Response Surface Methodology for Optimizing Hyper Parameters

I. Czogiel * K. Luebke C. Weihs

September 2005

Universität Dortmund
Fachbereich Statistik

Abstract

The performance of an algorithm often largely depends on some hyper parameter which should be optimized before its usage. Since most conventional optimization methods suffer from some drawbacks, we developed an alternative way to find the best hyper parameter values. Contrary to the well known procedures, the new optimization algorithm is based on statistical methods since it uses a combination of Linear Mixed Effect Models and Response Surface Methodology techniques. In particular, the Method of Steepest Ascent which is well known for the case of an Ordinary Least Squares setting and a linear response surface has been generalized to be applicable for repeated measurements situations and for response surfaces of order $o \leq 2$.

Keywords: repeated measurements, Random Intercepts Model, deterministic error terms, Method of Steepest Ascent, Support Vector Machine

*e-mail: czogiel@statistik.uni-dortmund.de

1 Introduction

It is a common situation that a (statistical) algorithm contains some hyper parameters the values of which are to be chosen by the operator before its usage. Since the goodness of a resulting output often largely depends on this choice, finding the optimal hyper parameter values is a ubiquitous challenge. At the moment this is mostly tackled using methods like the Nelder Mead algorithm or a Grid Search. These methods, however, are potentially impractical since the gradient of the objective function need not exist and evaluating a combination of hyper parameters can be very time intensive. The task therefore is to develop an alternative way for optimizing the hyper parameter values which preferably uses as few trials as possible. Problems of this kind are well known in the field of Response Surface Methodology (RSM). Therefore – to utilize existing knowledge – we developed an optimization algorithm which is based on RSM techniques. Since it is designed to cope with situations when the input of the algorithm of interest consists of repeated measurements of the same object, the incorporated RSM techniques have been modified to be compatible with the repeated measurements situation.

This paper contains a description of the newly developed optimizing algorithm. Its actual optimization procedure is explained in Section 4. Before this is presented, Sections 2 and 3 provide a detailed description of the incorporated statistical methods. In Section 2, the statistical model used to determine an approximation of the unknown function which characterizes the impact of the hyper parameter values on the goodness of a corresponding output is introduced. Section 3 contains a description of the implemented RSM techniques which were adapted to the specific nature of the problem at hand. In Section 5, the optimizing algorithm is applied to the Support Vector Machine (SVM) algorithm using the Gaussian radial basis function kernel. Here it is used to find the values for the two parameters of the algorithm, namely the bandwidth of the kernel and the misclassification cost, which result in the lowest misclassification rate. Section 6 contains a summary of the main results as well as some concluding remarks.

2 The Random Intercepts Model

As stated above, the regarded optimization method for the hyper parameters of an algorithm is based on (say m) repeated measurements which serve as input for the algorithm of interest. Therefore, considering one combination of the hyper parameters results in m outputs, the goodness of which can then be assessed by a quantitative performance criterion y. Since the properties of the outputs depend on the hyper parameters, y as well provides information about the goodness of the used combinations of the hyper parameters. It can therefore be used as the response variable in the model utilized for the optimization.

2.1 The Model Equation

In order to examine the impact of the hyper parameters on the goodness of an output, several combinations of hyper parameter values are used. Let n be the number of the considered parameter combinations, then the resulting values of y reveal the following structure:

$$\begin{array}{ccc} y_{1\,1} & \cdots & y_{1\,n} \\ \vdots & \ddots & \vdots \\ y_{m\,1} & \cdots & y_{m\,n}. \end{array}$$

That is, the use of repeated measurements induces some kind of blocking structure on the data. The corresponding block effect can be regarded as a known error term and should therefore be included as an independent variable. In contrast to the other independent variables (i.e. the values of the hyper parameters) the values of the block effect are not selectable by the user. Thus, the block effect should be treated as a random variable.

To incorporate this, a Random Intercepts Model, the simplest member of the class of Linear Mixed Effects Models (Laird & Ware, 1982), is used:

$$\mathbf{y}_i = \mathbf{X}\boldsymbol{\beta} + b_i\,\mathbf{1}_n + \mathbf{e}_i, \quad i = 1, \ldots, m, \tag{1}$$

where

$\mathbf{y}_i \in I\!R^{n\times 1}$	:	performance of the outputs based on the i-th measurement,
$\mathbf{X} \in I\!R^{n\times p}$	:	matrix of the used hyper parameter values settings, higher-order moments and interactions,
$\boldsymbol{\beta} \in I\!R^{p\times 1}$	:	unknown vector containing the impact of the hyper parameters on the goodness of the corresponding outputs,
$b_i \in I\!R$	:	measurement error resulting from the i-th measurement,
$\mathbf{e}_i \in I\!R^{n\times 1}$	:	measurement specific error vector.

In order to make statistical inference on basis of (1), the following assumptions are made:

- $\mathbf{e}_i \underset{i.i.d.}{\sim} \mathcal{N}\big(\mathbf{0}, \sigma_{\mathrm{e}}^2\,\mathbf{I}_n\big), \quad i = 1, \ldots, m,$
- $b_i \underset{i.i.d.}{\sim} \mathcal{N}\big(0, \sigma_{\mathrm{b}}^2\big), \qquad i = 1, \ldots, m,$
- $b_1, \ldots, b_m,\ \mathbf{e}_1, \ldots, \mathbf{e}_m$ mutually independent.

The primary aim in using (1) is to be able to draw conclusions about the impact of the hyper parameters on the goodness of an output independently from the m repeated measurements at hand. For that purpose, the marginal distribution of y is considered:

$$\mathbf{y} \sim \mathcal{N}\big(\mathbf{X}\boldsymbol{\beta}\,, \mathbf{V}\big), \tag{2}$$

where $\mathbf{V} = \sigma_{\mathrm{e}}^2\,\mathbf{I}_n + \sigma_{\mathrm{b}}^2\,\mathbf{1}_n\,\mathbf{1}_n'$. The form of $\mathbf{V}$ points out the variance decomposition which results from (1). The matrix $\sigma_{\mathrm{e}}^2\,\mathbf{I}_n$ denotes the variance within a measurement, and $\sigma_{\mathrm{b}}^2\,\mathbf{1}_n\,\mathbf{1}_n'$ represents the variance between the different measurements.

2.2 Estimating the Relevant Model Parameters

Estimating the Impact of the Hyper Parameters

Let $\boldsymbol{\eta}' = \big(\boldsymbol{\beta}', \sigma_{\mathrm{e}}, \sigma_{\mathrm{b}}\big)$ be the vector of all fixed model parameters and $\boldsymbol{\zeta}' = \big(\sigma_{\mathrm{e}}, \sigma_{\mathrm{b}}\big)$ be the vector of variance parameters. The likelihood function which follows from (2) is then given by

$$L_{\mathrm{ML}}\big(\boldsymbol{\eta}\big) \;:=\; (2\pi)^{-\frac{m\,n}{2}} |\mathbf{V}(\boldsymbol{\zeta})|^{-\frac{m}{2}} \tag{3}$$
$$\times \exp\Big(-\frac{1}{2}\sum_{i=1}^{m}\big(\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}\big)'\mathbf{V}(\boldsymbol{\zeta})^{-1}\big(\mathbf{y_i} - \mathbf{X}\boldsymbol{\beta}\big)\Big),$$

where $|\,.\,|$ denotes the determinate of a matrix. If the values of σ_{e} and σ_{b} are known, the maximum likelihood estimator of $\boldsymbol{\beta}$ is obtained by maximizing $L_{\mathrm{ML}}\big(\boldsymbol{\eta}\big)$ with respect to $\boldsymbol{\beta}$. The resulting estimator for the impact of the hyper parameters is then given by

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\zeta}) := \frac{1}{m}\Big(\mathbf{X}'\,\mathbf{V}^{-1}(\boldsymbol{\zeta})\;\mathbf{X}\Big)^{-1}\sum_{i=1}^{m}\mathbf{X}'\,\mathbf{V}^{-1}(\boldsymbol{\zeta})\;\mathbf{y}_i.$$

This estimator is equal to the weighted-least-squares solution and can be shown to be *BLUE* (*Best Linear Unbiased Estimator*) for $\boldsymbol{\beta}$ (Laird & Ware, 1982).

In practice, the vector $\boldsymbol{\zeta}$ is unknown. It is commonly replaced by its *MMLE* (*Marginal Maximum Likelihood Estimator*) or its *REMLE* (*Restricted Maximum Likelihood Estimator*), both of which have attractive statistical properties such as consistency, asymptotic normality and efficiency (e.g. Verbeke & Molenberghs, 2000, p.46). Whether the MML- or REML-estimator is more appropriate depends on the nature of the problem at hand. The gravest disadvantage of the MMLE is, that it does not take into account the loss in degrees of freedom resulting from the estimation of $\boldsymbol{\beta}$. The MMLE of $\boldsymbol{\zeta}$ is therefore biased downwards, and the bias grows with the dimension of $\boldsymbol{\beta}$ (e.g. Fahrmeir & Tutz, 1994, p.226). In the considered application, the primary aim is to estimate the impact of the hyper parameters. Since –

as will be shown later – the estimates of the variance parameters are not involved in any further calculation, this disadvantage exhibits no serious problem and there is no objection against the use of the MML-estimation method. The MMLE of $\boldsymbol{\beta}$ is obtained by replacing $\boldsymbol{\beta}$ in (3) by $\hat{\boldsymbol{\beta}}(\boldsymbol{\zeta})$. The resulting likelihood function only depends on the unknown variance parameters:

$$\begin{aligned} L_{\mathrm{MML}}\big(\boldsymbol{\zeta}\big) \;&:=\; (2\pi)^{-\frac{m\,n}{2}} |\mathbf{V}(\boldsymbol{\zeta})|^{-\frac{m}{2}} \\ &\qquad \times \exp\Big(-\frac{1}{2}\sum_{i=1}^{m}\big(\mathbf{y}_i - \mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\zeta})\big)'\mathbf{V}(\boldsymbol{\zeta})^{-1}\big(\mathbf{y_i} - \mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\zeta})\big)\Big). \end{aligned}$$

Maximizing this function leads to $\hat{\boldsymbol{\zeta}}$ which can then be used to obtain $\hat{\boldsymbol{\beta}} := \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\zeta}})$.

Estimating the Random Effects

Equation (1) combined with the assumption $b_i \sim \mathcal{N}\big(0, \sigma_{\mathrm{b}}^2\big)$ yields an hierarchical structure which enables the estimation of the random effects by means of Bayesian techniques. The above distribution serves as a prior distribution $\pi\big(b_i\big)$. Following the theory on General Bayesian Linear Models (Smith, 1973), the corresponding posterior distribution has the form

$$\big(b_i|\,\mathbf{y}_i\big) \sim \mathcal{N}\big(\sigma_{\mathrm{b}}^2\,\mathbf{1}_n'\,\mathbf{V}(\boldsymbol{\zeta})^{-1}\big(\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}\big), \boldsymbol{\Lambda}_i\big),$$

where $\boldsymbol{\Lambda}_i$ is a positive definite matrix which is not further specified in this paper. The mean of this posterior distribution can be used as an estimator for the random effects:

$$\hat{b}_i(\boldsymbol{\eta}) \;:=\; \sigma_{\mathrm{b}}^2\,\mathbf{1}_n'\,\mathbf{V}(\boldsymbol{\zeta})^{-1}\big(\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}\big).$$

It can be shown that the above estimator is BLUE for b_i (e.g. Harville, 1976).

Since the fixed model parameters are unknown in practice, they are commonly replaced by their estimates $\hat{\boldsymbol{\zeta}}$ and $\hat{\boldsymbol{\beta}}$. The resulting estimator $\hat{b}_i := \hat{b}_i(\hat{\boldsymbol{\eta}})$ is often called *Empirical Bayes Estimator* (*EB-Estimator*).

2.3 Variable Selection

To avoid the problem of overfitting, a selection of the relevant independent variables should be performed. Here, the independent variables are selected using a modification of goodness of fit statistic R^2 which is well known in the context of an Ordinary Least Squares (OLS) setting:

$$R^2_{\text{meta}} := 1 - \frac{\sum_{i=1}^{m} \big(\mathbf{y}_i - \hat{\mathbf{y}}_i\big)' \big(\mathbf{y}_i - \hat{\mathbf{y}}_i\big)}{\sum_{i=1}^{m} \big(\mathbf{y}_i - \bar{y}_{i.}\big)' \big(\mathbf{y}_i - \bar{y}_{i.}\big)},$$

where $\hat{\mathbf{y}}_i = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{1}_n \hat{b}_i$ and $\bar{y}_{i.} = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$. The value of R^2_{meta} represents the reduction in conditional variation of y accounted for by the fitted values $\hat{\mathbf{y}}_i$, over that accounted for by fitting only the conditional means $\hat{\mu}_i = \bar{y}_{i.}$ (Vonesh & Chinchilli, 1997, p.423f.). Similar to the usual R^2-statistics, R^2_{meta} increases with the dimension of $\boldsymbol{\beta}$. This can be avoided by calculating the adjusted values

$$\tilde{R}^2_{\text{meta}} := 1 - a\big(1 - R^2_{\text{meta}}\big),$$

where $a = \frac{m\,n}{m\,n-p}$. On basis of $\tilde{R}^2_{\text{meta}}$, two Linear Mixed Effects Models with different numbers of independent variables can be compared appropriately. Therefore, this quantity can be used as the criterion for a forward selection.

2.4 The Benefit of the Random Intercepts Model and Justification

If a model of order $o > 1$ is used, the expected value of (2) leads to a prediction equation which can be utilized for optimizing the values of the hyper parameters. The newly developed optimization algorithm fits a quadratic model to the given data. Then the predicted (overall) goodness of an output for a given set of hyper parameters $\mathbf{x}$ can be calculated as

$$\phi\big(\mathbf{x}\big) = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i\, x_i + \sum_{i=1}^{k}\sum_{j=1}^{k} \hat{\beta}_{ij}\, x_i x_j, \tag{4}$$

where k denotes the number of hyper parameters. The function ϕ can now be used as an approximation of the unknown function which describes the relationship of the hyper parameters and the goodness of an output of the algorithm of interest. If the values of the chosen performance criterion y grow with the goodness of an output, the maximum of ϕ can be regarded as the optimal parameter combination. Otherwise, ϕ is to be minimized in order to find the optimal set of hyper parameter values.

The above optimization approach needs to be justified because the response variable y reveals some unusual properties. In particular, for a given measurement i and a given set of hyper parameters j, the value y_{ij} is deterministic. The error term e_{ij} can therefore be seen as a pure approximation error. Since it does not comprise a stochastic component, the assumption $\mathbf{e}_i \sim \mathcal{N}\big(\mathbf{0}, \sigma_{\mathrm{e}}^2\,\mathbf{I}_n\big)$ is violated and the use of the maximum likelihood theory is questionable. However, a careful examination of the impact of this assumption on the optimization problem shows that it is maintainable despite its known violation: the vector $\hat{\boldsymbol{\beta}}$ is obtained by maximizing the likelihood function (3) with respect to $\boldsymbol{\beta}$. This maximization is equivalent to the minimization of

$$f^*\big(\boldsymbol{\beta}, \sigma_{\mathrm{e}}^2, \sigma_{\mathrm{b}}^2 |\ \mathbf{y}\big) := \sum_{i=1}^{m} \big(\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}\big)' \big(\sigma_{\mathrm{e}}^2\,\mathbf{I}_n + \sigma_{\mathrm{b}}^2 \mathbf{1}_n \mathbf{1}_n'\big)^{-1} \big(\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}\big) \tag{5}$$

for fixed values of σ_{e}^2 and σ_{b}^2. The normality assumption for the term $\mathbf{e}_i$ therefore implies that $\hat{\boldsymbol{\beta}}$ minimizes the weighted quadratic distances of the observed to the predicted performance values. Since this property is reasonable in all situations, the use of the assumption $\mathbf{e}_i \sim \mathcal{N}\big(\mathbf{0}, \sigma_{\mathrm{e}}^2\,\mathbf{I}_n\big)$ can be justified. Also the choice of the respective distribution parameters can shown to be reasonable. First, it is plausible that a batch of approximation error terms reveals an average of zero and second, the assumption $\mathrm{Cov}\big(\mathbf{e}_i\big) = \sigma_{\mathrm{e}}^2\,\mathbf{I}_n$ implies that the quadratic distances in (5) are weighted equally. That means, that the fit in every subregion of the considered Region of Interest is demanded to be equally good. Since in most applications, there is no need to require different fits in different subregions, the homogeneity assumption for the error term $\mathbf{e}_i$ is preferable over a heterogeneity assumption.

The preceding argumentation states that the violation of the assumption concerning the $\mathbf{e}_i$ comprises no identifiable disadvantage for the usage of (4). Butler & Louis (1992) showed that the same is true for a violation of the assumption $b_i \sim \mathcal{N}\big(0, \sigma_b^2\big)$. In order to get valid inferences, only the standard errors of all components of $\boldsymbol{\eta}$ need to be corrected. Since the variable selection implemented in the considered optimization algorithm is done by the descriptive criterion $\tilde{R}^2_{\text{meta}}$, no significance tests for the components of $\boldsymbol{\beta}$ are involved to obtain the function approximation (4). Therefore, the approximation goodness does not depend on the correct choice of random effects distribution and on the whole, the form of (4) can be shown to be very robust against most possible violations of the assumptions of model (1).

3 Response Surface Methodology for Linear Mixed Effects Models

In order to find the best combination of the hyper parameters in as few trials as possible, their values should be varied systematically. To accomplish that, RSM techniques are implemented in the optimization algorithm. The starting point is a Cartesian product $\big[x_{1l}, x_{1u}\big] \times \ldots \times \big[x_{kl}, x_{ku}\big]$ which comprises a conjecture about the location of the desired optimum. To cover this region, a Central Composite Design (CCD) with an axial distance of $\alpha = \sqrt{k}$ is used which results in all points being placed on a sphere with the radius $\sqrt{k}$, i.e. the coded values of the hyper parameters have the property $\sum_{i=1}^{k} x_{i_c}^2 = k$, where x_{i_c} denotes the coded value of the i-th hyper parameter. The coding is done such that $x_{il_c} = -\sqrt{k}\ \forall i$ and $x_{iu_c} = \sqrt{k}\ \forall i$.

As described earlier, a quadratic Random Intercepts Model is used to fit the resulting data and a prediction equation for the (overall) goodness of an output is generated. Using the coded values of the hyper parameters, it has the form

$$\hat{y} = \phi\big(\mathbf{x_c}\big) = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i\, x_{i_c} + \sum_{i=1}^{k} \sum_{j=1}^{k} \hat{\beta}_{ij}\ x_{i_c}\, x_{j_c}. \tag{6}$$

Since such an approximation is valid only locally, the search for an optimal combination of hyper parameter values should at first be restricted to combinations within the considered Region of Interest. That is, the function ϕ is to be optimized under the spherical constraint $\sum_{i=1}^{k} x_{i_c}^2 \leq k$. In order to make this solvable for the algorithms implemented in the prevalent software packages, we developed a method for transforming the above problem into an unconstraint optimization problem.

3.1 Determining the Optimum within the Region of Interest

Let S be the considered spherical Region of Interest, i.e. $S := \left\{\mathbf{x_c} \,|\, \sum_{i=1}^{k} x_{i_c}^2 \leq k\right\}$. In order to ensure that a quadratic prediction equation of the form (6) can be optimized without violating its local validity, the quantity

$$\Delta := \sqrt{\sum_{i=1}^{k} x_{i_c}^2} - \sqrt{k}$$

is defined. Its absolute value denotes the Euclidean distance of a point $\mathbf{x_c}$ from S. Moreover, its sign contains the additional information whether $\mathbf{x_c}$ lies inside (negative sign) or outside (positive sign) the sphere. On basis of Δ, the function ϕ is augmented by an additive penalty term which has a negative sign in maximization and a positive sign in minimization problems:

$$\phi_{\text{mod}}\big(\mathbf{x}_c\big) = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_{i_c} + \sum_{i=1}^{k}\sum_{j=1}^{k} \hat{\beta}_{ij} x_{i_c} x_{j_c} + \upsilon,$$

where

$$\upsilon = \begin{cases} \pm\big(\sum_{i=1}^{k} |\hat{\beta}_i|\, \Delta + \sum_{i=1}^{k}\sum_{j=1}^{k} |\hat{\beta}_{ij}| \left(2\sqrt{k}\Delta + \Delta^2\right) + \Delta\big) & , \Delta > 0, \\ 0 & , \text{else.} \end{cases}$$

Without loss of generality, let us consider a maximization problem to illustrate the benefit of υ: Let $\bar{S} := \left\{\mathbf{x_c} \,|\, \sqrt{\sum_{i=1}^{k} x_{i_c}^2} = \sqrt{k}\right\}$ denote the surface of the sphere of interest. It can be shown that

$$\forall\, \mathbf{x_c} \notin S \;\; \exists\, \tilde{\mathbf{x}}_\mathbf{c} \in \bar{S} \;\; : \phi_{\text{mod}}\big(\tilde{\mathbf{x}}_\mathbf{c}\big) > \phi_{\text{mod}}\big(\mathbf{x_c}\big). \tag{7}$$

For this, let $\mathbf{x_c} = \big(x_{c_1}, \ldots, x_{c_k}\big)'$ be an arbitrary but constant point outside the sphere and let $\mathbf{s} = \big(s_1, \ldots, s_k\big)'$ denote the point on the surface of S which minimizes the Euclidean distance to $\mathbf{x_c}$. Moreover, let $\boldsymbol{\delta} = \big(\delta_1, \ldots, \delta_k\big)'$ denote the straight connection between $\mathbf{x_c}$ and $\mathbf{s}$, i.e. $\boldsymbol{\delta} := \mathbf{x_c} - \mathbf{s}$, where $|\boldsymbol{\delta}|$ has the minimal possible value. Then

$$
\begin{aligned}
\phi_{\text{mod}}\big(\mathbf{x}_c\big) &= \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i \, x_{i_c} + \sum_{i=1}^{k}\sum_{j=1}^{k} \hat{\beta}_{ij} \, x_{i_c} x_{j_c} \\
&\quad - \Big(\sum_{i=1}^{k} |\hat{\beta}_i| \, \Delta + \sum_{i=1}^{k}\sum_{j=1}^{k} |\hat{\beta}_{ij}| \, (2\sqrt{k}\Delta + \Delta^2) + \Delta\Big) \\
&= \hat{\beta}_0 + \sum_{i=1}^{k} \big[\hat{\beta}_i \, (s_i + \delta_i) - |\hat{\beta}_i| \, \Delta\big] \\
&\quad + \sum_{i=1}^{k}\sum_{j=1}^{k} \big[\hat{\beta}_{ij} \, (s_i + \delta_i)(s_j + \delta_j) - |\hat{\beta}_{ij}| \, (2\sqrt{k}\Delta + \Delta^2)\big] - \Delta \\
&= \phi_{\text{mod}}\big(\mathbf{s}\big) + \underbrace{\sum_{i=1}^{k} \big[\hat{\beta}_i \, \delta_i - |\hat{\beta}_i| \, \Delta\big]}_{A} \\
&\quad + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{k} \big[\hat{\beta}_{ij} \, (s_i\delta_j + s_j\delta_i + \delta_i\delta_j) - |\hat{\beta}_{ij}| \, (2\sqrt{k}\Delta + \Delta^2)\big] - \Delta}_{B}
\end{aligned}
$$

Since $\Delta > 0$ in the considered case, in order to prove statement (7) it still has to be shown that $A + B \leq 0$. It is easy to see that $|s_i| \leq \sqrt{k} \;\forall i$ and $|\delta_i| \leq \Delta \;\forall i$, which can be used in the following case differentiations:

A:

$$
\left.
\begin{array}{lll}
\underline{\text{1. case}}: & \hat{\beta}_i \geq 0: & \hat{\beta}_i\delta_i - |\hat{\beta}_i| \, \Delta = \hat{\beta}_i(\underbrace{\delta_i - \Delta}_{\leq 0}) \leq 0 \;\; \forall i \\
\underline{\text{2. case:}} & \hat{\beta}_i < 0: & \hat{\beta}_i\delta_i - |\hat{\beta}_i| \, \Delta = \hat{\beta}_i(\underbrace{\delta_i + \Delta}_{\geq 0}) \leq 0 \;\; \forall i
\end{array}
\right\} \Rightarrow A \leq 0
$$

B:

$$\left.\begin{aligned}
&\underline{\text{1. case}}: \quad \hat{\beta}_{ij} \geq 0: \quad \hat{\beta}_{ij}(s_i\delta_j + s_j\delta_i + \delta_i\delta_j) - |\hat{\beta}_{ij}|(2\sqrt{k}\Delta + \Delta^2) \\
&\qquad\qquad = \hat{\beta}_{ij}\big(\underbrace{(s_i\delta_j + s_j\delta_i + \delta_i\delta_j) - (2\sqrt{k}\Delta + \Delta^2)}_{\leq 0}\big) \leq 0 \quad \forall i,j \\
&\underline{\text{2. case:}} \quad \hat{\beta}_{ij} < 0: \quad \hat{\beta}_{ij}(s_i\delta_j + s_j\delta_i + \delta_i\delta_j) - |\hat{\beta}_{ij}|(2\sqrt{k}\Delta + \Delta^2) \\
&\qquad\qquad = \hat{\beta}_{ij}\big(\underbrace{(s_i\delta_j + s_j\delta_i + \delta_i\delta_j) + (2\sqrt{k}\Delta + \Delta^2)}_{\geq 0}\big) \leq 0 \quad \forall i,j
\end{aligned}\right\} \Rightarrow B \leq 0$$

The above shows that the optimum of ϕ_{mod} lies within the considered Region of Interest S. Since the function ϕ and its modified version are identical within S, the unconstraint maximization of ϕ_{mod} is equivalent to maximizing the prediction equation (6) under the spherical constraint $\sum_{i=1}^{k} x_{i_c}^2 \leq k$. Moreover, the penalty term υ is chosen in a way that ϕ_{mod} is continuous and differentiable – in particular at the surface $\bar{S}$ of the Region of Interest. This permits the use of derivative based optimization methods to find the optimum of ϕ_{mod}.

If the optimum of ϕ_{mod} is located at the surface of the considered Region of Interest, it can be assumed that parameter combinations outside of S result in better outputs than those considered in the first optimization. The Region of Interest should therefore be relocated within the Operability Region.

3.2 Systematical Relocation of the Region of Interest

The relocating procedure we used in the optimization algorithm is based on the well known Method of Steepest Ascent which is designed to use as few trials as possible to find a combination of the considered factors which results in a value of the response which is as great as possible. If the objective is to minimize a response, the Method of Steepest Ascent can easily be adapted to minimization problems (Method of Steepest Descent). Due to this straight forward adaption, we reduce the following description to the case of a maximization problem.

The conventional Method of Steepest Ascent

In the literature (Box & Wilson, 1951), the Method of Steepest Ascent is described for the case of a linear response surface and an Ordinary Least Squares setting. It is usually applied when little is known about the process of interest. In such a situation, it is likely that the initially chosen operating conditions are remote from the optimum. Therefore, it usually suffices to fit a first-order model to the data and equation (6) reduces to

$$\hat{y} = \phi(\mathbf{x}_c) = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i \, x_{i_c}. \tag{8}$$

Since the surface of ϕ describes a (hyper-)plane, the contours of (8) are represented by a series of parallel lines such as shown in figure 1.

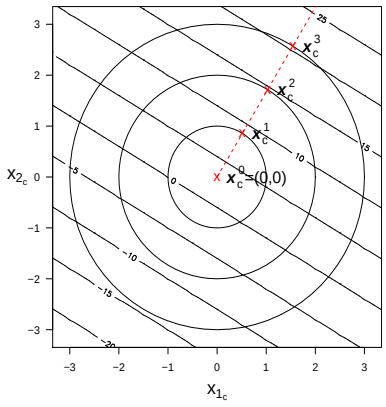

Figure 1: Path of Steepest Ascent for a virtual first-order response surface with $k = 2$ factors: The movement of the operating conditions is normal to the contour planes

The Method of Steepest Ascent is a procedure for moving the operating conditions sequentially along the direction in which $\hat{y}$ increases most rapidly. It is easy to see that this direction is normal to the contour planes. Since the Path of Steepest Ascent is usually taken as the line through the center of the Region of Interest, the coordinates along the path are proportional to the estimated regression coefficients in (8). To clarify this, let $\left\{\mathbf{x}_c^s\right\}$ denote the sequence of coded factor values which forms

the Path of Steepest Ascent. The s-th member of the sequence, $\mathbf{x}_c^s = \left(x_{1_c}^s, \ldots, x_{k_c}^s\right)'$, is then defined as the combination of coded factor values which maximizes (8) under the constraint

$$\sum_{i=1}^{k} x_{i_c}^2 = r_s^2, \tag{9}$$

where $r_s \in \mathbb{R}^+$ denotes the radius which defines the sphere $\mathbf{x}_c^s$ is restricted to lie on. The sequence of radii is chosen to satisfy $0 = r_0 < r_1 < r_2 < r_3 < \ldots$ Using the above definitions, the dependence the components of $\mathbf{x}_c^s$ on the estimated coefficients $\hat{\beta}_i$ can be formulated as

$$x_{1_c}^s = \rho_s \hat{\beta}_1, \quad x_{2_c}^s = \rho_s \hat{\beta}_2, \quad \ldots, \quad x_{k_c}^s = \rho_s \hat{\beta}_k, \tag{10}$$

where the proportionality constant has a value[1] of $\rho_s = \frac{r_s}{\sqrt{\sum_{i=1}^{k} \hat{\beta}_i^2}}$. As a result of this proportionality property, the determination of the Path of Steepest Ascent in practice reduces to the choice of the values $r_1, r_2, \ldots$

After calculating the Path of Steepest Ascent, experimental runs are conducted along it, i.e. the members of $\left\{\mathbf{x}_c^s\right\}$ are successively used as operating conditions. This procedure normally results in improving values of the response, but since moving the operating conditions along the path comprises an extrapolation of (8), at some region along the path the improvement will decline and eventually disappear. The best point found can then be chosen as a base for a new first-order design from which further advantage might be possible by again calculating the Path of Steepest Ascent.

Since the Method of Steepest Ascent only yields any benefit in regions where the linear effects dominate the interactions and higher-order terms it is not generally applicable. By exploiting the similarity of (9) and the problem of optimizing a quadratic prediction equation within a spherical Region of Interest (see Section 3.1) we developed a generalization of the conventional Method of Steepest Ascent which can be applied to response surfaces of order $o \leq 2$.

[1] If the objective is to calculate the Path of Steepest Descent, $\rho_s = -\frac{r_s}{\sqrt{\sum_{i=1}^{k} \hat{\beta}_i^2}}$ is chosen.

The Quadratic Method of Steepest Ascent

As described in Section 3.1, when a CCD with the axial distance $\alpha = \sqrt{k}$ is used to obtain a quadratic prediction equation, maximizing the function ϕ_{mod} enables to find the operating condition which maximizes equation (6) within the associated spherical Region of Interest S. If the found maximum lies at the surface of S, a method for systematically relocating the Region of Interest which is applicable for quadratic response surfaces would be desirable. Since in the case of a linear surface, the constraint (9) is equivalent to the constraint $\sum_{i=1}^{k} x_{i_c}^2 \leq r_s^2$, a modified version of the penalty term υ can be used to tackle this problem:

Let $\{d_s\}$ be a sequence of successively growing positive values and let S_s denote the sphere around the origin with the radius of $(\sqrt{k} + d_s)$, i.e. $S_s := \{\mathbf{x_c} \mid \sum_{i=1}^{k} x_{i_c}^2 \leq (\sqrt{k} + d_s)^2\}$. With the definition

$$\Delta_s := \sqrt{\sum_{i=1}^{k} x_{i_c}^2} - (\underbrace{\sqrt{k} + d_s}_{r_s}),$$

the quantity $|\,\Delta_s|$ denotes the Euclidean distance of a point $\mathbf{x}_c$ and S_s. Again, its sign contains the information whether $\mathbf{x}_c$ lies within (negativ) or outside (positiv) S_s. Maximizing

$$\phi_s(\mathbf{x}_c) = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_{i_c} + \sum_{i=1}^{k} \sum_{j=1}^{k} \hat{\beta}_{ij} x_{i_c} x_{j_c} + \upsilon_s, \tag{11}$$

where

$$\upsilon_s = \begin{cases} -\left(\sum_{i=1}^{k} |\hat{\beta}_i|\, \Delta_s + \sum_{i=1}^{k} \sum_{j=1}^{k} |\hat{\beta}_{ij}| \left(2 r_s \Delta_s + \Delta_s^2\right) + \Delta_s\right) & , \Delta_s > 0, \\ 0 & , \text{else}, \end{cases}$$

then enables to find the vector $\mathbf{x}_c^s \in S_s$ which – according to (6) – maximizes the predicted response. The sequence of maxima of the respective functions ϕ_s therefore has the same interpretation as the Path of Steepest Ascent calculated in the conventional way[2]. Since the prediction equation comprises quadratic terms, the members

[2] In minimization problems, the quadratic Path of Steepest Descent can be obtained by minimizing ϕ_s using a positive sign of υ_s.

of this sequence do not reveal the proportionality property (10). The quadratic Path of Steepest Ascent does therefore not necessarily describe a straight line.

After obtaining the quadratic Path of Steepest Ascent, the further proceeding is identical to that described for the linear case. The member of the path are successively used as operating conditions till no further improvement can be achieved. The best point found is then used as a center for a new experimental design.

Application of the Method of Steepest Ascent to Linear Mixed Effects Models

As mentioned earlier, the newly developed optimization algorithm for the hyper parameters of an algorithm is designed to cope with the situation when the input at hand consists of m repeated measurements. In order to implement the Method of Steepest Ascent, it therefore is necessary to generalize it for the context of Linear Mixed Effects Models. For both the conventional and the quadratic Method of Steepest Ascent, this generalization is straight forward since the form of the prediction equation (6) utilized for the optimization is identical to that of a prediction equation in an OLS setting. Obtaining the Path of Steepest Ascent therefore does not require any additional thinking. Also, incorporating the specific nature of the data (cf. page 3) when conducting the control experiments along the path, is rather simple. Let $\mathbf{y}_s = \left(y_{1s}, \ldots, y_{ms}\right)'$ denote the vector of the obtained values of the performance criterion y when the s-th member of the Path of Steepest Ascent is used as setting for the algorithm of interest. The average $\bar{y}_s$ of those values can then be used as an criterion which determines whether an additional step along the Path of Steepest Ascent should be conducted or not. As soon as $\bar{y}_{s+1} < \bar{y}_s$, the Path of Steepest Ascent is truncated and the uncoded version of the hyper parameter combination $\mathbf{x}_c^s$ can be used as the center for a new design.

4 The RSM Algorithm

Figure 2 shows how the methods described in Sections 2 and 3 are combined in order to find the best combination of hyper parameter values for an algorithm of interest. This procedure was automated using the software package **R** (R Development Core Team, 2004) and will from now on be called *RSM algorithm*.

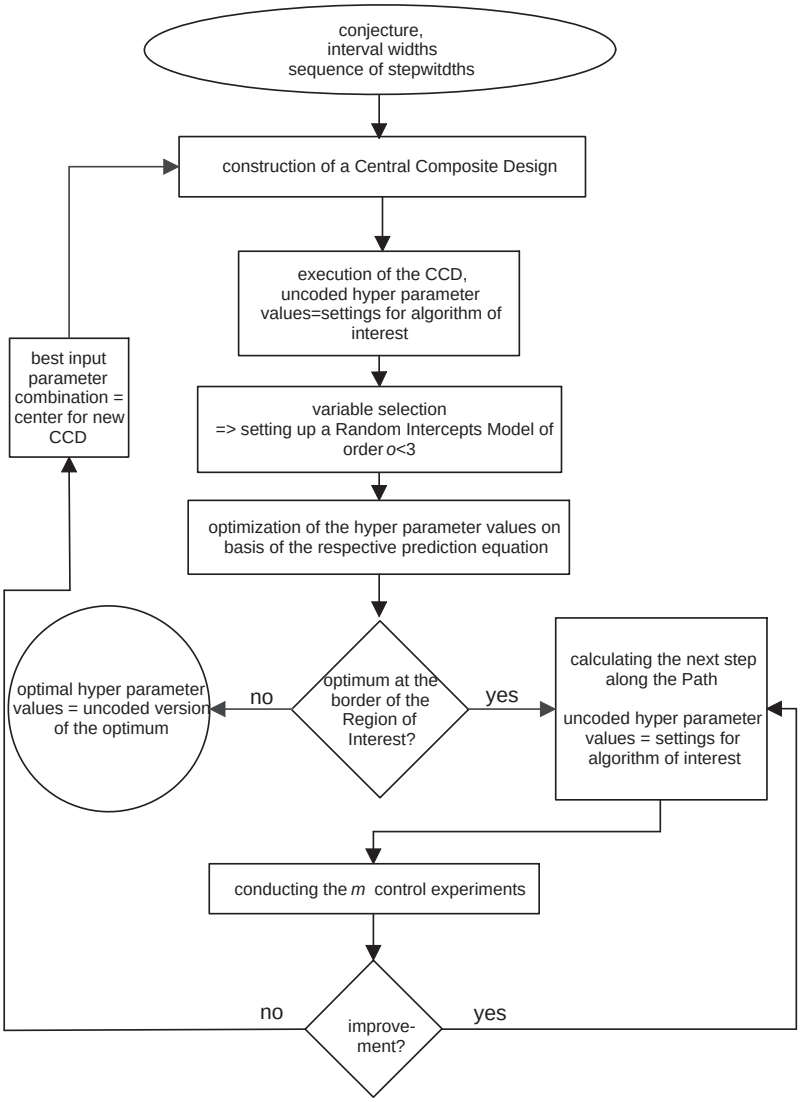

Figure 2: Flow Chart for Demonstrating the Mode of Operation of the RSM Algorithm

The input for the optimization algorithm consists amongst others of a user's conjecture about the best parameter combination for the algorithm of interest. The optimization algorithm uses this (mostly) vague estimate together with k parameter specific interval widths to determine the first Region of Interest. The default value of these widths is one for each parameter but they can independently from each other be varied by the user. Another part of the input are the m repeated measurements which serve as input for the algorithm of interest. Moreover, the sequence of step widths $\{d_s\}$ is to be specified. The default setting is $d_s = s \cdot \frac{\sqrt{k}}{2}$, i.e. the radius of the sphere in which the optimization is conducted, is at each step augmented by half of the original radius.

The first step of the optimization procedure consists of the determination of the first CCD which is chosen to have an axial distance of $\alpha = \sqrt{k}$. This results in $n = 2^k+2k+1$ hyper parameter combinations which are investigated in the first cycle. Now the algorithm of interest is run $m \cdot n$ times since each combination is applied to each of the m repeated measurements. The goodness of the resulting outputs is then assessed by the performance criterion y which is to be optimized. The $m \cdot n$ values of y together with the corresponding coded values of the hyper parameters form the data on basis of which the first Random Intercepts Model is set up. This model is found by means of a forward selection utilizing of the goodness of fit statistic $\tilde{R}^2_{\text{meta}}$. Since this variable selection is programmed to result in an at most quadratic model the corresponding prediction equation reveals an order of $o \leq 2$. Subsequently, the prediction equation is modified as described in Section 3.1 to ensure that the found optimum lies within the considered spherical Region of Interest S. The actual optimization is accomplished by the L-BFGS-B algorithm (Byrd *et al.*, 1995), a quasi-Newton method which is implemented in **R** and uses function values and gradients to build up a picture of the surface of the objective function. If the found optimum lies in $S \backslash \bar{S}$ (cf. Section 3.1), the optimization procedure is completed and the uncoded version of the optimum are returned as the best combination of hyper parameter values for the algorithm of interest.

If the found optimum is located at the surface of S, the quadratic Path of Steepest Ascent (Descent) is calculated and till $\bar{y}_{s+1} < \bar{y}_s$, at each step along the path a control experiment is conducted. The uncoded version of the s-th member of the Path of Steepest Ascent is then used as center of a new CCD with $\alpha = \sqrt{k}$. After executing this CCD and applying the variable selection, it is rechecked if the optimum lies within of the new Region of Interest. If so, the optimal combination of hyper parameter values is returned. Otherwise, the procedure described above is iterated till a Region of Interest which contains an optimum is found.

5 Optimizing the Hyper Parameters of a Nonlinear Support Vector Machine

Examples for algorithms containing some hyper parameters which are to be optimized in order to gain the best performance can be found throughout statistical literature. In particular, optimizing the involved parameters of a learning machine is a crucial step for obtaining the minimal error. In this paper we consider the problem of optimizing the hyper parameters of an SVM utilizing the Gaussian radial basis function (RBF) kernel

$$K\big(\mathbf{z}, \mathbf{z}'\big) = \exp\big(-\gamma ||\mathbf{z} - \mathbf{z}'||^2\big), \quad \gamma \geq 0,$$

where γ denotes the bandwidth of the kernel. Let the training data consist of l observations $(\mathbf{z}_i, y_i)'$, where $\mathbf{z}_i \in I\!R^d$ is the set of explanatory variables for the i-th observation and $y_i \in I\!R$ denotes the corresponding class membership. In the dichotomous case (i.e. $y_i \in \big\{-1, +1\big\}$) using a kernel-based SVM to find the optimal border between the classes is equivalent to maximizing the function

$$L(\boldsymbol{\alpha}) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K\big(\mathbf{z}_i, \mathbf{z}_j'\big), \tag{12}$$

where $\boldsymbol{\alpha} \in I\!R^d$ is a vector of Lagrangian multipliers which arises from the fact that (12) is the dual formulation of a constraint optimization problem (e.g. Vapnik, 2000,

p.133ff.). In order to get a valid result, the maximization has to be carried out under the constraints

$$\sum_i \alpha_i y_i = 0 \quad \text{and} \quad 0 \leq \alpha_i \leq C \; \forall i,$$

where $C \geq 0$ is a parameter which assigns a cost for misclassified observations.

With the kernel of an SVM fixed, there are only a few hyper parameters that are to be specified by the user. Since we decided to use the RBF kernel, optimizing the hyper parameters reduces to finding the best values of the kernel bandwidth γ and the cost parameter C. The optimal values of γ and C are data-specific. They therefore have to be found separately for each application which can be done by using the RSM algorithm presented in Section 4. To demonstrate this, we applied the non-linear SVM described above to the West German Business Cycle data (Heilemann & Münch, 1996) which is analyzed by the project B3 of the SFB475 (Collaborative Research Centre “Reduction of Complexity for Multivariate Data Structures"), supported by the Deutsche Forschungsgemeinschaft. This data set consists of 13 economic variables with $l = 157$ quarterly observations from 1955/4 to 1994/4. The German business cycle is to be classified in a four phase scheme: upswing, upper turning point, downswing and lower turning point. The considered time period reveals 6 complete cycles. In order to extend the SVM method to the given four-class situation, the four phases are split into six one-against-one test situations.

The performance of a combination of γ and C is assessed in terms of (in-)accuracy (e.g. Hand, 1997, p.99) which is estimated by the misclassification rate. For measuring this rate, $B = 200$ bootstrap samples were generated from the given data set. Each bootstrap sample is used to train a classification rule $\hat{f}^b(\mathbf{x})$ which then is applied to the subset of $\{\mathbf{x}_i\}$ not present in the b-th bootstrap sample ($b = 1, \ldots, B$). Let $\hat{e}_b$ denote the proportion of the observations misclassified by $\hat{f}^b(\mathbf{x})$, then considering n combinations of γ and C (and setting $B = m$) leads to a set of outputs which reveals the structure shown in Section 2.1. The RSM algorithm can therefore be applied directly to the problem of finding the optimal set of the involved hyper

parameters γ and C. To prevent the relocation procedure from leaving the feasible parameter space $[0, \infty) \times [0, \infty)$, we searched for the best combination of $a \in \mathbb{R}$ and $b \in \mathbb{R}$ which were then transformed by the bijective functions $\gamma = e^a$ and $C = 10^b$. The starting point for the RSM algorithm was chosen to be $(a, b) = (0, 0)$, i.e. $(\gamma, C) = (0, 0)$, and the default setting for the parameter specific interval widths (each equal to one) was used. The first (uncoded) Region of Interest was therefore determined by the inscribed circle of the Cartesian product $[a_l, a_u] \times [b_l, b_u] = [-0.5, 0.5] \times [-0.5, 0.5]$. After three relocations of the Region of Interest (the successive CCDs are displayed in Figure 3), an optimal combination of a and b was found. It resulted in an average error rate of $\hat{e} = 0.241$, where $\hat{e}$ is defined as the average error rate of the 200 drawn bootstrap samples.

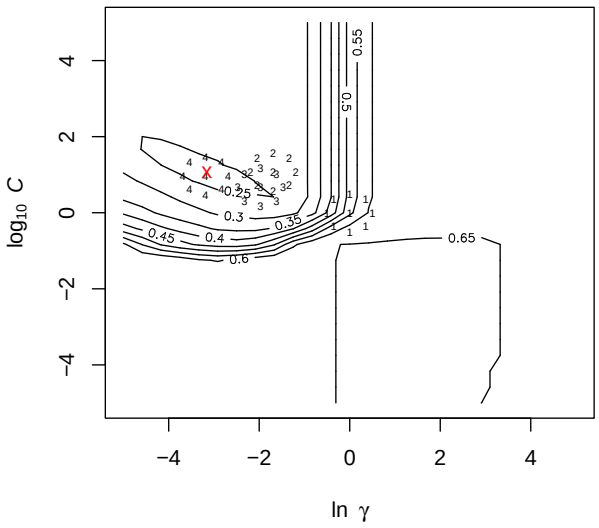

Figure 3: Response Surface of the error rate obtained by a grid search using 625 equidistantly spaced points in $[-5, 5] \times [-5, 5]$ and the CCDs successively executed by the RSM algorithm. The optimum found by the RSM algorithm is marked by an x.

To assess the performance of the RSM algorithm, we compared the error rate resulting from the optimal combination of γ and C found by the RSM algorithm with the corresponding error rates obtained by two well known optimization methods: the Nelder-Mead algorithm (Nelder & Mead, 1965) and a grid search. For the Nelder-Mead algorithms we used the starting point $(a, b) = (0, 0)$ and a grid search was

performed on a very fine grid (25×25 grid points) equidistantly spread over the space $[-5, 5] \times [-5, 5]$ (cf. Figure 3). Moreover, the number parameter combinations each optimization procedure tested to find an (almost) optimal combination of the hyper parameters was compared. During the Nelder-Mead algorithm and the grid search, testing one combination (a, b) comprises drawing 200 bootstrap sample and constructing the respective classification rules. The same holds for the RSM algorithm when an operating condition of a CCD is considered and when a control experiment is conducted along the Path of Steepest Descent, but since the intern optimization procedures are based on a prediction equation of the form (11), parameter combinations considered by the L-BFGS-B algorithm are much less computational expensive and are therefore neglected in our count. The results of the comparison are summarized in Table 1.

Table 1: Performances of the considered optimization algorithms measured with respect to the (in-)accuracy of the resulting classification rule and to speed

	$\hat{e}$	# evaluations
RSM Algorithm	0.241	52
Nelder-Mead	0.252	63
Grid Search	0.252	625

The RSM algorithm outperformed the conventional optimization methods with respect two both criteria, accuracy and computational effort.

6 Conclusion

Statistical methods have been shown to be a valuable tool for optimizing hyper parameters. By utilizing an extension of the Method of Steepest Ascent for the case of response surfaces of order $o \leq 2$, an optimizing algorithm for the hyper parameters of an algorithm of interest could be developed which offers a good compromise between

accuracy and computational effort. Moreover, the use of the quadratic Method of Steepest Ascent is not restricted to the RSM algorithm. It can be applied in every situation the task is to efficiently relocate the considered Region of Interest in order to find an optimal operating condition for a system.

Acknowledgment

This work has been supported by the Collaborative Research Center 475 'Reduction of Complexity for Multivariate Data Structures' of the German Research Foundation (DFG).

References

Box, G. E. P. & Wilson, K. B. (1951): On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society, Series B*, 13, 1–45.

Butler, S. M. & Louis, T. A. (1992): Random effects models with nonparametric priors. *Statistics in Medicine*, 11, 1981–2000.

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995): A limited memory algorithm for bound constrained optimization. *SIAM Journal of Scientific Computation*, 16, 1190–1208.

Fahrmeir, L. & Tutz, G. (1994): *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer Verlag.

Hand, D. J. (1997): *Construction and Assessment of Classification Rules*. New York: John Wiley & Sons.

Harville, D. A. (1976): Extension of the Gauss-**M**arkow Theorem to include estimation of random effects. *Annals of Statistics*, 4, 384–395.

Heilemann, U. & Münch, H. (1996): West german business cycles 1963-1994: A multivariate discriminant analysis. In: *CIRET-Conference in Singapore*, CIRET-Studien 50.

Laird, N. M. & Ware, H. (1982): Random-effects models for longitudinal data. *Biometrics*, 38, 963–974.

Nelder, J. & Mead, R. (1965): A simplex method for functional minimization. *Computer Journal*, 7, 308–313.

R Development Core Team (2004): *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Wien. URL `http://www.R-project.org`.

Smith, A. (1973): A general bayesian linear model. *Journal of the Royal Statistical Society, Series B*, 35, 67–75.

Vapnik, V. N. (2000): *The Nature of Statistical Learning Theory*. New York: Springer Verlag, 2nd edition.

Verbeke, G. & Molenberghs, G. (2000): *Linear Mixed Models for Longitudinal Data*. New York: Springer Verlag.

Vonesh, E. F. & Chinchilli, V. M. (1997): *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. New York: Marcel Dekker.