

# Modelling Conditional Variance Function in Industrial Data: A Case Study

Ilmari Juutilainen<sup>\*</sup>, Juha Röning

*Computer Engineering Laboratory, PO BOX 4500, 90014 University of Oulu,  
Finland*

---

## Abstract

We study the suitability of different modelling methods for joint prediction of mean and variance based on large data sets. We review the approaches to the modelling of conditional variance function that are capable of handling a problem where conditional variance depends on about 10 explanatory variables and training data set consists of 100000 observations. We present a promising approach for neural network modelling of mean and dispersion. We compare different approaches in predicting the mechanical properties of steel in two case data sets collected from the production line of a steel plate mill. As a conclusion we give some recommendations concerning the modelling of conditional variance in large data sets.

*Key words:* Joint modelling of mean and dispersion, Variance estimation, Heteroscedasticity

---

## 1 Introduction

Joint modelling of mean and dispersion is becoming a more and more commonly used approach in statistical prediction. In many real problems, not only the mean but also the variance and even other moments of the response variable depend on the explanatory variables. In these cases, dispersion modelling is needed to predict the conditional distribution realistically. When a prediction model is utilized in practice, the inference is based on the predicted conditional distribution. Incorporating a model for conditional variance takes into account heteroscedasticity and can often yield benefits in practice. Also in

---

<sup>\*</sup> Corresponding author, tel +358-8-5532994, fax +358-8-5532612  
*Email address:* ilmari.juutilainen@ee.oulu.fi, juha.roning@ee.oulu.fi.  
*URL:* <http://www.ee.oulu.fi/research/isg>.

optimization, an approach taking into account both conditional mean and conditional variance has become common [14]. A model for conditional variance has often been employed to make mean model estimation more efficient [1]. In many applications, including industrial quality improvement experiments [9], the variance function itself has been the focus of the interest.

A single observation does not give any information about variance, and many more observations are needed to estimate a model for variance than a model for mean. Although modelling of conditional variance has been applied in many fields, applications to large data sets seem to be lacking. Several possible response variables, model frameworks and estimation methods have been proposed for modelling of conditional variance. The different methods for modelling of conditional variance have not been compared to each other, and their prediction abilities and suitability to large data sets are rather unclear.

This paper gives insight into different methods which have been proposed for joint prediction of mean dispersion and which can be applicable in large data sets. In section 2, we review the literature of mean and variance modelling and describe the methods we use in our comparison. The case study of section 3 compares the methods for their accuracy in predicting the mean and variance of the mechanical properties of steel plates using two industrial data sets with about 25 explanatory variables and 100000 observations. Finally, section 4 draws up conclusions and recommendations that arise from the study.

## 2 Joint modelling of mean and dispersion

We denote the observations of the response variable with  $Y = (y_1, y_2, \dots, y_N)^T$  and let  $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$  denote the values of the  $p$  explanatory variables of the  $i$ th observation. We assume that  $y_i$  are normally, independently distributed  $y_i \sim N(\mu_i, \sigma_i^2)$ , with both the mean  $\mu_i = \mu(x_i)$  and the variance  $\sigma_i^2 = \sigma^2(x_i)$  depending on the explanatory variables.

Joint modelling of mean and dispersion can be divided into two tasks: estimation of the mean function and estimation of the variance function [1]. In the iterative estimation method, the mean function is estimated with weighted least squares by keeping the variance model fixed and by using weights proportional to the inverses of the predicted variances. The variance function is then estimated by keeping the mean model fixed [1]. There has been controversy as to the number of iterations needed. Sometimes good results have been obtained using only one iteration [26], and two iterations have often been considered best [1]. Estimating both the mean parameters and the variance parameters at the same time using for example maximum likelihood is computationally more complex and approximately same results can be achieved with

the iterative method [1]. Thus the described iterative method is employed in joint estimation of mean and variance functions for all models examined in the performed comparative study.

### 2.1 The choice of response for dispersion modelling

The selection of the response for dispersion model fitting is not obvious because direct measurements of variance cannot be made without replication. Natural measurement of the variance is provided by the squared residual  $\hat{\varepsilon}_i^2 = (y_i - \hat{\mu}(x_i))^2$ . One advantage of using response  $\hat{\varepsilon}_i^2$  is that when the variance is not functionally dependent on the mean, the described iterative estimation of mean and variance (pseudo-likelihood) converges to the maximum likelihood fit [1].

Fitting of the mean model biases the estimation of the variance function because the fitted model always adapts itself to the estimation data. This bias can be corrected by modifying the response: for example, in a regression context the response  $\hat{\varepsilon}_i^2/(1 - h_{ii})$ , where  $h_{ii}$  are the diagonal elements of the hat matrix, corresponds approximately to restricted maximum likelihood (REML) estimation [20]. If the fit can be expressed using a smoother matrix,  $\hat{Y} = SY$ , the expectation of a squared residual in the estimation data is  $E\hat{\varepsilon}_i^2 = \sigma_i^2 - 2S_{ii}\sigma_i^2 + \sum_{j=1}^N S_{ij}^2\sigma_j^2 + (\mu_i - E\hat{\mu}_i)^2$  [19]. Defining  $\Delta = \text{diag}(2S - SS^T)$  and assuming the fit to be conditionally unbiased, the result motivates the  $\Delta$ -corrected response

$$r_i = \hat{\varepsilon}_i^2/(1 - \Delta_i). \quad (1)$$

The  $\Delta$ -corrected response  $r_i$  has been proposed to be used for correcting the bias arising from the mean function estimation [19].

In addition to squared residuals, also absolute residuals  $|\hat{\varepsilon}_i|$  and logarithms of squared residuals  $\log \hat{\varepsilon}_i^2$  have been used as the response in variance function estimation. These alternatives are more robust to outliers [2]. If  $\varepsilon_i \sim N(0, \sigma_i^2)$  then  $E(\log \varepsilon_i^2) = \Psi(1/2) - \log(1/2) + \log \sigma_i^2 \approx -1.27 + \log \sigma_i^2$  where  $\Psi(\cdot)$  is the digamma function [7]. Thus when the dispersion model is fitted using the response  $\log \hat{\varepsilon}_i^2$ , it is suggested to predict the conditional variance with

$$\hat{\sigma}_i^2 = \exp(1.27 + \widehat{\log \varepsilon_i^2}) \quad (2)$$

where  $\widehat{\log \varepsilon_i^2}$  is the prediction of the fitted model. If  $\varepsilon_i \sim N(0, \sigma_i^2)$  then  $E(|\varepsilon_i|) = \sqrt{2\sigma_i^2/\pi}$ . Thus, it is suggested that when the dispersion model is

fitted using the response  $|\widehat{\varepsilon}_i|$ , the conditional variance is predicted with

$$\widehat{\sigma}_i^2 = \frac{\pi}{2} \widehat{|\varepsilon_i|}^2 \quad (3)$$

where  $\widehat{|\varepsilon_i|}$  is the prediction of the fitted model.

There are some methods for variance function estimation where a model for the mean does not need to be specified before estimating the variance function. When a model  $v(x_i)$  is fitted to predict the squared response  $y_i^2$ , the variance can be estimated with  $\widehat{v}(x_i) - \widehat{\mu}(x_i)^2$  [10]. Another possibility is to use squared pseudo-residuals as the response in variance function fitting [12]. A pseudo-residual  $p_i$  means a meaningfully weighted sum of neighbouring observations  $p_i = \sum_{j \in \mathcal{N}_i} w_j Y_j$ ,  $\sum w_j = 0$ ,  $\sum w_j^2 = 1$  [12]. These methods are, however, not very competitive with the approach where a model for mean is used [9] and are therefore not included in our comparative study.

## 2.2 The choice of loss function for dispersion modelling

The learning method, i.e. model type and estimation method, is another major selection problem in dispersion modelling. In principle, most of the learning methods can be used for modelling dispersion. If the residuals are normally distributed,  $\varepsilon_i \sim N(0, \sigma_i^2)$ , then the squared residuals are gamma distributed,  $\varepsilon_i^2 \sim \text{Gamma}(\sigma_i^2, 2)$ , and the fitting can be based on a gamma log-likelihood.

Models based on squared residuals are sensitive to outliers [2] and more robust methods for variance function estimation can be considered [1]. The variances of absolute residuals are proportional to their squared expectations  $\text{var}(|\varepsilon_i|) = E\varepsilon_i^2 - (E|\varepsilon_i|)^2 = (\pi/2 - 1)(E|\varepsilon_i|)^2$  and therefore the model for variance can be estimated by iteratively reweighted least squares using the response  $|\widehat{\varepsilon}_i|$  and weights  $\widehat{|\varepsilon_i|}^{-2}$  [1]. Here  $\widehat{|\varepsilon_i|}$  is the predicted response from previous iterations. Let us write  $\varepsilon_i^2 = u_i^2 \sigma_i^2$ ,  $u_i \sim N(0, 1)$ . Now the variance of the logarithms of squared residuals  $\text{var}(\log \varepsilon_i^2) = \text{var}(\log \sigma_i^2 + \log u_i^2) = \text{var}(\log u_i^2)$  is constant. Thus a model using the response  $\log \widehat{\varepsilon}_i^2$  is estimated using ordinary least squared [7].

The loss function can be penalized by model complexity using the techniques familiar from mean model estimation [17] [3]. Also, robust methods like iteratively weighted bounded-influence estimation can be applied to variance function modelling [6] [16].

### 2.3 The choice of modelled variance function structure

Because the estimation of the mean function and the variance function can be separated to its own tasks, a variety of statistical modelling frameworks can be used for dispersion modelling. We emphasize that the method used for the modelling of variance can be selected independently of the method used for the modelling of mean. Penalized estimation methods can be used for the estimation of the variance as well as for the estimation of the mean. In principle, all model frameworks able to handle large data sets could have been used for the modelling of variance in our data.

For the comparison we selected three methods most commonly discussed in the literature and a neural network model as an example of the other possible alternatives. The selected methods are discussed below in more detail. A part of the training data set was set aside as a validation data set to select the metaparameters of the methods (smoothing parameters, bandwidths, number of hidden neurons, etc).

### 2.4 Heteroscedastic regression

Heteroscedastic regression is a simple method, which can be easily applied to large data sets [20]. In heteroscedastic regression both mean and variance depend on inputs via a parametric function  $\mu_i = \mu(\beta, x_i)$ ,  $\sigma_i^2 = \sigma^2(\tau, x_i)$ . The most commonly used heteroscedastic regression model is

$$\begin{aligned} f(\mu_i) &= \tilde{z}_i^T \beta \\ g(\sigma_i^2) &= z_i^T \tau \end{aligned} \tag{4}$$

where the link functions  $f$  and  $g$  define the relationship between the linear predictors and the mean and variance, respectively. The input vectors  $\tilde{z}_i$  and  $z_i$  include transformations and product terms of the original explanatory variables to allow non-linear effects and interactions between the explanatory variables.

The most commonly used link function for variance is log-link  $g(x) = \log(x)$ . Log-link assures the positivity of predicted variance, which is an important property [15]. Log-link implies the assumption that the explanatory variables have a multiplicative effect on the conditional variance which is not always reasonable. As an alternative for log-link we successfully tested square root link  $g(x) = \sqrt{x}$  and identity link  $g(x) = x$ . These choices for link function concern all the models we discuss for variance modelling.

In our comparative study (Section 3) we used the identity link  $f(x) = x$  for the mean. We made the model selection manually based on the prediction accuracy in the validation data set. The selected mean models included about 110 terms for strength and 40 terms for elongation. The selected dispersion models had about 25 terms. The included terms were carefully selected transformations of the original input variables.

### 2.5 Mean and dispersion additive models

Generalized additive models are known to be able to handle large data sets pretty well [8]. Mean and dispersion additive models  $f(\mu_i) = \sum_{j=1}^p h_j(x_{ij})$ ,  $g(\sigma_i^2) = \sum_{j=1}^p k_j(x_{ij})$  have been proposed for joint modelling of mean and dispersion [17]. Splines are often used as basis functions  $h_j(\cdot)$  and  $k_j(\cdot)$  and the approach can be extended to include multivariate splines [24].

In our comparative study we used a mean and dispersion additive model that allows two-way interactions

$$\begin{aligned} f(\mu_i) &= \sum_{j=1}^p h_j(x_{ij}) + \sum_{j=1}^p \sum_{k=1}^p h_{jk}(x_{ik}, x_{ij}) \\ g(\sigma_i^2) &= \sum_{j=1}^p k_j(x_{ij}) + \sum_{j=1}^p \sum_{k=1}^p k_{jk}(x_{ik}, x_{ij}). \end{aligned} \quad (5)$$

The functions  $h_j(\cdot)$  and  $k_j(\cdot)$  were linear functions or univariate penalized regression splines with 10 knots. The 10 knots of each input variable were approximately equally spaced. The functions  $h_{ij}(\cdot)$  and  $k_{ij}(\cdot)$  were zero functions or two-dimensional penalized regression splines. The two-dimensional splines employed 10-dimensional basis that was reduced from the  $10 \times 10$  grid of knots using Wood's truncated eigenbasis approximation [23]. The smoothing parameters were estimated during iteratively reweighted least squares fitting by minimizing a generalized cross-validation criterion (GCV) [22]. The univariate spline terms of the models were selected to include all the input variables for which a spline transformation makes a significant improvement to the model fit. Then the non-zero interaction terms of the models (about 50 in the mean models and 15 in the variance models) were selected using a forward-stagewise algorithm, which goes through all possible two-dimensional splines and expands the model by adding the terms that improve most the model's log-likelihood in a validation data set. The details of the modelling methodology (like number of knots) were decided by taking account the model performance in validation data and the computational restrictions. The excessive growth of computational demands gave restrictions to the number of used knots, the degrees of freedom of truncated spline bases and the number of performed

iteration stages in the model selection algorithm.

### 2.6 Local linear regression for mean and dispersion

In local methods, the whole set of estimation data serves as the model, and prediction is based on the nearest neighbours of the query point. Local methods like kernel averaging [21] and local linear regression [19] have been proposed for joint modelling of mean and dispersion. The method was improved by proposing that the variance is estimated by minimizing the local gamma likelihood instead of the sum of squares [26]:

$$\begin{aligned}
\hat{\mu}_i &= \hat{a} \\
(\hat{a}, \hat{\beta}) &= \arg \min_{a, \beta} \sum_{j=1}^N (y_j - a - (x_j - x_i)^\top \beta)^2 K_1\left(\frac{\|x_j - x_i\|}{h_{1j}}\right) \\
\hat{\sigma}_i^2 &= g^{-1}(\hat{c}) \\
(\hat{c}, \hat{\tau}) &= \arg \min_{c, \tau} \sum_{j=1}^N \left[ \frac{\varepsilon_j^2}{g^{-1}(c + (x_j - x_i)^\top \tau)} + \log g^{-1}(c + (x_j - x_i)^\top \tau) \right] \\
&\quad \cdot K_2\left(\frac{\|x_j - x_i\|}{h_{2j}}\right). \tag{6}
\end{aligned}$$

Here,  $K_1$  and  $K_2$  are kernel functions and  $h_{1j}$  and  $h_{2j}$  are bandwidths. The suitability of local methods to high-dimensional problems has been questioned, because the distances between the neighbouring points grow rapidly with the number of dimensions and the local neighbourhood becomes too sparse [8].

In our comparative study we used the local likelihood method (Eq. 6) with the Epanechnikov quadratic kernel  $K_\lambda(x_0, x) = \frac{3}{4}(1 - |x - x_0|/\lambda)^2 I(|x - x_0| < \lambda)$ . We used a simple adaptive bandwidth, which gives positive weights to a constant number (few thousands) of estimation data instances. The model selection task was simplified to the selection of a suitable number of neighbours to be used in prediction, which was decided on the basis of performance in validation data.

### 2.7 Neural network modelling of mean and dispersion

Neural networks are known as a flexible modelling method with good predictive performance in large data sets [8]. The idea of using neural networks for modelling conditional variance function is not new [13] [5], but the iterative neural network modelling of mean and dispersion we propose below seems to be a novel approach.

In our comparative study mean and variance were modelled using separate single-hidden-layer perceptron models with skip-layer connections

$$\begin{aligned}
 f(\mu_i) &= \beta_{00} + \sum_{k=1}^p x_{ik}\beta_{0k} + \sum_{j=1}^h \beta_j f_j \left( \beta_{j0} + \sum_{k=1}^p x_{ik}\beta_{jk} \right) \\
 g(\sigma_i^2) &= \tau_{00} + \sum_{k=1}^p x_{ik}\tau_{0k} + \sum_{j=1}^h \tau_j g_j \left( \tau_{j0} + \sum_{k=1}^p x_{ik}\tau_{jk} \right)
 \end{aligned} \tag{7}$$

where the activation functions  $f_j(\cdot)$  and  $g_j(\cdot)$  are logistic  $e^{-t}/(1 + e^{-t})$ . The variance model was fitted by maximizing the penalized gamma log-likelihood related to the squared residuals of the mean model. Squared penalty terms  $\lambda_\mu \sum_j (\beta_j^2 + \sum_k \beta_{jk}^2)$  and  $\lambda_\sigma \sum_j (\tau_j^2 + \sum_k \tau_{jk}^2)$  were employed in the estimation to prevent overfitting. Model selection consisted of selecting the number of hidden neurons  $h$  and selecting the regularization parameters  $\lambda_\mu$  and  $\lambda_\sigma$ . Different models were tested and the model that worked best in the validation data was selected. We ended up modeling variance using networks with 10-15 hidden neurons and to model mean using networks with about 30 hidden neurons.

### 2.8 Other models for the conditional variance

Reproducing kernels have recently become one of the most discussed learning methods. A reproducing kernel method has been developed for joint modelling of mean and variance [3]. The proposed model is estimated by iterating mean and variance model estimation. The algorithm is not suitable for large data sets because it utilizes fully dense kernel expansion: A modification that utilizes sparse kernel approximation would be needed.

Altogether, over one hundred of papers concerning the estimation of multivariate variance functions have been written. Our review includes the methods that are most usable for analyzing large data sets. For a more complete listing, see [11].

A step ahead from joint modelling of mean and dispersion are the methods which model the conditional distribution more exactly and take account also third and even higher moments of conditional distribution. Quantile regression [25], GAMLSS (Generalized Additive Models for Location, Scale and Shape) [18], conditional density estimation and LMS-method [4] have been commonly used. However, in the case of several explanatory variables, it may be difficult to gain information about the dependence between the explanatory variables and the higher moments of the conditional distribution of the response.

### 3 A comparative case study on dispersion modelling methods

We compared heteroscedastic regression (HetReg), mean and dispersion additive models (MADAM), local linear regression for mean and dispersion (LL-RMD) and neural network modelling of mean and dispersion (NNMMD) in a real data set. We studied also the effects of response variable, link function and fitting method to the modelling of variance. The data were collected from an industrial process of steel plate production. Three response variables were measured from finished products; tensile strength, yield strength and elongation, all being approximately normally distributed (Figure 1). Information about the reasons affecting the variance in the strength of steel can be utilized in process optimization [13] [11].

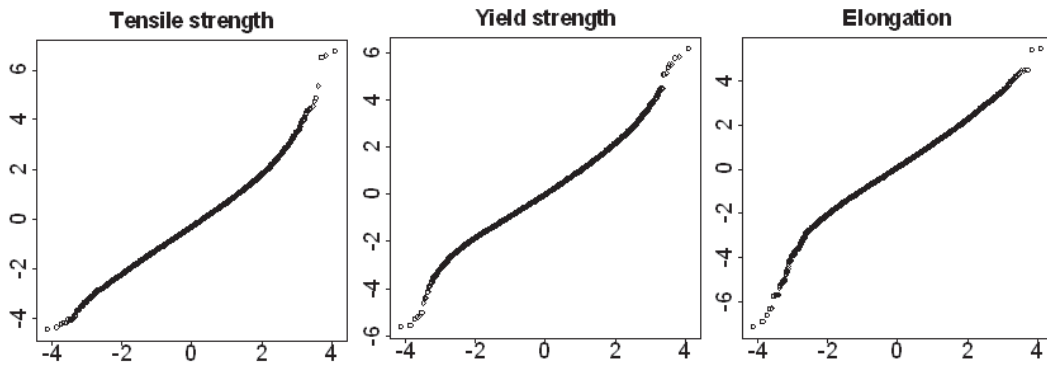


Fig. 1. The Q-Q plots of standardized residual of the test data.

#### 3.1 Data

The training data for tensile and yield strength consisted of 90 000 observations. In the modelling, 27 explanatory variables related to the steel plate production process and likely to have an effect on the responses were used. The explanatory variables were related to the concentrations of alloying elements, the thermomechanical treatments made during the process of production and the size and shape of the plate and the test specimen. In the modelling of variance, 12 of the explanatory variables with a likely effect on the conditional variance were used.

The 22 explanatory variables used in the modelling of elongation were mainly the same as used in the modelling of strength. The training data for elongation consisted of 80 000 observations. The number of input variables in the modelling of the variance of elongation was 13. For both strength and elongation, about 25 % of the training data was used as validation data based on which model terms and metaparameters were selected. Strength and elongation are

measured in a tensile test where a test bar is stretched apart until it breaks. Sometimes the test bar slips causing the results to be strongly erroneous. In our manual data pre-processing, about one hundred of strange observations were interpreted to be caused by a slipped test bar and were omitted from the analysis.

The test data set was collected from the production line after the training data set and consisted of 25 000 observations for strength and 40 000 observations for elongation. The prediction accuracies of the models were compared using the negative log-likelihood of the test data set under a Gaussian assumption. In the learning data set the predicted variance varied between 30 and 1500 for strength and between 1 and 20 for elongation. In the results for strength, the variance predictions smaller than 16 were transformed to 16. The predicted variance of strength was almost always above 16, but for some models there were a couple of near-zero variance predictions whose effect to log-likelihood was thousands of units so that without the truncation they would have dominated the results.

### 3.2 *Methods*

The fitting of models was accomplished using the iterative approach. First, the model for mean was fitted, and the variance model was then fitted based on the squared residuals from the mean model fit. In the optional second iteration, the mean model was weighted with the inverses of the predicted variances, and the variance model was fitted again. The parameters of the mean model were estimated via least squares. Four response variables were compared in the variance function estimation:  $\Delta$ -corrected squared residuals  $\varepsilon_i^2/(1 - \Delta_i)$ , squared residuals  $\varepsilon_i^2$ , absolute residuals  $|\varepsilon_i|$  and logarithms of squared residuals  $\log \varepsilon_i^2$ . The parameters of the variance model were estimated with the gamma log-likelihood (for  $\varepsilon_i^2/(1 - \Delta_i)$  and  $\varepsilon_i^2$ ), or least squares (for  $\varepsilon_i^2/(1 - \Delta_i)$  and  $\log \varepsilon_i^2$ ) or quasi-likelihood (for  $|\varepsilon_i|$ ). Penalized estimation were used for the models MADAM and NNMMMD. A linear link was used for the mean and a square root link or log link for the variance.

### 3.3 *Computational demands*

The computational requirements of modelling methods are a focus of interest when prediction is based on large data sets. We tested the computational needs using R software (<http://www.r-project.org/>) installed on a SunOS unix machine with 15 Gb of memory. The CPU power used in the computation was 900 MHz. R is known to be fast but to use memory inefficiently. The observed need for memory and computation time for fitting the model for strength

Table 1

The required computational resources for applying different methods to the strength of steel data.

	Fitting	Prediction	Model selection	Memory need (Mb)
HetReg	1 min	< 1 min	15 h	800 Mb
MADAM	70 min	< 1 min	12 h	3500 Mb
LLRMD	70 h	20 h	240 h	400 Mb
NNMMD	120 min	< 1 min	10 h	400 Mb

is shown in Table 1. The time needed to produce 25000 predictions for the test data set is also presented. We used a simple model selection practice for each case; the approximate computation times used by the model selection procedures are also presented in Table 1.

### 3.4 Results

We compared the prediction accuracy of joint modelling of mean and dispersion using the negative log-likelihood in the test data set  $T$

$$-\log\text{-lik} = \frac{1}{2} \sum_{i \in T} \ln 2\pi\hat{\sigma}_i^2 + \frac{1}{2} \sum_{i \in T} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\sigma}_i^2}. \quad (8)$$

It can be easily seen that the gamma log-likelihood of the dispersion model is equivalent to the likelihood of the whole model when the mean model is kept fixed. Thus, the comparison of dispersion models by keeping the mean model fixed can be based on the full likelihood. For the comparison of mean models, the root mean squared errors  $\text{rMSE} = \sqrt{\text{ave}(\hat{\varepsilon}_i^2)}$  are also presented.

Table 2 shows the achieved prediction accuracies of the different methods for joint modelling of mean and dispersion in the test data set. Neural network modelling of mean and dispersion predicts best the conditional distribution of the responses. The differences in the likelihood are strongly related to the differences in the accuracy of the mean model.

To compare the model frameworks of variance modelling, we fixed the mean models to the fitted neural network models and fitted the dispersion models using the squared residuals. The results are presented in Table 3. The results indicate that the choice of model type has a significant effect on the accuracy of prediction. The neural network model performed best also in this comparison.

Table 2

Prediction accuracy in the test data set.

model	<u>Tensile strength</u>		<u>Yield strength</u>		<u>Elongation</u>	
	rMSE	-log-lik	rMSE	-log-lik	rMSE	-log-lik
HetReg	9.25	95125	14.39	108399	1.94	73453
MADAM	9.67	95837	14.28	108172	1.73	71931
LLRMD	9.23	95468	14.09	107800	1.63	68789
NNMMD	8.95	94442	13.90	107482	1.58	68139

Table 3

The negative log-likelihoods (the smaller, the better) in the test data set when the mean model was kept fixed.

model	Tensile S.	Yield S.	Elongation
HetReg	94410	107646	68415
MADAM	94726	107623	71738
LLRMD	94593	107514	68582
NNMMD	94442	107482	68139

The basic method for fitting the dispersion model was to use the response  $\varepsilon_i^2/(1-\Delta_i)$  and the square root link function and to fit the model using gamma likelihood without iterating the mean model and variance model estimation. Some alternatives for this basic setting were tested: effects are presented in Table 4. If the parameters were estimated under a Gaussian likelihood instead of a gamma likelihood, the likelihood of the test data decreased significantly. The effect of a link function was moderate, and on average log-link and square root link worked equally well. The number of iterations in the joint modelling of mean and dispersion had a major but fluctuating effect on the results. Sometimes, the weighted estimation of the second iteration gave better results when measured using likelihood. The third iteration changed the results of the second iteration only slightly, and the differences in log-likelihood were about 10-30. The subsequent iterations had a very small effect on the results, the change in log-likelihood being about 1-4.

The results of comparison between different response variables of variance model fitting are presented in Table 5. Using the uncorrected squared residual  $\varepsilon^2$  had only a small effect on the results; prediction accuracy usually decreased. Using squared residuals as the response gave ostensibly better results than absolute residuals or logarithms of squared residuals. That was expected because the model goodness was measured using squared residual and models fitted using absolute residuals or especially the logarithms of squared residuals give, on average, smaller predicted variance. Thus, we can not conclude that squared residual would be the best choice for the response variable even in this data.

Table 4

The differences in test data log-likelihood between the standard fitting method and the alternatives. The plus sign means that the alternative gave better likelihood in the test data set.

model	<u>Tensile strength</u>			<u>Yield strength</u>			<u>Elongation</u>		
	LS	log	iter	LS	log	iter	LS	log	iter
HetReg	-56	-24	+61	-303	-6	+187	-380	+131	-306
MADAM	-2050	-375	+117	-643	+13	-665	-1546	+355	-79
LLRMD	-68	.	.	-73	.	.	-604	.	.
NNMMD	-350	+30	+251	-230	-185	-211	-348	+147	-160

LS: Model for variance estimated with least squares.

log: Log-link for the variance was used.

iter: The estimation was iterated twice.

Table 5

The differences in test data log-likelihood between the the alternative responses of variance model fitting. The minus sign means that the alternative gave worse likelihood in the test data set compared to the basic response  $\varepsilon_i^2/(1 - \Delta_i)$ .

model	<u>Tensile strength</u>			<u>Yield strength</u>			<u>Elongation</u>		
	$\varepsilon^2$	$ \varepsilon $	$\log \varepsilon^2$	$\varepsilon^2$	$ \varepsilon $	$\log \varepsilon^2$	$\varepsilon^2$	$ \varepsilon $	$\log \varepsilon^2$
HetReg	0	-150	-426	0	-107	-253	-7	-226	-502
MADAM	-36	-218	-604	+12	-109	-309	-32	-1310	-1245
LLRMD	-80	-57	-348	-27	-53	-128	-62	-573	-2591
NNMMD	.	-171	-242	.	-107	-200	.	-556	-1573

In neural network modelling, it was noticeable that a network with skip-layer connections was much better than an ordinary single-layer perceptron without skip-layer connections. For yield strength the difference in log-likelihood was 800 and for tensile strength 1300. The use of log-link for variance with local likelihood fitting caused convergence problems at several prediction points, and log-link could thus not be used. Constant bandwidth seemed to work poorly; the difference in log-likelihood with the adaptive bandwidth was about 2000. We did not try the iterated version of local linear modelling, because too time-consuming computations would have been needed.

## 4 Conclusion

The results on the predictive performance of the models in predicting the distribution of the mechanical properties of steel plates are presented. This case study is the first extensive comparison of the methods for modelling of conditional variance function in a real prediction problem. The study encourages us to give some recommendations, although the generalizability of results that are based on only three closely related data sets is not very wide. In this section, we conclude the arisen recommendations for the modelling of variance in large data sets.

Modification of the response in dispersion model fitting with  $\Delta$ -corrections to take into account the effect of estimating the mean model has a small effect on prediction. In a simple model with a large number of observations,  $\Delta$ -corrections have practically no impact, but the effect increases with the complexity of the model. We suggest that good results are obtained with an uncorrected response, but if the  $\Delta$ -corrections are easily available, the corrected response should be used.

The traditional log-link ensures the positivity of predicted variance, but it was not superior in our case study. Log-link implies that the explanatory variables have a multiplicative effect on variance, which is not necessarily a rational assumption. We suggest that a linear model for variance and a linear model for deviation should be also considered when selecting link function.

Iteration of mean model estimation and variance model estimation increases the computation time needed for model fitting. Our results agree well with the earlier results claiming that two iterations are needed, and the subsequent iterations have only a minor effect on the results. In our data sets, the first iteration also gave pretty good results. Our suggestion is to use two iterations.

We compared two loss functions in variance function estimation; least squares and gamma log-likelihood. Least squares yielded poor results, which was expected, as the distribution of squared residuals is far from normal. We did not try any robust estimation method which could sometimes be considered an alternative.

A wide variety of learning methods can be used for modelling dispersion, and the choice of the model type has a clear influence on the accuracy of the prediction. The results suggest that neural networks are included among the methods that provide a suitable model framework for joint prediction of mean and dispersion based on large data sets. The fitting of additive spline models to large data sets requires a huge amount of memory when multi-dimensional splines are employed. This makes additive splines difficult to use when there are interactions between the input variables. Local linear modelling is time-consuming,

and it may not be applicable to real-time applications. Heteroscedastic regression models are appropriate when simplicity and interpretability are required. We suggest that the model framework for the conditional variance should be chosen to best fulfill problem demands independently of the method used for the modelling the mean.

## Acknowledgements

I am grateful to Ruukki for providing data and partial funding for the research.

## References

- [1] R.J. Carroll, D. Ruppert, Transformation and Weighting in Regression, Chapman and Hall, New York, 1988.
- [2] M. Davidian, R.J. Carroll, Variance function estimation, *Journal of the American Statistical Association* 82 (1987) 1079-1091.
- [3] G.C. Cawley, N.L.C. Talbot, R.J. Foxall, S.R. Dorling, D.P. Mandic, Heteroscedastic kernel ridge regression, *Neurocomputing* 57 (2004) 105-124.
- [4] T.J. Cole, P.J. Green, Smoothing reference centile curves: the LMS method and penalized likelihood, *Statistics in Medicine* 11 (1992) 1305-1319.
- [5] S.R. Dorling, R.J. Foxall, D.P. Mandic, G.C. Cawley, Maximum-likelihood cost functions for neural network models of air quality data, *Atmospheric Environment* 37 (2003) 3435-3443.
- [6] D.M. Giltinan, R.J. Carroll, D. Ruppert, Some new estimation methods for weighted regression when there are possible outliers, *Technometrics* 28 (1986) 219-230.
- [7] A.C. Harvey, Estimating regression models with multiplicative heteroscedasticity, *Econometrica* 44 (1976) 461-465.
- [8] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer, New York, 2001.
- [9] S.K. Fan, A generalized global optimization algorithm for dual response systems, *Journal of Quality Technology* 32 (2000) 444-456.
- [10] W. Härdle, A.B. Tsybakov, Local polynomial estimators of the volatility function in nonparametric autoregression, *Journal of Econometrics*, 81 (1997) 223-242.
- [11] I. Juutilainen, *Modelling of Conditional Variance and Uncertainty using Industrial Process Data* (doctoral thesis), University of Oulu, Finland.

- [12] H-G. Müller, U. Stadtmüller, Estimation of heteroscedasticity in regression analysis, *The Annals of Statistics* 15 (1987) 610-625.
- [13] P. Myllykoski, A study on the causes of deviation in mechanical properties of thin steel sheets, *Journal of Materials Processing Technology* 79 (1998) 9-13.
- [14] K-J. Kim, D.K.J. Lin, Optimization of multiple responses considering both location and dispersion effects, *European Journal of Operational Research* 169 (2006) 133-145.
- [15] T.K. Mak, Modelling and estimating variances in regression, *Communications in Statistics - Theory and Methods* 31 (2002) 351-365.
- [16] R.A. Rigby, D.M. Stasinopoulos, Robust Fitting of an additive model for variance heterogeneity, In: R. Dutter, W. Grossmann (Eds.), *COMPSTAT 1994 - Proceedings in Computational Statistics*, Physica, Heidelberg, 1994 pp. 263-268.
- [17] R.A. Rigby, D.M. Stasinopoulos, A semi-parametric additive model for variance heterogeneity, *Statistics and Computing* 6 (1996) 57-65.
- [18] R.A. Rigby, D.M. Stasinopoulos, Generalized additive models for location, scale and shape, (with discussion), *Applied Statistics* 54 (2004) 507-554.
- [19] D. Ruppert, M.P. Wand, U. Holst, O. Hössjer, Local polynomial variance-function estimation, *Technometrics* 39 (1997) 262-273.
- [20] G.K. Smyth, A.V. Huele, A.P. Verbyla, Exact and approximate REML for heteroscedastic regression, *Statistical modelling* 1 (2001) 161-175.
- [21] U. Stadtmüller, A.B. Tsybakov, Nonparametric recursive variance function estimation, *Statistics* 27 (1995) 55-63.
- [22] S.N. Wood, mgcv: GAMs and generalized ridge regression for R. *R News* 1 (2001) 20-25.
- [23] S.N. Wood, Thin plate regression splines, *Journal of the Royal Statistical Society series B* 65 (2003) 95-114.
- [24] P. Yau, R. Kohn, Estimation and variable selection in nonparametric heteroscedastic regression, *Statistics and Computing* 13 (2003) 191-208.
- [25] K. Yu, Z. Lu, J. Stander, Quantile regression: applications and current research areas, *Journal of the Royal Statistical Society Series D* 52 (2003) 331-350.
- [26] K. Yu, M.C. Jones, Likelihood-based local linear estimation of the conditional variance function, *Journal of the American Statistical Association* 99 (2004) 139-144.