Philip Hougaard

Parameter Transformations in Multiparameter Nonlinear Regression Models

Preprint
February
1984

2

Institute of Mathematical Statistics
University of Copenhagen

Philip Hougaard

PARAMETER TRANSFORMATIONS IN MULTIPARAMETER NONLINEAR REGRESSION MODELS

ABBREVIATED TITLE: PARAMETERS IN NONLINEAR REGRESSION

Preprint 1984 No. 2

INSTITUTE OF MATHEMATICAL STATISTICS
UNIVERSITY OF COPENHAGEN

February 1984

Summary In a oneparameter nonlinear regression it is possible to find a parametrization, which has bias and skewness of lower order than usual, constant asymptotic variance and normal likelihood. In a multiparameter model a parametrization with these properties does not always exist. An easily applicable condition for the existence is found. Three classes of special models, for which it does exist, are examined. In the general case the nonlinearity is often marked and there is a need for finding a parameter, which reduces the nonlinearity. It is suggested to consider transformations of a single parameter because this is easier to interpret than general transformations. The solution will in general depend on which nonlinearity effect is considered most important.

American Mathematical Society 1980 subject classifications. Primary 62J02; Secondary 62F12.

Key words: asymptotic bias; asymptotic skewness; covariance stabilization; curvature; normal likelihood.

1. INTRODUCTION

Inference in statistical models should ideally be invariant under parameter transformations or we can say inference should depend on the model but not the way the model is parametrized. This is, however, not always possible and in practice an analysis of a nonlinear model is made using a chosen parametrization of the model. The purpose of the present paper is to examine how much can be achieved by transforming the parameter. In practice parameters are often chosen because they make physical sense or they are traditionally used in the subject area for the statistical analysis. The present approach should be viewed in this context.

We will focus on the nonlinear regression model, i.e. assume that the n-dimensional observation Y is distributed according to $N_n(\eta(\theta), \sigma^2 I)$, the n-dimensional normal distribution with mean vector $\eta(\theta)$, where $\eta(\cdot)$ is a known function of the p-dimensional parameter θ and the components are independent all with variance σ^2. An alternative description is $Y_i = \eta_i(\theta) + \varepsilon_i$, where the ε_i's are i.i.d. $N(0,\sigma^2)$. Often the model is described by means of covariates $x_1,\ldots,x_n \in \mathbb{R}^k$, such that there is a function $\xi(\cdot)$, with $\eta_i(\theta) = \xi(\theta;x_i)$. Later we will omit the index θ in the meanvalue and the derivatives thereof.

The calculations could be made for the more general case of a curved exponential family. However for the interpretation of the results and the comparison between properties the nonlinear regression is simpler, cf. Hougaard (1982) and Kass (1984).

If the parameter is onedimensional, i.e. $p = 1$, there is, cf. Hougaard (1982), a natural parameter, say β, for which the asymptotic variance is constant, the asymptotic skewness is 0, the asymptotic bias is 0 and the likelihood is approximately normal, cf. Sprott (1973). Using as well

Beale's (1960) measure of nonlinearity as Bates & Watts (1980) curvature, this parameter has zero parametrization curvature. The parameter has a geometrical interpretation as the arc length in the solution locus. The transformation $\beta = g(\theta)$ from the original parameter can be found from the differential equation $g'(\theta) = c\ (\dot{\eta}'\dot{\eta})^{1/2}$, where $\dot{\eta}$ is the vector of derivatives of η w.r.t. θ and c an arbitrary constant.

For multiparameter models such a nice parametrization does not exist in general. It was mentioned by Reeds in the discussion of Efron (1975) that if the Riemannian curvature vanishes identically a covariance stabilizing transformation exists. Holland (1973) examined the same problem and found a condition for existence when the parameter is twodimensional, but it involves the choice of a square root of the information matrix. In Section 2 we derive a simple condition for the existence in the nonlinear regression for arbitrary p. If such a parametrization exists, it is optimal in the sense that the measures of curvature, Beale (1960) and Bates & Watts (1980) has zero parameter effects, the asymptotic third central moments and the asymptotic biases are zero and the likelihood is normal, cf. Hougaard (1981). In some simple but common models a covariance stabilizing transformation exists, cf Section 4.

Most of this paper will consider the situation, when such a parametrization does not exist or is too complicated to handle or interpret. The question considered is: Is it possible to find a simple transformation, which, although it is not optimal, lowers the parametrization effects or makes some specific part of the parametrization effect vanish. We will try to find an "optimal" solution within a smaller class of transformations. In Section 3 we will discuss what should be meant by "optimal" and derive the corresponding solutions. Suppose θ_1 is the important parameter and $\theta_2,\ldots,\theta_p$ are nuisance parameters or alternatively that θ_1 has independent interest. In that case we can transform θ_1 alone or $\theta_2,\ldots,\theta_p$ alone, but we would not accept transformations like θ_1/θ_2, because they make it difficult to infer about θ_1. A transformation of the kind $\beta = g(\theta_1)$ might give a parametrization β, $\theta_2,\ldots,\theta_p$ for which inference about β is simple, e.g. asymptotic probability statements about $\hat{\beta}$ might be relatively precise. Thus it is simple to transform such statements to statements about $\hat{\theta}_1$. This idea was used by Sprott(1980) in a generalisation of normal likelihood to the situation of nuisance parameters.

In Section 4 three special cases and two examples are considered. In these cases it is possible to find transformations of the important parameter, which removes some nonlinearity effect, and these transformations are independent of the nuisance parameters.

As the choice of optimality criterion is important a comparison of the different criterias is discussed in Section 5.

2. EXISTENCE OF A COVARIANCE STABILIZING TRANSFORMATION

The question in this section is: Does there exist a covariance stabilizing transformation or more precisely does these exist a parametrization with constant information matrix. This can also be formulated as: Is the solution locus isometric to a Euclidean space? In the discussion of Efron (1975) Reeds mentioned that this is the case if the Riemannian curvature vanishes identically. Following this line we find the following result.

Theorem In a nonlinear regression with η three times continuously differentiable in a simply connected parameter set Θ, a covariance stabilizing transformation exists if and only if, for all i,j,k,l and all values of θ

$$\ddot{\eta}'_{ij}(I-P)\ddot{\eta}'_{kl} = \ddot{\eta}'_{il}(I-P)\ddot{\eta}_{kj},$$

where $\ddot{\eta}_{ij}$ is the vector of second derivatives $\partial^2\eta/\partial\theta_i\partial\theta_j$ and P the projection onto the tangent space for the solution locus in θ. If it exists all the functions $h(\theta_1,\ldots,\theta_p)$ giving a coordinate of the transformed parameter will satisfy the same differential equation

$$\ddot{h}_{ij} = \dot{h}\,(\dot{\eta}'\dot{\eta})^{-1}\dot{\eta}'\ddot{\eta}_{ij} \tag{2.1}$$

Proof Using g as the tensor corresponding to the information matrix, Sokolnikoff (1951, p.99) proves that there exists a parametrization having constant information if and only if the Riemann - Christoffel tensor is zero identically. Inserting the values for a nonlinear regression gives the desired equation for existence. For the transformed parameter the second derivatives of η should be orthogonal to the first derivatives. Expressing this in terms of h and the derivatives with respect to θ yields the differential equation.

Remarks For $p = 1$, the condition is trivially true and a variance stabilizing transformation exists. For $p = 2$ there is only one combination of i,j,k,l, for which the equation is not automatically true and that is

$$\ddot{\eta}'_{11}(I-P)\ddot{\eta}_{22} = \ddot{\eta}'_{12}(I-P)\ddot{\eta}_{12}$$

The condition in Holland(1973) is computationally more involved, but it can be proven to be identical to this condition for the nonlinear regression. A covariance stabilizing transformation will yield a parametrization with zero Bates & Watts (1980) parameter curvature and minimal Beale (1960) measure of nonlinearity, cf Hougaard (1981). In general it is only possible to obtain that in a single point. Some special cases, where a covariance stabilizing transformation exists are considered in Section 4.

Bates & Watts (1981) examined the equation (2.1) assuming that the coefficients of $\dot{h}$ were constant and equal to the value at $\hat{\theta}$. They found a condition for existence of solutions to this local equation. Usually there are no solutions to the equation in Special case 2, Section 4, showing an advantage of considering the original equation (2.1).

3. PARAMETER TRANSFORMATION RESULTS

As was found in Section 2 it is only in special cases possible to find a globally optimal parametrization. And if one exists, it might be complicated, mathematically intractable or impossible to interpret. In this section we consider transformations in a smaller class, giving parameters, which it is possible to interpret. Suppose θ_1 is an important parameter and $\theta_2,\ldots,\theta_p$ are nuisance parameters. If we want to make inference about θ_1, it is of little use to know that a smart parameter is sin θ_1/θ_2. If the smart parameter was log θ_1 we would be better off, because it is as easy to interpret as θ_1 and probability statements can easily be transformed back and forth. Two types of transformations are considered. Type 1, where only the important parameter is transformed, i.e. the new parameter has the form $(g(\theta_1),\theta_2,\ldots,\theta_p)$, where g is to be chosen optimally. Type 2, for which only results for $p = 2$ are reported, is a transformation of both parameters individually, i.e. the new parameter has the form $(g_1(\theta_1), g_2(\theta_2))$, where g_1 and g_2 are functions to be chosen optimally. A third type, where θ_1 is kept while the nuisance parameters are transformed, could also be considered, but it is limited what can be gained by such a transformation, since the distribution of the estimate of the first parameter is unchanged. For properties involving the distribution of a function of the first parameter it doesn't matter which of the two types is considered.

Using these transformations there are two theoretical problems. First what should be considered optimal? If a covariance stabilizing transformation exists, it will automatically have all parametrization effects zero to a low order, see Hougaard (1981). But in the smaller classes of transformations the two types are, it is important how the nonlinearity is evaluated. Below it will be shown how different definitions of "optimal" give different results.

Using these results it is possible to get a deeper understanding of how parametrization effects influences the inference. Secondly the optimal choice of transformation of θ_1 might depend on the value of $\theta_2,\ldots,\theta_p$. By choosing such a value we end up with a result, which unfortunately is only optimal on a line. This is,however,better than the general transformations, which in most cases, cf Section 2, only can be optimal in a given point. Even when it is not optimal it will often be much better than the original parametrization. This is further discussed in Section 5.

For making inference about θ_1, a natural starting point is the marginal distribution of $\hat{\theta}_1$, where $\hat{\theta}$ denotes the maximum likelihood estimator which is also the least squares estimator.

We consider the limit $\sigma^2 \to 0$ or equivalently $m \to \infty$, m being the number of repetitions of the whole experiment, since the averages in each group are sufficient for estimating θ and the vector $\bar{Y}_m$ of averages has the distribution $N_m(\eta(\theta),\sigma^2/m\ I)$. The asymptotic distribution is, following ordinary theory, normal

$$\hat{\theta}_1 \sim N(\theta_1,\ \sigma^2\gamma_{11}), \tag{3.1}$$

where $\gamma_{11} = \gamma_{11}(\theta)$ is the upper left element in the inverse matrix of $\dot{\eta}'\dot{\eta}$. In an Edgeworth expansion of the distribution of $\hat{\theta}_1$ there are two first-order correction terms to the asymptotic distribution above. The two terms correspond to the bias and the skewness respectively. For a Type 1 transformation $\beta = g(\theta_1)$ each of these first-order terms vanish if g satisfies a differential equation.

<u>Lemma 1. Bias of $\hat{\theta}_1$.</u> The asymptotic bias of $\hat{\beta}$ is of order $o(\sigma^2)$ if and only if

$$\frac{g''(\theta_1)}{g'(\theta_1)} = L_1 \ddot{\eta}_{11} + 2 L_1 \ddot{\eta}_{12} \gamma_{21} \gamma_{11}^{-1} + L_1 \ddot{\eta}_{22} \gamma_{22} \gamma_{11}^{-1}$$

If $p > 2$ the last term should be replaced by $\Sigma_{i,j\geq 2} L_1 \ddot{\eta}_{ij} \gamma_{ji} \gamma_{11}^{-1}$

Here L_1 is the first row in $(\dot{\eta}'\dot{\eta})^{-1}\dot{\eta}'$. The geometrical interpretation of this quantity is the linear function giving the coefficient to $\dot{\eta}_1$ in a projection onto the space spanned by $\dot{\eta}_1,\ldots,\dot{\eta}_p$, i.e. the tangent space.

Proof All proofs in this section are derived by similar methods taking known expressions and inserting the transformed values of the first and second derivatives. It is practical to derive the expressions for $h = g^{-1}$. Let us derive it for $p = 2$. In the new parameter (β,θ_2) given by $\theta_1 = h(\beta)$ the first and second derivatives expressed by the quantities in the original parametrization are

$$\frac{\partial\eta}{\partial\beta} = \dot{\eta}_1 \, h'(\beta) \quad , \quad \frac{\partial\eta}{\partial\theta_2} = \dot{\eta}_2$$

$$\frac{\partial^2\eta}{\partial\beta^2} = \ddot{\eta}_{11}\{h'(\beta)\}^2 + \dot{\eta}_1 \, h''(\beta)$$

$$\frac{\partial^2\eta}{\partial\beta\partial\theta_2} = \ddot{\eta}_{12} \, h'(\beta)$$

$$\frac{\partial^2\eta}{\partial\theta_2^2} = \ddot{\eta}_{22}$$

From this the inverse information $\sigma^2 \tilde{\gamma}$ can be derived

$$\tilde{\gamma}_{11} = \gamma_{11} \{h'(\beta)\}^{-2}, \; \tilde{\gamma}_{12} = \gamma_{12}\{h'(\beta)\}^{-1} \; , \; \tilde{\gamma}_{22} = \gamma_{22}$$

The bias is calculated by Box (1971, equation (2.20)). The bias of $\hat{\theta}_1$ is

$$- \tfrac{1}{2} \sigma^2 \Sigma_{i,j} L_1 \ddot{\eta}_{ij} \gamma_{ji}$$

Inserting the corresponding quantities for (β,θ_2) yields that the bias af $\hat{\beta}$ is zero if and only if (under the condition that $h'(\beta)\neq 0$)

$$-\frac{h''(\beta)}{h'(\beta)^2} = L_1\ddot{\eta}_{11} + 2\, L_1\ddot{\eta}_{12}\gamma_{21}\gamma_{11}^{-1} + L_1\ddot{\eta}_{22}\gamma_{22}\gamma_{11}^{-1}$$

For finding the expression in terms of g we find that $-h''(\beta)/h'(\beta)^2 = g''(\theta_1)/g'(\theta_1)$ giving the desired equation.

Remark Equations of the kind $g''(\theta)/g'(\theta) = k(\theta)$, k a known function, can always be expressed as an integral, namely $g(\theta) = \int^{\theta}\exp(\int^{x}k(u)du)dx$. The arbitrary constants correspond to affine transformations of g or equivalently to choices of lower limits in the integrals.

Lemma 2 Skewness The asymptotic skewness of $\hat{\beta}$ is of order $o(\sigma)$ if and only if

$$\frac{g''(\theta_1)}{g'(\theta_1)} = L_1\ddot{\eta}_{11} + 2\, L_1\ddot{\eta}_{12}\gamma_{12}\gamma_{11}^{-1} + L_1\ddot{\eta}_{22}\gamma_{21}^{2}\,\gamma_{11}^{-2}$$

If $p > 2$ the last term should be replaced by $\Sigma_{i,j\geq 2}\, L_1\ddot{\eta}_{ij}\gamma_{i1}\gamma_{j1}\gamma_{11}^{-2}$

Proof The proof is analogous to the proof of Lemma 1, using that from (3.2) in Hougaard (1981) it follows that asymptotically

$$E(\hat{\theta}_i - E\hat{\theta}_i)^3 = -3\sigma^4\, \Sigma_{qsr}\gamma_{iq}\gamma_{is}\gamma_{ir}\, \dot{\eta}'_q\ddot{\eta}_{sr}$$

Remarks It follows that if $\ddot{\eta}_{22} = 0$ or more generally $L_1\ddot{\eta}_{22} = 0$ parametrizations have zero bias and skewness at the same time. In fact there is then asymptotically a direct relationship

$$E(\hat{\theta}_1 - E\hat{\theta}_1)^3 = 6\cdot \mathrm{Var}(\hat{\theta}_1)\cdot E(\hat{\theta}_1 - \theta_1).$$

This is also true for $p > 2$, in which case $L_1\,\ddot{\eta}_{22} = 0$ should be considered as an array equation meaning that for all $i,j \geq 2$ $L_1\,\ddot{\eta}_{ij} = 0$. The condition $\ddot{\eta}_{22} = 0$ means that for known θ_1 the model is linear in $\theta_2,\ldots,\theta_p$.

In formula (3.1) above the asymptotic variance $\sigma^2\gamma_{11}$ is a function of θ. If we want to use Wald's test statistic $W = (\hat\theta_1 - \theta_1)^2/\{\hat\sigma^2\gamma_{11}(\hat\theta)\}$ it is preferable that γ_{11} is independent of θ, cf. Væth (1981). Instead of making the whole covariance matrix constant, as in Section 2, we only need to make the variance of $\hat\theta_1$ constant. This is, however, not possible using transformations of the two types suggested, because this variance is a function of θ_1 as well as $\theta_2,\dots,\theta_p$. What can be attained is e.g. that $\mathrm{Var}(\hat\theta_1)$ is independent of θ_1 for a given value of $\theta_2,\dots,\theta_p$. This can be formulated as $\partial\mathrm{Var}(\hat\theta_1)/\partial\theta_1 = 0$. As $\mathrm{Var}(\hat\theta_1)$ is a function of p variables, we can at least from a mathematical point of view as well choose another direction, say v, in which the derivative should be 0. More precisely let $\theta = \theta_0 + bv$, $b \in \mathbb{R}$, $v \in \mathbb{R}^p$, where θ_0 is the parameter point we consider. Then the condition can be formulated in terms of $d\mathrm{Var}(\hat\theta_1)/db$, which will be denoted $d\mathrm{Var}(\hat\theta_1)/dv$. As this definition of "constant variance" only involves one point, θ_0, we can actually to each point θ_0, choose a direction in which the derivative should be 0. These directions change with the linear part of the transformations. This should be accounted for, such that the properties are invariant under linear transformations of θ_1. The direction (v_1,v_2) in (θ_1,θ_2)-space transforms to $(g'(\theta_1)v_1,v_2)$ in (β,θ_2)-space using Type 1 transformations.

Lemma 3 Variance of θ_1 The variance of θ_1 has derivative 0 in direction $v = (v_1,v_2)$, where $v_1 \neq 0$, if and only if

$$\frac{g''(\theta_1)}{g'(\theta_1)} = L_1\ddot\eta_{11} + L_1\ddot\eta_{12}(\gamma_{12}/\gamma_{11} + v_2/v_1) + L_1\ddot\eta_{22}\gamma_{12}\gamma_{11}^{-1}v_2v_1^{-1}$$

If $p > 2$, v_2 is a $p-1$ dimensional vector, say $v_2 = (v_2,\dots,v_p)'$, and the last term should be replaced by $\Sigma_{i,j>2} L_1\ddot\eta_{ij}\gamma_{1i}v_j\gamma_{11}^{-1}v_1^{-1}$

$$d\mathrm{Var}(\hat\theta_1)/d\theta_1 = 0 \quad \text{if and only if}$$

$$\frac{g''(\theta_1)}{g'(\theta_1)} = L_1\ddot\eta_{11} + L_1\ddot\eta_{12}\gamma_{12}/\gamma_{11}$$

The derivative of $\mathrm{Var}(\hat{\theta}_1)$ in the direction $(\gamma_{11}, \gamma_{12})$ is 0 if and only if

$$\frac{g''(\theta_1)}{g'(\theta_1)} = L_1\ddot{\eta}_{11} + 2L_1\ddot{\eta}_{12}\gamma_{12}/\gamma_{11} + L_1\ddot{\eta}_{22}\gamma_{12}^2\gamma_{11}^{-2}$$

Remarks The last equation is the same as gives skewness 0.

Proof The same method as Lemma 1, using (3.1) in Hougaard (1981), which says that

$$\frac{d\gamma_{ij}}{d\theta_k} = - \Sigma_{qs}\ \gamma_{iq}\ \gamma_{sj}(\dot{\eta}_q'\ \ddot{\eta}_{sk} + \dot{\eta}_s'\ \ddot{\eta}_{qk})\ , \text{ and}$$

$$\mathrm{Var}(\hat{\theta}_i, \hat{\theta}_j) = \sigma^2\gamma_{ij} \quad \text{asymptotically.}$$

If the parameters are of equal importance we might instead consider another aspect of the variance matrix and try to keep that constant, e.g. the determinant, the socalled generalized variance. Jeffrey (1946) suggested in Bayesian theory to use a prior distribution with a density proportional to the square root of the determinant of the information matrix, If the determinant is constant this prior will be uniform. By a Type 1 transformation the derivative with respect to θ_1 can be made zero.

Lemma 4 Variance determinant For $p = 2$, $d|\mathrm{Var}(\hat{\theta})|/d\theta_1 = 0$ if and only if

$$\frac{g''(\theta_1)}{g'(\theta_1)} = L_1\ddot{\eta}_{11} + L_2\ddot{\eta}_{12}$$

Proof Similar to Lemma 1, using that if the determinant of the variance has derivative zero for θ_1, the same is true for the information matrix and $|\mathrm{Inf}| = \dot{\eta}_1'\dot{\eta}_1\dot{\eta}_2'\dot{\eta}_2 - (\dot{\eta}_1'\dot{\eta}_2)^2$. Thus the derivative is

$$d|\mathrm{Inf}|/d\theta_1 = 2\dot{\eta}_1'\ddot{\eta}_{11}\dot{\eta}_2'\dot{\eta}_2 + 2\dot{\eta}_1'\dot{\eta}_1\dot{\eta}_2'\ddot{\eta}_{12}$$

$$- 2\ \dot{\eta}_1'\dot{\eta}_2(\dot{\eta}_2'\ddot{\eta}_{11} + \dot{\eta}_1'\ddot{\eta}_{12}).$$

Anscombe (1964) suggested using a parametrization with normal likelihood, i.e. the third derivative of the logarithm of the likelihood is 0 at the maximum likelihood estimate, making the log-likelihood approximately a parabola. When the estimate $\hat{\theta}$ is not sufficient the transformation equation depends on the observations. This problem can be overcome by instead considering the mean of the third derivative at the true value θ_0, as suggested by Sprott (1973). Because the third derivative is linear as a function of the observations, this is the same as considering the theoretical values, i.e. assuming all observations equal their means. The corresponding multiparameter property is to make the log likelihood, say ℓ, approximately a second degree polynomium in the p parameters. This is, however, not necessarily possible. Actually it is equivalent to having zero skewness for all linear combinations of the parameters. There is a duality between normal likelihood and zero skewness. Apart from a factor including a power of σ^2 the third derivative of the log likelihood in any direction equals the asymptotic third central moment for a corresponding linear combination. This follows from (3.7) in Hougaard (1981). Sprott (1980) found that in the multiparameter case it is not relevant to consider the third derivative with respect to θ_1. Instead he suggested to make the third derivative of the logarithm of the likelihood maximized over $\theta_2,\dots,\theta_p$ vanish. That makes more sense, when θ_1 is important and $\theta_2,\dots,\theta_p$ nuisance parameters. However also in this case the parameter transformation equation will typically depend on the observations. Because the expression is more complicated it is not possible to overcome this problem by taking the mean of the third derivative, but instead the problem can be overcome by considering the theoretical values, i.e. inserting the means instead of the observations in the third derivative.

Lemma 5 Normal likelihood For a given direction $v \in \mathbb{R}^2$, where $v_1 = 0$ the $Ed^3\ell/dv^3 = 0$ if and only if

$$\frac{g''(\theta_1)}{g'(\theta_1)} =$$

$$\frac{\dot\eta_1'\ddot\eta_{11} + \dot\eta_2'\ddot\eta_{11}v_2v_1^{-1} + 2(\dot\eta_1'\ddot\eta_{12}v_2v_1^{-1} + \dot\eta_2'\ddot\eta_{12}v_2^2v_1^{-2}) + \dot\eta_1'\ddot\eta_{22}v_2^2v_1^{-2} + \dot\eta_2'\ddot\eta_{22}v_2^3v_1^{-3}}{\dot\eta_1'\dot\eta_1 + \dot\eta_1'\dot\eta_2v_2v_1^{-1}}$$

$E\partial^3\ell/d\theta_1^3 = 0$ if and only if

$$\frac{g''(\theta_1)}{g'(\theta_1)} = \frac{\dot\eta_1'\ddot\eta_{11}}{\dot\eta_1'\dot\eta_1}$$

$Ed^3\ell/dv^3 = 0$ for $v = (\gamma_{11},\gamma_{12})'$ if and only if

$$\frac{g''(\theta_1)}{g'(\theta_1)} = L_1\ddot\eta_{11} + 2\,L_1\ddot\eta_{12}\gamma_{12}\gamma_{11}^{-1} + L_1\ddot\eta_{22}\gamma_{12}^2\gamma_{11}^{-2}$$

Also $\frac{d^3}{d\beta^3} \max_{\theta_2} \ell(\theta_1,\theta_2) = 0$ in the true value (θ_1,θ_2) for points having $Y = \eta(\theta)$ if and only if

$$\frac{g''(\theta_1)}{g'(\theta_1)} = L_1\ddot\eta_{11} + 2\,L_1\ddot\eta_{12}\gamma_{12}\gamma_{11}^{-1} + L_1\ddot\eta_{22}\gamma_{12}^2\gamma_{11}^{-2}$$

Remark The last two equations are the same and identical to zero skewness for $\hat\beta$.

Proof. The first three equations are simple consequences of $\ell(\theta) = -\Sigma\{Y_i - \eta_i(\theta)\}^2/(2\sigma^2)$ + constant using metoods similar to Lemma 1. For the last one we consider a point (θ_1,θ_2). First we find the function of Y and θ_1, which gives $\hat\theta_2$ the value maximizing ℓ for given θ_1. This is done for general models using derivatives of the log likelihood function. By implicit differentiation in the point (θ_1,θ_2) it is found that

$$d\hat{\theta}_2/d\theta_1 = - \ddot{\ell}_{12}\ddot{\ell}_{22}^{-1}$$

$$d^2\hat{\theta}_2/d\theta_1^2 = - \ddot{\ell}_{22}^{-1} \{\dddot{\ell}_{112} + 2\, \dddot{\ell}_{122}\, d\hat{\theta}_2/d\theta_1$$

$$+ \dddot{\ell}_{222}(d\hat{\theta}_2/d\theta_1)^2\}$$

Also it is found that

$$\frac{d^3\ell_{\max}}{d\theta_1^3} = \dddot{\ell}_{111} + 3\, \dddot{\ell}_{112}\, d\hat{\theta}_2/d\theta_1 + 3\, \dddot{\ell}_{122}(d\hat{\theta}_2/d\theta_1)^2$$

$$+ \dddot{\ell}_{222}(d\hat{\theta}_2/d\theta_1)^3 + 3\, \ddot{\ell}_{12} d^2\hat{\theta}_2/d\theta_1^2 + 3\, \ddot{\ell}_{22}(d^2\hat{\theta}_2/d\theta_1^2)(d\hat{\theta}_2/d\theta_1)$$

$$+ \dot{\ell}_2\, d^3\hat{\theta}_2/d\theta_1^3 \;,$$

where $\ell_{\max}(\theta_1) = \max_{\theta_2} \ell(\theta_1,\theta_2)$. Inserting the values of $d\hat{\theta}_2/d\theta_1$ and $d^2\hat{\theta}_2/d\theta_1^2$ yields

$$\frac{d^3\ell_{\max}}{d\theta_1^3} = \dddot{\ell}_{111} - 3\, \dddot{\ell}_{112}\, \ddot{\ell}_{12}\, \ddot{\ell}_{22}^{-1} + 3\, \dddot{\ell}_{122}\, \ddot{\ell}_{12}^{2}\, \ddot{\ell}_{22}^{-2}$$

$$- \dddot{\ell}_{222}\, \ddot{\ell}_{12}^{3}\, \ddot{\ell}_{22}^{-3} + \dot{\ell}_2\, d^3\, \hat{\theta}_2/d\theta_1^3$$

In the nonlinear regression the mean or theoretical values of the derivatives are

$$E\, \dot{\ell}_i = 0, \quad E\, \ddot{\ell}_{ij} = - \sigma^{-2} \dot{\eta}_i' \dot{\eta}_j',$$

$$E\, \dddot{\ell}_{ijk} = - \sigma^{-2} (\dot{\eta}_i' \ddot{\eta}_{jk} + \dot{\eta}_j' \ddot{\eta}_{ik} + \dot{\eta}_k' \ddot{\eta}_{ij})$$

Inserting these values calculated for the transformed parameter into the expression for the third derivative, yields that the theoretical value of $d^3\ell_{\max}/d\beta^3$ is zero if and only if $g(\theta_1)$ satisfies the differential equation.

Bates & Watts (1980) suggested a measure of curvature, which was later discussed in Bates & Watts (1981), Hamilton, Watts & Bates (1982) and Hougaard (1981). To each direction v in the parameter space two measures of non-linearity were defined, one independent of the parametrization and one dependent of the parametrization. In this paper only the latter is considered. In order to reduce them to one number Bates & Watts (1980) suggested using the maximum over all directions. This maximun can be used for constructing conservative confidence regions. Hougaard (1981) found that also the directional curvatures can be used to judge how close the distributions of estimated parameters are to the asymptotic distribution.

It is not simple to minimize the maximal curvature by means of transformations of Type 1 and 2. However if a specific direction is important the corresponding curvature can be minimized using Type 1 as well as Type 2 transformations. Also in this case the direction might be a function of the parameter. Hougaard (1981, formula (3.7)) showed e.g. that the skewness of $\hat{\theta}_1$ is smaller than $3/\sqrt{p}$ times the curvature in the direction $(\gamma_{11},\gamma_{12})$. It is therefore natural to find the transformation, which minimizes the curvature in that direction, giving a parameter β with small skewness. The distribution and thereby the skewness of $\hat{\beta}$ is the same for Type 1 and Type 2 transformations, but the upper bound can be smaller for a Type 2 transformation than for Type 1. As will appear below in Lemma 6, the bound is minimized for the same transformation as makes the skewness 0. The equation for Type 1 involves only the projection onto the space spanned by $\dot{\eta}_1$, whereas for Type 2 it is the $\dot{\eta}_1$ - part of a projection onto the space spanned by $\dot{\eta}_1$ and $\dot{\eta}_2$.

<u>Lemma 6 Bates & Watts Curvature</u> For a given direction $v \in \mathbb{R}^2$, where $v_1 \neq 0$, Bates & Watts curvature in direction v is minimal over all transformations of θ_1, if and only if

$$\frac{g''(\theta_1)}{g'(\theta_1)} = \frac{\dot{\eta}'_1\ddot{\eta}_{11} + 2\,\ddot{\eta}'_1\ddot{\eta}_{12}v_2v_1^{-1} + \dot{\eta}'_1\ddot{\eta}_{22}v_2^2v_1^{-2}}{\dot{\eta}'_1\dot{\eta}_1}$$

In the direction of θ_1 this equation is

$$\frac{g''(\theta_1)}{g'(\theta_1)} = \frac{\dot{\eta}'_1\ddot{\eta}_{11}}{\dot{\eta}'_1\dot{\eta}_1}$$

For minimizing the curvature over all transformations of the form $(g_1(\theta_1), g_2(\theta_2))$ the equation for $g_1(\theta_1)$ is

$$\frac{g''_1(\theta_1)}{g'_1(\theta_1)} = L_1\,\ddot{\eta}_{11} + 2\,L_1\,\ddot{\eta}_{12}v_2v_1^{-1} + L_1\,\ddot{\eta}_{22}v_2^2v_1^{-2}$$

The equation for $g_2(\theta_2)$ is

$$\frac{g''_2(\theta_2)}{g'_2(\theta_2)} = L_2\,\ddot{\eta}_{11}v_1^2v_2^{-2} + 2\,L_2\,\ddot{\eta}_{12}v_1v_2^{-1} + L_2\,\ddot{\eta}_{22}$$

If $(v_1,v_2) = (\gamma_{11},\gamma_{12})$ the equation for g_1 is

$$\frac{g''_1(\theta_1)}{g'_1(\theta_1)} = L_1\,\ddot{\eta}_{11} + 2\,L_1\,\ddot{\eta}_{12}\gamma_{12}\gamma_{11}^{-1} + L_1\,\ddot{\eta}_{22}\gamma_{12}^2\gamma_{11}^{-2}$$

<u>Proof</u>. The squared curvature in direction v in the transformed parameter is of the form

$$d(a(g''/g')^2 + b\,g''/g' + c),$$

with $a = \dot{\eta}'_1\dot{\eta}_1v_1^4$ and $b = -\,2\,\{\dot{\eta}'_1\ddot{\eta}_{11}v_1^4 + 2\,\eta'_1\ddot{\eta}_{12}v_1^3v_2 + \dot{\eta}'_1\ddot{\eta}_{22}v_1^2v_2\}$. The quantities a,b,c and d are functions of θ, but not of the transformation. Thus the curvature is minimized for $g''/g' = -\,\frac{1}{2}\,b/a$ giving the desired equation. Similarly for the Type 2 transformation, the curvature is a quadratic form in g''_1/g'_1 and g''_2/g'_2 yielding minimum for the described solution.

For minimizing the measure of nonlinearity proposed by Beale (1960) the equations are more complicated. In Lemma 7 the result is reported for Type 1-transformations. The resuld for Type 2 transformations is much more complicated and therefore not reported here.

<u>Lemma 7 Beales measure of nonlinearity</u> Beales measure of nonlinearity is minimized over transformations of θ_1 if and only if

$$\frac{g''(\theta_1)}{g'(\theta_1)} = \frac{\dot{\eta}_1'\ddot{\eta}_{11} + 2\,\dot{\eta}_1\ddot{\eta}_{12}\gamma_{12}\gamma_{11} + \dot{\eta}_1'\ddot{\eta}_{22}\{\gamma_{22}/(3\gamma_{11}) + 2\gamma_{12}^2/(3\gamma_{11}^2)\}}{\dot{\eta}_1'\dot{\eta}_1}$$

<u>Proof</u> Similar to Lemma 6, Beales measure is also a quadratic function of $g''(\theta_1)/g'(\theta_1)$.

4. SPECIAL CASES AND EXAMPLES

Special case 1 Suppose $\ddot{\eta}_{12} = 0$ and $\ddot{\eta}_{22} = 0$, which is the case, when η is a sum of a nonlinear function of θ_1 and a linear function of θ_2, $\eta_i = f_i(\theta_1) + b_i\,\theta_2$. We will assume that all b_i's are 1, because that makes the formulae a little simpler. In this case the condition in the Theorem in Section 2 is trivially fulfilled, such that there exists a covariance stabilizing transformation, say $(\beta_1,\beta_2) = (g_1(\theta_1,\theta_2),g_2(\theta_1,\theta_2))$. Both of g_1 and g_2 satisfies the same differential equation (2.1), so we need two independent solutions g to the equation. It follows from the equation that $\ddot{g}_{12} = \ddot{g}_{22} = 0$, such that $g(\theta_1,\theta_2) = h(\theta_1) + a\,\theta_2$, for some function h and some constant a. In terms of h the differential equation is

$$h''(\theta_1) = h'(\theta_1)\mathrm{SPD}_{\dot{f},\ddot{f}}/\mathrm{SSD}_{\dot{f}} + a(\dot{f}'\dot{f}\ S_{\ddot{f}} - \dot{f}'\ddot{f}\ S_{\dot{f}})/(n\ \mathrm{SSD}_{\dot{f}}),$$

where $\mathrm{SSD}_{\dot{f}} = \dot{f}'\dot{f} - S^2_{\dot{f}}/n$, $\mathrm{SPD}_{\dot{f},\ddot{f}} = \dot{f}'\ddot{f} - S_{\dot{f}}S_{\ddot{f}}/n$ and $S_{\dot{f}} = \Sigma df_i(\theta_1)/d\theta_1$.

The complete solution to the differential equation is

$$h(\theta_1) = c_1 \int \sqrt{\mathrm{SSD}_{\dot{f}}} + a\ S_f/n + c_2\ ,$$

where c_1 and c_2 are arbitrary constants. A natural choice of g_1 and g_2 is then g_1 given by $g_1'(\theta_1) = \sqrt{\mathrm{SSD}_{\dot{f}}}$, which is independent of θ_2 and $g_2(\theta_1,\theta_2) = S_f/n + \theta_2$, which is also the average of the means of the observations, say $\bar{\eta} = \Sigma\eta_i/n$. The estimate of this parameter is $\bar{\bar{Y}}$, the grand average of all observations. We have then succeeded in finding a covariance stabilizing transformation.

Most of the transformations from Section 3 coincide, such that there essentially are two equations. One, which corresponds to the onedimensional case, i.e. θ_2 known, is $g''(\theta_1)/g'(\theta_1) = \dot{\eta}_1'\ddot{\eta}_{11}/(\dot{\eta}_1'\dot{\eta}_1)$. It can be simplified to $g'(\theta_1) = c\ \sqrt{(\dot{\eta}_1'\dot{\eta}_1)}$, where c is an arbitrary constant. In this parametrization $\beta = g(\theta_1)$ has constant information and normal likelihood in direction (1,0). Also Beales measure and Bates & Watts curvature in all directions are minimal under Type 1 transformations. The equation is independent of θ_2.

The other transformation is the same as the above covariance stabilizing transformation with $a = 0$. The equation is $g''(\theta_1)/g'(\theta_1) = L_1\ddot{\eta}_{11}$, which reduces to $g'(\theta_1) = \sqrt{}SSD_{\dot{f}}$. This parameter $\beta = g(\theta_1)$ has constant asymptotic variance for $\hat{\beta}$, the bias and skewness of $\hat{\beta}$ are 0 and the determinant of the variance matrix is constant. The likelihood is normal, as well in the sense of Sprott (1980) as in direction $(\gamma_{11},\gamma_{12})$. Finally Bates & Watts curvature in any direction is minimal among Type 2 transformations. The transformation equation corresponding to θ_2 known has the interpretation of $g'(\theta_1)$ as a constant times the square root of the sum of squares of derivatives of f. The second equation is similarly a constant times the square root of the sum of squared deviations of derivatives of f.

<u>Special case 2</u> Suppose $\eta_i(\theta) = \theta_2 \, f(\theta_1;x_i)$, which is a common example of nonlinear regression models. By differentiation $\dot{\eta}_1 = \theta_2 \dot{f}$, $\dot{\eta}_2 = f$, $\ddot{\eta}_{11} = \theta_2 \ddot{f}$, $\ddot{\eta}_{12} = \dot{f}$, $\ddot{\eta}_{22} = 0$. In particular $\ddot{\eta}_{12}$ is proportional to $\dot{\eta}_1$, from which it follows that the space spanned by the first and second derivatives has a dimension at most 1 higher than the space spanned by the first derivatives alone. It also follows that the condition in the theorem in Section 2 is satisfied, so there does exist a covariance stabilizing transformation. From inserting in the transformation equation we find that solutions have the form $g(\theta_1,\theta_2) = \theta_2 \, h(\theta_1) + a$. For convenience we can choose $a = 0$. The differential equation for h is then

$$h''(\theta_1) = \tfrac{1}{2} h'(\theta_1)Q'(\theta_1)/Q(\theta_1) + h(\theta_1)R(\theta_1)/Q(\theta_1), \qquad (4.1)$$

where $Q(\theta_1) = f'f \, \dot{f}'\dot{f} - (f'\dot{f})^2$ and $R(\theta_1) = \dot{f}'\dot{f} \, f'\ddot{f} - f'\dot{f} \, \dot{f}'\ddot{f}$ are known functions of θ_1. $Q(\theta_1)$ appears in a number of places, e.g. the determinant of $\dot{\eta}'\dot{\eta}$ is $\theta_2^2\,Q(\theta_1)$.

The covariance stabilizing transformations depends on both parameters. Now we will consider the simpler transformations of Type 1 and 2. It turns out in this case that the transformations of θ_1 are all independent of θ_2. Also the equations can be reduced to equations for $g'(\theta_1)$. The transformations corresponding to θ_2 known, i.e. for properties like constant information, normal likelihood in the direction (1,0)' and Bates & Watts curvature minimal in direction (1,0)' is $g'(\theta_1) = c\,\surd(\dot{f}'\dot{f})$, where c is an arbitrary constant.

In this case the transformations, which make the bias and the skewness 0 coincide, such that the distribution of this parameter is well approximated by the asymptotic distribution. The equation is $g'(\theta_1) = Q(\theta_1)^{\frac{1}{2}}/f'f$. This parameter also has normal likelihood in the sense of Sprott (1980) and normal likelihood in the direction $(\gamma_{11},\gamma_{12})$, the derivative of the variance of $\hat{\theta}_1$ in the direction $(\gamma_{11},\gamma_{12})$ is 0 and the Bates & Watts curvature in direction $(\gamma_{11},\gamma_{12})$ is minimized over Type 2 transformations.

The determinant of the covariance matrix is independent of β, when $g'(\theta_1) = \surd Q(\theta_1)$. By also transforming θ_2 to $\rho = \theta_2^2$ the determinant becomes constant. The derivative of the variance in direction (1,0)' is zero, when $g'(\theta_1) = \surd\{Q(\theta_1)/f'f\}$. Finally $g'(\theta_1) = \surd(\dot{f}'\dot{f})/f'f$ yields minimal Bates & Watts curvature in direction $(\gamma_{11},\gamma_{12})'$ among Type 1 transformations.

<u>Special case 3</u> Suppose $\eta_i(\theta) = \theta_2 f(\theta_1;x_i) + \theta_3$, which is the natural combination of the first two special cases. An example is the Gompertz function $\theta_2\theta_1^{x_i} + \theta_3$, used in actuarial science to model mortality rates, x_i being the age of group i. Also in this case a covariance stabilizing transformation exists. The solutions must have the form $g(\theta) = \theta_2 h(\theta_1) + a\theta_3 + b$. For convenience we can assume b to be 0. One solution to the equation is the average of all meanvalues, $g(\theta) = \theta_2\,\Sigma_i f(\theta_1;x_i)/n + \theta_3$. For the solution

with $a = 0$, h will satisfy a modified version of the equation in Special case 2, i.e. in Q and R sums of squares and sums of products of derivatives are replaced by sums of squared deviations respectively sums of products of deviations. Mathematically $Q(\theta_1) = SSD_f\, SSD_{\dot f} - SPD^2_{f,\dot f}$ and $R(\theta_1) = SSD_{\dot f}\, SPD_{f,\ddot f} - SPD_{f,\dot f}\, SPD_{\dot f,\ddot f}$, where for example $SSD_{\dot f} = \dot f'\dot f - S^2_{\dot f}/n$, $S_{\dot f} = \Sigma_i\, df(\theta_1;x_i)/d\theta_1$ and $SPD_{\dot f,\ddot f} = \dot f'\ddot f - S_{\dot f}S_{\ddot f}/n$.

Also for the transformations of θ_1 alone, we find equations similar to those from special case 2, only modified in the same way as above.

Example 1 Suppose $\eta_i(\theta) = \theta_2 \exp(-\theta_1 x_i)$, which for example appears in physical and biochemical applications, where x_i is the time since start of the experiment. Suppose first that θ_2 is known, say $\theta_2 = 1$. This model has been considered in Bates & Watts (1981), with a slightly different notation. In that paper it was found that the parametrization which removes the parameter-effects curvature is a solution to

$$g''(\theta_1)/g'(\theta_1) = -\Sigma_i\{x_i^3 \exp(-2\theta_1 x_i)\}/\Sigma_i\{x_i^2 \exp(-2\theta_1 x_i)\} .$$

This equation can be simplified to

$$g'(\theta_1) = c\{\Sigma_i x_i^2 \exp(-2\theta_1 x_i)\}^{\frac{1}{2}}$$

However the solution is not simple and it will depend on the design variables $x_1,\ldots,x_n$. But we might be able to find an approximate solution, depending less on the design variables. We can approximate the sum over i by an integral. Often in practice the x_i's are chosen equidistant from time 0 to x_n. We can then approximate

$$\Sigma_i\, x_i^2 \exp(-2\theta_1 x_i) \simeq c_1 \int_0^{x_n} x^2 \exp(-2\theta_1 x)\,dx$$

where c_1 is a constant depending on the time between two successive observations. We might adjust the integral limits by half of this time. Because of the arbitrary constant we need not care about c_1. The integral is

$$[1 - \exp(-2\theta_1 x_n)\{1 + 2\theta_1 x_n + 2(\theta_1 x_n)^2\}]/(4\theta_1^3)$$

However the square root of this is not simple either. In applications often $\theta_1 > 0$ and we continue observation until the mean is practically zero. Therefore we can approximate x_n by infinity, then having to solve only the integral

$$g(\theta_1) = c \int (4\theta_1)^{-3/2} d\theta \;= c \cdot \theta_1^{-1/2}$$

This final solution is independent of the design. We choose $\gamma = - \theta_1^{-1/2}$. Because of the way this approximate solution is found there are no guarantee that this parametrization is good. On Fig. 1 the skewness of the maximum likelihood estimator for this parametrization and the original one are compared, assuming $n = 10$, $x_i = i$, $\sigma = 0.01$. For values of θ_1 larger than 0.1 the absolute skewness is smaller for γ than for θ_1. For large values of θ_1, the skewness of $\hat{\gamma}$ is very small. For θ_1 less than 0.1, γ is worse in terms of skewness.

Consider now the two parameter model. It is then an example of Special case 2. The skewness removing transformation is independent of θ_2 and given by

$$g'(\theta_1) = c\, Q(\theta_1)^{1/2}/f'f,$$

where $$Q(\theta_1) = \Sigma x_i^2 \exp(-2\theta_1 x_i) \cdot \Sigma \exp(-2\theta_1 x_i)$$

$$- \{\Sigma x_i \exp(-2\theta_1 x_i)\}^2,$$

and $$f'f = \Sigma \exp(-2\theta_1 x_i)$$

This equation is even more complicated than the one for θ_2 known, but again we can approximate the sums by integrals and then insert ∞ for x_n. Again the time between successive measurements goes out as a factor which can be incorporated in c. Apart from the constant factors we approximate $Q(\theta_1) \simeq 2\theta_1^{-3}\,\theta_1^{-1} - (\theta_1^{-2})^2 = \theta_1^{-4}$ and $f'f \simeq \theta_1^{-1}$ which gives $g'(\theta_1) = c/\theta_1$, i.e. $g(\theta_1) = c \ln \theta_1$, say $\delta = \ln \theta_1$. This parametrization is compared with θ_1 and γ on Fig. 2 using the same design as above for Fig. 1, which also includes the skewness of $\hat{\delta}$. The value of the skewness is proportional to $1/\theta_2$. Fig. 2 corresponds to the value $\theta_2 = 1$. For values of θ_1 greater than 0.2 the absolute skewness of $\hat{\delta}$ is smaller than that of $\hat{\theta}_1$ and for all values in the range the absolute skewness is smaller for $\hat{\delta}$ than for $\hat{\gamma}$. In particular δ is a good parametrization for large values of θ_1.

In conclusion, we can find good parametrizations by approximating the sums by integrals. An important point is that the solution depends on the assumptions made about θ_2, whether it is known or unknown. This is of course not surprising, because it is similar to considering marginal or conditional variances, but earlier discussions of transformations have often involved approximations based on assuming the other parameters known. Although the parameters δ and γ are suggested assuming that the design consists of equidistant points, they might also be better than θ_1 when the design is not equidistant. We must then expect them not to be as good as in the equidistant case. If the design is not equidistant, it is possible to include a weight function in the integral, if we can suggest one, which approximate the true weights and for which the integral can be calculated.

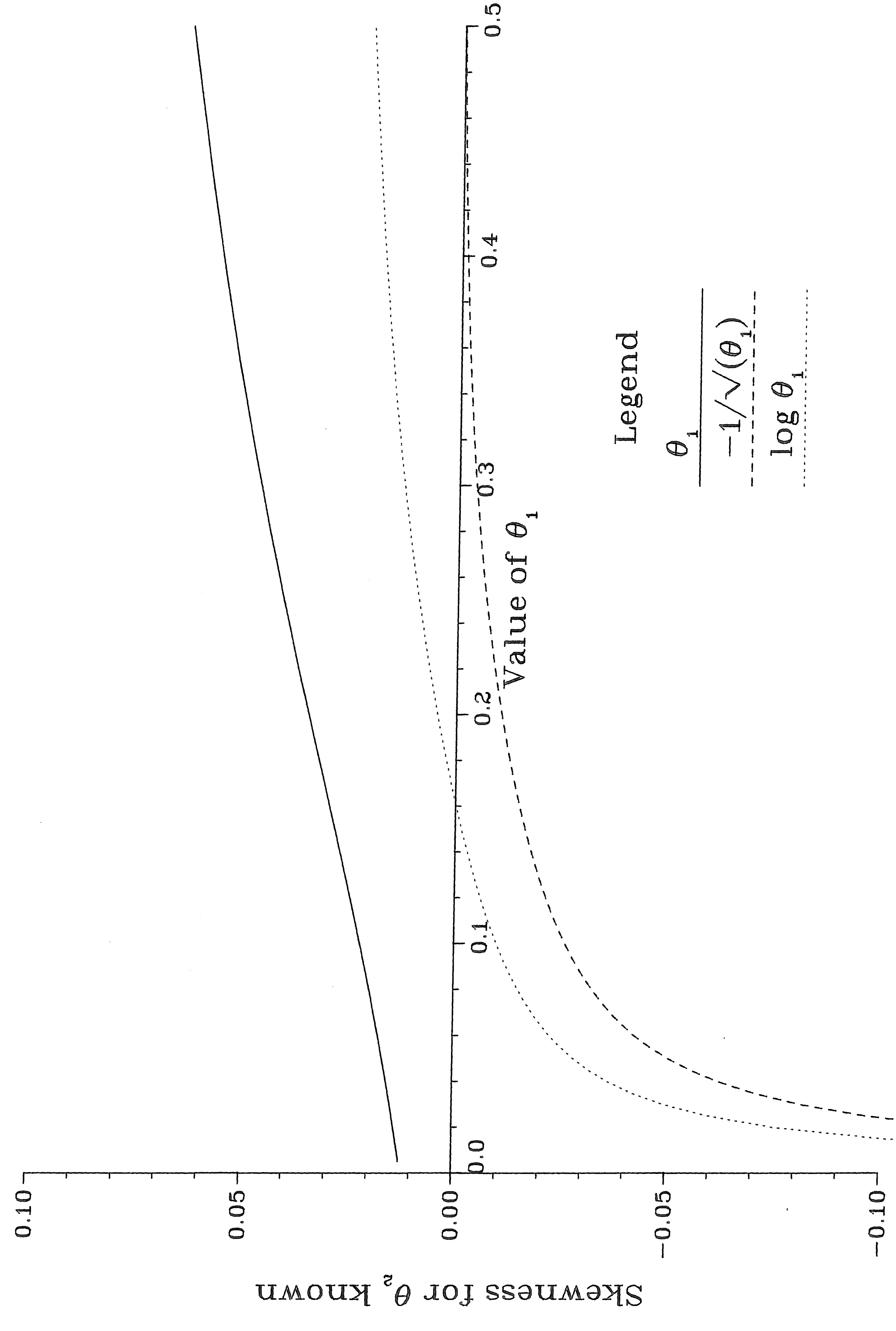
Skewness for θ_2 known
0.10
0.05
0.00
−0.05
−0.10
Value of θ_1
0.0
0.1
0.2
0.3
0.4
0.5
Legend
θ_1
$-1/\sqrt(\theta_1)$
$\log \theta_1$

Fig. 1.

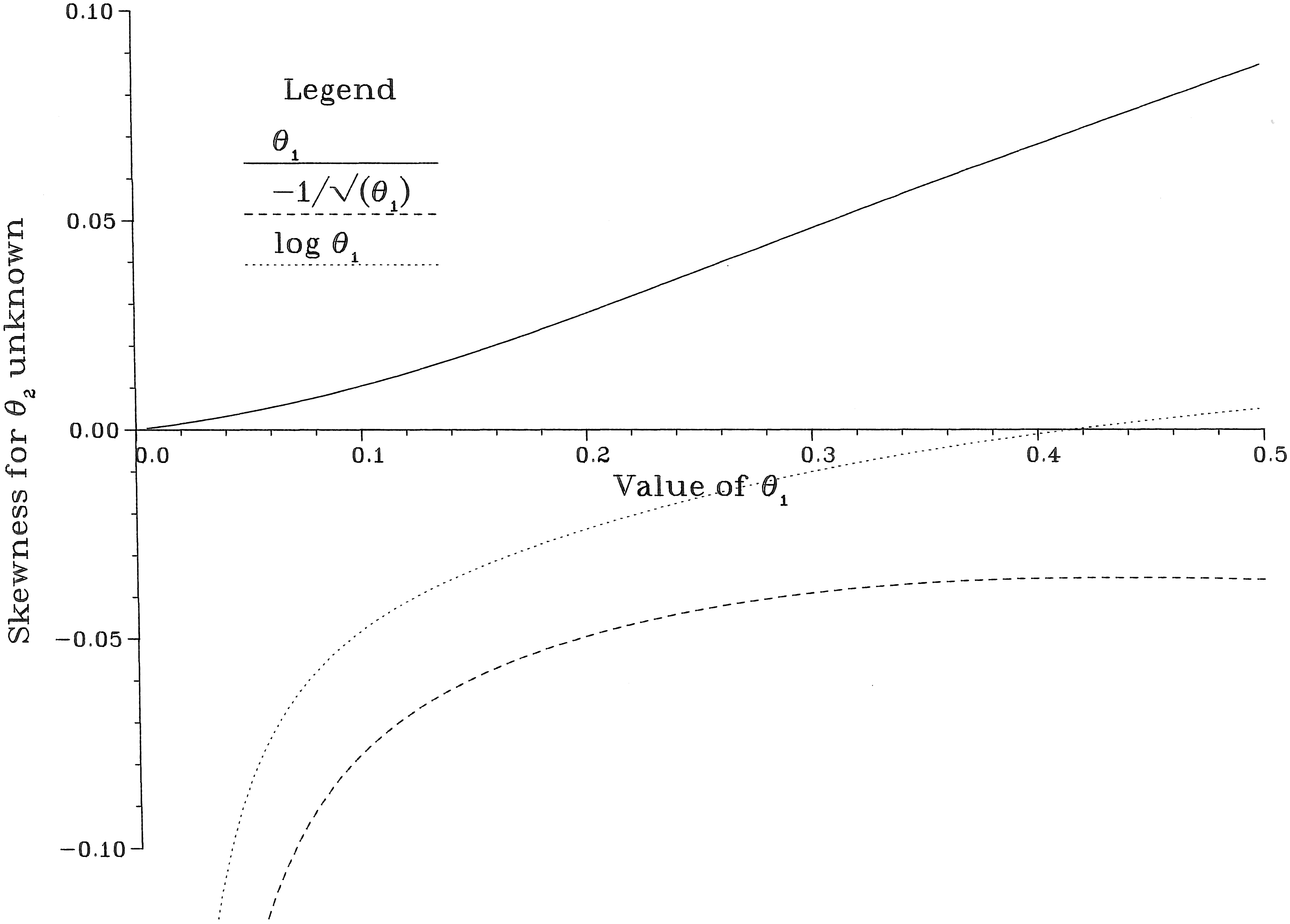
Legend
θ_1
$-1/\sqrt(\theta_1)$
$\log \theta_1$
0.10
0.05
0.00
−0.05
−0.10
Skewness for θ_2 unknown
0.0
0.1
0.2
0.3
0.4
0.5
Value of θ_1

Fig. 2.

Example 2 Errors on two variables Suppose there are $n = 2k$ random variables $X_1,\dots,X_k,Y_1,\dots,Y_k$, which are independent and normally distributed with variance σ^2 and the following meanvalues. $EX_i = \xi_i$, $EY_i = \alpha + \beta\,\xi_i$, $i = 1,\dots,k$. Of the $k + 2$ parameters, α and β are interesting parameters and $\xi_1,\dots,\xi_k$ nuisance parameters. Therefore we are mostly concerned with making inference about α and β. Of the second derivatives only a few are different from 0, i.e. $d^2EY_i/d\beta d\xi_i = 1$, $i = 1,\dots,k$. Because $(I-P)d^2EY_i/d\beta d\xi_i \neq 0$ for $k > 2$, no covariance stabilizing transformation exists. Because of the simple structure of the second derivatives, for known β it is a linear model, the transformation of β to remove bias is the same as the one to remove skewness. The equation for this transformation is $g''(\beta)/g'(\beta) = -\,2\beta/(1+\beta^2)$, i.e. we should instead use $\gamma = \text{Arctan}\ \beta$. Anderson (1976) also suggested this because the expansion of $\hat{\gamma}$ is simpler than that for $\hat{\beta}$ and because the distribution of $\hat{\gamma}-\gamma$ is invariant under rotations of the (X,Y) space. Also γ has normal likelihood, cf. Sprott (1980). This parameter does not have constant variance. The variance is $1/\{SSD(1+\beta^2)\}$, where $SSD = \Sigma\,\xi_i^2 - (\Sigma\xi_i)^2/k$, so this depends on β as well as the other parameters. It is, however, invariant under rotations of the (X,Y) space. For making the variance independent of β, one should, from Lemma 3, solve the equation $g''(\beta)/g'(\beta) = -\,\beta/(1+\beta^2)$, which has the solution $g(\beta) = \log\{\beta + \surd(1+\beta^2)\}$. The variance of $g(\hat{\beta})$ is thus $1/SSD$.

In this model the value of Beales (1960) measure of nonlinearity is quite simple. The value is

$$N = \sigma^2(k+4\beta^2)/\{SSD(k+4)\}$$

Transforming β to $\gamma = g(\beta)$ yields

$$N\gamma = \frac{\sigma^2}{4(k+4)} \left[\frac{3}{SSD} \left\{ \frac{g''(\beta)(1+\beta^2)}{g'(\beta)} + 2\beta \right\}^2 \right.$$

$$\left. + 3 \frac{\{g''(\beta)\}^2 (1+\beta^2)^2 \bar{\xi}^2 k}{\{g'(\beta)\}^2\ SSD^2} + \frac{4(k+\beta^2)}{SSD} \right]$$

where $\bar{\xi} = \Sigma\, \xi_i / k$. The measure is minimized for

$$\frac{g''(\beta)}{g'(\beta)} = - \frac{2\beta}{(1+\beta^2)(1+\varepsilon)} ,$$

where $\varepsilon = k \bar{\xi}^2 / SSD$. It can be reduced to $g'(\beta) = c(1+\beta^2)^{-1/(1+\varepsilon)}$. The equation depends on the other parameters through ε. For $\bar{\xi} = 0$ this transformation is the same as the zero bias - skewness transformation. Using this transformation yields the following minimal value of $N\gamma$

$$N_{\gamma,\min} = \frac{\sigma^2}{(k+4)SSD} \left\{ \frac{3\beta^2 \varepsilon}{1+\varepsilon} + k + \beta^2 \right\}$$

Transformation of β by the zero bias - skewness transformation yields the value

$$N\gamma = \frac{\sigma^2}{(k+4)SSD} \left\{ 3\ \beta^2\ \varepsilon + k + \beta^2 \right\}$$

By supplementing this with the nonlinear transformation of α to $\tilde{\alpha} = \alpha + \beta\, \bar{\xi}$, the measure is

$$N_{\gamma,\tilde{\alpha}} = \frac{\sigma^2}{(k+4)SSD} (k + \beta^2) .$$

This transformation of α is fairly natural in the light of the special cases 1 and 3. The parameter $\tilde{\alpha}$ is the average of the means of the Y-observations. Because Beales measure is invariant under linear transformations and the average of means of the X-observations is linear in the ξ-parameters, this is as good as using the average $(\tilde{\alpha} + \bar{\xi})\ /2$ of all mean values.

5. DISCUSSION

In an asymptotic expansion of the distribution of the estimate of a parameter the first parametrization dependent term is of higher order than the parametrization independent terms. In the oneparameter nonlinear regression there exists a parametrization in which the first parametrization dependent term is of lower order than in general, such that convergence to the asymptotic distribution is much faster. In the multiparameter model it is not in general possible to remove the first term by reparametrization. Therefore a more detailed consideration is necessary. Bates & Watts (1980) examined 24 published data sets and found that in 18 of these cases the parameter-effects were unacceptable large, showing a clear need for transformations of parameters. Choosing a good, but not necessarily optimal, transformation might make the nonlinearity acceptable. The present approach is designed to give parameters, which can be interpreted and where the computational problems in transforming back and forth are not too large. The main difficulties in the present approach are the following three. Firstly the solution depends on the property, which makes it necessary to give priorities to the different properties. Then it is also interesting to examine, when several properties have the same solution. The different properties are discussed below. Secondly the solution might depend on the value of the other parameters. In many cases, e.g. the three special cases and the examples mentioned in Section 4, the solution is actually independent of the other parameters. This might be true even if the distribution depends on the other parameters, e.g. in the errors in two variables example, the solution $\gamma = \arctan \beta$ is independent of other parameters, but the variance is $\sigma^2/\{SSD(1+\beta^2)\}$, and SSD is a function of the other parameters. If the solution depends on θ_2 we might choose a value of θ_2, because even if the transformation is only optimal for one value of θ_2, it will be better than the original, at least in a neighborhood of the θ_2-value.

Thirdly it might be difficult to find or express the solution to the differential equation or the solution might be too complicated for practical purposes. Also we would prefer the transformation not being too dependent on the design. It might still be an advantage to consider the equation because it can be used to suggest approximate solutions or there might be a specific θ_2-value for which the solution is simple. Because of these difficulties finding a good parametrization is still a matter of trying several and examining their performances in the data set at hand. The differential equations can be used to find the optimal transformation in simple cases and to suggest possible transformations in more complicated cases. How this can be done in practice was demonstrated in Example 1, where some sums could be approximated by integrals and inserting the approximation, it was possible to solve the differential equation.

In Example 1 and in the special cases it was also seen how the optimal transformation depends on whether other parameters are included, even if it does not depend on the value of the other parameters. Such a dependence is of course present in inference in general, but it is important to note that a transformation, which is good in a simple model, is not necessarily any good in more complicated models of the same structure.

Because of the dependence on properties an examination of the different properties is required. For the oneparameter case these properties were discussed in Hougaard (1982). It is important that the distribution of $\hat{\theta}_1$ is simple, i.e. the bias and skewness should preferably be small. If $\ddot{\eta}_{22} = 0$, the same transformation yields zero bias and zero skewness. However in general it will be two different transformations. If they are different skewness should be considered most important, because it is much easier to correct for bias and also because the zero skewness transformation has other properties.

It makes the likelihood normal in the multiparameter way suggested by Sprott (1980). He showed that if the likelihood is normal and a quantity F_4, derived from the fourth derivative of the likelihood function, is small, the test statistic $(\hat{\theta}_1 - \theta_1)^2/I^{11}(\hat{\theta})$ is a rather close approximation to the likelihood ratio test statistic. Here I^{11} is the upper left element of the inverse of $I(\hat{\theta})$, the socalled observed information, the matrix of second derivatives of the log likelihood function. Also this transformation minimizes the Bates & Watts curvature in a specific direction using a transformation of Type 2, i.e. (θ_1, θ_2) is transformed to $(g_1(\theta_1), g_2(\theta_2))$.

Væth (1981) examined Wald's test in oneparameter exponential families. Such a test is of the form $W = (\hat{\theta} - \theta)^2/\mathrm{Var}(\hat{\theta})$, where $\mathrm{Var}(\hat{\theta})$ is an estimate of the variance of $\hat{\theta}$. He showed that in some cases W converges to 0 for $\hat{\theta}$ converging to ∞, such that for each value of θ, the hypothesis is accepted for values of $\hat{\theta}$ large enough, values so extreme that the hypothesis ought to be rejected. This problem arises because the denominator depends more on $\hat{\theta}$ than the numerator. Using the variance stabilizing transformation that problem disappears and Wald's test is valid. As shown in Section 2 a covariance stabilizing transformation does not always exist. For testing a hypothesis about θ_1, it would suffice, if $\mathrm{Var}(\hat{\theta}_1)$ were constant. Unfortunately it is not possible to make it constant with a Type 1 or 2 transformation. All that can be obtained is that the derivative in some chosen direction is zero. $\mathrm{Var}(\hat{\theta}_1)$ is in general a real function of p variables. Therefore there will automatically be a $(p-1)$- dimensional subspace, where the derivatives are 0. A transformation of θ_1 changes this subspace, but the space will typically not increase in dimension, not even with the transformations making the derivative in some direction zero.

Also Beales (1960) measure of nonlinearity as well as Bates & Watts (1980) curvature can be reduced. However these properties give equations, which involve the projection onto the direction $\dot{\eta}_1$, whereas the other properties involves the $\dot{\eta}_1$-part of a projection onto the space spanned by $(\dot{\eta}_1,\dot{\eta}_2)$. For Bates & Watts curvature this can be changed by considering Type 2 transformations instead.

Ross (1970) suggested use of the socalled stable parameters, which only vary a little in the region of parameters fitting the data well. An example of a stable parameter is in our Special case 1 the average of all meanvalues. If the model is given by some design variables $x_1,\dots,x_n$, stable parameters can also be the meanvalues for some given values of the design variables. This usually works well in terms of reducing the nonlinearity. Bates & Watts (1981) examined the Michaelis-Menten reaction, with meanvalue function $\eta_i(\theta) = \theta_2 x_i/(\theta_1 + x_i)$, and found values r and s of the design variable, such that the parameter effects array vanishes in $\hat{\theta}$ for the parameter $\beta_1 = \theta_2\, r/(\theta_1 + r)$, $\beta_2 = \theta_2 s/(\theta_1 + s)$. As long as we only try to remove the parameter effects in a single point $\hat{\theta}$, there are many possiblities. We just need a function, which satisfies (2.1) in $\hat{\theta}$. For a model of the form $\eta_i(\theta) = \theta_2 f(\theta_1;\, x_i)$ (Special case 2), we can reduce it to a oneparameter problem by considering functions of the kind $\theta_2 h(\theta_1)$ and then choose two functions h_1,h_2, which satisfies (4.1) in $\hat{\theta}_1$, which will give a transformation with zero parameter effects for points on the line for which $\theta_1 = \hat{\theta}_1$. Choices as $h(\theta_1) = (\theta_1 + s)^\delta$, for δ a constant, $\delta \notin \{0,1\}$ and $h(\theta_1) = \exp(\theta_1 s)$, makes (4.1) to a second order equation in s, which if we are lucky has two solutions for our value of $\hat{\theta}_1$. We find Ross' (1970) solution by choosing h proportional as a function of θ_1 to f ; in the Michaelis-Menten example that is $\delta = -1$. The stable parameters have a very clear interpretation as means of observations for specific values of the design variables, but for testing hypothesis about θ_1, it is simpler to consider transformations of θ_1 alone.

ACKNOWLEDGEMENT

This project is supported by the Danish Natural Science Research Council.

The comments from Søren Johansen are greatly appreciated.

REFERENCES

1. Anderson, T.W. (1976). Estimation of linear functional relationships: Approximate distributions and connections with simultaneous equations in econometrics. J.R. Statist. Soc. B 38, 1 - 20.

2. Anscombe, F.J. (1964). Normal likelihood functions. Ann. Inst. Stat. Math., 16, 1 - 19.

3. Bates, D.M. & Watts, D.G. (1980). Relative curvature measures of nonlinearity. J.R. Statist. Soc. B 42, 1 - 25.

4. Bates, D.M. & Watts, D.G. (1981). Parameter transformations for improved approximate confidence regions in nonlinear least squares. Ann.Stat. 9, 1152-1167.

5. Beale, E.M.L. (1960). Confidence regions in non-linear estimation. J.R. Statist. Soc. B 22, 41 - 76.

6. Box, M.J. (1971). Bias in non-linear estimation. J.R. Statist. Soc B, 32, 171 - 201.

7. Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). Ann. Stat. 3, 1189 - 1242.

8. Hamilton, D.C., Watts, D.G. & Bates, D.M. (1982): Accounting for intrinsic nonlinearity in nonlinear regression parameter inference regions. Ann. Stat. 10, 386 - 393.

9. Holland, P.W. (1973). Covariance stabilizing transformations. Ann.Stat. 1, 84 - 92.

10. Hougaard, P. (1981). The appropriateness of the asymptotic distribution in a nonlinear regression model in relation to curvature. Research Report 81/9. Statistical Research Unit.Danish Medical and Social Science Research Councils. Copenhagen, Denmark.

11. Hougaard, P. (1982). Parametrizations of non-linear models. J.R. Statist. Soc. B 44, 244 - 252.

12. Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. Proc. Roy. Soc., A 196, 453 - 461.

13. Kass, R. (1984). Canonical parameterizations and zero parameter effects curvature. To appear in J.R. Statist. Soc. B.

14. Ross, G.J.S. (1970). The efficient use of function minimization in non-linear maximum likelihood estimation. Appl. Statist. 19, 205 - 221.

15. Sokolnikoff, I.S. (1951). Tensor Analysis. Wiley, New York.

16. Sprott, D.A. (1973). Normal likelihoods and their relation to large sample theory of estimation. Biometrika, 60, 457 - 465.

17. Sprott, D.A. (1980). Maximum likelihood in small samples: Estimation in the presence of nuisance parameters. Biometrika 67, 515 - 523.

18. Væth, M. (1981). On the use of Wald's test in exponential families. Research Report no 70, Department of Theoretical Statistics, Institute of Mathematics, University of Aarhus, Denmark.

PREPRINTS 1983

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK.

No. 1 Jacobsen, Martin: Two Operational Characterizations of Cooptional Times.

No. 2 Hald, Anders: Nicholas Bernoulli's Theorem.

No. 3 Jensen, Ernst Lykke and Rootzén, Holger: A Note on De Moivre's Limit Theorems: Easy Proofs.

No. 4 Asmussen, Søren: Conjugate Distributions and Variance Reduction in Ruin Probability Simulation.

No. 5 Rootzén, Holger: Central Limit Theory for Martingales via Random Change of Time.

No. 6 Rootzén, Holger: Extreme Value Theory for Moving Average Processes.

No. 7 Jacobsen, Martin: Birth Times, Death Times and Time Substitutions in Markov Chains.

No. 8 Hougaard, Philip: Convex Functions in Exponential Families.

PREPRINTS 1984

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK.

No. 1 Rootzén, Holger and Sternby, Jan: Consistency in Least Squares Estimation: A Bayesian Approach.

No. 2 Hougaard, Philip: Parameter Transformations in Multiparameter Nonlinear Regression Models.