ePub^WU Institutional Repository

Valery V. Fedorov and Peter Hackl and Werner Müller

Estimation and Experimental Design for Second Kind Regression Models

Working Paper

http://epub.wu.ac.at/

Estimation and Experimental Design for Second Kind Regression Models

Valery V. Fedorov, Peter Hackl, Werner Müller
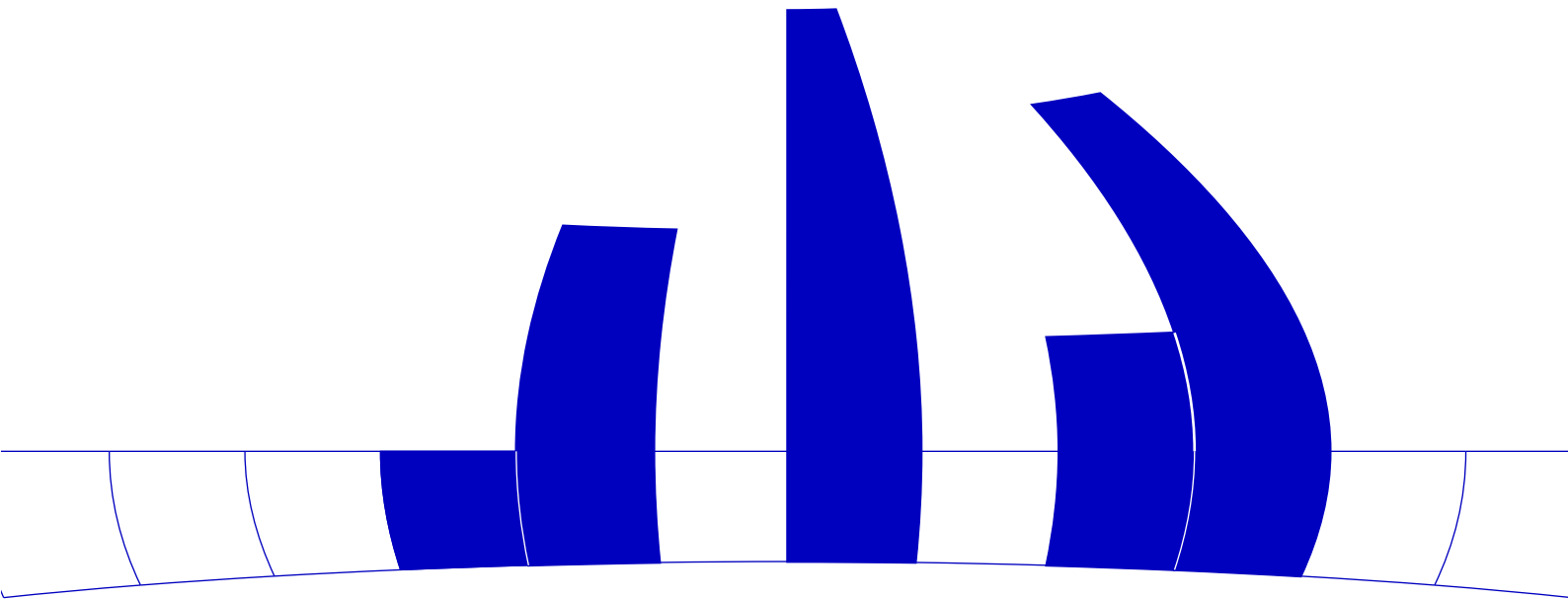
Estimation and Experimental Design for Second Kind Regression Models

by

Valery V. Fedorov[1]
Peter Hackl
Werner Müller[2]

Abstract: Estimation procedures and optimal designs for estimation of the individual parameters and of the global parameters are discussed under various conditions of prior knowledge. The extension to nonlinear parametrization of the response function is based on the asymptotical validity of the results for the linear parametrization. For the case where the error variance and the dispersion matrix are unknown, an iterative estimation procedure is suggested. An example based on dental plaque pH profiles demonstrates the improvement that is achieved (a) through using the optimal design or a design that is close to the optimal, and (b) through taking into account prior information.

Keywords: nonlinear regression, random regression coefficients, optimal design, dental plaque pH.

---

[1]Laboratory of Social and Economic Measurement, USSR Academy of Sciences, Moscow; visitor at the Department of Statistics from October till December 1990.
[2]Institute for Advanced Studies, Vienna, and lecturer at the Department of Statistics.

1   Statement of the Problem

In the following we consider observations y_{ij} that are taken from the j-th object $(j = 1,\ldots,r)$ under conditions x_i, $i = 1,\ldots,q$. In total, we have $r \cdot q$ observations. Let us assume that under condition x_i the responses y_{ij} of all objects have the same structure

$$y_{ij} = \eta(x_i, \theta_j) + \varepsilon_{ij}, \tag{1}$$

where ε_{ij} is the observation error and $\theta_j = (\theta_{1j},\ldots,\theta_{mj})^T$ is the vector of (individual) parameters that reflect the specific features of the j-th object. Let us further assume that the variability of θ_j can be described by some distribution with mean θ_0 (the global parameters) and dispersion matrix $\mathcal{D}$, that the observation errors are independent and have zero mean and variance σ^2, and that they are independent of θ_j. In the following we will use indices ε and θ for the expectation to indicate the respective distribution; expectation without index is related to the joint distribution.

The model described by (1) usually is called a regression model with random coefficients or a regression model of the second kind. Objectives of statistical analysis of such a model can be

- estimation of the individual parameters θ_j

- estimation of the global parameters θ_0

- estimation of σ^2 and $\mathcal{D}$

In Sections 2 and 3 we will discuss estimation procedures and optimal designs for these objectives under various conditions of prior knowledge, respectively. Section 3 illustrates the results of the earlier sections on the basis of dental plaque pH data.

2   Estimation Problems

We will start this section by considering the case of linear parametrization of the model, i.e., $\eta(x,\theta) = \theta^T f(x)$. This case has extensively been studied [see, e.g., Fedorov and Müller (1989), Fisk (1967), Gladitz and Pilz (1982), Rao (1973), Spjøtvoll (1977)]. The main results will be sketched in Section 2.1. In Sections 2.2 and 2.3 we will discuss the nonlinear model.

2.1   The linear case

Let us assume that $\eta(x,\theta) = \theta^T f(x)$. Estimation is discussed for the following cases:

- σ^2 and $\mathcal{D}$ are known, but θ_0 is unknown

- θ_0, σ^2, and $\mathcal{D}$ are known

- θ_0, σ^2, and $\mathcal{D}$ are unknown

2

2.1.1 Parameters σ^2 and $\mathcal{D}$ are known, but θ_0 is unknown

Let us first consider the linear, unbiased estimator

$$\hat{\theta}_j = M^{-1} Y_j \tag{2}$$

where $M = \sum_{i=1}^q f(x_i) f^T(x_i)$ and $Y_j = \sum_{i=1}^q f(x_i) y_{ij}$. Given true values θ_{jt} for the parameter vector θ_j of object j, we find the following conditional moments with respect to the error distribution:

$$
\begin{aligned}
\mathrm{E}_\epsilon\{\hat{\theta}_j\} &= \theta_{jt} \\
\mathrm{Var}_\epsilon\{\hat{\theta}_j\} &= \mathrm{E}_\epsilon\{(\hat{\theta}_j - \theta_{jt})(\hat{\theta}_j - \theta_{jt})^T\} = \sigma^2 M^{-1}.
\end{aligned}
$$

A natural estimator of θ_0 is

$$\hat{\theta}_0 = r^{-1} \sum_{j=1}^r \hat{\theta}_j = r^{-1} M^{-1} \sum_{j=1}^r Y_j. \tag{3}$$

The (unconditional) variance of $\hat{\theta}_0$ is

$$\mathrm{Var}\{\hat{\theta}_0\} = r^{-1}(\mathcal{D} + \sigma^2 M^{-1}). \tag{4}$$

We have two alternatives to calculate the estimate for θ_0:

(a) find $\hat{\theta}_j = \arg\min_\theta \sum_{i=1}^q [y_{ij} - \theta^T f(x_i)]^2$ for $j = 1, \ldots, q$, and then calculate the arithmetic mean $\hat{\theta}_0$;

(b) minimize the total sum of squared residuals:

$$\hat{\hat{\theta}}_0 = \arg\min_\theta \sum_{j=1}^r \sum_{i=1}^q [y_{ij} - \theta^T f(x_i)]^2. \tag{5}$$

Some algebra shows that $\hat{\hat{\theta}}_0 = M^{-1}\overline{Y}$ where $\overline{Y} = \sum_{i=1}^q f(x_i)\overline{y}_i$ and $\overline{y}_i = r^{-1}\sum_{j=1}^r y_{ij}$, and therefore $\hat{\hat{\theta}}_0 = \hat{\theta}_0$.

2.1.2 θ_0, σ^2, and $\mathcal{D}$ are known

In this case we wish to find an estimator $\tilde{\theta}_j$ that minimizes the unconditional variance

$$\tilde{\theta}_j = \arg\min_{\theta_j} \mathrm{E}\{(\theta_j - \theta_{jt})(\theta_j - \theta_{jt})^T\};$$

equivalently, this is

$$\tilde{\theta}_j = \arg\min_\theta \left\{ \sigma^{-2} \sum_{i=1}^q [y_{ij} - \eta(x_i, \theta)]^2 + (\theta - \theta_0)^T \mathcal{D}^{-1}(\theta - \theta_0) \right\}. \tag{6}$$

The estimator turns out to be

$$\tilde{\theta}_j = (\mathcal{D}^{-1} + \sigma^{-2}M)^{-1}(\mathcal{D}^{-1}\theta_0 + \sigma^{-2}Y_j). \tag{7}$$

It is biased in the sense that $E_\epsilon\{\tilde{\theta}_j\} = (\mathcal{D}^{-1} + \sigma^{-2}M)^{-1}(\mathcal{D}^{-1}\theta_0 + M\theta_{jt}) = \theta_{jt} + (\mathcal{D}^{-1} + \sigma^{-2}M)^{-1}\mathcal{D}^{-1}(\theta_0 - \theta_{jt})$. Nevertheless, the unconditional dispersion matrix of $\tilde{\theta}_j$ is

$$E\{(\tilde{\theta}_j - \theta_{jt})(\tilde{\theta}_j - \theta_{jt})^T\} = (\mathcal{D}^{-1} + \sigma^{-2}M)^{-1}. \tag{8}$$

Having in mind that $E\{(\hat{\theta}_j - \theta_{jt})(\hat{\theta}_j - \theta_{jt})^T\} = \sigma^2 M^{-1}$, we find that "in total" $\tilde{\theta}_j$ has a smaller dispersion matrix than $\hat{\theta}_j$:

$$(\mathcal{D}^{-1} + \sigma^{-2}M)^{-1} < \sigma^2 M^{-1}.$$

For comparing of $\tilde{\theta}_j$ with $\hat{\theta}_j$, it should also be noted that

$$E\{(\tilde{\theta}_j - \theta_0)(\tilde{\theta}_j - \theta_0)^T\} = \mathcal{D}^{-1} - (\mathcal{D}^{-1} + \sigma^{-2}M)^{-1} = \mathcal{D}(\mathcal{D} + \sigma^2 M^{-1})^{-1}\mathcal{D}. \tag{9}$$

2.1.3 θ_0, σ^2, and $\mathcal{D}$ are unknown

Taking into account that (2) and (3) do not depend upon θ_0, σ^2, and $\mathcal{D}$, and in order to stay in the class of quadratic unbiased estimators, we propose the following estimators:

$$\hat{\sigma}^2 = r^{-1}(q - m)^{-1} \sum_{j=1}^{r} \sum_{i=1}^{q} [y_{ij} - \hat{\theta}_j f(x_i)]^2, \tag{10}$$

$$\hat{\mathcal{D}} = (r - 1)^{-1} \sum_{j=1}^{r} (\hat{\theta}_j - \hat{\theta}_0)(\hat{\theta}_j - \hat{\theta}_0)^T - \hat{\sigma}^2 M^{-1} \tag{11}$$

[see, e.g., Spjøtvoll (1977)]. Unfortunately, (11) does not necessarily satisfy the natural condition for $\hat{\mathcal{D}}$ to be positive definite for any sample $\{y_{ij}\}_{1,1}^{q,r}$. In applications this fact can cause difficulties, e.g., if (11) is used as a substitute for $\mathcal{D}$ in numerical procedures to derive an optimal experimental design.

Having a sufficient number of observations in a learning sample we can use (3), (10) and (11) to substitute θ_0, σ^2, and $\mathcal{D}$, respectively, in (7) by their estimates. We could not find any studies about the "reliability" of this substitution for small samples; however, it is evident that asymptotically (for $r \to \infty$ and $q \to \infty$) in the learning sample, we will obtain the same results as if we used the "true" θ_0 and $\mathcal{D}$ (see also Section 2.3).

2.2   The nonlinear case: A quasilinear approach

The results of the last section can be generalized for the nonlinear parametrization of the response function $\eta(x, \theta)$. If we use

$$\hat{\theta}_j = \arg\min_{\theta} \sum_{i=1}^{q} [y_{ij} - \eta(x_i, \theta)]^2 \tag{12}$$

4

instead of (2) and the $\hat{\theta}_j$ are strongly consistent [see Jennrich (1969)], then all results from Section 2.1 are asymptotically (for $q \to \infty$) valid; in calculating M, the basic functions $f(x)$ have to be changed to $f(x, \hat{\theta}_j)$ where $f(x, \hat{\theta}) = \partial \eta(x, \theta)/\partial \theta|_{\theta = \hat{\theta}}$.

Application of this general rule to (4) requires some precaution, because (4) implicitly assumes that $E\{y_{ij}\} = \eta(x_i, \theta_0)$. This is fulfilled in the linear case but is not generally true. To illustrate this let us assume that

$$E_\theta\{(\theta_j - \theta_0)(\theta_j - \theta_0)^T\} = \mathcal{D} = \gamma^2 D$$

where D is the dispersion matrix of the corresponding normalized random values (i.e., all diagonal elements of D are 1). For any symmetric distribution of θ we get

$$
\begin{aligned}
E\{y_{ij}\} &= E_\theta\{\eta(x_i, \theta_j)\} \\
&= E_\theta\{\eta(x_i, \theta_0) + \gamma \operatorname{tr} \left. \frac{\partial \eta(x_i, \theta)}{\partial \theta} \right|_{\theta = \theta_0} (\theta - \theta_0)^T \\
&\quad + \frac{\gamma^2}{2} \operatorname{tr}(\theta - \theta_0)^T \left. \frac{\partial^2 \eta(x_i, \theta)}{\partial \theta \, \partial \theta^T} \right|_{\theta = \theta_0} (\theta - \theta_0) + \dots\} \\
&= \eta(x_i, \theta_0) + \frac{\gamma^2}{2} \operatorname{tr} D \left. \frac{\partial^2 \eta(x_i, \theta)}{\partial \theta \, \partial \theta^T} \right|_{\theta = \theta_0} + O(\gamma^4)
\end{aligned}
$$

If the first two terms in the last line are of similar size then in order to achieve the accuracy of the above expansion we should use

$$\hat{\hat{\theta}}_0 = \arg \min_\theta \sum_{j=1}^r \sum_{i=1}^q \sigma^{-2} \left[y_{ij} - \eta(x_i, \theta) - \frac{1}{2} \operatorname{tr} \mathcal{D} \frac{\partial^2 \eta(x_i, \theta)}{\partial \theta \, \partial \theta^T} \right]^2 \tag{13}$$

instead of (5). In the approach discussed in Section 2.1.1 we can use (6) but not (7) in the nonlinear case.

2.3 Maximum likelihood method and iterative estimators

In this section we assume that ε_{ij} and θ_j are normally distributed. Neglecting additive constants, we find

$$R(y, \Theta, \sigma^2, \theta_0, \mathcal{D}) = -2 \ln L(y, \Theta, \sigma^2, \theta_0, \mathcal{D}) = rq \ln \sigma^2 + r \ln |\mathcal{D}| \tag{14}$$
$$+ \sum_{j=1}^r \sum_{i=1}^q \sigma^{-2} [y_{ij} - \eta(x_i, \theta_j)]^2 + \sum_{j=1}^r (\theta_j - \theta_0)^T \mathcal{D}^{-1} (\theta_j - \theta_0),$$

where L is the likelihood function, $\Theta = (\theta_1, \dots, \theta_r)$, and y is the $q \times r$-matrix $\{y_{ij}\}_{1,1}^{q,r}$.

If σ^2, θ_0, and $\mathcal{D}$ are known, minimization of (14) immediately leads to (6), or to (7) if $\eta(x, \theta) = \theta^T f(x)$. More interesting results are obtained if some of the quantities σ^2, θ_0, and $\mathcal{D}$ are unknown.

2.3.2 σ^2, θ_0, and $\mathcal{D}$ are unknown

If σ^2, θ_0, and $\mathcal{D}$ are unknown, setting the derivatives of R to zero gives the estimators

$$\bar{\Theta} = \arg\min_{\Theta}[rq\ln\tilde{\sigma}^2(\Theta) + r\ln|\tilde{\mathcal{D}}(\Theta)|], \tag{20}$$

where

$$\tilde{\sigma}^2(\Theta) = (rq)^{-1}\sum_{j=1}^{r}\sum_{i=1}^{q}[y_{ij} - \eta(x_i,\theta_j)]^2$$

$$\tilde{\mathcal{D}}(\Theta) = r^{-1}\sum_{j=1}^{r}[\theta_j - \tilde{\theta}_0(\Theta)][\theta_j - \tilde{\theta}_0(\Theta)]^T$$

$$\tilde{\theta}_0(\Theta) = r^{-1}\sum_{j=1}^{r}\theta_j.$$

The fixed point method leads to

$$\bar{\Theta} = \lim_{s\to\infty}\Theta_s, \tag{21}$$

$$\theta_j(s+1) = \arg\min_{\theta}\left\{\tilde{\sigma}^{-2}(\Theta_s)\sum_{i=1}^{q}[y_{ij} - \eta(x,\theta)]^2 \right.$$
$$\left. + [\theta - \tilde{\theta}_0(\Theta_s)]^T\tilde{\mathcal{D}}^{-1}(\Theta_s)[\theta_j - \tilde{\theta}_0(\Theta_s)]\right\}.$$

Comparing (21) with (6) shows that the proposed estimator for θ_j is an iterative version of (6) and that the two-stage procedure discussed in Section 2.1.3 coincides with the first iteration of (21). Asymptotically (for $r \to \infty$) $\bar{\theta}_j$ has the same properties as (7) in the linear case, i.e., if $\eta(x,\theta) = \theta^T f(x)$. In the nonlinear case the asymptotical behavior of $\bar{\theta}_j$ coincides with that of (7). This is due to the strong consistency of $\tilde{\sigma}^2$ and $\tilde{\mathcal{D}}$ (for their consistency it is sufficient that $r \to \infty$ while q can be finite). For the consistency of $\bar{\theta}_j$ it is necessary that $q \to \infty$. These facts can be proved by means of the techniques used by Jennrich (1969), Malyutov (1982), and Wu (1981).

Moreover, this consistency can be proved—under rather mild conditions—for non-Gaussian distributions of ε_{ij} and θ_j (e.g., the distributions must have finite second moments; for the asymptotical normality they must have finite third moments). In the nonlinear case the most crucial assumption for the consistency of $\bar{\theta}_j$ is as follows:

The minimization in (17), (18), (20), and (21) has to be done over some compact Ω (i.e., $\theta_j \in \Omega$), the true value θ_{jt} is an interior point of Ω, and uniformly on Ω

$$\lim_{q\to\infty}q^{-1}\sum_{i=1}^{q}[\eta(x_i,\theta) - \eta(x_i,\theta_t)]^2 = v^2(\theta),$$

where the function $v^2(\theta)$ has a unique minimum at point θ_t.

7

2.3.1 Only σ^2 is unknown

Let us first discuss the case when only σ^2 is unknown. The maximum likelihood estimators for σ^2 and Θ are

$$\tilde{\sigma}^2 = (rq)^{-1} \sum_{j=1}^{r} \sum_{i=1}^{q} [y_{ij} - \eta(x_i, \theta_j)]^2$$

and

$$\tilde{\Theta} = \arg\min_{\Theta} \left\{ rq \ln(rq)^{-1} \sum_{j=1}^{r} \sum_{i=1}^{q} [y_{ij} - \eta(x_i, \theta_j)]^2 + \sum_{j=1}^{r} (\theta_j - \theta_0)^T \mathcal{D}^{-1} (\theta_j - \theta_0) \right\}, \quad (15)$$

respectively. At the first glance, (15) looks quite different from previously considered estimators. Applying the fixed point method leads to the following iterative procedure:

$$\tilde{\Theta} = \lim_{s \to \infty} \Theta_s \tag{16}$$

$$\Theta_s = \arg\min_{\Theta} \left\{ \sigma_s^{-2} \sum_{j=1}^{r} \sum_{i=1}^{q} [y_{ij} - \eta(x_i, \theta_j)]^2 + \sum_{j=1}^{r} (\theta_j - \theta_0)^T \mathcal{D}^{-1} (\theta_j - \theta_0) \right\}. \tag{17}$$

The optimization problem (17) can be splitted into r separated optimization problems: For $j = 1, \ldots, r$

$$\theta_j(s) = \arg\min_{\theta} \left\{ \sigma_s^{-2} \sum_{i=1}^{q} [y_{ij} - \eta(x_i, \theta)]^2 + (\theta - \theta_0)^T \mathcal{D}^{-1} (\theta - \theta_0) \right\}. \tag{18}$$

In (17) and (18)

$$\sigma_s^2 = (rq)^{-1} \sum_{j=1}^{r} \sum_{i=1}^{q} \{y_{ij} - \eta[x_i, \theta_j(s)]\}^2.$$

A possible choice of initial values for (16) is

$$\theta_j(0) = \hat{\theta}_j = \arg\min_{\theta} \sum_{i=1}^{q} [y_{ij} - \eta(x_i, \theta)]^2, \quad j = 1, \ldots, r. \tag{19}$$

Comparing (18) and (6) shows that (16) and (18) can be considered as an iterative version of (6). Thus, the maximum likelihood approach gives a hint how (6)—independently of whether ε_{ij} and θ_j follow a normal distribution or not—can be adjusted to the case where the variance of the observation errors is unknown. Malyutov (1982) discussed asymptotics of the iterative estimators for the regression model with constant parameters. The generalization to the second kind regression model is straightforward.

6

Inspite of its elegance, the m.l.e. or iterative estimators require rather extensive calculations. Even in simple cases and when the procedure is started from initial values (2), (3), (10), and (11), two to four iterations are needed in practice.

Iterative estimation based on generalized least squares and various applications of related algorithms are discussed in a surveying paper by del Pino (1989).

3 Design of Experiments

The problem of optimal choice of the regressors $\xi_q = (x_1, \ldots, x_q)$ is usually referred to as the design problem. For the various estimators defined in Section 2 we have to expect different solutions of the design problem since the optimum will be attained at

$$\xi_q^* = \arg\min_{\xi_q} \Psi[D(\xi_q)], \quad x_i \in \mathcal{X} \tag{22}$$

where $D(\xi_q)$ denotes the dispersion matrix for the respective estimation problem [see the comments next to (2), and (4) and (8)]. Here, $\mathcal{X}$ stands for the design region and Ψ is the design criterion that is used to reduce the multiobjective problem "minimization of $D(\xi_q)$" to a scalar task. Possible choices for Ψ are discussed by Silvey (1980). For the sake of simplicity we restrict ourselves to the popular D-criterion, i.e., $\Psi[D(\xi_q)] = \ln |D(\xi_q)|$.

According to the general theory [see Silvey (1980)] the supporting points of ξ^* have to be located at the maxima of the so-called dual function $\psi(x, \xi)$. For the situations of interest, this function is given in the respective section below [for more details, see Fedorov and Müller (1989)].

For second kind regression models the question of optimal design is more difficult than for the standard case. The additional difficulties come from the fact that we have more parameters (θ_0, θ_j, σ^2, and $\mathcal{D}$) to estimate. Analogously to Section 2, we will consider several situations of prior knowledge and start with the linear parametrization.

3.1 No prior information–initial experiments

3.1.1 Estimation of θ_j

Having no prior knowledge about θ_0, σ^2, and $\mathcal{D}$, it is evident from Sections 2.1.1 and 2.1.3 that in order to perform an optimal experiment we have to look for designs that minimize

$$\Psi[D(\xi_q)] = \ln |M^{-1}| = \ln |[\sum_{i=1}^{q} f(x_i) f^T(x_i)]^{-1}| . \tag{23}$$

This means that we look for the most accurate estimates for θ_j. The dual function used by the optimization algorithm is given by

$$\psi(x, \xi) = f^T(x) M^{-1} f(x) . \tag{24}$$

8

Of course, in later experiments information about θ_0, σ^2, and $\mathcal{D}$ that is gained in this "initial" or "starting" experiment will be used in deriving optimal designs.

3.1.2 Estimation of θ_0

Minimizing (23) also affects the quality of the estimator $\hat{\theta}_0 = r^{-1} \sum_{j=1}^r \hat{\theta}_j$. The dispersion matrix of $\hat{\theta}_0$,

$$E\{(\hat{\theta}_0 - \theta_0)(\hat{\theta}_0 - \theta_0)^T\} = r^{-1}(\mathcal{D} + \sigma^2 M^{-1}), \tag{25}$$

is a monotonically increasing function of M^{-1} in the sense of matrix ordering. As $\mathcal{D}$ and σ^2 are unknown nothing better can be proposed than to minimize $|M^{-1}|$, i.e., to use the criterion (23) for finding the optimal design. In the case of no prior information the optimal design is the same for estimating both θ_0 and θ_j.

3.2 Preliminary information available

3.2.1 Estimation of θ_0

Let us assume that reliable estimates for $\mathcal{D}$ and σ^2 are given, e.g., values obtained from (10) and (11). Then

$$\Psi[D(\xi_q)] = \ln |\hat{\mathcal{D}} + \hat{\sigma}^2 M^{-1}| \tag{26}$$

has to be used as appropriate design criterion. The corresponding dual function is

$$\psi(x, \xi) = f^T(x) M^{-1} [\hat{\mathcal{D}} + \hat{\sigma}^2 M^{-1}]^{-1} M^{-1} f(x). \tag{27}$$

3.2.2 Estimation of θ_j

Let us assume that estimates such as (3), (10), and (11)—or other reliable estimates— are available for θ_0, $\mathcal{D}$, and σ^2, respectively, and that we want to estimate θ_ℓ for a new object ℓ, $\ell \notin \{1, \ldots, r\}$. We have the alternatives to use an approach based on either (2) or (7). The difference between the two approaches is discussed in Section 2.1.2. If it is reasonable to believe that the new object belongs to the group of objects that is basis of the preliminary information on θ_0, $\mathcal{D}$, and σ^2, then using

$$\Psi[D(\xi_q)] = \ln |\hat{\mathcal{D}}^{-1} + \hat{\sigma}^{-2} M|^{-1} \tag{28}$$

makes the estimate $\tilde{\theta}_\ell$ most accurate [On the other hand, if the new object is studied in a "stand alone" mode, use (23).]

At the same time, the asymptotically minimal bias for the estimate of $\mathcal{D}$ is achieved if (28) is used. This follows from the fact that the asymptotical bias of the iterative estimator $\tilde{\mathcal{D}}$ is $(\mathcal{D}^{-1} + \sigma^{-2} M)^{-1}$ [cf. (9)]. The bias vanishes when q and therefore the

elements of M become very large. The rate of this convergence is determined by the locations $x_1, \ldots, x_q$, and (28) guarantees the best rate in the sense of the determinant value.

For criterion (28) we obtain the dual function

$$\psi(x, \xi) = f^T(x)[\hat{\mathcal{D}}^{-1} + \hat{\sigma}^{-2}M]f(x). \tag{29}$$

3.3 Nonlinear problems

In the light of the discussion in Sections 2.2 and 2.3, the design rules of the last two sections can be applied straightforwardly to nonlinear problems. A crucial point is that the basic function $f(x, \theta)$ depends on θ. If a reliable estimate $\hat{\theta}_0$ is known, $\partial \eta(x, \theta)/\partial \theta|_{\hat{\theta}_0}$ should be used for $f(x, \theta_0)$. To take the stochastic nature of the parameter estimate $\hat{\theta}_0$ into account, we can follow one of two approaches: (a) find the optimal designs for a variety of plausible values of θ_0 and then merge them to obtain an "average" design [see Fedorov et al. (1987)]; (b) use a Bayesian or a minimax procedure [see, e.g., Ford et al. (1989)].

4 An Application

Stephan (1940) already observed that if plaque on the tooth surface is exposed to carbohydrate the plaque pH rapidly decreases to a minimum and lateron increases slowly to approach the initial pH value; this pattern or profile of the dental plaque pH is called the "Stephan curve" (see Figure 1). Measurement of the Stephan curve is of particular interest as its characteristics are used, e.g., as indicators of the varying acidogenicity of various substances in our nutrition.

insert Figure 1

A usual approach of measuring the Stephan curve is to take six to ten pH values at regular time intervals of three to five minutes. We will demonstrate in the following that the use of the earlier discussed refinements in the estimation procedure as well as in the sampling design leads to considerably better estimates as the standard approach.

insert Table 1

A suitable model to describe the Stephan curve is

$$\text{pH}(t) = \theta_1 \left[1 + \exp\left(-\theta_2 t\right) - \exp\left(-\theta_2 t\right)\right] \tag{30}$$

Figure 1 shows measurements (cf. Table 1) taken from twelve individuals at constant time intervals of four minutes over a time-span of 28 minutes after rinsing with

sucrose, together with the fit of (30) to these data ($\hat{\theta}_1 = 7.01$, $\hat{\theta}_2 = 0.62$, $\hat{\theta}_3 = 0.39$). The figure indicates that the observed pH-values are considerably scattered around the mean curve. This suggests to use a regression model with stochastic parameters: For the j-th individual the mean value function is

$$\mathrm{pH}_j(t) = \eta(t, \theta_j) = \theta_{1j} \left[1 + e^{-\theta_{2j}t} - e^{-\theta_{3j}t} \right] ; \tag{31}$$

where $\theta_j = (\theta_{1j}, \theta_{2j}, \theta_{3j})^T$ is a random vector with $\mathrm{E}\{\theta_j\} = \theta_0$ and $\mathrm{Var}\{\theta_j\} = \mathcal{D}_0$. The increase in the explained variation due to the stochastic nature of the parameters is significant. Assuming normality of the deviations between observations and model, we find that the test statistic of the likelihood ratio test has the value 213.0, corresponding to a highly significant p-value in the table of the asymptotic $\chi^2(33)$-distribution.

To construct optimal designs for estimating θ_j and θ_0 by means of the criteria (23), (26), and (28) we use classical numerical algorithms [see, e.g., Fedorov et al. (1987)] that are based on the dual functions (24), (27), and (29), respectively. Our design space $\mathcal{X}$ is a grid of width 0.5 on the interval $[0, 8]$, corresponding to a time unit of four minutes. Comparisons of the performance of various designs ξ will be based on the design criterion $\ln |D(\xi)|$. In addition, the maximal variance of the response functions will be given. According to the equivalence theorem of optimal design theory the optimal design consists of those points where the variance function has its maxima, and for such a design both the design criterion and the maximal variance are minimized.

First we consider the equally spaced ("uniform") design of eight measurements that traditionally is used by dentists. It will serve as a reference design for preformance comparisons. We also use it as a starting design for the numerical algorithms. Estimation of θ_j and θ_0 on the basis of the uniform design without making use of prior information corresponds to a value 2.55×10^{-5} of the design criterion (23). The maximal variance of the response function is 0.45.

The optimal design for estimating θ_j and θ_0 without making use of prior information, i.e., the design obtained from minimizing (23), is shown in Figure 2. This figure— and Figures 3 to 5—are constructed as follows. The vertical bars indicate the design points. The height of each bar is proportional to the weight that is given to the observation at this point; the weight can be read from the scale on the left axis. The small circles indicate the variance of the response function at specific points; the maximum of the variance is given on the right axis.

$$\boxed{\text{insert Figure 2}}$$

The optimal design shown in Figure 2 does not consist of three supporting points with equal weights as it must be expected from optimal design theory: One of the points is splitted due to the roughness of the underlying grid. If we again take a sample of eight observations the value of criterion (23) is 0.78×10^{-5}, a reduction to

11

one third (30%) of the value for the uniform design. The maximal variance of the response function is 0.10, more than four times less the value for the uniform design.

Of course, in practice it is impossible to locate eight observations at four points (with unequal weights). A feasible design has to fulfill the restriction that the number of supporting points is finite, and they have to have equal weights. This affords the use of special algorithms, different from those usually found in literature. Such an algorithm for finding restricted optimal designs is described by Fedorov and Müller (1989).

For estimating θ_j or θ_0 without making use of prior information this algorithm yields the design shown in Figure 3. The value of criterion (23) is now 1.15×10^{-5}, i.e., 45% of the criterion value for the uniform design. The maximal variance of the response function is 0.19. The restricted optimal design is slighly deteriorated as compared to the unrestricted design but still considerably better than the reference design.

$$\boxed{\text{insert Figure 3}}$$

It should be noted that the unrestricted and the restricted optimal designs do not depend on how far we expand the design region. Although the variance function increases slightly at the right end and approaches the indicated maximal value, the expansion of the design region leaves the optimal design unchanged. We worked out but do not report corresponding results for the design region $[0, 21]$. The asymptotic behavior of the variance function, however, makes clear that the an unrestricted design region would lead to an additional supporting point very far ahead in time.

So far, we took advantage from improving the sampling design. Next we will show that further improvement can be gained by incorporating prior information (see Section 3.2). Figure 4 shows the restricted optimal design for estimating θ_j making use of estimates for the error variance σ^2 and the dispersion matrix $\mathcal{D}$, i.e., obtained from minimizing the design criterion (28). The value of the criterion (28) is 0.29×10^{-5}, that is 25% of the value of criterion (23) for the restricted design without use of prior information, and only 11% of the criterion value for the uniform design. The maximal variance of the response function is 0.11.

$$\boxed{\text{insert Figure 4}}$$

Finally, Figure 5 gives the restricted optimal design for estimating θ_0 making use of prior information, derived by minimizing criterion (26). The value of the design criterion is 0.44×10^{-5}, that is 38% and 17% of the values of criterion (23) for the restricted design without use of prior information and for the uniform design, respectively.

$$\boxed{\text{insert Figure 5}}$$

The example demonstrates the improvements that can be achieved (a) through using the optimal design or a design that is close to the optimal, and (b) through taking into account prior information. In practice the gain in the design criterion will lead to savings in the sample size. Figure 6 shows the optimality criteria for various sample sizes and different estimation schemes.

```
insert Figure 6
```

Similar improvements can be expected for the estimation of characteristics of the model such as the minimum of or the area under the profile.

References

del Pino, G.E. (1989). The unifying role of iterative generalized least squares in statistical algorithms. *Statistical Science*, 4, 394-408.

Fedorov, V.V., Leonov, S., and Pitovranov, S. (1988). Experimental design technique in the optimization of a monitoring network, pp. 165-175 in V.V. Fedorov and H. Läuter (Eds.), *Model-Oriented Data Analysis*. Berlin: Springer-Verlag.

Fedorov, V.V. and Müller, W. (1989). Comparison of two approaches of optimal design of an observation network. *Statistics*, 20, 339-351.

Fisk, P.R. (1967). Models of the second kind in regression analysis. *Journal of the Royal Statistical Society*, B-29, 266- 278.

Ford, I., Titterington, D.M., and Kitsos, C.P. (1989). Recent advances in nonlinear experimental design. *Technometrics*, 31, 49-60.

Gladitz, J. and Pilz, J. (1982). Construction of optimal designs in random coefficient regression models. *Mathematische Operationsforschung und Statistik, Serie Statistik*, 13, 371-385.

Jennrich, R.J. (1969). Asymptotic properties for nonlinear least squares estimators. *Annals of Mathematical Statistics*, 40, 633-643.

Malyutov, M.B. (1982). Asymptotical properties and applications of the IRGINA-estimator of parameters of generalized regression model, pp. 144-158 in *Stochastic Processes and Applications*. Moscow.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*. New York: John Wiley.

Silvey, S.D. (1980). *Optimal Design*. London: Chapman and Hall.

Spjøtvoll, E. (1977). Random coefficient regression models. A review. *Mathematische Operationsforschung und Statistik, Serie Statistik*, 8, 69-93.

Stephan, R.M. (1940), Changes in hydrogen-ion concentration in tooth surfaces and in caries lesions. *Journal of the American Dental Association*, 27, 718-722.

Wu, C.F. (1981). Asymptotic theory of nonlinear least squares estimation. *Annals of Statistics*, **9**, 501-513.

Table 1: Measurements of plaque pH values from twelve individuals after rinsing with sucrose.

	0 min	4 min	8 min	12 min	16 min	20 min	24 min	28 min
1	7.30	6.24	5.78	5.94	5.79	6.51	7.01	6.70
2	6.45	6.01	5.72	5.60	5.47	5.29	6.46	6.50
3	6.97	5.65	5.41	5.42	5.90	6.35	6.40	6.61
4	6.85	6.45	6.10	6.13	6.55	6.32	6.81	6.56
5	7.41	6.68	6.42	6.57	6.45	6.94	7.01	7.08
6	7.01	6.41	6.42	6.69	6.70	6.92	6.96	6.92
7	6.84	6.30	6.17	6.09	6.35	6.51	6.76	6.81
8	6.59	5.97	5.37	5.44	5.64	6.06	6.34	6.31
9	6.96	5.65	5.57	5.54	5.50	5.82	6.16	6.35
10	6.94	5.25	4.93	5.08	5.02	5.38	5.69	5.59
11	6.79	6.10	5.86	6.20	6.63	6.50	6.40	6.85
12	7.50	6.16	5.94	6.10	6.55	6.97	7.15	7.20
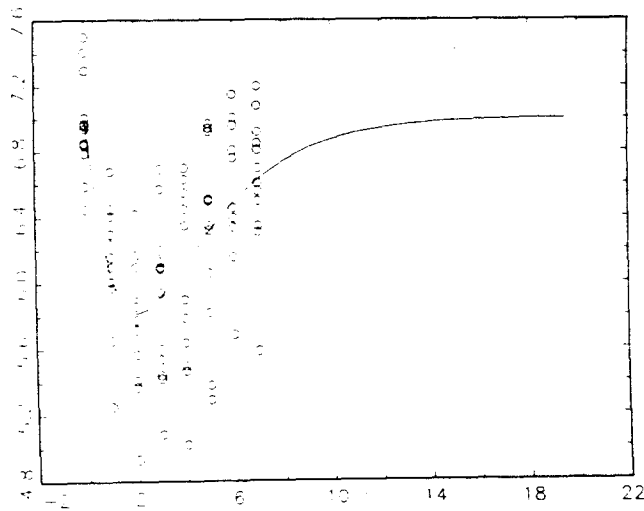
Figure 1: Measurements of plaque pH values after rinsing with sucrose, fitted curve (30) and error bands.
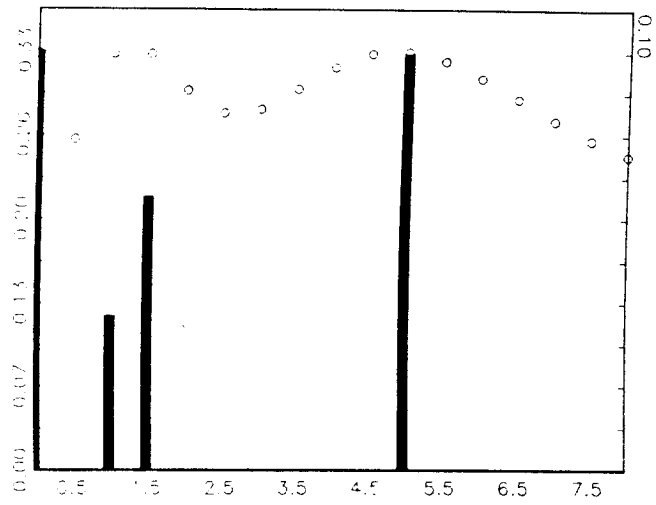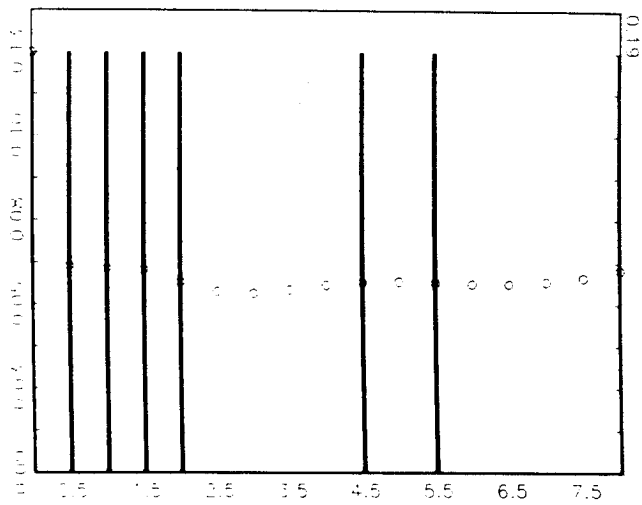
Figure 2: Optimal design for (23).

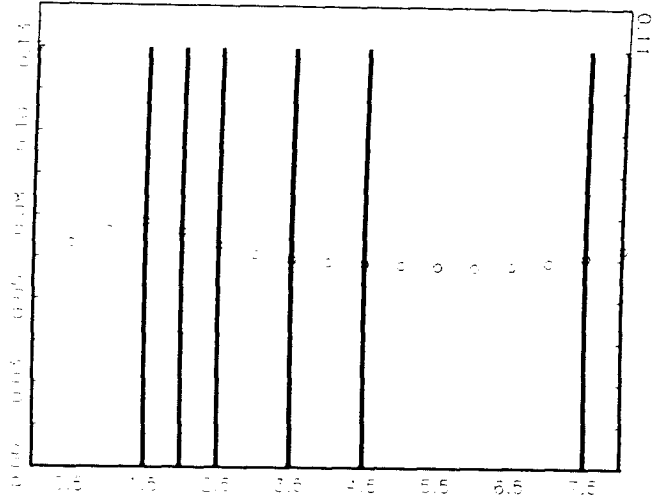

Figure 3: Restricted optimal design for (23).

16

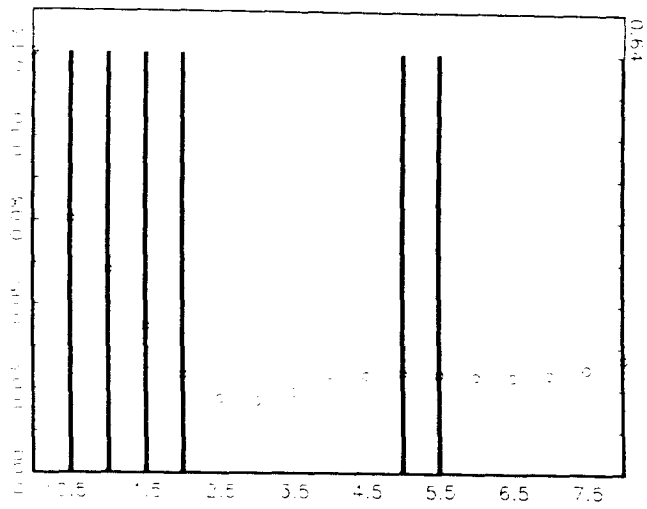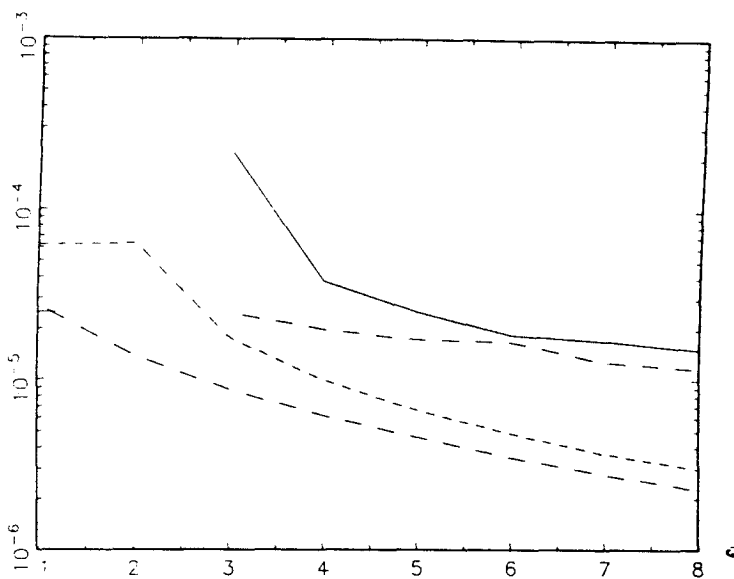*Figure 4:* Restricted optimal design for (28).

Figure 5: Restricted optimal design for (26).

17

Figure 6: Design criteria for estimating θ_j as a function of sample size q. The solid and wide dashed polygon is obtained for the uniform design without use and using prior information, respectively; The narrow and dot-dashed polygon is obtained for the optimal design without use and using prior information, respectively.