_i^2$ (Carroll & Ruppert 1988). Another possibility is to model the deviation using the relationship $E|\varepsilon_i| = \sigma_i$ and relax the assumption $E(\varepsilon_i^2) = \sigma_i^2$ (Welsh *et al.* 1994). Absolute residuals are more robust to large errors than squared residuals (Carroll & Ruppert 1988).

2.3.4 Quadratic forms

Because variance can be expressed as $\sigma^2 = E y^2 - (E y)^2$, it has been proposed that variance can be modelled with

$$\sigma_j^2 = \widehat{v}(x_j) - \widehat{\mu}(x_j)^2 \quad (16)$$

where $\widehat{v}(\cdot)$ is a fitted prediction model for y^2 and $\widehat{\mu}(\cdot)$ is a fitted prediction model for y . Different prediction models, like local polynomials (Härdle & Tsybakov 1997) and linear models (Gourieroux & Monfort 1992), have been used. However, this method can yield seriously biased and even negative estimates of variance and it has a large variance (Fan & Yao 1998).

To avoid the mean model estimation, variance modelling on the basis of the squared differences in the response variable has been discussed (Müller & Stadtmüller 1993, 1987). In difference-based methods, the model for variance is estimated using pseudo-residuals p_i , which are properly weighted differences in the measured response values in neighbouring observations \mathcal{N}_i ,

$$p_i = \sum_{j \in \mathcal{N}_i} \omega_j y_j, \quad \sum_{j \in \mathcal{N}_i} \omega_j = 0, \quad \sum_{j \in \mathcal{N}_i} \omega_j^2 = 1. \quad (17)$$

The efficiency of difference-based methods is not very good, and the methods have been used most commonly in univariate cases (Fan & Yao 1998).

2.3.5 Results for steel plate data

We compared different response variables in modelling the variance of steel strength. In the comparison, the squared residual proved to work clearly better than absolute residuals or logarithm transformed residuals. In agreement with the earlier results (Cawley *et al.* 2004), it seemed recommendable to use the corrected squared residual Eq. (10) to take into account the bias following from the mean model fitting; the details are given in original paper V.

2.4 Estimation methods of variance function

Gamma generalised linear models applied to the squared residuals of the mean model have been commonly used to estimate the variance function (Aitkin 1987, McCullagh & Nelder 1989). The method assumes that the response variable is Gaussian distributed and that variance is assumed to depend on the linear predictor via a monotone link function

$$g(\sigma_i^2) = z_i^\top \tau. \quad (18)$$

The vector z_i consists of transformations of explanatory variables found to best describe conditional variance. If the mean model were known, the model would be most efficiently estimated by maximising gamma log-likelihood

$$\widehat{\tau} = \max_{\tau} \sum_i \left(-\log \sigma_i^2 - \frac{\varepsilon_i^2}{\sigma_i^2} \right). \quad (19)$$

This section reviews the estimation techniques of the variance model $\sigma_i^2 = g(\tau, x_i, \mu_i)$. The mean function is often modelled using a linear model

$$\mu_i = x_i^\top \beta. \quad (20)$$

The discussed methods employ the squared residuals of the mean model; the estimation methods for possible response variables other than squared residuals have been discussed less in the literature.

If variance does not depend on the mean, the information matrix of the mean and variance parameters is block diagonal (Aitkin 1987). The iterative separate estimation of the mean and variance models yields the same results as joint estimation (Carroll & Ruppert 1988). Maximising the joint likelihood of the mean and variance parameters can be complicated, and thus iterative procedures have often been preferred, even though variance depends on the mean. In iterative weighted least squares (IWLS), the mean is estimated using weighted least squares with weights proportional to the inverses of the estimated variances from the previous variance estimation (Mak 1992). The variance parameters are updated in every iteration, and the discussed methods differ in the way the variance parameters are estimated. There has been discussion about the number of iterations needed. Some authors have suggested that the first iteration is enough (Yu & Jones 2004), but most often two iterations have been considered recommendable (Carroll & Ruppert 1988).

2.4.1 Maximum likelihood

The maximum likelihood estimator related to the linear mean model Eq. (20) maximises the log-likelihood

$$l(\beta, \tau; y) = -\frac{1}{2} \log |\Sigma_m| - \frac{1}{2} (y - X\beta)^\top \Sigma_m^{-1} (y - X\beta) \quad (21)$$

related to both the mean parameters β and the variance parameters τ . The covariance matrix of y , $\Sigma_m = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2)$, is assumed to be diagonal. Because of the loss in degrees of freedom that follows from estimating β , the estimator of the variance parameters is biased.

2.4.2 Restricted maximum likelihood

Restricted maximum likelihood (REML) is often preferable to ML, because it is adjusted to be unbiased (Verbyla 1993). In REML estimation, the marginal log-likelihood of τ is maximised by setting β to its conditional maximum likelihood estimates $\hat{\beta}(\tau)$ (Smyth 2002). The maximised log-likelihood related to the linear mean model Eq. (20) is

$$\begin{aligned} l_R(\tau; y) &= l(\hat{\beta}(\tau), \tau; y) - \frac{1}{2} \log |X^\top \Sigma_m^{-1} X| \\ &= -\frac{1}{2} (\log |\Sigma_m| + y^\top P_\Sigma y + \log |X^\top \Sigma_m^{-1} X|) \end{aligned} \quad (22)$$

where $P_\Sigma = \Sigma_m^{-1} - \Sigma_m^{-1} X (X^\top \Sigma_m^{-1} X)^{-1} X^\top \Sigma_m^{-1}$ and $l(\beta(\tau), \tau; y)$ is defined in Eq. (21). Because of the complexity of Eq. (22), the full REML estimation is slow for large data sets, although recently a faster algorithm has been developed (Smyth 2002).

Smyth *et al.* (2001) proposed an iterative method for REML estimation in the framework of the gamma generalised linear model Eq. (18). The proposed IWLS version of REML yields exactly the same parameter estimates and almost the same standard errors as the full REML. The correct response in the iterative REML is

$$\frac{\widehat{\varepsilon}_j^2}{(1-h_i)} \quad (23)$$

and the correct estimation prior weights are $w_i = 1 - h_i$ where $h_i = H_{ii}$ are the diagonal elements of the hat matrix, H , of the previous mean model fit $\widehat{Y} = HY$. (Smyth *et al.* 2001.) The iterative REML is easy to implement, fast and thus recommendable for dispersion parameter estimation in large data sets. The iterative REML estimation method has also been proposed for estimation of variance in double-generalised linear models (Smyth & Verbyla 1999).

2.4.3 Pseudo-likelihood

Maximisation of log-likelihood over part of the parameters while holding the rest of the parameters fixed is called the pseudo-likelihood method (Gong & Samaniego 1981). The pseudo-likelihood estimation of the variance function iteratively maximises the full Gaussian likelihood. At each iteration, the variance parameters are estimated from the log-likelihood function by fixing the mean parameters to their current values. The mean model parameters are then estimated by fixing the variance parameters to their current values. If the variance does not depend on the mean $\sigma_i^2 = g(\tau, x_i)$, the pseudo-likelihood estimator converges towards the ML estimator. (Carroll & Ruppert 1988.)

In the iterative procedure, the mean model is fitted using the weighted least squares with weights proportional to the inverses of the predicted variances.

$$\tilde{\beta} = \max_{\beta} \left[-\frac{1}{2} \sum_{i=1}^n \frac{(y_i - f(\beta, x_i))^2}{\widehat{\sigma}_i^2} \right]. \quad (24)$$

The log-likelihood maximised in variance model estimation has the form of gamma likelihood

$$\tilde{\tau} = \max_{\tau} \left[-\frac{1}{2} \sum_{i=1}^n \log g(z_i, \tau, \widehat{\mu}_i) - \frac{1}{2} \sum_{i=1}^n \frac{(y_i - \widehat{\mu}_i)^2}{g(z_i, \tau, \widehat{\mu}_i)} \right]. \quad (25)$$

The pseudo-likelihood method was motivated by the generalised least squares method without any assumption about the normality of the error distribution. If variance also depends on the mean, pseudo-likelihood has been preferred to ML because of its better robustness against a non-Gaussian distribution (Carroll & Ruppert 1988).

2.4.4 Extended quasi-likelihood

In quasi-likelihood estimation (Wedderburn 1974, Chiou & Müller 1999), the exact error distribution is not assumed, but only the form of the variance function $\text{var}(y_i) = V(\mu_i)$. It has been shown that the quasi-likelihood function

$$Q(y; \mu) = \int^{\mu} \frac{(y-s)}{V(s)} ds \quad (26)$$

has many properties similar to those of exact likelihood (Wedderburn 1974), and that it has connections to generalised linear models (McCullagh & Nelder 1989).

The quasi-likelihood estimation method has been extended to estimation of the conditional variance function, especially in the case where variance depends on the mean. In the proposed extended quasi-likelihood (EQL) approach, the variance parameters are estimated by maximising the extended quasi-likelihood function (Nelder & Pregibon 1987). Now, let the variance function be $\sigma_i^2 = V(\mu_i)g(z_i^T \tau)$ and

$$D_{\tau}(y; \mu) = -2[Q(y; \mu) - Q(y; y)] = -2 \int_y^{\mu} \frac{y-s}{V(s)} ds \quad (27)$$

be the quasi-deviance function. The EQL function allows a comparison of different variance functions by extending quasi-likelihood with a variance factor (Nelder & Pregibon 1987). The EQL function is the sum of the contributions of single observations

$$Q^+(Y; \beta, \tau) = \sum_i -\frac{1}{2} \log(2\pi V(y_i)g(z_i^T \tau)) - \frac{1}{2} \frac{D_{\tau}(y_i; \mu_i)}{g(z_i^T \tau)}. \quad (28)$$

EQL parameter estimates are obtained by IWLS (Nelder & Pregibon 1987). EQL estimation has been criticised for its inconsistency (Davidian & Carroll 1988).

2.4.5 Robust estimation

Errors in the data, a violation of assumptions or a model misspecification can have a major effect on statistical inference when least-squares or maximum likelihood methods are used (Huber 2004). Robust estimation is an alternative to very careful removal and downweighting of observations that are interpreted as erroneous. Inference on the basis of robust methods is more valid in the presence of outliers or in the violation of assumptions. In contrast, robust estimators are less efficient than standard maximum likelihood estimators when optimal conditions hold. An ideal robust estimator is efficient, but insensitive to deficiencies in the data (Huber 2004).

Variance is even more sensitive to outliers than the mean (Carroll & Ruppert 1988). Several robust algorithms have been proposed for estimating heteroscedastic models. An early proposal (Carroll & Ruppert 1982) bounds the influence of residual but not of leverage. The bounded-influence maximum likelihood and bounded-influence pseudo-likelihood also bind the influence of leverage (Giltinan *et al.* 1986). The properties of generalised M-estimators for robust estimation of variance are discussed in Bianco *et al.*

(2000). Also, the robust methods developed for generalised linear models (Cantoni 2004) can be useful in the framework of gamma generalised linear models.

The bounded-influence estimation of the variance model $\sigma_i^2 = g(z_i, \tau)$ presented in (Carroll & Ruppert 1988) can be written as follows: The variance model is estimated using estimation weights $w = (w_1, w_2, \dots, w_n)^\top$. The score vector, i.e. the gradient of log-likelihood with respect to model parameters, for the i th observation $\Psi_i = (\Psi_{i1}, \Psi_{i2}, \dots, \Psi_{iq})^\top$ is

$$\Psi_{ij} = \left(\frac{\varepsilon_i^2}{\widehat{\sigma}_i^2} - 1 \right) \frac{\partial}{\partial \tau_j} g(z_i, \tau). \quad (29)$$

Let A be $q \times q$ -matrix

$$\widehat{A} = \frac{1}{n} \sum_{i=1}^n w_i^2 \Psi_i \Psi_i^\top. \quad (30)$$

Then the weights

$$w_i = \min \left(1, \frac{a}{\sqrt{q \Psi_i^\top \widehat{A}^{-1} \Psi_i}} \right) \quad (31)$$

downweight the observations whose normed score is too large. The tuning parameter a controls the number of downweighted observations. The algorithm is initialised by unit weights, and the re-definition of weights and the estimation of parameters are iterated until convergence.

2.4.6 Results for steel plate data

In original paper V, the number of iterations needed in pseudo-likelihood estimation is studied empirically. The results agreed with earlier results (Carroll & Ruppert 1988) indicating that two iterations is recommendable, but also the first iteration gave pretty good results. Fitting of the variance model by least squares yielded poor results.

Robust estimation of the variance function was also studied using the steel plate data set. The bounded influence estimation of variance model Eqs. (29), (30) and (31) did not seem to work very well. Robust prediction was poor particularly for elongation: The average negative log-likelihood in the test data set was 1.93 when the model was estimated using the robust method, and 1.80 using the usual pseudo-likelihood. The results were calculated using the tuning constant $a = 7$, which downweighted 0.55 % of the observations. The results got worse when the tuning parameter was decreased to increase robustness. The bounded influence estimator downweighted observations with high absolute residuals, which resulted in low predicted variances. In this case, the bounded influence method predicted variances to be 5 % smaller, on average, than variances predicted with the pseudo-likelihood method. It seems that the bias of the tested robust variance function estimator may be unacceptably large. The results for Rm and ReH had the same tendency, but the differences between the estimation methods were not significant.

2.5 Models for joint modelling of mean and variance

The earliest models for variance were linear, generalised linear and non-linear parametric models. In parametric modelling, the results depend on the parametrisation of the model. It is often difficult to find correctly transformed explanatory variables for the model. Non-parametric methods for variance estimation have been developed since (Carroll 1982), but non-parametric methods became common in multivariate variance modelling years later.

Nonparametric modelling diminishes the risk of model misspecification, but increases the complexity of the model and decreases the interpretability of the model. The smoothness of a nonparametric fit is controlled with a penalty on complexity, and the optimal complexity level has to be explored in each application. The task of nonparametric variance function estimation is to estimate the smooth variance function $\sigma_j^2 = g(z_j, \mu_j)$. Non-parametric variance modelling in multivariate regression has been discussed, for example, in (Stadtmüller & Tsybakov 1995, Yau & Kohn 2003, Ruppert *et al.* 1997, Pan & Wang 2000).

This section reviews the models of variance that are applicable to multiple input variables. Many of the dispersion modelling methods assume that the error term is normally distributed and the focus is also here in that case. Variance is considered a function of explanatory variables, the mean or both.

2.5.1 Heteroscedastic regression

The concept of heteroscedastic regression refers broadly to regression models with non-constant variance. Most commonly the variance function is estimated on the basis of squared residuals.

The normal linear heteroscedastic regression model is defined as

$$\begin{aligned} y_j &\sim N(\mu_j, \sigma_j^2) \\ \mu_j &= x_j^T \beta \\ \sigma_j^2 &= g(z_j, \tau, \mu_j) \end{aligned} \quad (32)$$

where the function g defines the form of the variance function and x_j and z_j are vectors consisting of input variable observations and their transformations, which are selected to describe the modelled dependence. Some common variance functions are

$$\begin{aligned} \sigma_j^2 &= \phi \mu_j^\theta \\ \sigma_j^2 &= \exp(z_j^T \tau) \\ \sigma_j^2 &= \mu_j^\theta \exp(z_j^T \tau). \end{aligned} \quad (33)$$

Non-linearity can be achieved by using transformations of the original explanatory variables in the input vectors x_j and z_j . Interactions between the input variables can be taken into account by using product terms. Generalisation to non-linear regression $\mu_j = f(x_j, \beta)$ is straightforward. A practical alternative is a link-linear model

$$f(\mu_j) = x_j^T \beta. \quad (34)$$

2.5.2 Generalised linear models in dispersion modelling

In the family of generalised linear models (GLM), the additivity of effects is assumed to hold on a scale transformed by a monotone link function $f(\mu_j) = x_j^\top \beta$, and variance is a function of the mean $\sigma_j^2 = \phi V(\mu_j)$. The distribution of the error term is assumed to be included in the exponential family of distributions. (McCullagh & Nelder 1989.)

Many of the recent theoretical results concerning estimation of the parameters in heteroscedastic regression are related to models where variance is independent of the mean and the effect of explanatory variables is additive in a transformed scale $g(\sigma_j^2) = z_j^\top \tau$ (Smyth 2002). In this case, variance modelling can be performed in the GLM framework. It is never necessary to include the mean in the input variables of the variance model, because it is possible to include all the explanatory variables of the mean model, instead. However, this may make the variance model quite complicated. The independence between the mean and variance can often be achieved by transforming the response Box (1988).

An usual choice for the variance link function is the log-link

$$\log \sigma_j^2 = z_j^\top \tau \quad (35)$$

which guarantees the positivity of predicted variance. However, selection of the link function should be based on the data, and often some other link function can yield a significantly better model. For example, a linear link $\sigma_j^2 = z_j^\top \tau$ and a square root link $\sigma_j = z_j^\top \tau$ have been used, but negative linear predictors can cause problems. Mak (2002) proposed selecting the link function from the modified Box-Cox family

$$\begin{aligned} \sigma_j^2 &= (1 + |\lambda z_j^\top \tau|)^{\text{sign}(\lambda z_j^\top \tau)/\lambda} \quad \text{for } \lambda \neq 0, \\ \sigma_j^2 &= \exp(z_j^\top \tau) \quad \text{for } \lambda = 0. \end{aligned} \quad (36)$$

In double generalised linear models $f(\mu_j) = x_j^\top \beta$ and dispersion depends on the explanatory variables in a link-linear way $g(\phi_j) = z_j^\top \tau$ (Smyth 1989). The model parameters can be solved using ML estimation (Smyth & Verbyla 1999).

2.5.3 Local methods in variance function estimation

Local methods, i.e. kernel smoothing and local polynomials, use the entire data set as the prediction model. The prediction is given by a local model fitted in the neighbourhood of the query point. In kernel smoothing the model is a weighted average of the response values. In local linear modelling the model is a linear model and in local quadratic modelling the model is a second order polynomial regression model. The suitability of local methods in high dimensions has been questioned (Hastie *et al.* 2001).

Kernel smoothing of the squared residuals of the mean model has been commonly used for variance estimation (Carroll & Ruppert 1988, Vining & Bohn 1998). Let T denote the

model data set. The prediction model becomes

$$\hat{\sigma}_i^2 = \frac{\sum_{j \in T} K\left(\frac{\|x_i - x_j\|}{h_i}\right) \hat{\varepsilon}_j^2}{\sum_{j \in T} K\left(\frac{\|x_i - x_j\|}{h_i}\right)} \quad (37)$$

where the bandwidth h controls the smoothness of the model. The explanatory variables x should be standardised or properly scaled. The smoothing kernel function K is a unimodal density function with its mode at the origin (Hastie *et al.* 2001). The usual choices for K are

$$\begin{aligned} \text{Epanechnikov } K(s) &= \frac{3}{4}(1-s^2)I(|s| < 1) \\ \text{tri-cube } K(s) &= (1-s^3)^3I(|s| < 1) \\ \text{Gaussian } K(s) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}s^2\right). \end{aligned} \quad (38)$$

A multidimensional density function can also be used as the smoothing kernel. The mean can be modelled parametrically or kernel smoothing can be applied with a different bandwidth (Vining & Bohn 1998). Recursive kernels (Stadtmüller & Tsybakov 1995) and kernel smoothing on the logarithm of the squared residuals (Kuk 1999) have also been proposed.

Ruppert *et al.* (1997) and Fan & Yao (1998) proposed to use local polynomials to estimate the variance function

$$\begin{aligned} \hat{\mu}_i &= \hat{\beta}_0 \\ (\hat{\beta}_0, \hat{\beta}) &= \arg \min_{\beta_0, \beta} \sum_{j=1}^N [y_j - \beta_0 - (x_j - x_i)^\top \beta]^2 K_1\left(\frac{\|x_j - x_i\|}{h_1}\right) \\ \hat{\sigma}_i^2 &= \hat{\tau}_0 \\ \hat{\varepsilon}_j &= (y_j - \hat{\mu}_j) \\ (\hat{\tau}_0, \hat{\tau}) &= \arg \min_{\tau_0, \tau} \sum_{j=1}^N [\hat{\varepsilon}_j^2 - \tau_0 - (z_j - z_i)^\top \tau]^2 K_2\left(\frac{\|x_j - x_i\|}{h_2}\right). \end{aligned} \quad (39)$$

Ruppert *et al.* (1997) proposed correcting the squared residuals to take into account the biasing effect of mean model estimation by using the Eq. (10). Yu & Jones (2004) proposed using local likelihood for joint estimation of the mean and variance. Their approach differs from Eq. (39) in that variance is estimated by locally maximising the gamma log-likelihood instead of minimising the sum of squares, and a link function for variance is introduced.

2.5.4 Mean and dispersion additive models

Additive models are a flexible family for nonparametric or semiparametric estimation of the regression function. The usual form of a generalised additive model is

$$f(\mu_i) = \beta_0 + h_1(x_{i1}) + h_2(x_{i2}) + \dots + h_p(x_{ip}) \quad (40)$$

where f is a monotone link function and h_i are univariate smooth functions, for example cubic spline functions or parametric regression functions. Multivariate functions can be included in the model to allow interactions between the explanatory variables. Additive models are estimated using the backfitting algorithm. (Hastie *et al.* 2001.)

Mean and dispersion additive models (MADAM) (Rigby & Stasinopoulos 1996) are a general class of models proposed for joint modelling of the mean and variance. The model can be written as

$$\begin{aligned} f(\mu_i) &= \sum_{j=1}^p h_j(x_{ij}) \\ \sigma_i^2 &= V(\mu_i)\phi_i \\ g(\phi_i) &= \sum_{j=1}^q k_j(z_{ij}) \end{aligned} \quad (41)$$

and the conditional density function of y_i is $p(\mu_i, \phi_i)$ (Stasinopoulos & Rigby 2000). The model is estimated by maximising the penalised log-likelihood

$$l_p = \sum_{i=1}^n \log p(\mu_i, \phi_i) - \frac{1}{2} \sum_{j=1}^p \lambda_{1j} \int_{-\infty}^{\infty} h_j''(s)^2 ds - \frac{1}{2} \sum_{j=1}^q \lambda_{2j} \int_{-\infty}^{\infty} k_j''(s)^2 ds. \quad (42)$$

It can be shown that the functions h_j and k_j that maximise Eq. (42) are natural cubic splines. Let $N_i(s) = (a_{i0} + a_{i1}s + a_{i2}s^2 + a_{i3}s^3)I(\xi_{i-1} < s < \xi_i)$, $i = 1, \dots, n^*$ be piecewise polynomials, $\xi_0 = -\infty$, $\xi_1 < \xi_2 < \dots < \xi_{n^*-1}$ be the ordered distinct values of the explanatory variable in question and $\xi_{n^*} = \infty$. Then a function of the form $\sum_{i=1}^{n^*} N_i(s)$ having continuous second derivatives and being linear when $s < \xi_1$ or $s > \xi_{n^*-1}$ is a natural cubic spline. MADAM includes generalised additive models and double generalised linear models as special cases. In this framework, the mean and dispersion can be modelled independently using parametric models or nonparametric cubic splines. (Rigby & Stasinopoulos 1996.)

2.5.5 Reproducing kernels

Support vector machines and reproducing kernel methods have been one of the most discussed learning methods during the recent years (Schölkopf & Smola 2002). These methods apply a linear model in a transformed, high-dimensional space and the model parameters pay a squared penalty in the model estimation. The transformed feature space is induced by a continuous, symmetric and positive definite reproducing kernel function $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}}$. The reproducing kernel function is the inner product of the Hilbert space $\mathcal{H} = \{f | f : \mathcal{X} \rightarrow \mathfrak{R}\}$. Here \mathcal{X} is the data space, often \mathfrak{R}^p and $\phi : \mathcal{X} \rightarrow \mathcal{H}$, $\phi(x) = K(x, \cdot)$ is the mapping in the feature space. The reproducing property is useful because the model can be expressed in terms of reproducing kernel functions in the original space. The solution to the problem $\min_{f \in \mathcal{H}} \sum_{i=1}^n L(y_i, f(x_i)) + \lambda_p \|f\|_{\mathcal{H}}$ is of the form

$$f(x) = b_0 + \sum_{i=1}^n \alpha_i \phi(x_i) = b_0 + \sum_{i=1}^n \alpha_i K(x, x_i) \quad (43)$$

where $L(y_i, f(x_i))$ is the loss paid from the prediction error. Two common reproducing kernel functions are

$$\begin{aligned} \text{Gaussian } K(x_i, x_j) &= \exp\left(-\frac{\|x_i - x_j\|}{h}\right) \\ \text{polynomial } K(x_i, x_j) &= (x_i^\top x_j + \theta_0)^\theta. \end{aligned} \quad (44)$$

Cawley *et al.* (2004) developed a reproducing kernel method for joint modelling of the mean and variance. The proposed kernel ridge regression model is

$$\begin{aligned} \mu(x) &= \beta_0 + \sum_{i=1}^n \alpha_i^\mu K^\mu(x, x_i) \\ \log \sigma(x) &= \tau_0 + \sum_{i=1}^n \alpha_i^\sigma K^\sigma(x, x_i). \end{aligned} \quad (45)$$

The model is estimated using IWLS by iteratively maximising the penalised weighted least squares of the mean model and the penalised gamma log-likelihood of the variance model, correspondingly using the penalty terms $\lambda_\mu \sum_{i=1}^n \sum_{j=1}^n \alpha_i^\mu K^\mu(x_i, x_j) \alpha_j^\mu$ and $\lambda_\sigma \sum_{i=1}^n \sum_{j=1}^n \alpha_i^\sigma K^\sigma(x_i, x_j) \alpha_j^\sigma$. The hyperparameters λ_μ and λ_σ control the complexity of the model.

2.5.6 Bayesian methods

Some authors have applied Bayesian methods to inference about conditional variance. Shao (1992) proposed an empirical Bayes estimation of variance heteroscedasticity. His estimator is a mixture of prior information, within-group variation and smoothed squared residuals. In the approach of Yau & Kohn (2003), the variance and mean functions are estimated using penalised splines and log-link for variance. The authors assume Gaussian distributed response and propose an MCMC sampling scheme for simultaneous estimation and variable selection of the mean and variance. Bayesian inference about the hyperparameters of the model has been proposed for heteroscedastic kernel ridge regression (Cawley & Talbot 2005).

2.5.7 Neural networks

Neural networks have been used for joint modelling of the mean and dispersion (Bishop 1995, Williams 1996, Dorling *et al.* 2003). In these proposals, a multi-layer perceptron with an additional output unit for variance is used. The model parameters are estimated by maximising Gaussian log-likelihood. The single hidden-layer network proposed by Boyd & White (1994) uses separate networks for both the mean and variance. The model can be written as

$$\begin{aligned}
\mu_i &= \beta_0 + \sum_{j=1}^{m_\mu} A \left(\beta_{j0} + \sum_{k=1}^p x_{ik} \beta_{jk} \right) \beta_j \\
v_i &= \tau_0 + \sum_{j=1}^{m_\sigma} A \left(\tau_{j0} + \sum_{k=1}^p x_{ik} \tau_{jk} \right) \tau_j.
\end{aligned} \tag{46}$$

The number of hidden nodes m_μ and m_σ controls the complexity of the network. The authors used logistic function $A(s) = 1/(1 + e^{-s})$ as the activation function. The proposed link functions were $\sigma_i^2 = \exp(v_i)$ and $\sigma_i^2 = e + v_i^2$ where e is a small constant ensuring positivity. The parameters β and τ are estimated jointly by minimising the objective function

$$\frac{1}{2n} \left[\sum_{i=1}^n \log \sigma_i^2 + \frac{(y_i - \mu_i)^2}{\sigma_i^2} \right]. \tag{47}$$

2.5.8 Results for steel plate data

Heteroscedastic linear models, mean and dispersion additive models, local linear regression for mean and dispersion and neural network modelling of mean and dispersion were compared in predicting variances in the steel plate data set. In the proposed novel neural network approach, separate multi-layer perceptron models for the mean and variance were estimated iteratively using the pseudo-likelihood method. Neural networks proved to be a suitable method for variance modelling on large industrial data sets. Additive models had problems with interacting explanatory variables, and their fitting required a huge amount of memory. Local linear modelling was time-consuming and may not be applicable in real-time applications. Heteroscedastic linear models seemed to be a comparable alternative, especially when interpretability is required. The results suggest that the learning method of variance can be selected independently of the learning method of the mean. Details are given in original paper V.

2.6 Modelling of conditional distribution

The conditional distribution function determines the mean, variance and different quantiles and probabilities. There are three possible approaches to modelling conditional distribution. The stages are illustrated in Fig. 1. The usual method is to make a distributional assumption, estimate a model for the mean and draw the inferences using the predicted mean and the distributional assumption. In the second stage, the mean and dispersion are modelled jointly as a function of the input variables. The inference is based on the estimated mean, the estimated variance and a less restrictive assumption about the distribution of residuals. The third stage is the most general. The conditional distribution or its parameters, like kurtosis and skewness, are modelled as a function of the input variables.

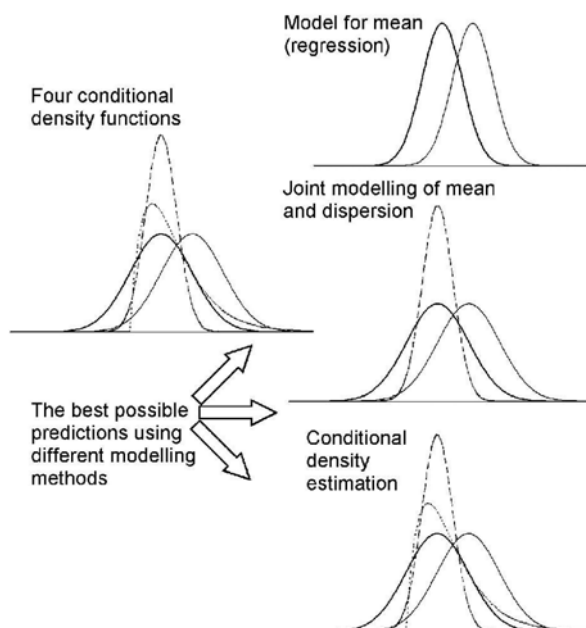


Fig. 1. The three stages of modelling.

In joint modelling of mean and dispersion, the conditional distribution depends on the explanatory variables only through the first two moments: the mean and variance. A review of methods for constructing confidence intervals on the basis of the modelling of conditional distribution was given by Wright & Royston (1997). This section gives an overview of the methods for flexible modelling of the conditional distribution function.

2.6.1 Conditional density estimation

Several methods have been proposed for estimation of the conditional density function. The methods successfully applied to multivariate data are kernel smoothing (Davis & Hwang 1998), local linear models (Fan *et al.* 1996), neural networks (Husmeier 1999, Sarajedini *et al.* 1999) and mixture models. Mixture models, including mixture density networks (Bishop 1995), model the conditional density function as a weighted sum of several Gaussian densities. All of these methods are computationally quite complex and may prove problematic in high dimensions.

2.6.2 *Quantile regression*

In quantile regression, regression models are used to predict the conditional quantiles of the response. The conditional α -quantile function is

$$q_\alpha(x) = \inf\{s : P(y|x < s) \geq \alpha\}. \quad (48)$$

For quantile regression, modelled quantile functions have to be selected. For example, it could be decided that the 2.5 %, 50 % and 97.5 % quantiles are modelled. Models have to be fitted for a large number of quantiles in order to describe the conditional distribution completely. Quantile regression has been often applied in the construction of confidence intervals.

Regression quantiles were introduced by Koenker & Bassett (1978). The authors used a separate parametric linear regression model for each of the quantiles of interest. Non-parametric and semiparametric modelling of regression quantiles has also been a popular topic of discussion. In nonparametric modelling, the conditional quantiles are allowed to depend quite freely on explanatory variables (Yu & Jones 1998). Several reviews on regression quantile estimation have been written recently (Wright & Royston 1997, Yu *et al.* 2003, Buchinsky 1998). Nonparametric quantile regression methods often have problems with high dimensionality (de Gooijer & Zerom 1999). Quantile regression has been widely used at least in medical and econometric applications. Apparently, quantile regression has been rarely applied to multivariate industrial data.

The quantile function is the inverse of the cumulative distribution function, and thus the problem of approximating the conditional quantile function is essentially the same as approximating the conditional cumulative distribution function (Peracchi 2002). Some methods, like local linear models, have been proposed for estimating the conditional cumulative distribution function (Hall *et al.* 1999).

2.6.3 *Modelling the parameters of distribution*

An easily applicable method for constructing conditional densities is to use separate prediction models for the parameters of the distribution. In addition to the mean and dispersion, models can be fitted for other parameters of the distribution. The LMS method (Cole & Green 1992) models skewness by estimating conditional skewness with an input-dependent Box-Cox transformation parameter. Generalised additive models for location, scale and shape (GAMLSS) (Rigby & Stasinopoulos 2005) use additive models to model the parameters of conditional distribution

3 Adaptive modelling of a variance function

In the earlier literature, modelling and estimation of a time-varying variance function $\sigma_t^2 = g(\tau_t, z_t)$ has not been discussed. In the notation, conditional variance depends on explanatory variables via a variance function g . The parameters of the variance function τ_t are not constant, but vary between the observations. In non-adaptive variance modelling it is assumed that $\tau_t = \tau \forall t$. In this chapter, statistical methods for modelling time-varying parameters are reviewed and considered for modelling of conditional variance.

3.1 Recursive estimation

Re-estimation of the model after every new observation using all the available data is always possible in principle, but in practice it is too time-consuming. Recursive, i.e. sequential estimation methods recalculate the model parameters after each observation on the basis of the current parameters and the new observation. A recursive method for updating the regression coefficients of the familiar linear regression $E y_i = x_i^T \beta$ after every new observation with

$$\begin{aligned} P_{t+1} &= P_t - \frac{(P_t x_{t+1})(P_t x_{t+1})^T}{1 + x_{t+1}^T P_t x_{t+1}} \\ \hat{\beta}_{t+1} &= \hat{\beta}_t + P_{t+1} x_{t+1} (y_{t+1} - x_{t+1}^T \hat{\beta}_t) \end{aligned} \quad (49)$$

was proposed by Plackett (1950). The updated formula of the inverse Hessian

$$P_t = \left(\sum_{j=1}^t x_j x_j^T \right)^{-1} \quad (50)$$

is based on the matrix equality

$$(A + BB^T)^{-1} = A^{-1} - (A^{-1}B)(I + B^T A^{-1}B)^{-1}(A^{-1}B)^T. \quad (51)$$

The formulas offer a way to compute the updated least squares estimator with a low computational cost. Later, Kalman proposed his filtering method for estimating more general state space models, which shared much the same ideas (Kalman 1960).

Computationally feasible recursive estimation methods have been suggested for several model types, such as reproducing kernel methods (Kivinen *et al.* 2004) and local methods (Schaal *et al.* 2002). A comprehensive treatment of recursive estimation in linear regression, linear difference equation models and state space models is given by Ljung & Söderström (1983). Sequential parameter estimation can also be approached from a Bayesian point of view. The recursive least squares algorithm can also be derived from the Bayesian approach to recursive estimation (Ljung & Söderström 1983).

Although the parameters change after each observation, the recursive models are not really adaptive. Behind the original formulas was the idea that the regression function is static and the estimated regression coefficients converge to the true ones as more observations are added (Pollock 2003). In a long run, the regression coefficients converge to the mean of the stochastic process generating them. Rolling regression and discounted regression using forgetting factors can be seen as adaptive variations of the recursive estimation scheme (Pollock 2003).

3.1.1 Forgetting factors

The recursive regression model can be made adaptive by gradually downweighting the older observations with a forgetting factor (Pollock 2003). Let $\gamma \in [0, 1[$ be a forgetting factor giving estimation weights γ^{t-j} for the observations $j = 1, 2, \dots, t$. (Pollock 2003) provided a computationally simple formula for updating of parameter estimates

$$\begin{aligned} P_{t+1} &= \frac{1}{\gamma} \left(P_t - \frac{(P_t x_{t+1})(P_t x_{t+1})^T}{\gamma + x_{t+1}^T P_t x_{t+1}} \right) \\ \widehat{\beta}_{t+1} &= \widehat{\beta}_t + P_t x_{t+1} (\gamma + x_{t+1}^T P_t x_{t+1})^{-1} (y_{t+1} - x_{t+1}^T \widehat{\beta}_t). \end{aligned} \quad (52)$$

3.1.2 Rolling regression

In the rolling regression scheme, the model is estimated based on the w previous observations, while the oldest observations are completely discarded. After each new observation, the model is re-estimated based on the observations in a moving time window: this approach has been commonly referred to with the term 'moving window'. The formulas used in recursive regression can be modified to make updating of parameter estimates computationally easy. First the observation $t - w$ is removed from the model using the formulas

$$\begin{aligned} P_{t+1}^* &= P_t + \frac{(P_t x_{t-w})(P_t x_{t-w})^T}{-1 + x_{t-w}^T P_t x_{t-w}} \\ \widehat{\beta}_{t+1}^* &= \widehat{\beta}_t - P_{t+1}^* x_{t-w} (y_{t-w} - x_{t-w}^T \widehat{\beta}_t) \end{aligned} \quad (53)$$

to obtain intermediate estimates P_{t+1}^* and $\widehat{\beta}_{t+1}^*$. Then the formulas Eq. (49) are applied to P_{t+1}^* and $\widehat{\beta}_{t+1}^*$ to add the new observation. (Pollock 2003.) A combination of rolling and recursive regression is proposed in (Clark & McCracken 2004).

Moving window modelling has been commonly applied in many kinds of models, although the updating formulas Eqs. (53) and (49) hold only for linear regression. In practice, the model can be re-estimated only occasionally, using the most recent data. Because of its simplicity, moving window modelling seems to be one of the most popular adaptive modelling methods in many fields.

3.2 On-line learning

An approach where training is based on a fixed training data set is called batch learning. When new data are continuously measured, on-line learning is an attractive alternative. In on-line learning, only one observation at a time is given to the learner, which adjusts the model parameters accordingly. After learning, the observations are forgotten and only stored information is the model. While the usual learning algorithm steps change the parameter estimates in some direction depending on the average gradient over the training data set, on-line learning algorithms change the parameter estimates according to the direction depending on the gradient of a single, new observation.

The original ideas behind on-line learning were proposed by Robbins & Monro (1951). The terms 'stochastic approximation' and 'on-line learning' have often been used in the context of stochastic gradient descent and related methods. In the update step of a typical on-line learning algorithm, the parameter estimates are moved on the basis of the gradient of the loss function at the current observation

$$\widehat{\beta}_{t+1} = \widehat{\beta}_t + \eta_t C_t(\widehat{\beta}_t) \frac{\partial}{\partial \beta} L \left[y_{t+1}, f(x_{t+1}, \widehat{\beta}_t) \right]. \quad (54)$$

The positive definite matrix $C_t(\widehat{\beta}_t)$ controls the search directions and is often related to the inverse Hessian. The matrix C_t can be updated using a separate update formula.

On-line learning has been studied in the frameworks of statistical learning theory (Murata 1998) and statistical physics (Biehl & Caticha 2001). Theoretical results (Oppen 1996) show that a properly implemented on-line learning algorithm learns the regression function asymptotically as efficiently as the best off-line algorithm.

On-line learning methods are especially practical when the modelled dependence varies with time, but they are also useful when the true regression function remains constant. The convergence properties of the algorithm depend on the learning rate η_t . It is known that the learning rate $\eta_t = c/t$ is optimal when the modelled phenomenon is not changing, because it guarantees local convergence of the algorithm under some regularity conditions. It is known that when a sufficiently rapidly shrinking learning rate $\eta_t = O(1/t)$ is used, the model is not able to adapt to changes in the regression function. Thus, on-line learning with a constant learning rate has been used in adaptive modelling. (Murata *et al.* 2002.) There have also been several proposals for an adaptive learning rate: The basic idea is that the learning rate is increased when the model predicts poorly and decreased to c/t when the smallest achievable prediction error level is found (Sompolinsky *et al.* 1995, Orr & Leen 1996).

3.2.1 On-line quasi-Newton algorithm

An on-line quasi-Newton algorithm for a squared error loss function was presented by Bottou (1998). The algorithm updates the parameter estimates after each new observation, like in the ordinary quasi-Newton algorithm step, but it uses the gradient at the new observation instead of the average gradient of the training data set. Let the non-linear regression function be $Ey_i = f(\beta, x_i)$. The author approximates the Hessian matrix by the Gauss-Newton outer product estimator

$$\sum_{i=1}^n \left[\frac{\partial}{\partial \beta} f(\beta, x_i) \right] \left[\frac{\partial}{\partial \beta} f(\beta, x_i) \right]^T. \quad (55)$$

The inverse approximate Hessian was proposed to be updated using the matrix formula Eq. (51). Let $\delta_t = \frac{\partial}{\partial \beta} f(\beta, x_t)$. The algorithm can be written as

$$\begin{aligned} P_{t+1} &= P_t - \frac{(P_t \delta_{t+1})(P_t \delta_{t+1})^T}{1 + \delta_{t+1}^T P_t \delta_{t+1}} \\ \hat{\beta}_{t+1} &= \hat{\beta}_t + P_{t+1} \delta_{t+1} \left[y_{t+1} - f(x_{t+1}, \hat{\beta}_t) \right]. \end{aligned} \quad (56)$$

The author proves that the algorithm converges to a local minimum of the squared error loss function. (Bottou 1998.) The method is not adaptive, because the step size is proportional to $1/n$: $P_n = O(1/n)$.

3.2.2 On-line neural networks

Much of the recent work in on-line learning has been done in the field of neural networks. Saad (1998) discusses the on-line learning of neural networks from many viewpoints. The learning of neural networks is challenging because of local minima and plateaus in the loss function (Bishop 1995). Despite that, on-line neural networks seems to be a well-established method for adaptive modelling of a time-varying regression function: Saad (1998) discusses the optimality and convergence properties of on-line neural networks and presents successful applications. Several on-line learning schemes, like on-line gradient descent

$$\hat{\beta}_{t+1} = \hat{\beta}_t + \eta_t \frac{\partial}{\partial \beta} f(x_{t+1}, \hat{\beta}_t) \left[y_t - f(x_{t+1}, \hat{\beta}_t) \right] \quad (57)$$

and quasi-Newton type algorithms, have been proposed for on-line learning of neural networks.

3.2.3 Reproducing adaptive kernels

On-line learning has been suggested for kernel methods where linear regression is performed in a transformed, high dimensional feature space. The major problems are how

to treat regularisation, the number of kernel functions and the high dimensionality of the feature space. In recent work, recursive least squares (Engel *et al.* 2004) and stochastic gradient descent (Kivinen *et al.* 2004) have been modified to be applicable in the reproducing kernel Hilbert space. The on-line support vector regression of Kivinen *et al.* (2004) can be used to model a time-varying regression function.

3.2.4 Real-time lazy learning

In the framework of lazy learning, the training data set works as a model: When a prediction is queried, the most similar observations are searched to construct the prediction. The most common methods in this framework are nearest neighbour methods, kernel smoothing and local linear regression. In the lazy learning framework, model updating is extremely easy, the update step consists of simply adding the new observation to the database. However, such a simple approach is problematic and not very adaptive: The prediction time grows linearly with the size of the training data set, and new and old observations cannot be distinguished. These problems have been addressed, for example, by Schaal *et al.* (2002). A related method is recursive calculation of kernel smooths, but that method is not adaptive either (Krzyzak 1992).

3.3 Time-varying parameter regression

Several methods for modelling time-varying parameters in the regression context have been developed (Riddington 1993). The methods used in time-varying parameter regression include change point models, functional coefficient models and stochastic coefficient models. Let $\{\beta(t)\}$ be the time-continuous series of the regression coefficient vector and let the true parameter at time t be β_t .

3.3.1 Change point models

Change point models assume that the regression coefficients are piecewise constant. There are break points B_1, B_2, \dots where a structural change in the modelled phenomenon happens

$$\beta_t = \beta_0 + b_1 I(t > B_1) + b_2 I(t > B_2) + \dots \quad (58)$$

The majority of the articles related to change point modelling discuss finding the break points and testing the existence of break points. These kinds of 'all data on hand' approaches are not of interest when adaptivity based on continuously measured new observations is needed.

Leisch *et al.* (2000) and Zeileis *et al.* (2005) propose to sequentially test for structural change based on recursive or moving window regression parameter estimates. These

papers do not discuss about prediction, but obviously the results of their tests could be utilised to time the refitting of adaptive models.

A recent research topic is prediction using change-point models with an unspecified number of change points. The model is capable of handling a continuous flow of observations in real time. In the proposed approach, the model is estimated using Bayesian methodology. New observations are predicted using the MCMC algorithm. (Koop & Potter 2004.)

3.3.2 Stochastic coefficient models

Stochastic coefficient models assume that the regression parameters form a stochastic process. At every time step, usually after each observation, the regression parameters are assumed to change because of stochastic innovations (Rosenberg 1972). A common approach is to assume that the regression parameters form a random walk

$$\beta_t = \beta_{t-1} + v_t, \quad v_t \sim N(0, G_v), v_t \text{ i.i.d.} \quad (59)$$

where the covariance matrix G_v can often be assumed to be diagonal. Cooley & Prescott (1976) proposed assuming that the parameter vector forms the process where changes in the parameters can be divided into permanent and trajectory changes

$$\begin{aligned} \beta_t &= \beta_t^* + u_t \quad u_t \sim N(0, G_u), u_t \text{ i.i.d.} \\ \beta_t^* &= \beta_{t-1}^* + v_t \quad v_t \sim N(0, G_v), v_t \text{ i.i.d.} \end{aligned} \quad (60)$$

By adjusting the magnitude of G_u and G_v , a model between the random coefficient model and the random walk model can be chosen. A more general approach is to assume that the regression parameters form an ARMA process (Liu & Hanssens 1981). Stochastic coefficient models can be estimated using a Bayesian approach (Shively & Kohn 1997) or a Kalman filter (Harvey 1989).

A special case of stochastic coefficient models is the random coefficient model, where the distribution of β_t is identical $\forall t$ (Hildreth & Houck 1968)

$$\beta_t = u_t, \quad u_t \sim N(0, G_u), u_t \text{ i.i.d.} \quad (61)$$

Random coefficient models can be used to introduce structural heteroscedasticity to the prediction. Prediction with random coefficient models is discussed in more detail by Beran (1995).

3.3.3 Functional coefficient models

Functional coefficient models are of the form

$$y_i = \sum_{j=1}^p \beta_j(U_i) x_{ij} + \varepsilon. \quad (62)$$

The coefficients of regressors depend on some covariates U_i . The coefficient functions $\beta_j(t)$ are usually estimated using a smoothing method (Fan & Zhang 1999.) Time-varying coefficient models are easily included in the framework when the coefficients are functions of time $\beta_j(t)$. In principle, functional coefficient models are not meant for adaptive modelling, but the purpose is to find out the past behaviour of the parameters. Functional coefficient models have also been applied in time series analysis (Huang & Shen 2004).

3.4 Models with time-varying variance

Several models allowing time-varying variance that does not depend on explanatory variables have been commonly applied in econometrics. The approaches assume that variance is a one-dimensional stochastic process, and the methods differ in the way the variance-generating process is formulated and estimated. However, none of these suggestions allow the variance to depend on the explanatory variables.

3.4.1 Conditional autoregressive heteroscedasticity

Autoregressive conditional heteroscedasticity (ARCH) models (Engle 1982) are defined as follows,

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t \sqrt{h_t} \\ h_t &= \alpha_0 + \zeta_1 \varepsilon_{t-1}^2 h_{t-1} + \zeta_2 \varepsilon_{t-2}^2 h_{t-2} + \dots + \zeta_r \varepsilon_{t-r}^2 h_{t-r} \end{aligned} \quad (63)$$

where ε_t are i.i.d. random variables with $E\varepsilon_t = 0$ and $\text{var}(\varepsilon_t) = 1$. Generalized autoregressive conditional heteroscedasticity (GARCH) models (Bollerslev 1986) are of the form

$$h_t = \alpha_0 + \zeta_1 \varepsilon_{t-1}^2 h_{t-1} + \dots + \zeta_r \varepsilon_{t-r}^2 h_{t-r} + \iota_1 h_{t-1} + \dots + \iota_{-m} h_{t-m}. \quad (64)$$

The estimated model parameters describe how variance depends on lagged squared errors and variances. Although the predicted variances are changing, the model describing the autoregressive structure is static. The models can be estimated by ML, and several modifications of the basic method have been used (Li *et al.* 2002). Recurrent mixture density networks (Schittenkopf *et al.* 2000) employ a similar approach, but the dependence of conditional variance on the lagged variances is modelled by a neural network model.

3.4.2 Stochastic volatility models

The principle of a stochastic volatility model is to assume that conditional variance forms a stochastic process. The approach can be seen as an application of time-varying coefficient models to the modelling of variance. Variance is usually examined on a transformed

scale to ensure its positivity

$$\begin{aligned} y_t &= \mu_t + \sqrt{h_t} \varepsilon_t \\ \log h_t &= \theta_1 + \theta_2 \log h_{t-1} + \nu_t \end{aligned} \quad (65)$$

where ε_t and ν_t are i.i.d., often Gaussian processes.

The most popular approaches to model estimation are probably Bayesian MCMC inference (Jacquier *et al.* 1994), the method of moments (Andersen *et al.* 1999) and quasi-maximum likelihood (Ruiz 1994). Some methods do not make restrictive assumptions about the form of the variance generating process: Andreou & Ghysels (2006) applied change point models to the modelling of volatility and Mercurio & Spokoiny (2004) applied a local linear method for predicting time-varying variance.

3.5 Adaptive modelling of conditional variance

This section proposes that the methodology related to on-line learning and time-varying parameter regression can be used to model a time-varying variance function. In original paper VIII, two methods, namely moving window estimation and adaptive on-line quasi-Newton, are proposed for adaptive modelling of conditional variance.

The process of squared residuals of the mean model fit can be used as the response variable in adaptive modelling of conditional variance. The difficulty is that many of the adaptive regression methods assume that the response is normally distributed, but squared residuals are gamma distributed. Zeileis & Hornik (2002) propose using quantile residuals of generalised linear models in change point modelling. A quantile residual is defined as the quantile of a standard normal distribution that corresponds to the value of the assumed cumulative distribution function of the observation (Dunn & Smyth 1996)

$$q_i = \Phi^{-1} \left[F(y_i; \hat{\mu}_i, \hat{\phi}) \right]. \quad (66)$$

It could be possible to utilise quantile residuals of the variance model in adaptive modelling of variance.

McGilchrist & Matawie (1998) developed a method for computing recursive parameter estimates in a GLM family. Let $\hat{\mu}_t = f(x_t^\top \hat{\beta}_{t-1})$ and $\sigma_t^2 = \phi V(\mu_t)$. Their recursive formulas can be written as

$$\begin{aligned} \hat{\beta}_t &= \hat{\beta}_{t-1} + \frac{\left[f'(x_t^\top \hat{\beta}_{t-1}) \right]^{-1} (y_t - \hat{\mu}_t) P_{t-1} x_t}{V(\hat{\mu}_t) \left[f'(x_t^\top \hat{\beta}_{t-1}) \right]^{-2} + x_t^\top P_{t-1} x_t} \\ P_t &= P_{t-1} - \frac{(P_{t-1} x_{t-1}) (P_{t-1} x_{t-1})^\top}{V(\hat{\mu}_t) \left[f'(x_t^\top \hat{\beta}_t) \right]^{-2} + x_t^\top P_{t-1} x_t}. \end{aligned} \quad (67)$$

The method can be interpreted as the on-line version of the Fisher scoring algorithm. The authors prove that the recursive formula approximately follows the exact recursive parameter estimates. The approach of Eq. (67) could be applied to recursive estimation of variance function parameters under the gamma generalised linear model.

3.5.1 Variance function estimation with recursive smoothing

Stadtmüller & Tsybakov (1995) proposed a method for recursively estimating the variance function $\sigma^2(x_i)$, $x_i \in \mathfrak{R}^p$. They employed recursive kernel smoothing of squared residuals to expand the variance function after each new observation

$$\widehat{\sigma}_t^2(x) = \widehat{\sigma}_{t-1}^2(x) + \eta_t [\widehat{\varepsilon}_t^2 - \widehat{\sigma}_{t-1}^2(x_t)] h_t^{-p} K\left(\frac{x_t - x}{h_t}\right). \quad (68)$$

The positive sequences h_t and η_t go to zero and $\sum_{i=1}^{\infty} \eta_i = \infty$, $th_t^p \rightarrow \infty$. The mean model is estimated simultaneously with a similar recursive kernel principle. Stadtmüller & Tsybakov (1995) assume that the variance function is static and prove the convergence properties of the estimator. Their estimation method does not adapt to changes in the variance function, because the old observations are never discounted or removed from the model.

3.5.2 Moving window modelling of conditional variance

In original paper VIII it is proposed that the moving window method can be used to estimate a variance function with time-varying parameters $\sigma_t^2 = g(x_t, \tau_t)$. The model is occasionally re-estimated by maximising the gamma log-likelihood related to the w previously squared residuals

$$\widehat{\tau}_t = \max_{\tau} \sum_{i=t-w}^t \left[-\log g(\tau, x_i) - \frac{\widehat{\varepsilon}_i^2}{g(\tau, x_i)} \right]. \quad (69)$$

Re-estimation of the model after each new observation is computationally expensive. In practical application the model can be re-estimated less frequently, for example at certain time intervals. In addition, discounting of older observations can be considered to make the model behave more smoothly.

3.5.3 Adaptive on-line quasi-Newton algorithm

Estimation of time-varying conditional variance by applying an adaptive version of the on-line quasi-Newton algorithm to the squared residuals is proposed in original paper VIII. Let $\widehat{\sigma}_{t+1}^2 = g(\widehat{\tau}_t, x_{t+1})$, and let

$$\delta(\tau, x_t) = (\partial/\partial\tau)g(\tau, x_t) \quad (70)$$

denote the vector of partial derivatives. The matrix

$$\mathcal{J}_t = \sum_{i=1}^t [\delta(\widehat{\tau}_i, x_i)/\widehat{\sigma}_i^2] [\delta(\widehat{\tau}_i, x_i)/\widehat{\sigma}_i^2]^T \quad (71)$$

denotes the first-order approximation of Hessian cumulated from the gamma distributed observations at $i = 1, \dots, t$. The matrix $P_t \approx \mathcal{J}_t^{-1}$ denotes the approximate inverse Hessian

used in the quasi-Newton algorithm. The parameter estimates are updated after each new observation by

$$\widehat{\tau}_{t+1} = \widehat{\tau}_t + \eta(t+1)P_{t+1} \left(\frac{\widehat{\varepsilon}_{t+1}^2}{\widehat{\sigma}_{t+1}^2} - 1 \right) \frac{\delta(\widehat{\tau}_t, x_{t+1})}{\widehat{\sigma}_{t+1}^2}. \quad (72)$$

The constant learning rate η is tuned separately in each application. The inverse Hessian is kept up-to-date by applying the matrix equality Eq. (51):

$$P_{t+1} = P_t - \frac{[P_t \delta(\widehat{\tau}_t, x_{t+1}) / \widehat{\sigma}_{t+1}^2] [P_t \delta(\widehat{\tau}_t, x_{t+1}) / \widehat{\sigma}_{t+1}^2]^T}{1 + [\delta(\widehat{\tau}_t, x_{t+1})]^T / \widehat{\sigma}_{t+1}^2} P_t [\delta(\widehat{\tau}_t, x_{t+1}) / \widehat{\sigma}_{t+1}^2]. \quad (73)$$

The algorithm is initialised by fitting the model using a sufficient number of early observations. New and old observations have equal contributions to P_t . The matrix P_t reflects the accumulated information, but the model has forgotten the oldest information. The proposed adaptive on-line quasi-Newton algorithm differs from the non-adaptive versions (McGilchrist & Matawie 1998, Bottou 1998) in that the learning steps are proportional to tP_t instead of the non-adaptive steps P_t .

3.5.4 Results in steel plate data

For the steel plate data set, the adaptive models performed better than the non-adaptive model. The differences between the models are significant, but the non-adaptive model seems fairly adequate. The adaptive quasi-Newton algorithm outperformed the moving window method. In two groups of steel plate products, variance had slightly decreased with time. The time paths of the model parameters and predicted variances indicated that major changes had not taken place in the conditional variances during the study period. Details are given in original paper VIII.

3.6 Simultaneous adaptive modelling of mean and dispersion

When the true mean model remains unchanged over time, a natural approach to adaptive modelling of the mean and variance is to update the mean model recursively. Because the accuracy of the mean model increases with time, estimation of time-varying variance should be based on the squared residuals of the latest, most accurate mean model.

In the case of the time-varying mean model $Ey_t = x_t^T \beta_t$, it is much more difficult to assume that the estimated mean model is the true model. Joint modelling of the time-varying mean and dispersion is balancing between bias and variance. Distinguishing the error due to bias and variance may often be difficult. A highly adaptive mean model estimator has a large variance and results in a strong correlation between the observations and the fit, so that often $E\varepsilon_i^2 < \sigma_i^2$. Often it may make more sense to avoid overfitting and to accept that the true mean model can not be satisfactorily estimated without a time delay. As a slowly changing mean model estimator has low variance, its bias has to be taken into

account in the modelling of variance. Because of the bias, it often holds that $E\epsilon_i^2 > \sigma_i^2$. The following moving window framework is an example of the latter approach.

In original paper VIII a moving window framework for adaptive heteroscedastic linear regression is sketched. In the proposed method, the true parameters $\{\beta_t\}$ are assumed to form a continuous time Lévy process

$$E(\beta_{t_i} - \beta_{t_a}) = 0, \quad \text{cov}(\beta_{t_i} - \beta_{t_a}) = B|t_i - t_a|. \quad (74)$$

In the notation, the i th observation is observed at time t_i , when the true parameter value is β_{t_i} . Updating of the model begins with an estimation of the mean model parameters using a moving window. The resulting estimate $\hat{\beta}^w$ is used to construct the squared residuals. Assuming that the estimator $\hat{\beta}^w$ equals the true parameters at time t_a , $\hat{\beta}^w = \beta_{t_a}$, the expected squared error is

$$E(y_i - x_i^T \hat{\beta}^w)^2 = \sigma^2(x_i, \tau_{t_i}) + |t_i - t_a| x_i^T B x_i. \quad (75)$$

The result can be utilised in the estimation of the variance function: An additional offset variable

$$q_i = |t_i - t_a| x_i^T B x_i \quad (76)$$

is employed in the variance model fitting. The problem is then to estimate B and to determine t_a . The result Eq. (75) is employed also when future predictions are produced.

Another possible approach to simultaneous adaptive modelling of the mean and dispersion is to apply an adaptive on-line learning method to the joint log-likelihood of the mean and variance parameters. When variance does not depend on the mean, the information matrix is block diagonal and the mean and variance can be treated separately.

4 Industrial process data in predictive modelling

A large amount of data are measured from industrial production processes. The data contain information that is useful in controlling the production process. The problem is to extract information from the data. This has commonly been done by developing statistical prediction models, which are then utilised in process control, process planning and product planning (Khattree & Rao 2003).

Industrial process data sets are usually large. Both the number of variables and the number of observations are high. Process data are often clustered: there are clusters of observations and sparse or empty regions between the clusters. The clusters occur in the regions of normal process settings. Statistical analysis of large data sets has been commonly discussed under the topic of data mining (Giudici 2003).

A successful industrial plant has to develop its processes. The process changes as it is developed, which means the relationships between the process variables vary. The relationship described by the model can also change. As a consequence, the need for model updating is obvious. Adaptive prediction models that are fit for using process data have been commonly applied in industry (VanDoren 2002).

These special features characterise industrial process data for the purpose of regression modelling. In this chapter, the results related to process data-based regression modelling are presented.

4.1 Dispersion modelling using process data

Modelling of variance in large data sets has not been specifically discussed in earlier literature. Recent industry-oriented textbooks about statistical modelling (Khattree & Rao 2003, Giudici 2003) do not consider modelling of conditional variance. However, it is not unusual that not only the mean but also the deviation of the response variable depends on the process variables. In this case, a model of the deviation gives additional information and makes more efficient utilisation of the models possible. Previously, dispersion modelling methods have rarely been applied to industrial process data, but some examples include (Myllykoski 1998, Smyth 2002). A related topic is modelling of model uncertainty (Tjärnström 2002). In original paper V, variance modelling methods suitable

for several explanatory variables are surveyed and their suitability for large data sets is discussed.

Variance is often sensitive to even small changes in the process. Changes and variation in the facilities and practices of the production line cause heteroscedasticity and also time-dependent variation in the conditional variance function. In spite of that, methods for adaptive modelling of the variance function have not been discussed earlier. In original paper VIII, methods for adaptive modelling of conditional variance are proposed and applied to steel plate data.

4.2 Model selection

Expected squared prediction error is the sum of irreducible variance, squared model bias and model variance

$$E \left[y_0 - \widehat{f}(x_0) \right]^2 = \sigma_{\varepsilon}^2 + \left[E\widehat{f}(x_0) - f(x_0) \right]^2 + E \left[\widehat{f}(x_0) - E\widehat{f}(x_0) \right]^2. \quad (77)$$

Model complexity increases model variance, but decreases model bias. Optimal complexity is determined in the model selection phase. For example, in linear regression, model variance is

$$E(x_j^T \widehat{\beta} - x_j^T E\widehat{\beta})^2 = x_j^T \text{cov}(\widehat{\beta}) x_j. \quad (78)$$

Model variance depends on the location of the observation x_j . If x_j is located so that y_j has a large influence on model estimation, model variance around x_j is large. Model variance usually increases when the distance from the query point to the estimation data set grows.

The task of model selection includes selection of the structure of the model, the input variables and their transformations and the model estimation method. Validation, cross-validation, bootstrap validation and information criteria are the most common methods used for model selection. In validation the training data set is split into two parts: The prediction accuracy of the fitted model is measured in the data which are not used for learning. In K -fold cross validation, the data are split into K parts. The model is fitted K times using $K - 1$ parts for fitting and one part for validation at a time. The model is selected using the average prediction error in the K validation data sets. Bootstrap validation works similarly, except that the training data sets are obtained by re-sampling the original data. Information criteria consist of the log-likelihood of the fitted model in the training data and a penalty term for the complexity of the models. (Hastie *et al.* 2001.)

In original paper II, it is suggested that validation methods are preferable in model selection for large process data sets. The real predictive performance of the models can be measured most reliably in the validation data set (Purushottam & Ibrahim 1995). If the data are split between validation and training using a split in time, more emphasis is given to the interpolation ability of the models, because the estimation and the validation data sets are more dissimilar.

In original paper II, an approach to model selection for joint modelling of the mean and dispersion is presented. The variance model is selected on the basis of the validation deviance of squared residuals. The proposed procedure for model selection is

1. Specify preliminary models for the mean and dispersion.
2. The model for the mean is selected by minimising the weighted sum of squared errors in the validation data set V . Unit weights or weights proportional to the inverses of the predicted variances are used.
3. Squared residuals are calculated and used for variance model fitting and validation. The selected variance model minimises validation deviance

$$D_V = 2 \sum_{i \in V} \left[-\log \frac{\hat{\varepsilon}_i^2}{\hat{\sigma}_i^2} + \frac{\hat{\varepsilon}_i^2 - \hat{\sigma}_i^2}{\hat{\sigma}_i^2} \right]. \quad (79)$$

4. The whole model is re-estimated.

Some authors have emphasised the role of model generalisation ability in model selection Busemeyer (2000). In original paper VI, model selection that attempts to respond to the need for well-generalising models is proposed. The proposed distance-weighted validation criterion gives more weight to the validation data observations whose distance to the training data set is large. However, the experiments did not confirm the hypothesis that the proposed distance-weighted validation criterion selects a model with better interpolation capability than the unweighted validation criteria.

4.3 Utilisation of industrial prediction models

The aim of industrial data analysis is to produce information and models that can really be utilised to improve the production process. When prediction models are developed for industrial use, the utility of the models lies in their applicability. Prediction models can be utilised in planning products and the flow of production, in planning and optimising process settings, and in process control.

In joint modelling of the mean and variance, prediction of conditional distribution is constructed on the basis of the conditional mean and variance. Sometimes it may be better to construct conditional distribution using the expected squared error

$$E(y_i - \hat{y}_i)^2 \quad (80)$$

rather than the error variance σ_i^2 . It can be thought that the mean model is first specified to be as accurate and unbiased as possible. Then the uncertainty of the mean model is modelled so that the expected likelihood of future observations becomes maximised. Error variance usually forms the major part of the expected squared prediction error, but model bias and model variance also have their own effects, as presented in Eq. (77). When the model is used for interpolation, both model bias and variance increase and the expected squared prediction error grows rapidly.

Predictions can be needed in different stages of the production process. At early stages, the realised values of some explanatory variables are not necessarily known exactly. When the predictions are made only target values can be given for some of the explanatory variables. Optimally, the expectations of the realised values agree with the target values, but the realised values have some variance. Lacking information about the explanatory variables decreases prediction accuracy: The expected squared prediction error increases as

the dispersion between the target values and measured values of the explanatory variables grows. If predictions are queried from the same model in different process stages in which the amount of uncertainty about the input variables differ, then the effect of the uncertainty should be taken into account in the predicted variance. This approach has not been given much attention in the literature. The alternative is to construct different models to be used in the different process stages.

In original paper IV, it is proposed that model uncertainty is predicted as a sum of three components:

1. Error variance. Error variance can depend on the explanatory variables.
2. The effect of uncertainty about the input variables used for prediction. The use of target values in the early stages of the process increases the prediction error.
3. Model variance. The model is less accurate in the regions with sparse data.

The effect of uncertainty about the input variables is approximated by the error propagation formula

$$E \left[y_i - f(\hat{\beta}, x_i) \right]^2 = \hat{\sigma}_i^2 + \text{var} \left[f(\hat{\beta}, x_i) \right] + \sum_{j=1}^p \sigma_{(j)}^2 \left[\frac{\partial f(\hat{\beta}, x)}{\partial x_{(j)}} \right]_{(x=x_i)}^2 \quad (81)$$

where each observation x_i is a realisation of the input vector $x = (x_{(1)}, x_{(2)}, \dots, x_{(p)})$ and $\sigma_{(j)}^2$ is the expected squared difference between the target value and the realised value of $x_{(j)}$. The approximation employs the first order Taylor expansion of the fitted model around the query point.

4.4 Model maintenance

The long-term utilisation of industrial process models requires systematic analysis of the data that aims at maintaining of prediction accuracy. Because of the evolution of the processes and changes in the relationships described by the models, there is a need for maintenance of industrial prediction models. In model maintenance, the model is regularly re-fitted to correspond to the current state of the production process. The development of a completely new prediction model can be laborious and it can be made easier by utilising the experience from the previous models. In adaptive modelling, model re-fitting is performed automatically. If the variance depends on the inputs, the variance model may need to be re-fitted regularly.

A system for model maintenance has to systematically analyse and store the production line measurements. If a large amount of data are measured, it is not feasible to store all the data or, at least, not all of the data can be used for modelling. Algorithms for data pre-processing and selection are needed in model maintenance systems. To detect the need for model updating actions, the performance of the models in the process data should be monitored. A model maintenance system should also include algorithms for re-fitting of models. An illustration is given in Fig. 2.

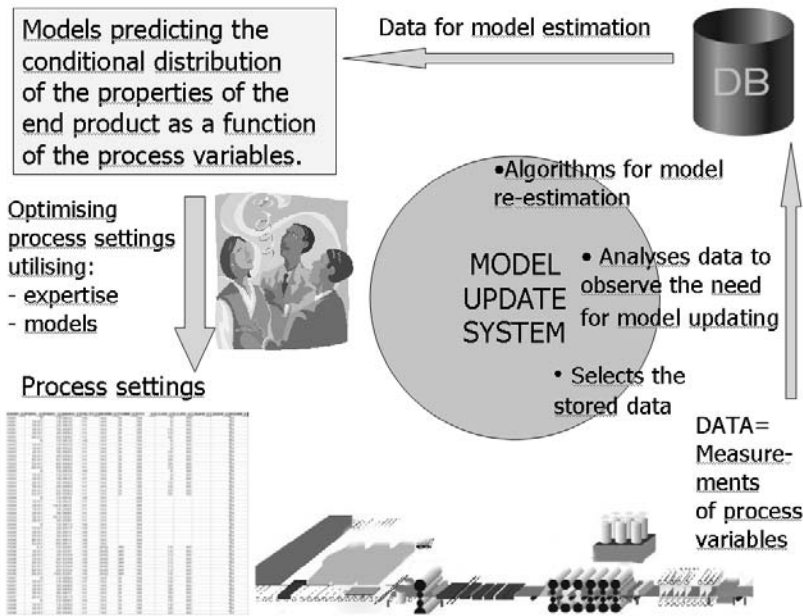


Fig. 2. Illustration of a system for maintaining process data-based models.

4.5 Measuring distance between data point and training data set

A measure of the distance between the training data set and a new query point or observation could be useful in assessing the uncertainty of prediction and in finding outliers. Measuring the distance from a training data set has not been proposed earlier, although quite similar approaches have been used. Standard errors of predictions measure uncertainty with variance, but they do not take bias into account. In local methods, distances between observations are utilised to find the nearest observations of a query point (Wettschereck *et al.* 1997). In clustering and prototype methods, several measures of the distance between a single observation and a set of observations, like average pairwise Euclidean distance $\sum_{i=1}^n \|x_0 - x_i\|/n$, Mahalanobis distance $(x_0 - m_x)^T S_x^{-1} (x_0 - m_x)$ and Euclidean distance to the cluster centroid $\|x_0 - m_x\|$, have been used (Kaufman & Rousseeuw 1990). Here S_x is the empirical covariance matrix of x and m_x is the average vector of x . In novelty detection, the aim is to detect abnormal observations. The usual approach has been to construct a model for a joint density function of input variables and judge observations with a density below a given threshold as novel (Markou & Singh 2003). Also several other approaches have been suggested, such as measuring the minimum distance $\min_i \|x_0 - s_i\|$ to a vector in a set of prototypes s_1, s_2, \dots, s_k (Alander *et al.* 1991).

An interesting approach was given by Angiulli & Pizzuti (2005), who used the sum of

Euclidean distances to the k nearest observations

$$\sum_{i=1}^k d_{(i)} \quad (82)$$

to rule out outliers. Mahamud & Hebert (2003) discussed optimal distance measures. The optimal distance measure in 1-nearest-neighbour prediction minimises the expected loss function $EL(y_0, y')$ where y' is the measured response at x' , which is the nearest neighbour of x_0 using distance measure d . The authors showed that the distance measure equaling the expected loss function $d(x_0, x_i) = EL(y_0, y_i)$ is optimal and they proposed to use a distance measure that approximates the expected loss of nearest neighbour prediction.

In original paper VI, a method for measuring the distance between a single observation and a training data set is proposed. The distance is measured as the harmonic sum of the distances to the k nearest observations in the data set

$$d = d(x_0, T) = \frac{1}{\sum_{i=1}^k \frac{1}{d_{(i)}}}. \quad (83)$$

The distances between single observations

$$d(x_0, x_j) = \hat{\sigma}^2 + \sum_{i=1}^p \kappa_i (x_{ji} - x_{0i})^2 \quad (84)$$

are squared Euclidean distances plus a constant $\hat{\sigma}^2$, the estimated error variance. In the formula, $d_{(1)}, d_{(2)}, \dots, d_{(k)}$ are the ordered k smallest distances $d(x_0, x_j)$ between x_0 and training data observations x_1, x_2, \dots, x_n and thus $d_{(i)} = d(x_0, x_j)$ for some j . The variables that have a large effect on the response have a large weight on the distance measure. In original paper VI, a data-driven procedure is proposed for defining the scalings κ_i of each variable in such a way that they are proportional to the variable's contribution to the fitted regression function. The proposed distance measure can be seen as a linear function of the approximated expected squared prediction error when the new observation is predicted by distance weighted k -nearest-neighbour. The expected squared prediction error is approximated under the condition that the distances and directions to the k nearest training data points and the measured response in these points are given.

Details, the mathematical reasoning behind the above definitions and the results for simulated data sets and for the steel plate data set are given in original paper VI. The results show that prediction accuracy depends on the distance to the training data: When the distance to the training data is large, the model gives a poor prediction with a high probability.

4.6 Interpolation ability of models

Often the predictive power at the boundaries of the data currently on hand is an important property in the practical utilisation of models. However, research on the interpolation ability of models at the boundaries of the data region has not been published previously.

The problem is that the generalisation ability of prediction models outside the data region is very difficult to measure. Intuitively, it is reasonable to assume that models behaving smoothly at the boundaries predict reliably in a wider volume around the data region, i.e. they interpolate better. A model that generalises well would be very useful in process optimisation problems and in finding novel improvements to the process. Unfortunately, all statistical modelling techniques have major problems in predicting reliably outside the training data. There are probably differences between the generalisation abilities of various learning methods, but no methods for gauging the merits of different learning methods have yet been proposed.

The upper bounds of the generalisation errors of learning methods have been derived under the assumption that the distribution of input variables does not change (Vapnik 1998). However, in industrial practice the distribution of inputs changes because of improvements in the process. A model with good generalisation ability is useful in the changing environment of industrial processes.

In original paper VII, it is proposed that the interpolation ability of models can be measured using the prediction accuracy of the validation data set observations whose distance from the training data set is large. The distance measure developed in original paper VI is utilised in measuring interpolation capability. In original paper VII, the interpolation capabilities of quadratic regression, local linear regression, additive spline models, multi-layer perceptron and support vector regression are compared. Quadratic regression and local linear regression interpolated poorly compared with the other methods, both for simulated data sets and for the steel plate data set. The result can be explained by the unstable boundary behavior of the quadratic terms and local fitting. The effect of model complexity and meta-parameters on interpolation capability was also examined. It seemed that model complexity does not have a strong effect on interpolation capability.

5 Application to the mechanical properties of steel plates

The methods presented in the thesis have been applied and developed for use at Ruukki's steel plate mill in Raahe. The research was carried out in three projects. In the first project, a product planning tool for planning the strength margins of steel plate products was developed. In the second project, a model for elongation was added to the developed tool. In addition, a tool for assuring the fulfillment of the requirements of the mechanical properties in the positioning of orders was developed for used in production planning. In the third, currently ongoing project, a maintenance system for the developed prediction models will be developed.

Both developed tools employ the same prediction models for tensile strength, yield strength and elongation. The functionality of the developed tools is based on predicting the probability of rejection in tensile testing. The probabilities are predicted on the basis of models for the conditional mean and conditional variance. In the Ruukki application, the term 'rejection' means the result of a single tensile test is below the specified minimum value. The research was based on a large data set collected from the production process of steel plates.

5.1 Modelling the probability of rejection in a qualification test

Joint modelling of the mean and dispersion has been commonly applied to the construction of confidence intervals (Rigby & Stasinopoulos 2000, Wright & Royston 1997). There are several proposals concerning how the confidence intervals are constructed. Let $z_{\alpha/2}$ denote the $\alpha/2$ quantile of the standard normal distribution. The basic method is to assume normal distribution and construct the confidence interval accordingly,

$$CL_{1-\alpha}(y_i) = [\hat{\mu}_i - z_{\alpha/2}\hat{\sigma}_i, \hat{\mu}_i + z_{\alpha/2}\hat{\sigma}_i]. \quad (85)$$

Rigby & Stasinopoulos (2000) proposed finding a Box-Cox transformation, Eq. (7), so that the error distribution of the transformed variable $y^* = L_\lambda(y)$ is Gaussian. The modelling is performed in the transformed scale. Mak (2002) proposed a sampling method for estimating variance in the original metric. The $100(1 - \alpha)\%$ confidence interval for y_i is

$$CL_{1-\alpha}(y_i) = [L_\lambda^{-1}(\hat{\mu}_i^* - z_{\alpha/2}\hat{\sigma}_i^*), L_\lambda^{-1}(\hat{\mu}_i^* + z_{\alpha/2}\hat{\sigma}_i^*)] \quad (86)$$

where $\widehat{\mu}_i^*$ and $\widehat{\sigma}_i^*$ are the predicted mean and variance of the transformed response. Akritas & van Keilegom (2001) proposed employing the empirical distribution of the standardised residuals

$$\widehat{F}_e(s) = \frac{1}{n} \sum_{i=1}^n I \left(\frac{y_i - \widehat{\mu}_i}{\widehat{\sigma}_i} \leq s \right) \quad (87)$$

to construct the confidence interval

$$CL_{1-\alpha}(y_i) = [\widehat{\mu}_i + \widehat{F}_e^{-1}(\alpha/2)\widehat{\sigma}_i, \widehat{\mu}_i + \widehat{F}_e^{-1}(1 - \alpha/2)\widehat{\sigma}_i]. \quad (88)$$

Confidence intervals constructed on the basis of the predicted mean and deviation have been employed in medical statistics (Stasinopoulos & Rigby 2000), for example. Also different conditional probabilities can be constructed by assuming a distribution for the standardised residuals. In original paper III, an approach to predicting the probability of rejection in a qualification test using joint modelling of the mean and dispersion is proposed. Let $[Y_{\min}, Y_{\max}]$ be the acceptance region of the qualification test. Then the probability of rejection is predicted with

$$P_r = F \left(\frac{Y_{\min} - \mu_i}{\widehat{\sigma}_i} \right) + 1 - F \left(\frac{Y_{\max} - \widehat{\mu}_i}{\widehat{\sigma}_i} \right) \quad (89)$$

where $F(s)$ is the assumed cumulative distribution function of the standardised residual. In original paper III, it is proposed to use the approximated empirical distribution function. The approximation consists of two halves of a cumulative normal distribution whose tails are replaced with exponential functions

$$\begin{aligned} F(s) &= C_l e^{-s\eta_l}, \text{ when } s < l_c \\ F(s) &= \Phi \left(\frac{s - m_c}{\sigma_l} \right), \text{ when } s \in [l_c, m_c] \\ F(s) &= \Phi \left(\frac{s - m_c}{\sigma_u} \right), \text{ when } s \in [m_c, u_c] \\ F(s) &= 1 - C_u e^{-s\eta_u}, \text{ when } s > u_c. \end{aligned} \quad (90)$$

Later it was observed that a better fit with the data can be achieved using a Box-Cox transformed response variable. The probability of rejection is predicted by transforming the requirements Y_{\min} and Y_{\max} and comparing them to the cumulative standard normal distribution

$$P_r = \Phi \left[\frac{L_\lambda(Y_{\min}) - \mu_i^*}{\widehat{\sigma}_i^*} \right] + 1 - \Phi \left[\frac{L_\lambda(Y_{\max}) - \widehat{\mu}_i^*}{\widehat{\sigma}_i^*} \right]. \quad (91)$$

5.2 Steel plate data set

The steel plate data set utilised in the research consists of tensile test results, measurements of the composition of steel, the dimensions of steel plates and slabs, the thermo-mechanical treatments of the production process and the shape of the test bar. About 50 variables were measured and about 30 of them have an effect on the response variables. About 225000 observations made between July 2001 and October 2005 were utilised.

In tensile testing, the test bar is drawn until it breaks. Three important mechanical properties are measured: tensile strength (R_m), yield strength (R_{eH}) and elongation (A_5) (Fig. 3). Elongation correlates negatively with the strengths, which are positively correlated. Tensile strength is considered the most important mechanical property of steel. All steel plate products have a minimum requirement for R_m . Most of the products have minimum requirements for R_{eH} and A_5 and a maximum requirement for R_m . Tensile testing is used to assure fulfillment of the requirements.

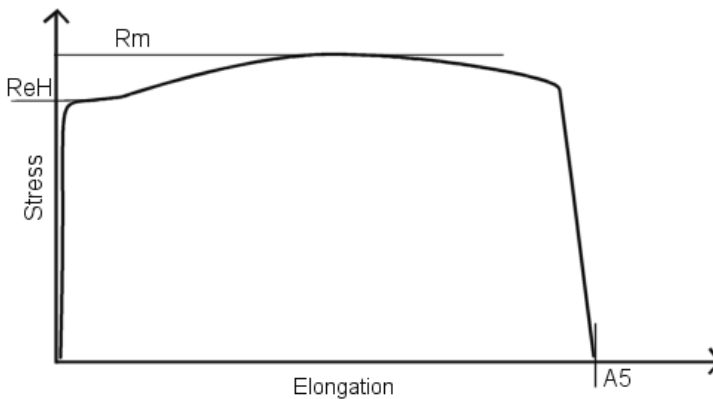


Fig. 3. Measurement of mechanical properties from a tensile test curve.

Steels are the most widely used metallic material because of the wide variety of properties that can be produced at relatively low cost for different purposes. Steel plates are produced in large-scale rolling mills. At Ruukki's steel plate mill the focus is on producing low-alloyed hot-rolled steels with a ferritic-pearlitic microstructure. The modelling done in this research aims at predicting the properties of these types of steels.

Many variables and mechanisms affect the mechanical properties of hot-rolled steel plates (Honeycombe 1981). These mechanisms are controlled with alloying as well as with heating, working, and cooling operations (thermomechanical treatments). In many cases complicated and expensive treatments are needed to obtain the required properties for the final product. Modelling of steel properties is important because many of the

required properties are achieved only if the interactions between composition and thermo-mechanical treatments can be predicted and controlled reliably. Modelling of the strength of steel is commonly applied in steel mills, and many models for strength have been created (Hodgson & Gibbs 1992, Dumortier & Leheret 1999). One model can usually cover only a small subset of all known steel grades.

5.3 Planning the strength margins of steel plates

Before this research work was done, planning of strength at Ruukki's steel plate mill had been based on an old planning model that predicts tensile strength with a linear combination of the composition of steel. The first motivation for this research was the need to update the model to predict better under the developed production process. The development of a new prediction model for tensile strength is reported in original paper I. Carefully specified regression models predicted similarly to neural network models, and it was decided to implement the regression model because of its better interpretability.

The variance of tensile strength depends strongly on several explanatory variables. It was observed that the variance heterogeneity of strength should be taken into account in planning the process settings for steel plate products. A large planning margin is needed between the expected strength and the strength requirements when the variance of strength is large. In original paper IV, an approach to planning strength margins on the basis of the predicted probability of rejection in a tensile test is presented.

The proposed approach to planning tensile strength margins seemed to work well, and a similar model was developed for yield strength. Both developed models were heteroscedastic linear regression models Eq. (32). Original paper II discusses the model selection procedure used in the development of the models. The models were implemented in a planning tool. The purpose of the planning tool is to help determine process settings for the products in product planning. The developed planning tool is in everyday use in Ruukki, and the number of rejections has significantly decreased after the model was introduced. Modelling deviation in addition to the mean significantly improves the model's ability to prevent rejections. Details are given in original paper IV. The decrease in rejections yields economical benefits to the steel plate mill.

One of the tasks of product planning is to define process settings separately for each product. This is a demanding task, as there are thousands of products and every product has many requirements for mechanical properties, alloying and the delivery condition. On the other hand, the flexibility of the production process is important because it allows more products to be produced in a shorter time, permits smaller orders to be accepted and decreases the need for storage slabs. Needless alloying and expensive treatments also have to be avoided. When the variance of mechanical properties is decreased, these goals can be achieved more easily. The situation is controlled by determining default values and acceptable limits for the values of the process variables for each product separately. Figure 4 illustrates planning of steel plate products using the developed planning tool.

The prediction models are used in two process phases: before casting the slabs, where only the target values of the composition of steel are available, and after casting, where the composition is already measured. Uncertainty about the composition increases the

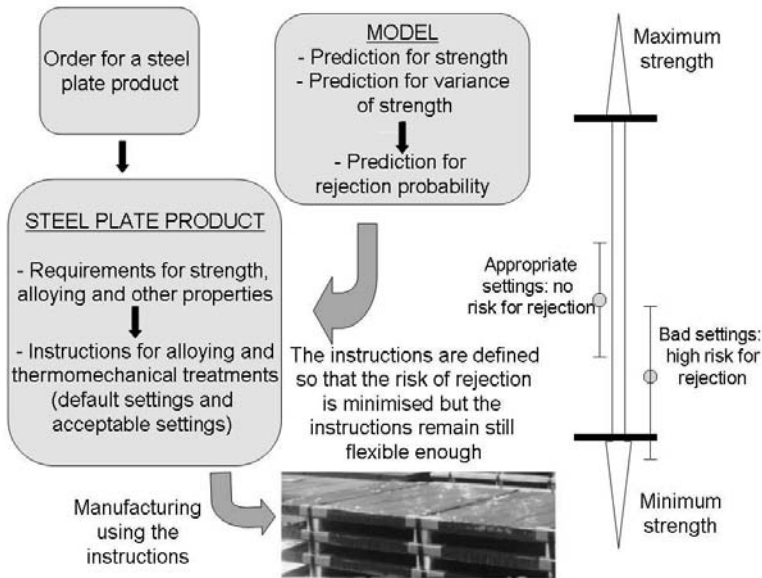


Fig. 4. Planning of steel plate products.

variance of the mechanical properties. The deviation of the mechanical properties is about 5-20 % larger before casting, depending on the input variables. When the same prediction model is used in both situations, the effect of the uncertainty about the composition before casting has to be added to the prediction of the variance model. A method for estimating the effect of uncertainty on the expected squared prediction error using the first order Taylor approximation is proposed in original paper IV. An alternative is to fit two separate models; the one used before casting fitted on the basis of target composition and the other fitted on the basis of the measured composition.

5.4 Modelling the mean and variance of elongation

Because the developed models for strength were considered useful, it was decided to expand the developed planning tool by also introducing a model for elongation. At the same time, the models for strength were updated. All three models were link-linear heteroscedastic regression models, Eq. (34). The models employ a transformed response variable to achieve a normally distributed error term. The link functions for the mean and variance were selected to maximise the fit with the data. In addition, the predicted mean was included in the explanatory variables of the variance model. All the models have about 120 terms in the mean model and about 30 terms in the variance model.

The functionality of the developed models is illustrated by using the developed model

to predict the mean and variance of elongation. Figure 5 gives an overview of the variance heterogeneity of elongation. Variance is highest when the steel plates are water-cooled or rolled at a low temperature. Variance is also high when the plates are relatively high-alloyed, but not very thin, and a normalising heat treatment is not applied. Variance is lowest for steel plates to which a normalising heat treatment is applied and whose predicted elongation is about 30.

The accuracy of the model in predicting rejections is illustrated in Fig. 6. Figure 7 is another illustration of the accuracy of the predicted probability in detecting rejections. Only products with over 30 measurements are included. Sensitivity means the proportion of correctly predicted rejections. Specificity means the proportion of correctly predicted acceptances. All of the presented results are obtained from an independent test data set.

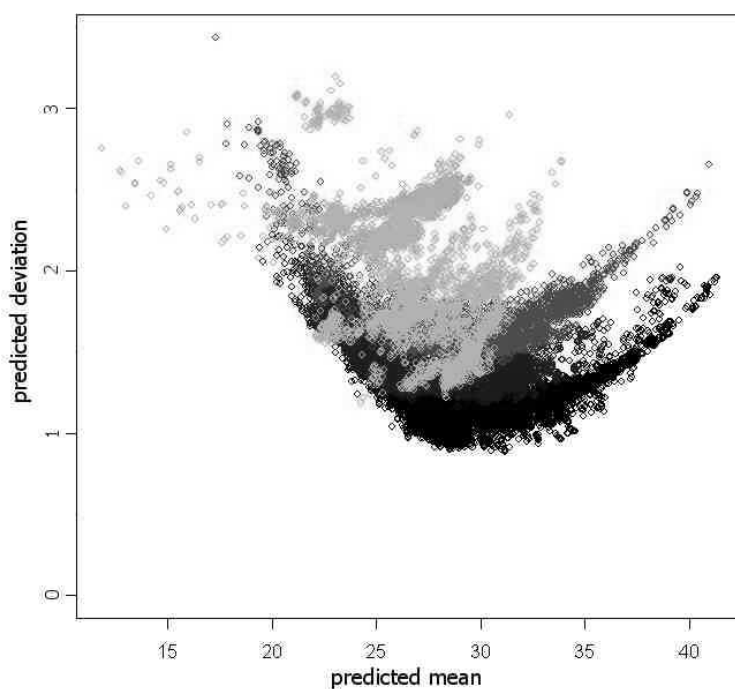


Fig. 5. Predicted deviation plotted against the predicted mean. The colours differentiate the different production method branches.

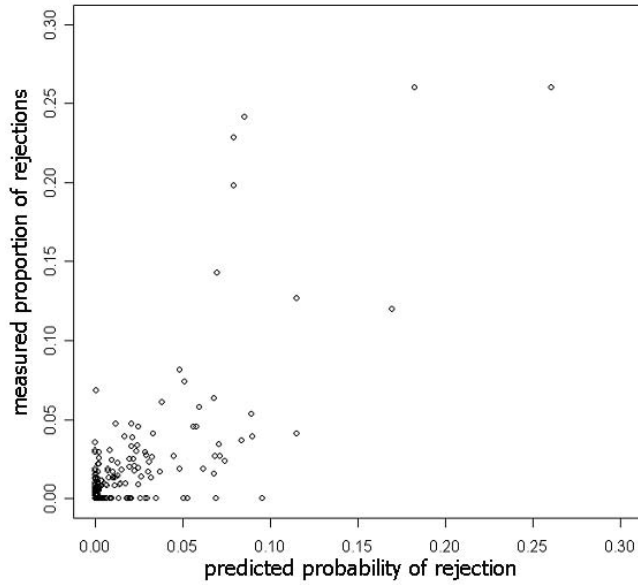


Fig. 6. Predicted and observed proportions of rejections. A point presents the average of a relatively homogeneous steel plate product.

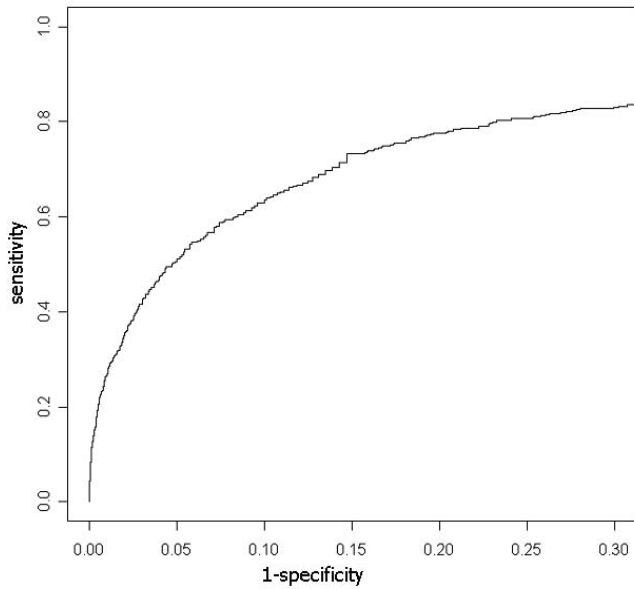


Fig. 7. ROC curve of the developed model in finding rejections in a tensile test.

6 Discussion

The methods presented in this thesis were developed for the purpose of predictive regression modelling for large industrial data sets. The presented ideas were tested using a large data set collected from a steel plate mill, but use of the methodology is not restricted to the steel industry. One data set does not give much evidence about the general importance of variance heteroscedasticity in analysing industrial process data. However, the various publications about quality improvement experiments, which are based on modelling of conditional variance, prove that variance heteroscedasticity is common in industrial processes. Analysis of designed experiments is a topic with a long history, but process data-based modelling has been established only recently. It seems probable that the developed methodology for joint modelling of the mean and dispersion can improve modelling accuracy in many kinds of industrial applications. The study demonstrates the potential need for modelling of conditional variance for industrial model developers.

When only the mean and variance of the response are modelled, it is assumed that the other moments of the distribution of the response are independent of the explanatory variables. In most applications, the conditional distribution of the response can be satisfactorily described by modelling only the first two moments. The advantage, compared with the usual approach of modelling only the mean, is that the conditional distribution can be modelled more accurately. The more accurate model can then be utilised in the applications. The advantage compared with modelling of skewness is simplicity: The danger of overfitting decreases and estimation of models is more efficient when additional parameters are not estimated. It seems probable that joint modelling of the mean and dispersion is the optimal solution in many cases: A model for conditional variance is needed, but a model for conditional skewness is unnecessary. Sometimes this is not enough, especially when the number of explanatory variables is small, and the conditional distribution function has to be modelled in more detail. Of course, in many cases constant variance is the most rational assumption and variance modelling is not needed.

The optimal variance model is often less complex than the optimal mean model, because observations give more information about the mean than about variance. A large number of observations is needed to fit a variance function that depends on several input variables. The presented methods focused on the case where there are at least a couple of input variables: In the case of only one or two input variables, even more detailed models can be utilised.

The application to the planning of strength margins shows that process data-based joint modelling of the mean and dispersion is a useful approach that can yield economical benefits when properly applied to industrial practice. The proposed approach to utilising the predicted probability of rejection in a qualification test could have plenty of applications, as the properties of final products have to be controlled in many industries.

6.1 Variance function estimation

Variance function fitting and mean model fitting can be accomplished separately using iterative methods. Thus, the form of the mean model does not constrain the learning method used in variance function estimation. The results of the study support the approach of independently selecting the learning method used in the estimation of the conditional variance function. It is not necessary to limit the proposed methods to joint modelling of the mean and dispersion; modern learning methods can also be utilised in the modelling of conditional variance. Surely, the optimal learning method depends on the application, and usually several alternatives give good results. Only a few extensive publications about variance modelling methods have been written (Dadidian & Carroll 1987, Carroll & Ruppert 1988, Welsh *et al.* 1994). Thus, the reviews written in this study are a relevant complement to the existing literature.

Variance is even more sensitive to outliers than is the mean. Real but exceptional instances of the process can contain very important information, which may lead to new understanding. Selective removal of true observations from the data leads to biased models, especially when variance is modelled. Thus, removal of exceptional true observations must be avoided. Another simple approach is to downweight observations that have been interpreted as outliers. The problems with this approach are the same, although more acceptable. Because the problem of outliers is emphasised in variance modelling, it seems natural to develop robust estimators for variance. The topic has been, however, rarely discussed. In our study, the bounded influence estimator Eqs. (29), (30) and (31) predicted too small variances because of systematic downweighting of the largest squared errors. Development of robust variance function estimators seems to be an interesting topic for further studies.

6.2 Process data-based modelling

A model is only an approximation of complex relationships. Models are based on data, a random instance produced by the underlying relationships. The data were collected in the past, but the model is used to improve the future.

In model development, special attention has to be paid to the applicability of models. Before models are implemented into the production line, the economic benefit of applying the models have to be assured. Cumbersome or needless implementations cause complexity and costs instead of benefits. Often, the payback of model application can not be measured directly using the average prediction accuracy of the model, such as MSE or log-likelihood. The utility value of models should be taken into account already in

the model development phase. Single bad predictions given by a model that usually predicts well can cause costly mistakes. The model should be accompanied with information about the reliability of prediction.

Expected response values when using familiar settings are often known even without a model. The model can be most valuable in predicting the properties of rarely manufactured products on the basis of rare process settings. Major improvements may be realised by using completely novel settings. Such settings can rarely be tested in the production line, because failures would be too costly. The validity of a model cannot be assumed in the empty parts of the data space. Optimisation of process settings has to be restricted to the data region where the model is reliable. A model with great interpolation ability would be very valuable in optimisation.

Because the process is usually run with familiar settings, it is quite difficult to obtain new information using the process data. An alternative to process data-based modelling is designed experiments. Designed experiments can give data from process settings that have not yet been tried in production. The drawback of designed experiments is their cost. A large number of expensive experiments would be needed to obtain an extensive data set, especially to model variance using many design factors. The advantage of process data-based modelling is that the data are cheap, as additional arrangements are not needed to obtain the data for modelling. The large amount of process data facilitates more accurate and more complex models than could be obtained using designed experiments.

6.3 Adaptive modelling

The maintenance of prediction models should be taken into account already in the implementation of the model applications. Implementation of adaptive models may sometimes be risky, because then the prediction model can change uncontrollably. A static model with an easy possibility for updating may often be a better alternative.

When a large amount of data is measured, it is not feasible to store all the data. The data can include millions of measurements from the same, typical process settings. Not all of them are needed in model fitting, and the whole data set cannot even be handled by the learning algorithm. On the other hand, measurements using rare and experimental process settings can be valuable for expanding the reliable prediction region of the model. An interesting approach to model maintenance is to refit the models using exceptional input observations for a very long time period, but only the newest observations for the usual settings. The practical functionality of the idea needs to be studied further.

Time-dependent changes in the modelled relationships of industrial processes are difficult to formalise in a statistical framework. The changes can be related to developments in the process, wearing of facilities or imperceptible changes in process practice. It may be difficult to describe a stochastic process that could explain the time-varying parameters of an industrial process model. Some of the changes are permanent, but some are temporary. Some of changes are slow shifts, but some can be sudden changes. Because of the permanent changes, the assumption about stationarity does not often hold. In spite of that, assuming stationarity may often be rational in order to simplify adaptive modelling.

Joint modelling of the time-varying mean and time-varying dispersion is a complicated problem. It is difficult to distinguish which part of the prediction error is due to mean

model bias and which is due to error variance. When a time-varying phenomenon is modelled in real time, the current parameter values are usually inaccurately estimated. Old parameter values can be estimated more accurately, because from their point of view also future observations are available for estimation. Obviously, differences in the accuracy of the mean model should be taken into account in the modelling of time-varying conditional variance. The topic is raised the first time in this thesis, and more studies are needed to develop more efficient methods.

Industrial processes are developed, and it is clear that the variance structure sometimes changes at the same time. However, it is unclear how important time-dependent variation is in the conditional variance function in practical applications. If the changes in the conditional variance function are small, the advantages of time-varying modelling can be smaller than the cost of implementation. Adaptive modelling of variance decreases the need for model updates. It seems probable that the proposed methods for adaptive modelling of the variance function are useful in some cases, but not very often.

6.4 Uncertainty of prediction

Basically, it seems rational to fit the model using the same explanatory variables that are available in the prediction phase. When only target values are available in prediction, target values should also be used in model fitting. Then the model answers the right question "*How is the response distributed when these target values are given?*". However, if the same model is needed in several process stages, it seems attractive to fit only one model using the measured values and to estimate the effect of using target input values on conditional variance. In these cases, the proposed method for estimating the effect of uncertainty about explanatory variables on variance can be utilised.

Information about the reliability of prediction is important when a model is utilised in decision-making. The developed distance measure seems most useful when considering the query points at the boundaries of the training data set. If the query point is clearly outside the data, then the prediction is not reliable. If the query point is in a dense data region, then the prediction is reliable. But, in cases where the query point is at the boundaries of the data, the distance can give valuable information about the reliability of prediction. In very large data sets, the proposed distance measure is, however, computationally too expensive for real-time applications.

The interpolation ability of models is difficult to compare in the model selection phase, because validating model performance in the empty regions of the input space is impossible. Thus, information about the interpolation capability of learning methods is very interesting when the model application needs prediction accuracy at the boundaries of the data. The steel plate data set is very suitable for comparing the interpolation capabilities of learning methods: The data include observations about rare products and settings, and several new products were introduced during the research period. In the study conducted on interpolation capabilities, the results from the simulated data sets and the steel plate data set supported each other. It can be concluded that the study gave a strong indication of the differences in the interpolation capabilities of the learning methods. The topic is interesting and important, and it would be useful to conduct further studies with other large data sets.

7 Summary

This thesis combined two known statistical methodologies: modelling of a conditional variance function and industrial process data-based prediction. The thesis presented methods, applications and approaches to utilising process data by means of joint modelling of the mean and dispersion.

A comprehensive review of the methods used to model conditional variance was given. The suitability of the methods for predictive modelling based on large industrial data sets was evaluated and compared. It was suggested that modern statistical learning methods can be applied to modelling of conditional variance.

The thesis highlighted the need and possibility of adaptive modelling of a time-varying variance function. Two methods for adaptive modelling of conditional variance were proposed, and the background for the methods was presented in detail.

Methods for estimating the uncertainty of a model's prediction at a query point were discussed. Variance modelling was utilised in a method developed for evaluation of the uncertainty of prediction at early process stages, when all the explanatory variables are not known. A novel measure of the distance between a training data set and a query point was proposed. The method approximates the model uncertainty based on the density of the data around the query point. The distance measure was utilised in a method developed for measuring the interpolation ability of models. A comparison of the interpolation capabilities of different learning methods was presented. The most important result was that local methods seem to interpolate poorly.

The developed methods were applied to a large data set related to the mechanical properties of steel plates. An approach to planning working allowances based on the predicted probability of rejection in a qualification test was proposed. The probability of rejection is predicted using joint modelling of the mean and dispersion. The approach was implemented in a planning tool that has been successfully used in a steel plate mill to optimise process settings. Modelling of the conditional variance of strength has yielded economic benefits for the steel plate mill.

References

- Aitkin, M. (1987) Modelling variance heterogeneity in normal regression using GLIM. *Applied Statistics* 36: 332–339.
- Akritas, M.G. & van Keilegom, I. (2001) Non-parametric estimation of the residual distribution. *Scandinavian Journal of Statistics* 28: 549–567.
- Alander, J.T., Frisk, M., Holmström, L., Hämmäläinen, A. & Tuominen, J. (1991) Process error detection using self-organizing feature maps. In: Kohonen, T., Mäkisara, O., Simula, O. & Kangas, J. (eds), *Artificial Neural Networks*. Elsevier, Amsterdam, 2: 1229–1232.
- Andersen, T.G., Chung, H-J. & Sørensen, B.E. (1999) Efficient method of moments estimation of a stochastic volatility model: A Monte Carlo study. *Journal of Econometrics* 91: 61–87.
- Andreou, E. & Ghysels, E. (2006) Monitoring for disruptions in financial markets. *Journal of Econometrics* 135: 77–124.
- Angiulli, F. & Pizzuti, C. (2005) Outlier mining in large high-dimensional data sets. *IEEE Transactions on Knowledge and Data Engineering* 17: 203–215.
- Bartlett, M.S. & Kendall, D.J. (1946) The statistical analysis of variance-heterogeneity and the logarithmic transformation. *Journal of the Royal Statistical Society: Series B* 8: 128–138.
- Beran, R. (1995) Prediction in random coefficient regression. *Journal of Statistical Planning and Inference* 43: 205–213.
- Bianco, A., Boente, G. & di Rienzo, J. (2000) Some results for robust GM-based estimators in heteroscedastic regression models. *Journal of Statistical Planning and Inference* 89: 215–242.
- Biehl, M. & Caticha, N. (2001) Statistical mechanics of generalization. In: Arbib, M.A. (ed), *Handbook of Brain Theory and Neural Networks*. MIT Press, Cambridge, MA, 922–924.
- Bishop, C.M. (1995) *Neural Networks for Pattern Recognition*. Clarendon press, Oxford, UK.
- Bollerslev, T. (1986) Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics* 31: 307–327.
- Bottou, L. (1998) Online learning and stochastic approximations. In: Saad, D. (ed), *On-Line Learning in Neural Networks*. Cambridge University Press, Cambridge, UK, 9–42.
- Box, G. (1988) Signal-to-noise ratios, performance criteria, and transformations. *Technometrics* 30: 1–17.
- Boyd, J. & White, H. (1994) Estimating data dispersion using neural networks. In: *Proceedings of IEEE World Congress of Artificial Intelligence*. Orlando, 4: 2175–2178.
- Buchinsky, M. (1998) Recent advances in quantile regression models. *The Journal of Human Resources* 33: 88–126.
- Busemeyer, J.R. (2000) Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology* 44: 171–189.
- Cantoni, E. (2004) Analysis of robust quasi-deviances for generalized linear models. *Journal of Statistical Software* 10(4): 1–9.
- Carlyle, W.M., Montgomery, D.C. & Runger, G.C. (2000) Optimization problems and methods in quality control and improvement. *Journal of Quality Technology* 32: 1–17.

- Carroll, R.J. (1982) Adapting for heteroscedasticity in linear models. *The Annals of Statistics* 10: 1224–1233.
- Carroll, R.J. & Ruppert, D. (1982) Robust estimation in heteroscedastic linear models. *The Annals of Statistics* 10: 429–441.
- Carroll, R.J. & Ruppert, D. (1988) *Transformation and Weighting in Regression*. Chapman and Hall, New York.
- Cawley, G.C. & Talbot, N.L.C. (2005) Constructing Bayesian formulations of sparse kernel learning methods. *Neural Networks* 18: 674–683.
- Cawley, G.C., Talbot, N.L.C., Foxall, R.J., Dorling, S.R. & Mandic, D.P. (2004) Heteroscedastic kernel ridge regression. *Neurocomputing* 57: 105–124.
- Chan, L.K. & Mak, T.K. (1995) A regression approach for discovering small variation around the target. *Applied Statistics* 44: 369–377.
- Chan, L.K. & Mak, T.K. (2001) Heteroscedastic regression models and applications to quality control. *Scandinavian Journal of Statistics* 28: 445–454.
- Chiou, J-M. & Müller, H-G. (1999) Nonparametric quasi-likelihood. *The Annals of Statistics* 27: 36–64.
- Clark, T.E. & McCracken, M.W. (2004) Improving forecast accuracy by combining recursive and rolling forecasts. Research Working Paper RWP 04-10, Federal Reserve Bank of Kansas City. Cited February 22nd 2006 from <http://ideas.repec.org/p/fip/fedkrw/rwp04-10.html>.
- Cole, T.J. & Green, P.J. (1992) Smoothing reference centile curves: The LMS method and penalized likelihood. *Statistics in Medicine* 11: 1305–1319.
- Cooley, T.F. & Prescott, E.C. (1976) Estimation in the presence of stochastic parameter variation. *Econometrica* 44: 167–84.
- Davidian, M. & Carroll, R.J. (1987) Variance function estimation. *Journal of American Statistical Association* 82: 1079–1091.
- Davidian, M. & Carroll, R.J. (1988) A note on extended quasi-likelihood. *Journal of the Royal Statistical Society: Series B* 50: 74–82.
- Davis, D.T. & Hwang, J-N. (1998) Expanding Gaussian kernels for multivariate conditional density estimation. *IEEE Transactions on Signal Processing* 46: 269–275.
- de Gooijer, J.G. & Zerom, D. (1999) On additive conditional quantiles with high-dimensional covariates. *Journal of the American Statistical Association* 98: 135–146.
- Dorling, S.R., Foxall, R.J., Mandic, D.P. & Cawley, G.C. (2003) Maximum-likelihood cost functions for neural network models of air quality data. *Atmospheric Environment* 37: 3435–3443.
- Dumortier, C. & Leher, P. (1999) Statistical modelling of mechanical tensile properties of steels by using neural networks and multivariate data-analysis. *ISIJ International* 39: 980–985.
- Dunn, P.K. & Smyth, G.K. (1996) Randomized quantile residuals. *Journal of Computational and Graphical Statistics* 5: 236–244.
- Engel, Y., Mannor, S. & Meir, R. (2004) The kernel recursive least-squares algorithm. *IEEE Transactions on Signal Processing* 52: 2275–2285.
- Engle, R.F. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of UK inflation. *Econometrica* 50: 987–1007.
- Fan, J. & Yao, Q. (1998) Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* 85: 645–660.
- Fan, J. & Zhang, W. (1999) Statistical estimation in varying coefficient models. *The Annals of Statistics* 27: 1491–1518.
- Fan, J., Yao, Q. & Tong, H. (1996) Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* 83: 189–206.
- Fan, S.K.A. (2000) Generalized global optimization algorithm for dual response systems. *Journal of Quality Technology* 32: 444–456.
- Giltinan, D.M., Carroll, R.J. & Ruppert, D. (1986) Some new estimation methods for weighted regression when there are possible outliers. *Technometrics* 28: 219–230.
- Giudici, P. (2003) *Applied Data Mining: Statistical Methods for Business and Industry*. Wiley, New York.
- Gong, G. & Samaniego, F.J. (1981) Pseudo maximum likelihood estimation: Theory and applications. *The Annals of Statistics* 9: 861–869.

- Gourierouxa, C. & Monfort, A. (1992) Qualitative threshold ARCH models. *Journal of Econometrics* 52: 159–199.
- Hall, P., Wolff, R.C.L. & Yao, Q. (1999) Methods for estimating a conditional distribution function. *Journal of the American Statistical Association* 94: 154–163.
- Hamada, M. & Nelder, J.A. (1997) Generalized linear models for quality-improvement experiments. *Journal of Quality Technology* 29: 292–304.
- Harvey, A.C. (1976) Estimating regression models with multiplicative heteroscedasticity. *Econometrica* 44: 461–465.
- Harvey, A.C. (1989) *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge, UK.
- Hastie, T., Tibshirani, R. & Friedman, J. (2001) *The Elements of Statistical Learning*. Springer, New York.
- Hildreth, C. & Houck, J. (1968) Some estimators for a linear model with random coefficients. *Journal of the American Statistical Association* 63: 584–595.
- Hodgson, P. & Gibbs, R. (1992) A mathematical model to predict the final properties of hot rolled C-Mn and microalloyed steels. *ISIJ International* 32: 1329–1338.
- Honeycombe, R. (1981) *Steels: Microstructure and Properties*. Edward Arnold Ltd, London.
- Härdle, W. & Tsybakov, A. (1997) Local polynomial estimators of the volatility function in non-parametric autoregression. *Journal of Econometrics* 81: 223–242.
- Huang, J. & Shen, H. (2004) Functional coefficient regression models for nonlinear time series: A polynomial spline approach. *Scandinavian Journal of Statistics* 31: 515–534.
- Huber, P.J. (2004) *Robust Statistics*. Wiley, New York.
- Husmeier, D. (1999) *Neural Networks for Conditional Probability Estimation: Forecasting beyond Point Predictions*. Springer, New York.
- Jacquier, E., Polson, N.G. & Rossi, P.E. (1994) Bayesian-analysis of stochastic volatility models. *Journal of Business and Economic Statistics* 12: 371–389.
- Jobson, D. & Fuller, W.A. (1980) Least squares estimation when the covariance matrix and parameter vector are functionally related. *Journal of the American Statistical Association* 75: 176–181.
- Kalman, R.E. (1960) A new approach to linear filtering and prediction problems. *Transactions of the ASME - Journal of Basic Engineering series D* 82: 35–45.
- Kaufman, L. & Rousseeuw, P.J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.
- Khattree, R. & Rao, C.R. (eds) (2003) *Statistics in industry - Handbook of statistics* 22. Elsevier, Amsterdam.
- Kim, K. & Lin, D. (2006) Optimization of multiple responses considering both location and dispersion effects. *European Journal of Operational Research* 169: 133–145.
- Kivinen, J., Smola, A.J. & Williamson, R.C. (2004) Online learning with kernels. *IEEE Transactions on Signal Processing* 52: 2165–2176.
- Köksoy, O. & Dognaksoy, N. (2003) Joint optimization of mean and standard deviation using response surface methods. *Journal of Quality Technology* 35: 239–252.
- Koenker, R. & Bassett, G. (1978) Regression quantiles. *Econometrica* 46: 33–46.
- Koop, G.M. & Potter, S.M. (2004) Forecasting and estimating multiple change-point models with an unknown number of change-points. *Discussion Papers in Economics* 04/31, Department of Economics, University of Leicester. Cited February 22nd 2006 from <http://ideas.repec.org/p/lec/leecon/04-31.html>.
- Krzyzak, A. (1992) Global convergence of the recursive kernel regression estimates with applications in classification and nonlinear system estimation. *IEEE Transactions on Information Theory* 38: 1323–1338.
- Kuk, A. (1999) Nonparametrically weighted least squares estimation in heteroscedastic linear regression. *Biometrical Journal* 41: 401–410.
- Leisch, F., Hornik, K. & Kuan, C.M. (2000) Monitoring structural changes with the generalized fluctuation test. *Econometric Theory* 16: 835–854.
- Li, W.K., Ling, S. & McAleer, M. (2002) Recent theoretical results for time series models with GARCH errors. *Journal of Economic Surveys* 16: 245–269.
- Liu, L. & Hanssens, D. (1981) A Bayesian approach to time-varying cross-sectional regression models. *Journal of Econometrics* 15: 341–356.

- Ljung, L. & Söderström, T. (1983) *Theory and Practice of Recursive Identification*. MIT Press, Cambridge, MA.
- Mahamud, S. & Hebert, M. (2003) Minimum risk distance measure for object recognition. In: *Proceedings of IEEE International Conference on Computer Vision*. Nice, France, 1: 242–248.
- Mak, T.K. (1992) Estimation of parameters in heteroscedastic linear models. *Journal of the Royal Statistical Society: Series B* 54: 649–655.
- Mak, T.K. (2002) Modelling and estimating regression variances. *Communications in Statistics - Theory and Methods* 31: 351–365.
- Mak, T.K. & Nebebe, F. (2003) Minimizing a general loss function in on-line quality control. *Applied Stochastic Models in Business and Industry* 18: 75–85.
- Markou, M. & Singh, S. (2003) Novelty detection: A review. *Signal Processing* 83: 2481–2521.
- McCullagh, P. & Nelder, J.A. (1989) *Generalized Linear Models*. Chapman and Hall, New York.
- McGilchrist, C.A. & Matawie, K.M. (1998) Recursive residuals in generalised linear models. *Journal of Statistical Planning and Inference* 70: 335–344.
- Mercurio, D. & Spokoiny, V. (2004) Statistical inference for time-inhomogeneous volatility models. *Annals of Statistics* 32: 577–602.
- Müller, H-G. & Stadtmüller, U. (1987) Estimation of heteroscedasticity in regression analysis. *The Annals of Statistics* 15: 610–625.
- Müller, H-G. & Stadtmüller, U. (1993) On variance function estimation with quadratic forms. *Journal of Statistical Planning and Inference* 35: 213–231.
- Montgomery, D.C. (1999) *Experimental design for product and process design and development*. *The Statistician* 48: 159–177.
- Murata, N. (1998) A statistical study of on-line learning. In: Saad, D. (ed), *On-line learning in Neural Networks*. Cambridge University Press, Cambridge, UK, 63–92.
- Murata, N., Kawanabe, M., Ziehe, A., Müller, K.R. & Amari, S. (2002) On-line learning in changing environments with applications in supervised and unsupervised learning. *Neural Networks* 15: 743–760.
- Myers, R.H. (1999) Response surface methodology - current status and future directions. *Journal of Quality Technology* 31: 30–44.
- Myers, R.H. & Carter, W.H. (1973) Response surface techniques for dual response systems. *Technometrics* 15: 301–317.
- Myers, R.H., Khuri, A.I. & Vining, G. (1992) Response surface alternatives to the Taguchi robust parameter design approach. *American Statistician* 46: 131–139.
- Myllykoski, P. (1998) A study on the causes of deviation in mechanical properties of thin steel sheets. *Journal of Materials Processing Technology* 79: 9–13.
- Nair, V.N. (1992) Taguchi's parameter design: A panel discussion. *Technometrics* 34: 127–204.
- Nair, V.N. & Pregibon, D. (1988) Analyzing dispersion effects from replicated factorial experiments. *Technometrics* 30: 247–257.
- Nelder, J.A. & Lee, Y. (1991) Generalized linear models for the analysis of Taguchi-type experiments. *Applied Stochastic Models and Data Analysis* 7: 107–120.
- Nelder, J.A. & Pregibon, D. (1987) An extended quasi-likelihood function. *Biometrika* 74: 221–232.
- Opper, M. (1996) On-line versus off-line learning from random examples: General results. *Physical Review Letters* 77: 4671–4674.
- Orr, G.B. & Leen, T.K. (1996) Using curvature information for fast stochastic search. In: Mozer, M.C., Jordan, M.J. & Petsche, T. (eds), *Advances in Neural Information Processing Systems* 9. MIT press, Cambridge, MA, 606–612.
- Pan, Z. & Wang, X. (2000) A wavelet-based nonparametric estimator of the variance function. *Computational Economics* 15: 79–87.
- Park, R.E. (1966) Estimation with heteroscedastic error terms. *Econometrica* 34: 888–888.
- Peracchi, F. (2002) On estimating conditional quantiles and distribution functions. *Computational Statistics and Data Analysis* 38: 433–447.
- Plackett, R.L. (1950) Some theorems in least squares. *Biometrika* 37: 149–157.
- Pollock, D.S.G. (2003) Recursive estimation in econometrics. *Computational Statistics and Data Analysis* 44: 37–75.

- Purushottam, W.L. & Ibrahim, J.G. (1995) Predictive model selection. *Journal of the Royal Statistical Society: Series B* 57: 247–262.
- Riddington, G.L. (1993) Time varying coefficient models and their forecasting performance. *Omega* 21: 573–583.
- Rigby, R.A. & Stasinopoulos, D.M. (1996) A semi-parametric additive model for variance heterogeneity. *Statistics and Computing* 6: 57–65.
- Rigby, R.A. & Stasinopoulos, D.M. (2000) Construction of reference centiles using mean and dispersion additive models. *The Statistician* 49: 41–50.
- Rigby, R.A. & Stasinopoulos, D.M. (2005) Generalized additive models for location, scale and shape. *Applied Statistics* 54: 1–48.
- Robbins, H. & Monro, S. (1951) A stochastic approximation method. *Annals of Mathematical Statistics* 22: 400–407.
- Robinson, T.J., Connie, M.B. & Myers, R.H. (2004) Robust parameter design: A review. *Quality and Reliability Engineering International* 20: 81–101.
- Rosenberg, B. (1972) The estimation of stationary stochastic regression parameters re-examined. *Journal of the American Statistical Association* 67: 650–654.
- Ruiz, E. (1994) Quasi-maximum likelihood estimation of stochastic volatility models. *Journal of Econometrics* 63: 289–306.
- Ruppert, D., Wand, M.P., Holst, U. & Hössjer, O. (1997) Local polynomial variance-function estimation. *Technometrics* 39: 262–273.
- Rutemiller, H.C. & Bowers, D.A. (1968) Estimation in a heteroscedastic regression model. *Journal of American Statistical Association* 63: 552–557.
- Saad, D. (ed) (1998) *On-line Learning in Neural Networks*. Cambridge University Press, Cambridge, UK.
- Sarajedini, A., Hecht-Nielsen, R. & Chau, P.M. (1999) Conditional probability density function estimation with sigmoidal neural networks. *IEEE Transactions on Neural Networks* 10: 231–238.
- Schaal, S., Atkeson, C.G. & Vijayakumar, S. (2002) Scalable techniques from nonparametric statistics for real-time robot learning. *Applied Intelligence* 17: 49–60.
- Schittenkopf, C., Dorffner, G. & Dockner, E.J. (2000) Forecasting time-dependent conditional densities: A semi-non-parametric neural network approach. *Journal of Forecasting* 19: 355–374.
- Schölkopf, B. & Smola, A.J. (2002) *Learning with Kernels: Support Vector Machines, Regularization, Optimization and beyond*. MIT press, Cambridge, MA.
- Shao, J. (1992) Empirical Bayes estimation of heteroscedastic variances. *Statistica Sinica* 2: 495–518.
- Shively, T.S. & Kohn, R. (1997) A Bayesian approach to model selection in stochastic coefficient regression models and structural time series models. *Journal of Econometrics* 76: 39–52.
- Shoemaker, A.C., Tsui, K-L. & Wu, J. (1991) Economical experimentation methods for robust design. *Technometrics* 33: 415–427.
- Smyth, G.K. (1989) Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society: Series B* 51: 47–60.
- Smyth, G.K. (2002) An efficient algorithm for REML in heteroscedastic regression. *Journal of Computational and Graphical Statistics* 11: 836–847.
- Smyth, G.K. & Verbyla, A.P. (1999) Adjusted likelihood methods for modelling dispersion in generalized linear models. *Environmetrics* 10: 696–709.
- Smyth, G.K., Huele, A.V. & Verbyla, A.P. (2001) Exact and approximate REML for heteroscedastic regression. *Statistical Modelling* 1: 161–175.
- Sompolinsky, H., Barkai, N. & Seung, H.S. (1995) On-line learning of dichotomies: Algorithms and learning curves. In: Oh, J.H., Cho, S.Z. & Kwon, C.L. (eds), *Neural networks: The statistical mechanics perspective*. World Scientific Publishing, Singapore, 105–130.
- Stadtmüller, U. & Tsybakov, A.B. (1995) Nonparametric recursive variance estimation. *Statistics* 27: 55–63.
- Stasinopoulos, D.M. & Rigby, R.A. (2000) Modelling rental guide data using mean and dispersion additive models. *The Statistician* 49: 479–493.
- Taguchi, G. (1986) *Introduction to Quality Engineering : Designing Quality into Products and Processes*. Asian Productivity Organization, Tokyo.

- Taguchi, G. (1987) *System of Experimental Design : Engineering Methods to Optimize Quality and Minimize Costs*. Unipub/Kraus International Publications, New York.
- Tang, L.C. & Xu, K.A. (2002) Unified approach for dual response surface optimization. *Journal of Quality Technology* 34: 437–447.
- Tjärnström, F. (2002) *Variance Expressions and Model Reduction in System Identification*. Ph.D. thesis, Linköping Studies in Science and Technology 730.
- VanDoren, V. (ed) (2002) *Techniques for Adaptive Control*. Elsevier, Burlington, MA.
- Vapnik, V. (1998) *Statistical Learning Theory*. Wiley, New York.
- Verbyla, A.P. (1993) Modelling variance heterogeneity: Residual maximum likelihood and diagnostics. *Journal of the Royal Statistical Society: Series B* 55: 493–508.
- Vining, G. & Bohn, L.L. (1998) Response surfaces for the mean and variance using a nonparametric approach. *Journal of Quality Technology* 30: 282–291.
- Vining, G. & Myers, R.H. (1990) Combining Taguchi and response surface philosophies: A dual response approach. *Journal of Quality Technology* 22: 15–22.
- Wedderburn, R.W.M. (1974) Quasi-likelihood functions, generalized linear models, and Gauss Newton-method. *Biometrika* 61: 439–447.
- Welsh, A.H., Carroll, R.J. & Ruppert, D. (1994) Fitting heteroscedastic regression models. *Journal of the American Statistical Association* 89: 100–116.
- Wettschereck, D., Aha, D.W. & Mohri, T. (1997) Review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artificial Intelligence Review* 11: 273–314.
- Williams, P.M. (1996) Using neural networks to model conditional multivariate densities. *Neural Computation* 8: 843–854.
- Wright, E.M. & Royston, P.A. (1997) Comparison of statistical methods for age-related reference intervals. *Journal of the Royal Statistical Society: Series A* 160: 47–69.
- Yau, P. & Kohn, R. (2003) Estimation and variable selection in nonparametric heteroscedastic regression. *Statistics and Computing* 13: 191–208.
- Yu, K. & Jones, M.C. (1998) Local linear quantile regression. *Journal of the American Statistical Association* 93: 228–237.
- Yu, K. & Jones, M.C. (2004) Likelihood-based local linear estimation of the conditional variance function. *Journal of American Statistical Association* 99: 139–144.
- Yu, K., Lu, Z. & Stander, J. (2003) Quantile regression: Applications and current research areas. *The Statistician* 52: 331–350.
- Zeileis, A. & Hornik, K. (2002) Testing for structural change in generalized linear models - Implementation in R and application. In: *Statistical Computing 2002*. Schloss Reinsburg, Germany. Cited February 22nd 2006 from <http://www.dkfz.de/biostatistics/Reinsburg2002/Beitraege/Zeileis-2002.pdf>.
- Zeileis, A., Leisch, F., Kleiber, C. & Hornik, K. (2005) Monitoring structural change in dynamic econometric models. *Journal of Applied Econometrics* 20: 99–121.

Original communications

- I Juutilainen I, Röning J, Myllykoski L (2003): *Modelling the Strength of Steel Plates Using Regression Analysis and Neural Networks*. In: Proceedings of International Conference on Computational Intelligence for Modelling, Control and Automation (CIMCA' 2003), Vienna, Austria, 681–691.
- II Juutilainen I, Röning J (2004): *Heteroscedastic Linear Models for Analysing Process Data*. WSEAS Transactions on Mathematics 2: 179–187.
- III Juutilainen I, Röning J (2004): *Modelling The Probability of Rejection in a Qualification Test Based on Process Data*. In: Proceedings of 16th Symposium of IASC (COMPSTAT 2004). Prague, Czech Republic, 1271–1278, ©Physica-Verlag Heidelberg 2004. Reprinted with kind permission of Springer Science and Business Media.
- IV Juutilainen I, Röning J (2006): *Planning of Strength Margins Using Joint Modelling of Mean and Dispersion*. Materials and Manufacturing Processes 21: 367 – 373. Copyright 2006. Reproduced by permission of Taylor & Francis Group, LLC., <http://www.taylorandfrancis.com>.
- V Juutilainen I, Röning J (2005): *A Comparison of Methods for Joint Modelling of Mean and Dispersion*. In: Proceedings of International Symposium on Applied Stochastic Models and Data Analysis (ASMDA 2005), Brest, France, 1499–1506.
- VI Juutilainen I, Röning J (forthcoming) *A Method for Measuring Distance from a Training Data Set*. Communications in Statistics (accepted for publication). Preprinted by permission of Taylor & Francis Group, LLC., <http://www.taylorandfrancis.com>.
- VII Juutilainen I, Röning J, Laurinen P (2005): *A Study on the Differences in the Interpolation Capabilities of Models*. In: Proceedings of IEEE Mid-Summer Workshop on Soft Computing in Industrial Applications (SMCia/05), Espoo, Finland, 202–207. Copyright ©2005 IEEE. Reprinted from SMCia/05 Workshop Proceedings.
- VIII Juutilainen I, Röning J (2006): *Adaptive Modelling of Conditional Variance Function*. In: Proceedings of 17th Symposium of IASC (COMPSTAT 2006), Rome, Italy, 1517–1524. ©Physica-Verlag Heidelberg 2006. Reprinted with kind permission of Springer Science and Business Media.

Original publications are not included in the electronic version of the dissertation.

245. Özer-Kemppainen, Özlem (2006) Alternative housing environments for the elderly in the information society. The Finnish experience
246. Laurinen, Perttu (2006) A top-down approach for creating and implementing data mining solutions
247. Jortama, Timo (2006) A self-assessment based method for post-completion audits in paper production line investment projects
248. Remes, Janne (2006) The development of laser chemical vapor deposition and focused ion beam methods for prototype integrated circuit modification
249. Kinnunen, Matti (2006) Comparison of optical coherence tomography, the pulsed photoacoustic technique, and the time-of-flight technique in glucose measurements *in vitro*
250. Iskanius, Päivi (2006) An agile supply chain for a project-oriented steel product network
251. Rantanen, Rami (2006) Modelling and control of cooking degree in conventional and modified continuous pulping processes
252. Koskiaho, Jari (2006) Retention performance and hydraulic design of constructed wetlands treating runoff waters from arable land
253. Koskinen, Miika (2006) Automatic assessment of functional suppression of the central nervous system due to propofol anesthetic infusion. From EEG phenomena to a quantitative index
254. Heino, Jyrki (2006) Harjavallan Suurteollisuuspuisto teollisen ekosysteemin esimerkkinä kehitettäessä hiiliteräksen ympäristömyönteisyyttä
255. Gebus, Sébastien (2006) Knowledge-based decision support systems for production optimization and quality improvement in the electronics industry
256. Alarousu, Erkki (2006) Low coherence interferometry and optical coherence tomography in paper measurements
257. Leppäkoski, Kimmo (2006) Utilisation of non-linear modelling methods in flue-gas oxygen-content control
258. Juutilainen, Ilmari (2006) Modelling of conditional variance and uncertainty using industrial process data
259. Sorvoja, Hannu (2006) Noninvasive blood pressure pulse detection and blood pressure determination

Book orders:
OULU UNIVERSITY PRESS
P.O. Box 8200, FI-90014
University of Oulu, Finland

Distributed by
OULU UNIVERSITY LIBRARY
P.O. Box 7500, FI-90014
University of Oulu, Finland

S E R I E S E D I T O R S

A
SCIENTIAE RERUM NATURALIUM
Professor Mikko Siponen

B
HUMANIORA
Professor Harri Mantila

C
TECHNICA
Professor Juha Kostamovaara

D
MEDICA
Professor Olli Vuolteenaho

E
SCIENTIAE RERUM SOCIALIUM
Senior Assistant Timo Latomaa

E
SCRIPTA ACADEMICA
Communications Officer Elna Stjerna

G
OECONOMICA
Senior Lecturer Seppo Eriksson

EDITOR IN CHIEF
Professor Olli Vuolteenaho

EDITORIAL SECRETARY
Publications Editor Kirsti Nurkkala

ISBN 951-42-8261-2 (Paperback)

ISBN 951-42-8262-0 (PDF)

ISSN 0355-3213 (Print)

ISSN 1796-2226 (Online)

