A Comparison of Methods for Joint Modelling of Mean and Dispersion

Ilmari Juutilainen and Juha Röning

Computer Engineering Laboratory
PO BOX 4500
90014 University of Oulu, Finland
(e-mail: ilmari.juutilainen@ee.oulu.fi, juha.roning@ee.oulu.fi)

Abstract. We considered the suitability of the methods for joint modelling of mean and dispersion for prediction based on large data sets under the assumption of normally distributed errors. Methods that seemed capable of handling a problem with 25 explanatory variables and 100000 observations were compared in predicting the strength of steel in a real data set collected from the production line of a steel plate mill. A neural network model for mean and dispersion gave the best prediction. The results indicate that neural networks are suitable for joint modelling of mean and dispersion in large data sets.
Keywords: Joint modelling of mean and dispersion, Heteroscedasticy.

1 Introduction

Joint modelling of mean and dispersion is a common problem in statistics. In many real problems, not only mean but also variance and even other moments of the conditional distribution of the response variable depend on the explanatory variables. In these cases, dispersion modelling is needed to predict the conditional distribution realistically. The variance model has often been employed to make mean model estimation more efficient. In many applications, including industrial quality improvement experiments, 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 joint modelling of mean and dispersion has been applied in many fields, applications to large data sets seem to be lacking. The different methods for joint modelling of mean and dispersion 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 the suitability of different methods proposed for joint prediction of mean dispersion based on large data sets. The models are compared for their accuracy in predicting the mean and variance of the strength of steel plates using a real data set with about 25 explanatory variables and 100000 observations.

2 Joint modelling of mean and dispersion

We denote the observations of the response variable with $Y = (y_1, y_2, \ldots, y_N)^{\mathrm{T}}$ and let $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ denote the values of the p explanatory variables of the ith observation. We assume that y_i are normally, independently distributed $y_i \mathrm{sim} N(\mu_i, \sigma_i^2)$, with both the mean $\mu_i(x_i)$ and the variance $\sigma_i^2(x_i)$ 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 [Carroll and Ruppert, 1988]. 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 [Carroll and Ruppert, 1988]. There has been controversy as to the number of iterations needed. Sometimes good results have been obtained using only one iteration [Yu and Jones, 2004], and two iterations have often been considered best [Carroll and Ruppert, 1988]. Simple models can be estimated without iteration using full maximum likelihood or restricted maximum likelihood (REML).

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 $\hat{\varepsilon}_i^2 = (y_i - \widehat{\mu}(x_i))^2$. 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 $\varepsilon_i^2/(1 - h_{ii})$, where h_{ii} are the diagonal elements of the hat matrix, corresponds to the REML estimation and leads to unbiased fitting [Smyth *et al.*, 2001]. If the fit can be expressed using a smoother matrix, $\widehat{Y} = SY$, the expectation of a squared residual in the estimation data is $E\hat{\varepsilon}_i^2 = \sigma_i^2 - 2S_{ii}\sigma_i^2 + \sum_{j=1}^{N} S_{ij}^2\sigma_j^2 + (\mu_i - E\widehat{\mu}_i)^2$ [Ruppert *et al.*, 1997]. Defining $\Delta = \mathrm{diag}(2S - SS^{\mathrm{T}})$ and assuming the fit to be conditionally unbiased, the result motivates the Δ-corrected response $r_i = \hat{\varepsilon}_i^2/(1 - \Delta_i)$.

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, $\varepsilon_i \mathrm{sim} N(0, \sigma_i^2)$, then the squared residuals are gamma distributed, $\varepsilon_i^2 \mathrm{simGamma}(\sigma_i^2, 2)$, and the fitting can be based on gamma log-likelihood. For most other possible responses (e.g. $|e_i|$ or $\log|e_i|$) no such helpful result is available, and the least squares method has been commonly used.

3 Methods

Heteroscedastic regression (HetReg), mean and dispersion additive models (MADAM), local linear regression for mean and dispersion (LLRMD) and

neural network modelling of mean and dispersion (NNMMD) were compared in a real data set. The estimation data were collected from an industrial process of steel plate production and consisted of 90 000 observations. Two response variables were measured from finished products; tensile strength and yield strength, both being approximately normally distributed. 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. [Myllykoski, 1998] has studied the reasons affecting the variance in the strength of thin steel sheets.

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 corrected or uncorrected 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 with the least squares, and the parameters of the variance model were estimated with the gamma log-likelihood or least squares. For the models MADAM and NNMMD the likelihoods were penalised. A linear link was used for the mean and a square root link or log link for the variance.

The test data set was collected from the production line after the training data set and consisted of 25 000 observations. The prediction accuracies of the models were compared using the negative log-likelihood of the test data set under a gaussian assumption. Variance predictions smaller than 16 (including negative predictions) were transformed to 16; otherwise, single bad predictions could have blurred the results.

Heteroscedastic linear regression is a simple method, which can be easily applied to large data sets [Smyth *et al.*, 2001]. We used a heteroscedastic regression model of the form

$$\begin{aligned} f(\mu_i) &= \tilde{z}_i^{\mathrm{T}}\beta \\ g(\sigma_i^2) &= z_i^{\mathrm{T}}\tau \end{aligned} \tag{1}$$

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. We made the model selection manually based on the prediction accuracy in the validation data set. The selected mean models included about 110 terms and the dispersion models about 25 terms. The model estimation was carried out using the iterative REML of [Smyth *et al.*, 2001].

Generalised additive models are known to be able to handle large data sets pretty well [Hastie *et al.*, 2001]. [Rigby and Stasinopoulos, 1996] pro-

posed mean and dispersion additive models for joint modelling of mean and dispersion. We used an additive model resembling the model of [Yau and Kohn, 2003] and allowing two-way interactions

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

The functions $h_j(\cdot)$ and $k_j(\cdot)$ were linear functions or univariate penalised regression splines with 10 knots. The functions $h_{ij}(\cdot)$ and $k_{ij}(\cdot)$ were zero functions or two-dimensional penalised regression splines with 10 knots selected out of 100 candidates. The estimation of the smoothing parameters of the different terms was accomplished using generalised cross-validation criteria. The non-zero terms of the models (about 50 in the mean models and 15 in the variance models) were selected using a simple algorithm, which expands the model by adding terms that improve the model's performance significantly in a validation data set.

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 linear regression was proposed for joint modelling of mean and dispersion by [Ruppert *et al.*, 1997]. [Yu and Jones, 2004] improved the method by proposing that the variance is estimated by minimising the local gamma likelihood instead of the sum of squares. They also used a link function $g(t) = \log(t)$ for variance in local estimation, leading to

$$\begin{aligned} \widehat{\mu}_i &= \widehat{a} \\ (\widehat{a}, \widehat{\beta}) &= \arg\min_{a,\beta} \sum_{j=1}^{N} (y_j - a - (x_j - x_i)^{\mathrm{T}}\beta)^2 K_1\Big(\frac{||x_j - x_i||}{h_1}\Big) \\ \widehat{\sigma}_i^2 &= g^{-1}(\widehat{c}) \\ (\widehat{c}, \widehat{\tau}) &= \arg\min_{c,\tau} \sum_{j=1}^{N} \Big[\frac{\varepsilon_j^2}{g^{-1}(c + (x_j - x_i)^{\mathrm{T}}\tau)} + \log g^{-1}(c + (x_j - x_i)^{\mathrm{T}}\tau)\Big] \\ &\qquad \cdot K_2\Big(\frac{||x_j - x_i||}{h_2}\Big). \end{aligned} \tag{3}$$

Here, K_1 and K_2 are kernel functions and the bandwidths h_1 and h_2 are chosen independently. 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 [Hastie *et al.*, 2001]. We used the local likelihood method of [Yu and Jones, 2004] with the Epanechnikov quadratic kernel $K_\lambda(x_0, x) = \frac{3}{4}(1 - |x - x_0|/\lambda)^2 I(|x - x_0| < 1)$. A simple adaptive bandwidth, which gives positive weights to a constant number (few thousands) of estimation data instances, was used. 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.

Neural networks are known as a flexible modelling method with good predictive performance in large data sets [Hastie *et al.*, 2001]. We fitted neural network models for mean and dispersion. The idea is not completely new, see [Myllykoski, 1998]. We used single-layer perceptron model with skip-layer connections of the form

$$f(\mu_i) = x_i^{\mathrm{T}}\beta + \sum_{j=1}^{h} f_j(x_i^{\mathrm{T}}\beta_j)$$
$$g(\sigma_i^2) = x_i^{\mathrm{T}}\tau + \sum_{j=1}^{h} g_j(x_i^{\mathrm{T}}\tau_j) \tag{4}$$

where the activation functions $f_j(\cdot)$ and $g_j(\cdot)$ are logistic $e^{-t}/(1+e^{-t})$. We fitted the variance model by maximising the penalised gamma log-likelihood related to squared residuals of the mean model. Model selection consisted of selecting the number of hidden neurons h and selecting the smoothing parameter. Different models were tested and the model that worked best in the validation data was selected. We modelled variance using single-layer perceptrons with 10 and 15 hidden neurons.

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

$$\text{-log-lik} = \frac{1}{2}\sum_{i\in T} \ln 2\pi\widehat{\sigma}_i^2 + \frac{1}{2}\sum_{i\in T}\frac{(y_i-\widehat{\mu}_i)^2}{\widehat{\sigma}_i^2}. \tag{5}$$

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 $\text{rMSE} = \sqrt{\text{ave}(\widehat{\epsilon}^2)}$ are also presented.

Table 1 shows the achieved prediction accuracies of the different methods for joint modelling of mean and dispersion in the test data set. To compare especially the dispersion models, 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 2.

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 the basic setting were tested: effects are

model	Tensile strength		Yield strength	
	rMSE	-log-lik	rMSE	-log-lik
HetReg	9.25	95125	14.39	108399
MADAM	9.67	95837	14.28	108172
LLRMD	9.23	95468	14.09	107800
NNMMD	8.95	94442	13.90	107482

Table 1. Prediction accuracy in the test data set.

model	Tensile S.	Yield S.
HetReg	94410	107646
MADAM	94726	107623
LLRMD	94593	107514
NNMMD	94442	107482

Table 2. The negative log-likelihoods (the smaller, the better) in the test data set when the mean model was kept fixed.

model	Tensile strength				Yield strength			
	ε^2	gaussian	log-link	weighted	ε^2	gaussian	log-link	weighted
HetReg	0	-56	-24	+61	0	-303	-6	+187
MADAM	-36	-2050	-375	+117	+12	-643	+13	-665
LLRMD	-80	-68	·	·	-27	-73	·	·
NNMMD	·	-350	+30	+251	·	-230	-185	-211

Table 3. 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 reduced test data set.

presented in Table 3. Using the response e^2 had only a small effect on the results; prediction accuracy usually decreased. If the parameters were estimated under gaussian likelihood instead of gamma likelihood, the likelihood of the test data decreased significantly. The effect of a link function was moderate, in most cases log-link for the variance function gave worse results. The number of iterations in the joint modelling of mean and dispersion had a major but fluctuating effect on the results. Usually, the weighted estimation of the second iteration gave better results when measured using likelihood but worse results when rMSE was used. The differences in rMSE were 0, -0.10 and -0.02 for tensile strength and +0.04, -0.57 and -0.10 for yield strength (in the same order as in Table 3). The third iteration changed the results of the second iteration only slightly, and the differences in log-likelihood were about 10-20. The subsequent iterations had a very small effect on the results, the change in log-likelihood being about 1-4.

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 weighted version of local linear modelling, because too large computations would have been needed.

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 are shown in Table 4. The time needed to produce 25000 predictions for the test data set is also presented. We used a simple model selection algorithm for each case; the approximate computation times used by the model selection procedures are also presented in Table 4.

	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

Table 4. The required computational resources for applying different methods to the strength of steel data.

5 Discussion

The results on the predictive performance of the models in predicting the distribution of the strength of steel plates are presented. This is the first extensive comparison of the methods for joint modelling of mean dispersion in a real prediction problem.

Modification of the response in dispersion model fitting with Δ-corrections to take into account the effect of estimating the mean model has a small effect on prediction. In heteroscedastic regression with a large number of observations, Δ-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 Δ-corrections are easily available, the corrected response should be used.

The traditional log-link ensures the positivity of predicted variance, but it did not perform very well 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 set, 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.

A wide variety of learning methods can be used for modelling dispersion, and the choice of the model type has a great 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, which makes them difficult to use. 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.

References

[Carroll and Ruppert, 1988]R.J. Carroll and D. Ruppert. *Transformation and Weighting in Regression*. Chapman and Hall, New York, 1988.

[Hastie *et al.*, 2001]T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

[Myllykoski, 1998]P Myllykoski. A study on the causes of deviation in mechanical properties of thin steel sheets. *Journal of Materials Processing Technology*, pages 9–13, 1998.

[Rigby and Stasinopoulos, 1996]R.A. Rigby and D.M. Stasinopoulos. A semi-parametric additive model for variance heterogeneity. *Statistics and Computing*, pages 57–65, 1996.

[Ruppert *et al.*, 1997]D. Ruppert, M.P. Wand, U. Holst, and O. Hössjer. Local polynomial variance-function estimation. *Technometrics*, pages 262–273, 1997.

[Smyth *et al.*, 2001]G.K. Smyth, A.V. Huele, and A.P. Verbyla. Exact and approximate reml for heteroscedastic regression. *Statistical Modelling*, pages 161–175, 2001.

[Yau and Kohn, 2003]P Yau and R Kohn. Estimation and variable selection in nonparametric heteroscedastic regression. *Statistics and Computing*, pages 191–208, 2003.

[Yu and Jones, 2004]K. Yu and M.C. Jones. Likelihood-based local linear estimation of the conditional variance function. *Journal of the American Statistical Association*, pages 139–144, 2004.