Chapter 4

First order asymptotic theory for sequences of analytic models

1 Introduction

In this section we consider first order asymptotic theory for sequences of analytic models defined on the same parameter space. It will be assumed that the sequence of indices of the models at a given point tend to zero, and this condition essentially guarantees the usual first order asymptotic results to hold; for example, the asymptotic normality and efficiency of the local maximum likelihood estimator, cf. Section 3, and the asymptotic chi-squared distribution of some commonly used test statistics, cf. Section 4. No assumptions concerning independence or identical distributions are needed. Such assumptions may instead be used to prove that the index tends to zero as the number of observations tend to infinity. In particular, we show a number of examples in Sections 6–11 and 12–13 of sequences of independent observations for which we derive sufficient conditions for the asymptotic results to hold. These examples are within the frameworks of the generalized linear models, described in Section 5, and the generalized non-linear models, described in Section 12. In Section 2 we outline the general setup for the chapter and prove some auxiliary results.

The results proved in Sections 3 and 4 are well known from the theory of first order asymptotics for independent replications; it is the assumptions needed to prove these — partly their generality and partly their simplicity — that is of interest here. Therefore no great effort is spent on going through the wide spectrum of such asymptotic results, only a few basic ones are included. Instead we go through a number of examples in Sections 5 to 14 to demonstrate the applicability of the results to models for random variables that are independent, but not identically distributed.

In fact, the condition that the index tends to zero is sufficiently strong to prove much more refined results, such as higher order stochastic expansions of various statistics. To turn such expansions into expansions of their distributions requires, however, one further and rather awkward condition, namely a condition that generalizes Cramér's condition concerning a bound for the 'tail' of the characteristic function. This is the condition that, i.a., excludes discrete distributions from the classical Edgeworth expansions of densities. Discrete models are not excluded from the present framework of analytic models, and if we wanted higher order expansions for general sequences of analytic models we would need to impose such a generalized Cramér condition on the sequence of models. While this would definitely be feasible, the condition would usually be almost impossible to verify unless the models were within a much more restricted class. Such a class might be a class of generalized linear models with absolutely continuous densities, for which higher order expansions then might be proved by use of theorems of Edgeworth expansions for independent, but not necessarily identically distributed, random variables. Such results may be found in Bhattacharya and Rao (1976) or in Skovgaard (1986b). The crucial simplification for these models is that the characteristic functions for the individual, independent, observations all belong to the same family of distributions. Therefore the Cramér condition can be imposed on these characteristic functions, corresponding to single observations, instead of on the characteristic function for the entire vector of observation.

The higher order asymptotic theory is deferred to the next chapter where we restrict ourselves to the simpler class of models for independent and identically distributed random variables. Thus, the more complicated development of higher order expansions for more general sequences will not be considered.

2 Sequences of analytic models

In this section we describe the setup of sequences of models considered together with a few basic technical results. The setup and notation described here will be used throughout the chapter.

We consider a sequence of models

$$\{(E^{(n)},\nu^{(n)}); f^{(n)}(\cdot\,;\beta); \beta \in B \subseteq V\}, \qquad n \in \mathsf{N}, \tag{2.1}$$

on measurable spaces $E^{(n)}$ with underlying measures $\nu^{(n)}$. Notice that the densities $f^{(n)}$ are assumed to be parametrized by the same parameter space B for all $n \in \mathsf{N}$. This parameter space is assumed to be a subspace of a finite-dimensional real vector space V on which we shall consider a fixed (but arbitrary) inner product denoted $\langle\,\cdot\,,\,\cdot\,\rangle$ and the corresponding Euclidean norm $\|\cdot\|$. No relations are assumed between the sample spaces $E^{(n)}$. For applications it would be reasonable to assume that each sample space was an augmentation of the previous one and that the models similarly were models for more and more observations, but we have no need for that assumption in the general theory.

The differentials of the log-likelihood functions at the point $y^{(n)} \in E^{(n)}$, cf. (2.2.1), are denoted

$$D_k^{(n)}(\beta) = D^k \log f^{(n)}(y^{(n)};\beta) \in \mathrm{Lin}_k(V;\mathbf{R}), \tag{2.2}$$

whenever they exist. The cumulants of these differentials, from (2.3.15), are

$$\chi^{(n)}_{k_1\cdots k_m}(\beta)(v_1^{k_1},\ldots,v_m^{k_m}) = \mathrm{cum}_\beta\left\{D_{k_1}^{(n)}(\beta)(v_1^{k_1}),\ldots,D_{k_m}^{(n)}(\beta)(v_m^{k_m})\right\} \tag{2.3}$$

for $v_j \in V$, $k_j \in \mathbb{N}$, and $m \in \mathbb{N}$. In particular the Fisher information in the nth model at $\beta \in B$ is

$$I^{(n)}(\beta) = \chi_{11}^{(n)}(\beta) \tag{2.4}$$

as in (2.3.16).

We shall be concerned with asymptotic properties of sequences of statistics with distributions corresponding to a fixed point $\beta_0 \in \operatorname{int}(B)$. In the sequel we adopt, without further mentioning, the previously used convention that if the argument or subscript β is omitted, then evaluation at $\beta = \beta_0$ is understood. That the argument β_0 is sometimes included, despite this convention, is either for typographical reasons or for ease of reference in theorems, etc.

The norm induced by the Fisher information, $I^{(n)}$, at β_0 will be denoted briefly by $\|\cdot\|_n$. Thus,

$$\|v\|_n^2 = I^{(n)}(v^2) \tag{2.5}$$

for $v \in V$, corresponding to (2.5.2). Finally we recall Definition 2.5.1 of the index of an analytic model. For the nth model this will be denoted $\lambda^{(n)}(\beta)$.

In the following sections we shall be referring to the conditions listed below, the first two of which will be the bases of any of the theoretical results derived in Sections 3 and 4.

Conditions 2.1.

(A) The models in (2.1) are analytic at the point $\beta_0 \in \operatorname{int}(B)$.

(B) The indices $\lambda^{(n)} = \lambda^{(n)}(\beta_0)$ *of the models in (2.1) at the fixed point* β_0 *satisfy*

$$\lambda^{(n)} \to 0 \quad \text{as} \quad n \to \infty. \tag{2.6}$$

(C) The models in (2.1) are analytic in the neighbourhoods $U^{(n)}(\beta_0)$, *satisfying*

$$\inf\{\, \|\beta - \beta_0\|_n : \beta \notin U^{(n)}(\beta_0)\,\} \to \infty \quad \text{as} \quad n \to \infty. \tag{2.7}$$

(D) For some $n_0 \in \mathbb{N}$ *and all* $n \geq n_0$, *the Fisher information* $I^{(n)} = I^{(n)}(\beta_0)$ *is positive definite.*

(E) All eigenvalues of the Fisher information, relative to the fixed inner product on V^* *induced by the fixed inner product on* V, *tend to infinity as* n *tends to infinity.*

Obviously (C) implies (A). Whenever (A) and (B) hold, the condition (D) does not impose any genuine restriction on the sequence of models considered, because Lemma 2.5.9 shows that if the index is finite we can always parametrize the model by the projection of the parameter onto a linear subspace of V on which the Fisher information is positive definite, without reducing the class of probability measures considered. Thus, condition (D) simply states that the model is not over-parametrized.

In relation to condition (C) it is helpful to introduce the sets

$$B_n(\delta) = \{\, \beta \in U^{(n)}(\beta_0) : \|\beta - \beta_0\|_n < \delta \rho_n^{-1}\,\}, \tag{2.8}$$

and their relative closures

$$\overline{B}_n(\delta) = \{\beta \in U^{(n)}(\beta_0) : \|\beta - \beta_0\|_n \leq \delta \rho_n^{-1}\}, \tag{2.9}$$

for $n \in \mathbb{N}$ and $\delta > 0$, where $\rho_n = 2(e\sqrt{p})\lambda^{(n)}$ is the 'inverse radius of convergence' from Corollary 2.5.4. The set $B_n(\delta)$ is identical to the set $U_a(\beta_0)$ in (2.3.1) with $a = \delta\rho_n^{-1}$ for the nth model, except that we use the Fisher information norm in (2.8). The set $B_n(\delta)$ is a sphere in V, restricted to the domain of analyticity of the model. The condition (C) above guarantees that for any $\delta > 0$, any sphere with fixed radius in the Fisher information norm will eventually be contained in $B_n(\delta)$.

This assumption will be required for any kind of asymptotic result concerning likelihood based inference in the neighbourhood of β_0 because the region in which the likelihood function needs to be approximated by a Taylor series expansion around β_0 must eventually include any point within a fixed distance in terms of the Fisher information norm. In fact, condition (B) ensures the radius of convergence of such expansions to tend to infinity, while the condition (C) guarantees the functions to equal their Taylor series expansions in regions increasing beyond any boundaries, still in terms of the Fisher information norm.

The condition (C) will be satisfied if all the models, for sufficiently large n, are analytic in some fixed neighbourhood of β_0 in V and the condition (E) holds. Notice that in condition (E) it does not matter which fixed inner product on V is chosen, the important thing is that it does not change with n, so the eigenvalues in this condition may be computed in the familiar way as the eigenvalues of the Fisher information matrix with respect to a fixed coordinate system. Although condition (E) is very weak in view of the asymptotic results that are to be proved, such as the asymptotic normality of the local maximum likelihood estimator, it will not be necessary in the general theory developed in Sections 3 and 4 to assume that it holds. Thus, the sequence of models considered may converge towards a linear normal model with limited information, in which case the asymptotic results hold without condition (E) being fulfilled. The convergence to a normal linear model will essentially be guaranteed by condition (B). Note that the estimator will usually not be consistent for such a sequence of models, but despite that its asymptotic normality may be proved. For the examples of the generalized linear or non-linear type considered later in this chapter the situation is different. Here the condition (E) will almost always be required to hold and the problem is usually what extra conditions are needed, if any, for condition (B) to hold also.

A natural question is to what extent Conditions 2.1 depend on the particular choice of parametrization. We consider only analytic reparametrizations of the form (2.6.1), i.e., including analytically parametrized sub-models. Furthermore we restrict the parametrizations to be of maximal rank, i.e., the first differential of the mapping in (2.6.1) must map the new parameter space, W say, onto a subspace of V of dimension equal to $\dim W$. We know from Theorem 2.6.1 that the models in the new parametrization are analytic if the original models are. Thus, condition (A) implies the same condition to hold in the new parametrization also. The same is trivially true for each of the conditions (D) and (E). For the more crucial condition (B) it is seen from (2.6.9) that we cannot expect the same condition to hold in the new parametrization unless the quantity R from (2.6.8) tends to zero as n tends to infinity. This will be the case if condition (E) holds, as is easily seen from (2.6.8). Thus, if (A), (B) and (E) all hold then (B) holds for the new parametrization. Similarly, conditions (C) and (E) together imply that (C) holds in the new parametrization. In conclusion, for the majority of cases, namely whenever condition (E) holds, the parametrization does not matter for the asymptotic results that we are going to prove in subsequent sections, because it will always be assumed that the models considered are analytic at β_0, and then each of the conditions will hold for the new parametrization if it holds for the original one.

To prove the convergence of the distribution of the score function or of the local maximum likelihood estimator to a normal distribution we need some kind of standardization of the variances. An estimator, $\hat{\beta}_n$, say, takes values in the space $B \subseteq V$ which is independent of n, but its asymptotic variance, usually $I^{(n)}(\beta_0)^{-1}$, depends on n and the distribution typically converges to a single point. The usual method is to transform the statistic, $\hat{\beta}_n - \beta_0$ say, by a square root of the inverse variance, e.g., by $I^{(n)}(\beta_0)^{1/2}$. Although this method may be shown to be equivalent to the approach chosen below, it is, in some sense, unsatisfactory because the square root is not unique unless V is one-dimensional. It is more in line with the development in Chapter 2 to consider $\hat{\beta} - \beta_0$ as a vector in the Euclidean space V equipped with the inner product $I^{(n)}(\beta_0)$. The standardized normal distribution on this space has variance $I^{(n)}(\beta_0)^{-1}$ and we want to consider 'convergence' to those standardized distributions as n tends to infinity. Since these standardized normal distributions depend on n we need a modification of the concept of convergence in distribution. A method would be to identify these Euclidean spaces with a fixed Euclidean space through a sequence of linear mappings, but that involves arbitrary choices of linear mappings, equivalent to the choices of square roots above. Instead we shall define the convergence through the convergence of any such sequence of linear transformations for which the asymptotic variance becomes constant. Thus, our definition of asymptotic normality below will replace the usual definition of convergence of the standardized distribution towards a normal distribution. First, however, we include a definition of the normal distribution, primarily to emphasize that we allow the variance to be singular. In particular, the normal distributions on $\mathbf{R}$ with variance zero are defined as the one-point distributions.

Definition 2.2. *For any $\mu \in V$ and positive semi-definite $\Gamma \in \mathrm{Sym}_2(V^*;\mathbf{R})$, we define the normal distribution $N(\mu,\Gamma)$ on the finite-dimensional real vector space V as the unique distribution with moment generating function*

$$t \mapsto \exp\{t(\mu) - \frac{1}{2}\Gamma(t^2)\}, \quad t \in V^*. \tag{2.10}$$

In the following definition we use the identification of an element in $\mathrm{Lin}_2(V^*;\mathbf{R})$, namely the variance Γ from above, with an element in $\mathrm{Lin}(V^*;V)$, cf. (1.1.11). We do not distinguish notationally between these two mappings.

Definition 2.3. *The distributions of a sequence of random variables X_n on finite-dimensional real vector spaces V_n is said to be asymptotically normal with* *asymptotic expectations $\mu_n \in V_n$ and asymptotic variances $\Gamma_n \in \mathrm{Lin}(V_n^*; V_n)$, if the distributions of $A_n(X_n - \mu_n)$ converges to the normal distribution $N(0, \Gamma)$ on W for any sequence of linear mappings $A_n \in \mathrm{Lin}(V_n; W)$ into some finite-dimensional real vector space W satisfying the condition that for all sufficiently large n the 'transformed asymptotic variance'*

$$\Gamma = A_n \Gamma_n A_n^T \in \mathrm{Lin}(W^*; W) \tag{2.11}$$

does not depend on n.

As noticed earlier this definition is equivalent to the convergence in distribution towards a standardized normal distribution for any particular sequence of standardized random variables $\Gamma_n^{-1/2}(X_n - \mu_n)$, provided that the spaces V_n are of the same dimension and that the asymptotic variances Γ_n are positive definite for sufficiently large n. Thus, the choices of square roots has no impact on this convergence.

An important well-known fact is that to prove the asymptotic normality for a sequence X_n according to the definition above, it is sufficient to consider sequences of real mappings, i.e., to take $W = \mathbf{R}$ in the definition. We shall use this fact without further comments in proofs in Section 3.

We conclude this section with a technical result of some independent interest, concerning the behaviour of the log-likelihood differentials. Recall from Corollary 2.5.4 and Lemma 2.4.1 that if the model is analytic at β_0 then

$$\|\chi_k^{(n)}\|_n \le k!\, \rho_n^{k-2} I_{\{k>1\}}, \tag{2.12}$$

and

$$\|D_k^{(n)} - \chi_k^{(n)}\|_n \le k!\, \rho_n^{k-1} H_n, \tag{2.13}$$

where $\rho_n = 2(e\sqrt{p})\lambda^{(n)}$, and $H_n = H_n(Y^{(n)}; \beta_0)$ from (2.4.2) and (2.4.3) is a non-negative real random variable satisfying

$$\mathrm{E}_{\beta_0} \exp\{\delta H_n\} < \gamma(p) \exp\{2p(e\delta)^2/(1 - \delta\rho_n)\} \tag{2.14}$$

for any $\delta < \rho_n^{-1}$, where $\gamma(p)$ is a constant depending only on $p = \dim V$.

Lemma 2.4. *Consider any sequence of random variables H_n, satisfying (2.14) above for which the sequence ρ_n is bounded. Then the β_0-distribution of H_n is tight, i.e.,*

$$\mathrm{P}_{\beta_0}\{\, |H_n| \ge A\} \to 0 \tag{2.15}$$

as $A \to \infty$, uniformly in n. Furthermore, for any $m \in \mathbf{N}$, the sequence of random variables H_n^m is uniformly integrable, i.e.,

$$\mathrm{E}_{\beta_0} \left\{ |H_n|^m I_{\{|H_n| \ge A\}} \right\} \to 0 \tag{2.16}$$

as $A \to \infty$, uniformly in n. Also the sequence $\exp\{sH_n\}$ *is uniformly integrable for sufficiently small $s > 0$.*

Proof. From (2.14) and Lemma 1.4.13 we get

$$\mathrm{P}\{H_n \geq A\} \leq \gamma(p)\exp\{2p(e\delta)^2/(1-\delta\rho_n)\}\exp(-A\delta)$$

for any $\delta < \rho_n^{-1}$. If we keep δ fixed with $\delta < a\rho_n^{-1}$ for all n, where $0 < a < 1$, and let A tend to infinity, then the right hand side tends to zero uniformly in n because ρ_n is bounded as a function of n.

Next, let $s > 0$ and $\delta > 0$ be such that $\delta + s < a\rho_n^{-1}$ for all n, where still $0 < a < 1$. Then

$$\begin{aligned}\int_A^\infty \exp(sH_n)\,dP_{\beta_0}(y) &\leq exp(-A\delta)\int_A^\infty \exp\{(s+\delta)H_n\}\,dP_{\beta_0}(y)\\ &\leq exp(-A\delta)\gamma(p)\exp\{2pe^2(\delta+s)^2/(1-(\delta+s)\rho_n\}\end{aligned}$$

which also tends to zero as A tends to infinity, uniformly in n. This shows the uniform integrability of $\exp(sH_n)$. The uniform integrability of H_n^m for $m \in \mathbb{N}$ follows from the inequality

$$H_n^m \leq m!\exp(sH_n)s^{-m}.$$

∎

In the bounds (2.13) it would have been nice if the random variables H_n could be chosen independently of n and still possess finite exponential moments. The lemma states that this is "nearly so" in the sense that bounds for tail probabilities and 'tail moments', even in exponential form, hold independently of n. The basic result showing this is, of course, the inequality (2.14) taken from Lemma 2.4.1. With this result it is easy to provide uniform bounds for the error arising from the truncation of Taylor series expansions of the log-likelihood function when this is used to derive distributional approximations, as we shall see in the following section.

3 Asymptotic normality of the local maximum likelihood estimator

For sequences of analytic models for which the indices tend to zero we demonstrate the asymptotic normality of the local maximum likelihood estimator in this section. By 'local' we mean the maximum likelihood estimator for the model restricted to some neighbourhood of β_0; a precise formulation is given below. Since the assumptions concern only the behaviour of the model in such a neighbourhood there is no way that these assumptions can lead to results concerning the behaviour of the global maximum likelihood estimator, i.e., the maximum likelihood estimator for the full model. Thus, quite different techniques have to be used to investigate whether the local maximum likelihood estimator agrees with the global one. That subject will not be pursued here for the general case, but for the case of independent replications the consistency of the maximum likelihood estimator is discussed in Section 5.4.

The basic argument for the local maximum likelihood estimator $\hat{\beta}_n$, a precise version of which follows later in this section, is the following well-known simple approximation. The first differential, $D_1^{(n)}(\hat{\beta}_n)$, of the log-likelihood function at $\hat{\beta}_n$, is zero. An expansion around β_0 yields

$$0 = D_1^{(n)}(\beta_0) + D_2^{(n)}(\beta_0)(\hat{\beta}_n - \beta_0) + R_n(\hat{\beta}_n), \tag{3.1}$$

where $R_n(\hat{\beta}_n)$ is the error term of this Taylor series approximation to $D_1^{(n)}(\hat{\beta}_n)$. In terms of the Fisher information norm, $R_n(\hat{\beta}_n)$ is of order $O(\lambda^{(n)})$ and so is $D_2^{(n)}(\beta_0) + I^{(n)}(\beta_0)$. Thus, to first order we have

$$0 \sim D_1^{(n)}(\beta_0) - I^{(n)}(\beta_0)(\hat{\beta}_n - \beta_0), \tag{3.2}$$

which implies that

$$\hat{\beta}_n - \beta_0 \sim I^{(n)}(\beta_0)^{-1}\left(D_1^{(n)}(\beta_0)\right). \tag{3.3}$$

The asymptotic normality of the score statistic $D_1^{(n)}(\beta_0)$ follows from the fact that all its cumulants, except the variance, tend to zero in terms of the Fisher information norm as n tends to infinity, because of the assumption that the index tends to zero. Then the asymptotic normality of the local maximum likelihood estimator follows from (3.3). We shall first prove the asymptotic normality of the score statistic.

Lemma 3.1. *For a sequence of models fulfilling Conditions 2.1 (A) and (B) the β_0-distributions of the sequence of score statistics $D_1^{(n)}(\beta_0) \in V^*$ is asymptotically normal with asymptotic expectations 0 and asymptotic variances $I^{(n)}(\beta_0)$.*

Proof. Consider a sequence $A_n \in \mathrm{Lin}(V^*; \mathbf{R})$ of real linear mappings, such that

$$A_n I^{(n)}(\beta_0) A_n^T = A,$$

say, is independent of n. According to Definition 2.3 we need to show that the distribution of $A_n\{D_1^{(n)}(\beta_0)\}$ converges to $N(0, A)$. The mapping A_n may be identified with a vector in V, and A_n^T with a vector in V^*. Thus, A may be identified with a non-negative real number, namely

$$A = I^{(n)}(\beta_0)(A_n^2) = \|A_n\|_n^2,$$

which is the variance, in the traditional sense, of $D_1^{(n)}(A_n)$. The mth cumulant of this random variable satisfies

$$\begin{aligned}\left|\text{cum}_m\left\{D_1^{(n)}(A_n)\right\}\right| &= |\chi_{1\cdots1}(A_n^m)| \\ &\leq m!\,\lambda^{(n)}(\beta_0)^{m-2}\|A_n\|_n^m \\ &= m!\,\lambda^{(n)}(\beta_0)^{m-2}A^{m/2} \qquad (3.4)\end{aligned}$$

for any $m \geq 2$. From (3.4) we see that for any $m \geq 3$, the mth cumulant tends to zero as n tends to infinity, because $\lambda^{(n)}$ tends to zero, while the means are identically zero for all n and the variance is constantly equal to A. Thus, all cumulants converge to the cumulants of the normal distribution $N(0, A)$. Since the normal distribution is characterized by its cumulants, also in the case $A = 0$, the convergence in distribution towards $N(0, A)$ follows. ∎

Notice that since any analytic reparametrization induces a linear transformation of the score statistic, the conclusion of the lemma holds for any reparametrized model, even if the index does not tend to zero in the reparametrized model.

For a more precise version of the argument given in the beginning of this section we now give a precise definition of the local maximum likelihood estimator.

Definition 3.2. *Any measurable function $\hat{\beta}_n(\delta) : E^{(n)} \to \overline{B}_n(\delta)$, where $\overline{B}_n(\delta)$ is the set defined in (2.9), which maximizes the likelihood function*

$$\beta \mapsto f^{(n)}(y^{(n)}; \beta)$$

as a function of $\beta \in \overline{B}_n(\delta)$, is called a local maximum likelihood estimator (LMLE).

Notice that when Conditions 2.1 (B) and (C) hold, then for sufficiently large n, the set $\overline{B}_n(\delta)$ is closed in V and with probability one an LMLE exists, but it may not be unique and it may be on the boundary of $\overline{B}_n(\delta)$. The following theorem excludes those two possibilities with probability tending to one, and demonstrates the asymptotic normality of the LMLE for sufficiently small δ.

Theorem 3.3. *Assume that Conditions 2.1 (B), (C) and (D) are satisfied. Then, for some sufficiently small δ the following assertions are true:*

(1) With probability tending to one as n tends to infinity, there is a unique solution in $B_n(\delta)$ to the likelihood equation $D_1(\beta) = 0$, uniquely maximizing the likelihood function on $\overline{B}_n(\delta)$.

(2) Any LMLE sequence $\hat{\beta}_n(\delta)$ is asymptotically normal with asymptotic expectations β_0 and asymptotic variances $I^{(n)}(\beta_0)^{-1}$.

(3) Any LMLE sequence $\hat{\beta}_n(\delta)$ satisfies

$$\|I^{(n)}(\beta_0)^{-1}(D_1^{(n)}) - (\hat{\beta}_n - \beta_0)\|_n \xrightarrow{\mathrm{P}} 0 \tag{3.5}$$

as $n \to \infty$.

Proof. The technique of proof follows that of Theorem 4.1 in Lehmann (1983), Chapter 6. There the result was proved in a somewhat more restricted setting, namely for independent identically distributed observations and for efficient maximum likelihood estimators, but the bounds in (2.12) and (2.13) allow essentially the same proof to hold.

Let $(a_n, n \in \mathbb{N})$ be a sequence of positive numbers satisfying the conditions

$$a_n\rho_n < \delta < \frac{1}{3} \qquad \text{and} \qquad a_n \to \infty \quad \text{as} \quad n \to \infty. \tag{3.6}$$

We first want to show that with probability tending to one, the likelihood function on $\overline{B}_n(\delta)$ is smaller anywhere outside the sphere $\{\|\beta - \beta_0\| < a_n\}$ than at β_0. To do this write the log-likelihood function on $\overline{B}_n(\delta)$ in the form

$$\begin{aligned} \ell^{(n)}(\beta) &= \log f^{(n)}(y^{(n)};\beta) \\ &= \ell^{(n)}(\beta_0) + D_1^{(n)}(\beta - \beta_0) - \frac{1}{2}I^{(n)}(\beta - \beta_0)^2 + R_1^{(n)}(\beta) \end{aligned} \tag{3.7}$$

where it follows from (2.12) and (2.13) that

$$\begin{aligned} |R_1^{(n)}(\beta)| &= \left|\sum_{k=2}^{\infty} \frac{1}{k!}(D_k^{(n)} - \chi_k^{(n)})(\beta - \beta_0)^k + \sum_{k=3}^{\infty} \frac{1}{k!}\chi_k^{(n)}(\beta - \beta_0)^k\right| \\ &\le \sum_{k=2}^{\infty} \rho_n^{k-1} H_n \|\beta - \beta_0\|_n^k + \sum_{k=3}^{\infty} \rho_n^{k-2} \|\beta - \beta_0\|_n^k \\ &< \frac{\delta}{1-\delta} H_n \|\beta - \beta_0\|_n + \frac{\delta}{1-\delta} \|\beta - \beta_0\|_n^2, \end{aligned} \tag{3.8}$$

where $\rho_n = 2(e\sqrt{p})\lambda^{(n)}$ and H_n is the random variable from (2.13). Also, we have

$$|D_1^{(n)}(\beta - \beta_0)| \le H_n \|\beta - \beta_0\|_n \tag{3.9}$$

and hence

$$\ell^{(n)}(\beta) \le \ell^{(n)}(\beta_0) - \frac{1}{2}\|\beta - \beta_0\|_n^2 + \frac{\delta}{1-\delta} \|\beta - \beta_0\|_n^2 + \frac{1}{1-\delta} H_n \|\beta - \beta_0\|_n. \tag{3.10}$$

From Lemma 2.4 we know that H_n/a_n tends to zero in probability, and therefore that for any $\beta \in \overline{B}_n(\delta)$ with $\|\beta - \beta_0\|_n \geq a_n$ we have

$$\ell^{(n)}(\beta) - \ell^{(n)}(\beta_0) \leq \left(-\frac{1}{2} + \frac{\delta}{1-\delta} + \frac{H_n}{a_n}\right) \|\beta - \beta_0\|_n^2.$$

The probability that the quantity in the parenthesis on the right is negative tends to one as $n \to \infty$. Condition 2.1 (C) shows that there is some sequence of a_n's, satisfying (3.6), such that

$$\{\beta \in V : \|\beta - \beta_0\|_n \leq a_n\} \subseteq B_n(\delta), \tag{3.11}$$

for all $n \in \mathbb{N}$, i.e., such that the range of analyticity of the nth model contains the closed sphere on the left in (3.11). For such a sequence it appears from above that with limiting probability one the likelihood function on $\overline{B}_n(\delta)$ will be maximized somewhere in the interior of the sphere $\{\|\beta - \beta_0\|_n \leq a_n\}$, because the log likelihood function is smaller anywhere on the boundary of, and outside this sphere than at its center. The maximum must be a solution to the likelihood equation

$$0 = D_1^{(n)}(\beta) = D_1^{(n)} - I^{(n)}(\beta - \beta_0) + R_2^{(n)}(\beta),$$

where, for any $v \in V$,

$$\begin{aligned} |\{R_2^{(n)}(\beta)\}(v)| &= \Bigg| \sum_{k=2}^{\infty} \frac{1}{(k-1)!}(D_k^{(n)} - \chi_k^{(n)})\left((\beta-\beta_0)^{k-1}, v\right) \\ &\qquad\qquad + \sum_{k=3}^{\infty} \frac{1}{(k-1)!}\chi_k^{(n)}\left((\beta-\beta_0)^{k-1}, v\right)\Bigg| \\ &\leq \sum_{k=2}^{\infty} k\rho_n^{k-1} H_n \|\beta - \beta_0\|_n^{k-1} \|v\|_n + \sum_{k=3}^{\infty} k\rho_n^{k-2} \|\beta - \beta_0\|_n^{k-1} \|v\|_n \\ &< \left\{\frac{2}{(1-\delta)^2}\rho_n a_n H_n + \frac{3}{(1-\delta)^2}\rho_n a_n \|\beta - \beta_0\|\right\} \|v\|_n. \end{aligned} \tag{3.12}$$

If the sequence (a_n) satisfies $a_n^2 \rho_n \to 0$ as $n \to \infty$, still with $a_n \to \infty$, then the factor in curly brackets in the last expression in (3.12) tends to zero in probability as n tends to infinity. Consequently any solution, $\hat{\beta}_n$ say, to the likelihood equation, satisfying $\|\hat{\beta}_n - \beta_0\| \leq a_n$ fulfils the relation

$$\|I^{(n)}(\hat{\beta}_n - \beta_0) - D_1^{(n)}\|_n \xrightarrow{\mathrm{P}} 0 \tag{3.13}$$

as $n \to \infty$, where we have used $\|\cdot\|_n$ to denote the $I^{(n)}(\beta_0)^{-1}$ norm on V^*, cf. (1.1.17). The result (3.13) is equivalent to (3.5) because the mapping $I^{(n)}$ from V to V^* is isometric with respect to the two Fisher information inner products, $I^{(n)}(\beta_0)$ on V and $I^{(n)}(\beta_0)^{-1}$ on V^*.

Next we want to show that the solution to the likelihood equation is unique on $\overline{B}_n(\delta)$ with probability tending to one. We show this by showing that the second differential of the log-likelihood function is negative definite on $\overline{B}_n(\delta)$ with limiting probability one. Analogously to the computations above we see that for any $v \in V$, this second differential satisfies the inequalities

$$\begin{aligned}
D_2^{(n)}(\beta)(v^2) &\le -I^{(n)}(v^2) + \left| \sum_{k=2}^{\infty} \frac{1}{(k-2)!}(D_k^{(n)} - \chi_k^{(n)})\left((\beta-\beta_0)^{k-2}, v^2\right) \right| \\
&\quad + \left| \sum_{k=3}^{\infty} \frac{1}{(k-2)!} \chi_k^{(n)} \left((\beta-\beta_0)^{k-2}, v^2\right) \right| \\
&\le \left\{ -1 + \sum_{k=2}^{\infty} k(k-1)\rho_n^{k-1} H_n \|\beta-\beta_0\|_n^{k-2} \right. \\
&\quad \left. + \sum_{k=3}^{\infty} k(k-1)\rho_n^{k-2} \|\beta-\beta_0\|_n^{k-2} \right\} \|v\|_n^2 \\
&< \left\{ -1 + \frac{2}{(1-\delta)^3}\rho_n H_n + \left(\frac{2}{(1-\delta)^3} - 2 \right) \right\} \|v\|_n^2. \qquad (3.14)
\end{aligned}$$

Since $\rho_n H_n \to 0$ in probability as $n \to \infty$, the probability that this expression is negative for all $v \neq 0$ tends to one if δ is chosen such that $(1-\delta)^3 > 2/3$. This concludes the proof of (1) and shows that any δ satisfying $\delta < \sqrt[3]{2/3}$ is sufficiently small. The assertion (3) was proved above and this together with Lemma 3.1 trivially implies (2). ∎

Notice that the regions $B_n(\delta)$ on which we can prove the unique maximization of the likelihood function, increase at the rate of ρ_n^{-1} in the scale of the standardized distribution of the LMLE, i.e., in the Fisher information norm. For the case of independent identically distributed observations, ρ_n^{-1} will be proportional to $\sqrt{n}$ and the regions $B_n(\delta)$ will be equal to a fixed set in V.

Unlike Lemma 3.1, the result of Theorem 3.3 may well depend on the particular choice of parametrization. This is easily seen from the case when the model for any n is identical to a fixed linear normal model. This constant sequence of models satisfies the conditions of the theorem, but a non-linear reparametrization will not in general lead to a normally distributed estimator. However, if Condition 2.1 (E) holds, it follows from the discussion of reparametrizations following Conditions 2.1 that the conditions for the theorem will also be fulfilled in any analytically reparametrized model, provided that the reparametrization is of maximal rank, so for those cases the parametrization is not important for the asymptotic results.

In the assertions (2) and (3) in Theorem 3.3 we might replace the Fisher information $I^{(n)}(\beta_0)$ anywhere by $I^{(n)}(\hat{\beta}_n)$, or by the 'observed Fisher information' $-D_2^{(n)}(\hat{\beta}_n)$, or by $-D_2^{(n)}(\beta_0)$, without affecting the results. Since these alternatives are random it does not make sense, according to Definition 2.3 to use them as inverse asymptotic variances, but the statement in (2) might be modified to read that the standardized version of $\hat{\beta}_n - \beta_0$, e.g., $I^{(n)}(\hat{\beta}_n)^{1/2}(\hat{\beta}_n - \beta_0)$ converges to a standard normal distribution for any sequence of square roots. That we may use minus the second differential instead of its expectation follows from (2.13) along the lines of the majorizations used in the proof of the theorem. That we may replace β_0 by $\hat{\beta}_n$ as the argument for the Fisher information follows from the fact that $(\hat{\beta}_n - \beta_0)$ converges to zero in probability in the Fisher information norm, combined with Lemma 2.5.6 and the assumption that the index tends to zero as n tends to infinity. By use of these results we conclude that

$$I^{(n)}(\beta_0)^{-1} \circ I^{(n)}(\hat{\beta}_n)$$

converges in probability to the identity mapping on V, where we have identified the two informations with mappings in $\mathrm{Lin}(V; V^*)$ a described in (1.1.11). The possibility of replacing the information at β_0 by any of its three estimates mentioned above also applies to its use in the definition of the n-norm used in (3) in the Theorem.

4 Asymptotic distributions of test statistics

It is easy on the basis of the results from the previous section to derive the limiting chi-squared distributions of some test statistics for the simple hypothesis $\beta = \beta_0$. Thus, in the framework of Section 3, we may consider the three test statistics

$$T_1 = \|D_1^{(n)}\|_n^2 = I^{(n)}(\beta_0)^{-1}\left(D_1^{(n)}(\beta_0)^2\right), \tag{4.1}$$

$$T_2 = \|\hat{\beta}_n - \beta_0\|_n^2 = I^{(n)}(\beta_0)\left(\hat{\beta}_n - \beta_0\right)^2, \tag{4.2}$$

$$T_3 = 2\left\{\log f^{(n)}(Y^{(n)}; \hat{\beta}_n) - \log f^{(n)}(Y^{(n)}; \beta_0)\right\}, \tag{4.3}$$

referred to in the sequel as the score test, Wald's test, and the log likelihood ratio test, respectively. In (4.2) and (4.3) we have omitted the δ from the local maximum likelihood estimator $\hat{\beta}_n(\delta)$ from Section 3. It follows immediately from Lemma 3.1 that if Conditions 2.1 (A), (B) and (D) hold, then the score test statistic T_1 converges in distribution to a chi-squared distribution on p degrees of freedom, where $p = \dim V$. If also the condition (C) holds then Theorem 3.3 implies the limiting chi-squared distribution of Wald's test statistic T_2. In fact, it follows that the difference $T_1 - T_2$ tends to zero in probability, so the two test statistics are asymptotically equivalent to first order. This equivalence also comprises the log likelihood ratio test statistic T_3, because it follows from Theorem 3.3 that the LMLE $\hat{\beta}_n$ satisfies

$$\log f^{(n)}(Y^{(n)}; \hat{\beta}_n) - \log f^{(n)}(Y^{(n)}; \beta_0)$$

$$
\begin{aligned}
&= D_1^{(n)}(\hat{\beta}_n - \beta_0) - \frac{1}{2} I^{(n)}(\hat{\beta}_n - \beta_0)^2 + o_p(1) \\
&= \frac{1}{2} I^{(n)}(\hat{\beta}_n - \beta_0)^2 + o_p(1), \qquad (4.4)
\end{aligned}
$$

where $o_p(1)$ is a term that tends to zero in probability as n tends to infinity. Thus, all three test statistics are asymptotically equivalent and have limiting chi-squared distributions on p degrees of freedom.

Moreover, the asymptotic equivalences mentioned at the end of Section 3 imply that in (4.1) and (4.2) we may replace $I^{(n)}(\beta_0)$ by $-D_2^{(n)}(\beta_0)$, by $-D_2^{(n)}(\hat{\beta}_n)$, or by $I^{(n)}(\hat{\beta}_n)$, without affecting the first order asymptotic behaviour of the test statistics.

In the remaining part of this section we shall generalize these results to composite hypotheses. As the results above, these results are well-known from likelihood theory for regular models in the case of independent identically distributed observations. The generalization to the present setting is quite trivial on the basis of Theorem 3.3 and the proofs are therefore given here in a somewhat abbreviated form.

We consider first linear hypotheses. The notation introduced below follows that in Section 2.6 in relation to Theorem 6.1. Thus, let

$$\beta : A \to \mathrm{int}(B) \qquad (4.5)$$

denote a mapping from $A \subseteq W$ into the parameter space B, where W is a real vector space with $\dim W = q < \infty$, and assume that the mapping is a linear mapping of full rank, i.e., that $\beta(\alpha_1 - \alpha_2) = 0$ implies $\alpha_1 - \alpha_2 = 0$. Beside the sequence of models described in Section 2, parametrized by $\beta \in B$, we consider the sequence of submodels

$$\{\tilde{f}^{(n)}(y^{(n)}; \alpha); \alpha \in A \subseteq W\}, \qquad (4.6)$$

parametrized by α and with the densities defined by

$$\tilde{f}^{(n)}(y^{(n)}; \alpha) = f^{(n)}(y^{(n)}; \beta(\alpha)). \qquad (4.7)$$

The common misuse of the notation β as a parameter and as a mapping should not give rise to ambiguities.

Quantities in the submodel (4.6) are denoted with a tilde, e.g.,

$$\tilde{D}_k^{(n)}(\alpha) = D^k \log \tilde{f}^{(n)}(y^{(n)}; \alpha). \qquad (4.8)$$

Let $\alpha_0 \in \mathrm{int}(A)$ be a fixed point with $\beta_0 = \beta(\alpha_0) \in \mathrm{int}(B)$. Again an omitted argument implies evaluation at α_0 or at β_0.

The Fisher information semi-norm on W is denoted $||\cdot||_n$, which is distinguished from the norm $||\cdot||_n$ on V by the fact that the argument is a vector in W. Thus, for $w \in W$ we have

$$\|w\|_n^2 = \tilde{I}^{(n)}(w^2) = I^{(n)}\left(\{D\beta(w)\}^2\right) = \|D\beta(w)\|_n^2. \tag{4.9}$$

Although $D\beta = D\beta(\alpha_0)$ is the notation for the differential of β at α_0, it should be noticed that presently $D\beta(\alpha)$ is independent of $\alpha \in A$ and equals the mapping β.

In the submodel (4.6) we denote the local maximum likelihood estimator by $\tilde{\alpha}_n = \tilde{\alpha}_n(\delta)$ and let $\tilde{\beta}_n = \beta(\tilde{\alpha}_n)$. We still let $\hat{\beta}_n = \hat{\beta}_n(\delta)$ denote the LMLE in the original model.

Lemma 4.1. *Assume that Conditions 2.1 (B), (C) and (D) hold and consider the submodel (4.6) induced by the mapping β, which is assumed to be a linear mapping of full rank. Then the Conditions 2.1 (B), (C) and (D) also hold for the submodel, and for sufficiently small $\delta > 0$ the LMLE $\tilde{\alpha}_n = \tilde{\alpha}_n(\delta)$ satisfies*

$$\|(\tilde{\alpha}_n - \alpha_0) - \tilde{I}_n(\alpha_0)^{-1}\left(\tilde{D}_1^{(n)}\right)\|_n \xrightarrow{\mathrm{P}} 0 \quad \text{as} \quad n \to \infty. \tag{4.10}$$

Furthermore, $\tilde{\beta}_n - \beta_0 = \beta(\tilde{\alpha}_n) - \beta(\alpha_0)$ and $\hat{\beta}_n - \tilde{\beta}_n$ are asymptotically independent, and

$$\|(\hat{\beta}_n - \tilde{\beta}_n) - I^{(n)}(\beta_0)^{-1}\left(D_1^{(n)}(\tilde{\beta}_n)\right)\|_n \xrightarrow{\mathrm{P}} 0 \quad \text{as} \quad n \to \infty. \tag{4.11}$$

Proof. It follows directly from Theorem 2.6.1 that the submodel (4.6) is analytic at α_0 and that the index at α_0 satisfies the inequality $\tilde{\lambda}^{(n)}(\alpha_0) \leq \lambda^{(n)}(\beta_0)$ since the condition (2.6.8) holds with $R = 0$ for a linear mapping. Hence Condition 2.1 (B) holds for the submodel. By use of the same argument at any other point of analyticity, and by use of the equality

$$\|\alpha - \alpha_0\|_n = \|\beta(\alpha) - \beta_0\|_n$$

which follows from (4.9), we see that also Condition 2.1 (C) holds. Finally, (D) is seen from (4.9) to be satisfied because $(D\beta)$ is assumed to be of full rank.

Since all the conditions hold for the submodel also, the results from Lemma 3.1 and Theorem 3.3 apply. In particular, (4.10) is the submodel version of (3.5). Now, (4.11) follows from the computation

$$\begin{aligned} D_1^{(n)}(\tilde{\beta}_n) &= D_1^{(n)}(\beta_0) - I^{(n)}(\tilde{\beta}_n - \beta_0) + o_p(1) \\ &= D_1^{(n)}(\beta_0) - I^{(n)}(\hat{\beta}_n - \beta_0) + I^{(n)}(\hat{\beta}_n - \tilde{\beta}_n) + o_p(1) \\ &= I^{(n)}(\hat{\beta}_n - \tilde{\beta}_n) + o_p(1), \end{aligned}$$

where $o_p(1)$ refers to a vector for which the $\|\cdot\|_n$-norm tends to zero in probability. This estimation of the remainder follows the line of proofs from Section 3.

The asymptotic independence between $\tilde{\beta}_n - \beta_0$ and $\hat{\beta}_n - \tilde{\beta}_n$ follows from Lemma 3.1 together with the asymptotic equivalence of these two vectors with two orthogonal projections, with respect to $I^{(n)}$, of $D_1^{(n)}$ onto the subspace spanned by

$(D\beta)$ and its orthogonal complement, respectively. This asymptotic equivalence with the orthogonal projections will not be proved here, but is given in the proof of Theorem 4.2 below. ▌

We want to consider the hypothesis $\beta \in \beta(A)$, i.e., that the parameter β belongs to the subspace parametrized by $\alpha \in A$. We shall establish the asymptotic equivalence and the limiting chi-squared distribution of the three test statistics

$$\|D_1^{(n)}(\tilde{\beta}_n)\|_n^2 = I^{(n)}(\tilde{\beta}_n)^{-1}\left\{\left(D_1^{(n)}(\tilde{\beta}_n)\right)^2\right\}, \tag{4.12}$$

$$\|\hat{\beta}_n - \tilde{\beta}_n\|_n^2 = I^{(n)}(\tilde{\beta}_n)\left\{(\hat{\beta}_n - \tilde{\beta}_n)^2\right\} \tag{4.13}$$

$$2\left\{\log f^{(n)}(Y^{(n)};\hat{\beta}_n) - \log f^{(n)}(Y^{(n)};\tilde{\beta}_n)\right\}, \tag{4.14}$$

which are natural generalizations of (4.1), (4.2) and (4.3).

Theorem 4.2. *Under the conditions of Lemma 4.1 the three test statistics in (4.12), (4.13) and (4.14) are asymptotically equivalent, i.e., their pairwise differences tend to zero in probability. Their (common) limiting distribution as n tends to infinity, is a chi-squared distribution on $p - q$ degrees of freedom.*

Proof. Consider first the modified forms of (4.12) and (4.13) obtained by replacing $I^{(n)}(\tilde{\beta}_n)$ by $I^{(n)}(\beta_0)$. These two tests will be referred to as T_1 and T_2 in the sequel. The asymptotic equivalence of T_1 and T_2 follows immediately from (4.11) in Lemma 4.1.

Define the linear mapping $P^{(n)} : V^* \to V^*$ by

$$P^{(n)} = I^{(n)}(\beta_0) \circ (D\beta) \circ \tilde{I}^{(n)}(\alpha_0)^{-1} \circ (D\beta)^T,$$

which is the orthogonal projection with respect to the inner product $I^{(n)}(\beta_0)^{-1}$ onto the subspace spanned by $(D\beta)$, and notice that

$$\begin{aligned} P^{(n)}\left(D_1^{(n)}\right) &= I^{(n)}(\beta_0) \circ (D\beta) \circ \tilde{I}^{(n)}(\alpha_0)^{-1}\left(\tilde{D}_1^{(n)}\right) \\ &= I^{(n)}(\beta_0)(\tilde{\beta}_n - \beta_0) + o_p(1), \end{aligned}$$

where $o_p(1)$ here, as in any vector equation in the sequel, denotes a term that tends in probability to zero in the norm $\|\cdot\|_n$. On combination with (3.5) we see that

$$D_1^{(n)} - P^{(n)}\left(D_1^{(n)}\right) = I^{(n)}(\beta_0)(\hat{\beta}_n - \tilde{\beta}_n) + o_p(1), \tag{4.15}$$

and hence that

$$\begin{aligned} &2\left\{\log f^{(n)}(Y^{(n)};\hat{\beta}_n) - \log f^{(n)}(Y^{(n)};\tilde{\beta}_n)\right\} \\ &\quad = 2D_1^{(n)}(\hat{\beta}_n - \tilde{\beta}_n) - I^{(n)}(\beta_0)(\hat{\beta}_n - \beta_0)^2 + I^{(n)}(\beta_0)(\tilde{\beta}_n - \beta_0)^2 + o_p(1) \end{aligned}$$

$$= 2I^{(n)}(\beta_0)^{-1}\left\{D_1^{(n)}, D_1^{(n)} - P^{(n)}\left(D_1^{(n)}\right)\right\} - I^{(n)}(\beta_0)^{-1}\left\{D_1^{(n)}\right\}^2$$

$$+ I^{(n)}(\beta_0)^{-1}\left\{P^{(n)}\left(D_1^{(n)}\right)\right\}^2 + o_p(1)$$

$$= I^{(n)}(\beta_0)^{-1}\left\{D_1^{(n)} - P^{(n)}\left(D_1^{(n)}\right)\right\}^2 + o_p(1) \tag{4.16}$$

because $P^{(n)}$ is a projection with respect to $I^{(n)}(\beta_0)^{-1}$. Since $D_1^{(n)}$ is asymptotically normal with asymptotic expectation 0 and asymptotic variance $I^{(n)}(\beta_0)$ it follows that the last expression in (4.16), except for the error term, as the squared length of the orthogonal projection of $D_1^{(n)}$ onto a $p-q$ dimensional subspace, is asymptotically chi-squared distributed with $p-q$ degrees of freedom. Thus, the limiting distribution of the log likelihood ratio test (4.14) is established. Its asymptotic equivalence with T_2 follows directly from (4.15) together with (4.16). Finally, the fact that

$$\lambda^{(n)}\,||\tilde{\beta}_n - \beta_0||_n \xrightarrow{\mathrm{P}} 0 \quad \text{as} \quad n \to 0$$

together with Lemma 2.5.6 shows that T_1 and T_2 are asymptotically equivalent to the test statistics in (4.12) and (4.13). ∎

In the test statistics (4.12) and (4.13) we may replace $I^{(n)}(\tilde{\beta}_n)$ by either of the quantities $I^{(n)}(\beta_0)$, $I^{(n)}(\hat{\beta}_n)$, $-D_2^{(n)}(\beta_0)$, $-D_2^{(n)}(\tilde{\beta}_n)$, or $-D_2^{(n)}(\hat{\beta}_n)$ without affecting the asymptotic properties. The first of these equivalences has already been used in the proof and the others are easily verified.

For the more general problem of testing a non-linear hypothesis, or more precisely a hypothesis given by a submodel as (4.6), generated from an analytic mapping of the form (4.5), we need stronger conditions to prove the results in Lemma 4.1 and Theorem 4.2. Obviously, Conditions 2.1 (B), (C) and (D) are not sufficient because these are satisfied by a sequence of linear normal models with bounded information. We might, for example, consider a constant sequence consisting of a fixed linear normal model, the same for each n. A non-linear hypothesis in this model would not, in general, lead to the results of the lemma and the theorem. To obtain these results, we only need the extra condition that the index $\lambda^{(n)}(\alpha_0)$ for the submodel tends to zero as n tends to infinity. Since this is assumed to be so for the full model, a comparison with the inequality (2.6.9) from Theorem 2.6.1 shows that it will be sufficient to demonstrate the existence of a sequence of constants $R_n \geq 0$ tending to zero, such that

$$||D^k\beta(\alpha_0)(w^k)||_n \leq k!\,R_n^{k-1}||D\beta(\alpha_0)(w)||_n^k \tag{4.17}$$

for all $k \geq 2$. The natural supplementary condition, namely that all eigenvalues of the Fisher information, relative to a fixed inner product, tend to infinity, is not sufficient to guarantee the existence of such a sequence. The problem is that the eigenvalues may tend to infinity at different rates, e.g., such that for some fixed $w \in W$ the ratio

$$||D^2\beta(\alpha_0)(w^2)||_n/||D\beta(\alpha_0)(w)||_n^2$$

is unbounded. A sufficient, but far from necessary, condition that ensures R_n, and hence $\lambda^{(n)}(\alpha_0)$, to tend to zero, is that all eigenvalues of the Fisher information tend to infinity at the same rate, i.e., in such a way that their ratios stay bounded. With that extra condition it is easy to modify the proofs above and deduce that the results in Lemma 4.1 and Theorem 4.2 hold for the non-linear hypothesis. Notice that this extra condition is not related to which particular non-linear hypothesis is being tested. While the condition of the same rate of increase of the eigenvalues may seem quite strong in the present framework, it is certainly weak enough to cover the case of independent replications, in which case all the eigenvalues are proportional to n. We shall not go through the proof of this modified theorem, the basis of which is established in Section 3.

The results in this and the previous section are closely related to the fact that if the index of a sequence of analytic models tends to zero, then the sequence of experiments converge, in a neighbourhood of that point, towards a Gaussian shift experiment, in the sense of LeCam (1986). More precisely, it is not hard to verify that the LAN condition in Definition 1 from LeCam (1986, Chapter 11, Section 7) holds at the point considered. Our conditions, and the type of convergence obtained here, is, however, stronger than those considered by LeCam. Thus, the convergence of the index to zero allows an expansion of not only the log-likelihood, but also of its derivatives. While certain sequences of experiments are therefore excluded from the theory given here, the conditions permit the derivation of stronger results, in particular in the sense of higher-order expansions as we shall see in the next chapter. The fact that the LAN condition from LeCam (1986) holds is, in turn, partly related to the fact that the two sequences of probability measures with densities

$$f^{(n)}(y^{(n)};\beta_0), \qquad f^{(n)}(y^{(n)};\beta_0+v_n),$$

respectively, are mutually contiguous (LeCam, 1986, Chapter 6, Section 3, Definition 4) whenever $I^{(n)}(v_n^2)$ is bounded, still provided that the index at β_0 tends to zero. This may easily be seen from the bounds in (2.12) and (2.13) together with the tightness of the sequence $H_n(Y^{(n)})$ proved in Lemma 2.4, and by use of Proposition 4 in LeCam (1986, Chapter 6, Section 3) which states the mutual contiguity of any two sequences for which any convergence to zero in probability of a sequence of random variables with respect to one of the sequences of probability measures implies the same convergence to hold with respect to the other sequence of probability measures.

5 Generalized linear models

This section contains the basic notations and results for the class of generalized linear models based on an analytic model. The notation introduced in this section will be used in the following sections which contain examples of asymptotic results for such models. The class is somewhat bigger than is usually implied by the notion of generalized linear models, cf. McCullagh and Nelder (1983).

Let $g(y;\psi,\phi)$ be a family of densities parametrized by two (vector) parameters $\psi \in \Psi \subseteq W$ and $\phi \in \Phi \subseteq V_2$. Assume that $Y^{(n)} = (Y_1,\ldots,Y_n)$ where $Y_1,\ldots,Y_n$ are mutually independent and the density of Y_i with respect to some underlying measure on the sample space is

$$f(y_i;\theta,\phi) = g(y_i;\psi_i,\phi) \tag{5.1}$$

where

$$\psi_i = a_i(\theta) \tag{5.2}$$

is a known linear function of the vector parameter θ. Typically the linear mapping a_i is given in terms of a matrix $[a_i]_{jk}$ of covariates for the ith observation Y_i. In that case the coordinate version of (5.2) might be written

$$[a_i(\theta)]_j = [a_i]_{jk}[\theta]_k \tag{5.3}$$

where sub- and superscripts outside square brackets refer to coordinates, and summation over k on the right side is implied by the summation convention.

Often, and in all our examples, the parameter ψ is one-dimensional and the index j then disappears from the expression above. We may then write

$$a_i(\theta) = [a_i]_1[\theta]_1 + \cdots + [a_i]_r[\theta]_r \in \mathbf{R} = W, \tag{5.4}$$

where $\theta = ([\theta]_1,\ldots,[\theta]_r)$ is a coordinate representation of θ and $\dim V_1 = r$. Typically the first covariate $[a_i]_1$ will be 1 such that $[\theta]_1$ is the intercept in the model. For this case of one-dimensional ψ-parameter we let A_n denote the design matrix given by

$$A_n = \begin{pmatrix} [a_1]_1 & \cdots & [a_1]_r \\ \vdots & \vdots & \vdots \\ [a_n]_1 & \cdots & [a_n]_r \end{pmatrix} \tag{5.5}$$

and $[A]_i$ the row vector equal to the ith row of A_n.

The parameter ϕ is the part of the parameter which is common to all of the observations. Typically this is a one-dimensional parameter such as the variance for models based on the normal distribution. In some cases there is no ϕ-parameter, e.g., for log-linear models based on the Poisson distribution. Although such models might be incorporated into the framework of (5.1) by allowing vector spaces of dimension zero we prefer the more natural approach of simply omitting the ϕ-parameter. Some obvious modifications of notations and results in the sequel are required for this case.

As described here any parametric model may be formulated as a generalized linear model in a trivial sense with $n = 1$, but the concept is, of course, useful only when it leads to a simplification of the problem at hand. For an infinite sequence of independent random variables $Y_1, Y_2, \ldots$ the class of generalized linear models is a genuine restriction of the class of parametric models that permits the derivation of asymptotic results.

For our purpose we restrict the class further by assuming that the model

$$\{g(y;\psi,\phi);(\psi,\phi) \in \Psi \times \Phi \subseteq W \times V_2\}, \tag{5.6}$$

which we shall denote the *reference model*, is analytic on $\Psi \times \Phi$, and that

$$a_i : \Theta \to \Psi, \tag{5.7}$$

where $\theta \in \Theta \subseteq V_1$, and V_1, V_2 and W are finite-dimensional vector spaces. It then follows immediately from Theorem 2.6.1 and Theorem 2.5.3 that the model

$$\{f^{(n)}(y^{(n)};\theta,\phi);(\theta,\phi) \in \Theta \times \Phi \subseteq V_1 \times V_2\} \tag{5.8}$$

is analytic at any point of its domain, where

$$f^{(n)}(y^{(n)};\theta,\phi) = \prod_{i=1}^{n} f(y_i;\theta,\phi) = \prod_{i=1}^{n} g(y_i;a_i(\theta),\phi). \tag{5.9}$$

If it is also assumed that the index of the reference model in (5.6) is finite throughout its domain then the same two theorems imply that the index of the model (5.8) is finite everywhere.

A somewhat more important problem is whether the indices from a sequence of models of the form (5.8) tend to zero as n tends to infinity. This problem is explored in the following theorem.

We denote the index of the reference model by $\lambda(\psi,\phi)$ while $\lambda^{(n)}(\theta,\phi)$ denotes the index of the model for $Y^{(n)}$. Similarly, let $I(\psi,\phi)$ denote the Fisher information in the reference model, $I_i(\theta,\phi)$ the Fisher information for the model for the single observation Y_i, and for $(v_1,v_2) \in V_1 \times V_2$ we then have

$$\begin{aligned} I^{(n)}(\theta,\phi)(v_1,v_2)^2 &= \sum_{i=1}^{n} I_i(\theta,\phi)(v_1,v_2)^2 \\ &= \sum_{i=1}^{n} I(\psi_i,\phi)\{a_i(v_1),v_2\}^2, \end{aligned} \tag{5.10}$$

which is the Fisher information in the model for $Y^{(n)}$.

Theorem 5.1. *Assume that the reference model in (5.6) is analytic with a finite index $\lambda(\psi,\phi)$ on $\Psi \times \Phi$. Then the index $\lambda^{(n)}(\theta,\phi)$ of the model in (5.8) for*

$Y_1, \ldots, Y_n$ tends to zero as n tends to infinity if both of the following conditions hold:

(1) The smallest eigenvalue of the Fisher information $I^{(n)}(\theta, \phi)$, relative to a fixed inner product on $V_1 \times V_2$, tends to infinity as n tends to infinity.

(2) Uniformly in $(v_1, v_2) \in V_1 \times V_2$, $(v_1, v_2) \neq (0,0)$, we have

$$\lambda^2(\psi_n, \phi)\{I_n(\theta, \phi)(v_1, v_2)^2\}/I^{(n)}(\theta, \phi)(v_1, v_2)^2 \to 0 \qquad (5.11)$$

as $n \to \infty$, where $\psi_n = a_n(\theta)$, cf. (5.2).

Provided that (1) holds, then the condition (2) holds if the sequence $\lambda(\psi_i, \phi)$, $i = 1, 2, \ldots$, is bounded and the following condition holds:

(3) Uniformly in $(v_1, v_2) \in V_1 \times V_2$, $(v_1, v_2) \neq (0,0)$, we have

$$I_n(\theta, \phi)(v_1, v_2)^2/I^{(n)}(\theta, \phi)(v_1, v_2)^2 \to 0 \qquad (5.12)$$

as $n \to \infty$.

Proof. From Theorem 2.6.1 it is known that the index of the model for a single observation Y_i, say, is bounded by $\lambda(\psi_i, \phi)$. Let $v = (v_1, v_2) \neq (0,0)$ and define

$$x_i(v) = \lambda^2(\psi_i, \phi)I_i(\theta, \phi)(v_1, v_2)^2$$

and

$$S_n(v) = I^{(n)}(\theta, \phi)(v^2).$$

Condition (2) states that there is a sequence (ϵ_n), say, such that for all v we have

$$x_n(v)/S_n(v) \leq \epsilon_n \to 0$$

as $n \to \infty$. We want to show that

$$a_n^2 = \sup\{\, x_i(v)/S_n(v) : v \in V, 1 \leq i \leq n \,\}$$

tends to zero as n tends to infinity, in which case the result that the index tends to zero will follow from Theorem 2.5.3. Note that a_n here as throughout the proof is the quantity from (2.5.14), not related to the sequence of linear mappings from (5.2). Let $\delta > 0$ be given and choose m such that $\epsilon_j < \delta$ for all $j \geq m$. Next, choose n such that

$$S_m(v)/S_n(v) < \delta$$

for all $v \in V$. It follows from condition (1) that this is possible. Then,

$$\begin{aligned} x_i(v)/S_n(v) &\leq \epsilon_i S_i(v)/S_n(v) \\ &\leq \sup\{\, \epsilon_j : j \in \mathbb{N}\,\}S_m(v)/S_n(v) + \delta S_i(v)/S_n(v) \\ &\leq \sup\{\, \epsilon_j : j \in \mathbb{N}\,\}\delta + \delta \end{aligned}$$

from which it is seen that $a_n \to 0$ as $n \to \infty$.

The statement that condition (2) follows from the conditions (1) and (3) when $\lambda(\psi_i, \phi)$ is bounded is trivially verified. ∎

Although the last version of the theorem with $\lambda(\psi_i, \phi)$ bounded and conditions (1) and (3) fulfilled is a trivial consequence of the first version, it is useful in some situations because it reduces the problem to an investigation of the Fisher information only. For this all eigenvalues must tend to infinity, which is definitely a weak requirement if asymptotic results are to be proved, and the condition (3) bears some similarity to the Lindeberg condition in the sense that it states that no single observation may contribute an asymptotically non-negligible amount to the variance of the score statistic. The condition that the indices $\lambda(\psi_i, \phi)$ are bounded is, for example, fulfilled either if the covariates take values in a compact subset of the parameter space, such that the sequence ψ_i also stays within a compact set, or if the model is a location model, with or without an unknown scale parameter, and ψ is the location parameter, as we saw in Sections 3.4 and 3.6. The condition may also be satisfied for more general transformation models but we shall not explore such possibilities here.

The theorem provides conditions for the index of the model to tend to zero as n tends to infinity. If all eigenvalues of the Fisher information tend to infinity, i.e., if the condition (1) of the theorem, or equivalently Condition 2.1 (E), holds, then the Conditions 2.1 (A)–(E) all hold and all of the asymptotic results in Sections 3 and 4 apply. Thus, the main difficulty in the application of these results is to derive conditions in terms of the covariates, implying that the conditions (1) and (2) of the theorem hold.

Notice that if the Fisher information $\{I^{(n)}(\theta, \phi)\}(v_1, v_2)^2$ tends to infinity for any fixed non-zero vector (v_1, v_2) in $V_1 \times V_2$ then it follows that all eigenvalues of this Fisher information tend to infinity. This follows from Dini's theorem which, briefly speaking, states that monotone convergence of continuous functions on a compact space, here the unit sphere, is uniform.

6 Linear normal models with known variance

In the setup of the generalized linear models from Section 5, consider the case where the reference distribution is normal, i.e.,

$$g(y;\psi) = (2\pi\sigma^2)^{-1/2}\exp\{-\frac{1}{2}[(y-\psi)/\sigma]^2\}, \qquad y \in \mathbf{R}, \tag{6.1}$$

where $\sigma > 0$ is known, $\psi \in \mathbf{R}$, and there is no ϕ-parameter. This is, of course, a usual linear normal model with known variance, for which we know that the distributional results for estimators and test statistics considered in Sections 3 and 4 are exact. This type of model is included here for comparison with other models.

The Fisher information in the reference model is $I(\psi) = \sigma^{-2}$ and for the model for $Y^{(n)}$ it is

$$I^{(n)}(\theta)(v^2) = \sigma^{-2}\sum_{i=1}^{n} a_i(v)^2, \tag{6.2}$$

where $v \in V_1$. In terms of the design matrix, the matrix representation of $I^{(n)}$ is

$$\sigma^{-2}A_n^T A_n,$$

where A_n^T denotes the transpose of A_n. The log-likelihood function is concave and the maximum likelihood estimator $\hat{\theta}$ is the unique solution to the likelihood equation, provided that the Fisher information is non-singular. The distribution of $\hat{\theta}$ is then exactly normally distributed with mean θ_0 and variance $I^{(n)}(\theta_0)^{-1}$. This well-known result may be deduced from the fact that the index of the model is exactly zero. More precisely, $\lambda(\psi) = 0$ for all ψ, and since the mapping from θ to ψ_i is linear it follows from Theorem 2.6.1 and Theorem 2.5.3 that also $\lambda^{(n)}(\theta) = 0$.

For the estimator to be consistent we need the requirement that all eigenvalues of the Fisher information, or equivalently of the matrix $A_n^T A_n$, tend to infinity as n tends to infinity.

For other generalized linear models this condition will be a minimal requirement for the derivation of any asymptotic result, as may be expected from Theorem 5.1, and we therefore state it as a condition below.

Condition 6.1. *The sequence of linear mappings $a_i : V_1 \to \mathbf{R}$ satisfies*

$$\inf\left\{\sum_{i=1}^{n}(a_i(v))^2 : v \in V_1, \|v\| = 1\right\} \to \infty$$

as $n \to \infty$. In terms of the design matrix this means that all eigenvalues of $A_n^T A_n$ tend to infinity as n tends to infinity.

7 Linear normal models with unknown variance

Consider again the model (6.1) but now with unknown variance $\phi = \sigma^2 > 0$. Observe that since the model is still a linear exponential family, when suitably reparametrized, the maximum likelihood estimator is again the unique solution to the likelihood equation and agrees therefore with the local maximum likelihood estimator from Section 3. For the reference model we have the Fisher information

$$I(\psi, \phi)(w, v_2)^2 = w^2/\phi + v_2^2/(2\phi^2), \tag{7.1}$$

where $(w, v_2) \in W \times V_2 = \mathbf{R} \times \mathbf{R}$. In matrix notation this information takes the more familiar form

$$\begin{pmatrix} \phi^{-1} & 0 \\ 0 & \frac{1}{2}\phi^{-2} \end{pmatrix}. \tag{7.2}$$

The main feature of this information, viewed in the general context, is that it does not depend on the parameter ψ, while the orthogonality of the ψ and ϕ parameter makes things a little easier. Direct computations along the lines shown in Section 9 for the Gamma distribution show that the index $\lambda(\psi, \phi)$ is constant as a function of both parameters, and that (less important)

$$2\sqrt{2} \leq \lambda(\psi, \phi) \leq 3\sqrt{2}. \tag{7.3}$$

It is conjectured that, in fact, the lower bound applies, but that has not been proved.

For the generalized linear model based on this reference model the Fisher information is given by

$$\left\{I^{(n)}(\theta, \phi)\right\}(v_1, v_2)^2 = \left\{\sum_{i=1}^{n} a_i(v_1)^2/\phi\right\} + \frac{1}{2}nv_2^2/\phi^2, \tag{7.4}$$

where $v_1 \in V_1$ and $v_2 \in V_2 = \mathbf{R}$. In terms of block matrices this is written

$$\begin{pmatrix} \frac{1}{\phi}A_n^T A_n & 0 \\ 0 & \frac{n}{2\phi^2} \end{pmatrix}. \tag{7.5}$$

Thus, all eigenvalues of the information tend to infinity if and only if Condition 6.1 is satisfied. To apply Theorem 5.1 we furthermore need the following condition. Recall the notation $[A]_i$ for the row vector of covariates for the ith observation.

Condition 7.1. *The linear mappings in (5.4) satisfy*

$$[A]_n(A_n^T A_n)^{-1}[A]_n^T \to 0 \tag{7.6}$$

as $n \to \infty$.

It is easily verified that this condition implies the condition (3) in Theorem 5.1 which then can be applied to verify that all the asymptotic results from Sections 3 and 4 hold. To summarize, this conclusion holds under the assumption that Conditions 6.1 and 7.1 hold for the covariates.

The condition 7.1, or equivalently the condition (3) in Theorem 5.1, states that in the limit no non-negligible contribution to the Fisher information stems from a single observation. From the well-established theory for the normal models we know that this condition is, in fact, not necessary here; the asymptotic normality of the maximum likelihood estimator is implied by the fact that the number of observations tends to infinity, even when the variance is unknown. For the consistency of the estimator we need Condition 6.1. Thus, the asymptotic results are exactly the same as for the case of known variance, although the distribution of the estimator is no longer exactly normal. That Condition 7.1 is not necessary here, although it is needed to apply the general results, is related to the special structure of this model in terms of the 'directional index' as defined in Section 2.5, cf. the comments following Corollary 2.5.4. This directional index is zero in any θ-direction, i.e., in the direction of any vector with ϕ-coordinate equal to zero, due to the fact that the linear normal model with known variance has index zero. That the index is not zero in the ϕ-direction does not matter because the Fisher information for the ϕ-parameter tends to infinity independently of the covariates. A precise version of this argument, which requires more careful considerations regarding mixed cumulants for the two kinds of 'directions', would require a much more elaborate development of the theory involving directional indices. This development has been avoided here because it does not seem to be of much use in 'realistic' examples, i.e., in examples where the answer is not known beforehand.

8 Location models with fixed unknown scale

In this section we consider generalized linear models for which the variable parameter ψ is a location parameter and the fixed parameter ϕ is a scale parameter. Thus, the reference model (5.6) has densities

$$g(y;\psi,\phi) = \frac{1}{\phi}h\left(\frac{y-\psi}{\phi}\right), \qquad \psi \in \mathbf{R}, \phi > 0, \tag{8.1}$$

on $\mathbf{R}$, where h is assumed to be a strictly positive density function, cf. (3.6.11). Conditions were given in Section 3.6 for such a location and scale model to be analytic. In the present section we work from the assumption that it is analytic; otherwise no restrictions will be imposed on the density function h.

From Lemma 3.6.4 we know that the index $\lambda(\psi,\phi)$ of the reference model is constant. From Theorem 2.6.1 it then follows that the same is true for the index, $\lambda_i(\theta,\phi)$ say, of the model for the single observation Y_i, since this is a linearly reparametrized version of the reference model.

As in Section 3.6 we introduce the standardized variable

$$U = \frac{Y-\psi}{\phi} \tag{8.2}$$

which has a distribution that is independent of the parameters when the distribution of Y is the reference distribution based on (ψ,ϕ). In terms of this random variable we may write the Fisher information matrix for the reference model as

$$I(\psi,\phi) = \frac{1}{\phi^2}\begin{pmatrix} -\mathrm{E}\{D^2\log h(U)\} & -\mathrm{E}\{UD^2\log h(U)\} \\ -\mathrm{E}\{UD^2\log h(U)\} & -\mathrm{E}\{U^2D^2\log h(U)\} - 1 \end{pmatrix}, \tag{8.3}$$

as is easily seen by differentiation of the log-density. We know from Lemma 3.6.3 that this Fisher information is positive definite and hence defines a norm on $\mathbf{R}^2$. Since any two norms on a finite-dimensional vector space are equivalent there exist two constants, c and C say, such that for any vector $(s,t) \in \mathbf{R}^2$ we have

$$c\,||(s,t)||^2 \leq I(\psi,\phi)(s,t)^2 \leq C\,||(s,t)||^2 \tag{8.4}$$

where $\|\cdot\|$ here denotes the usual Euclidean norm on $\mathbf{R}^2$. From (5.10) we then see that for any $v_1 \in V_1$ and $v_2 \in V_2 = \mathbf{R}$ we have

$$\frac{I_n(\theta,\phi)(v_1,v_2)^2}{I^{(n)}(\theta,\phi)(v_1,v_2)^2} \leq \left(\frac{C}{c}\right) \frac{\{a_n(v_1)\}^2 + v_2^2}{\sum\{a_i(v_1)\}^2 + nv_2^2} \tag{8.5}$$

where the notation I_n is recalled to denote the information for the nth observation alone.

It is now immediate to apply Theorem 5.1, the version based on the conditions (1) and (3) to conclude that the asymptotic results of Sections 3 and 4 hold for this type of model, if Condition 6.1 and 7.1 both hold. Thus, for example, the asymptotic normality of the LMLE for any of the models considered here follows under these two conditions on the covariates. It does not follow for this type of models that the LMLE agrees with the global maximum likelihood estimator, and it is indeed not true in general as is known from the example based on the Cauchy distribution.

9 The Gamma distribution with fixed unknown shape

Consider again a generalized linear model as described in Section 5, this time with the reference distribution being the Gamma distribution with densities

$$g(y;\psi,\phi) = \Gamma(\phi)^{-1}\psi^{-\phi}y^{\phi-1}\exp\{-y/\psi\}, \qquad y > 0, \tag{9.1}$$

where the parameters ψ and ϕ are both positive. Since the mean of this distribution is $\phi\psi$ we are considering models for which the means are linear functions of unknown parameters, while the shape parameter is fixed and unknown. This type of model is described in McCullagh and Nelder (1983), Section 7.3. Apart from being of interest in themselves these models occur, i.a., in connection with variance components models.

For once, in this example, we go fairly thoroughly through the calculations leading to conclusions concerning the index of the reference model. Once this is done, the step to the asymptotic results in the generalized linear models based on this distribution is quite small.

From (9.1) we have

$$\log g(y;\psi,\phi) = -\log\Gamma(\phi) - \phi\log\psi + (\phi-1)\log y - y/\psi, \tag{9.2}$$

the differentials of which, with respect to the two parameters, are

$$D_\psi \log g(y;\psi,\phi) = -\frac{\phi}{\psi} + \frac{y}{\psi^2} = \frac{1}{\psi}(-\phi + u) \tag{9.3}$$

and

$$D_\phi \log g(y;\psi,\phi) = -D\log\Gamma(\phi) + \log(y/\psi) = D\log\Gamma(\phi) + \log u, \tag{9.4}$$

where $u = y/\psi$. The higher order differentials, with $k \geq 2$, are given by

$$\begin{aligned} D_\psi^k \log g(y;\psi,\phi) &= (-1)^k (k-1)!\,\psi^{-k}\phi + (-1)^{k-1}k!\,\psi^{-(k+1)}y \\ &= -(-\psi)^{-k}\{k!\,u - (k-1)!\,\phi\} \end{aligned} \tag{9.5}$$

and

$$D_\phi^k \log g(y;\psi,\phi) = -D^k \log\Gamma(\phi), \tag{9.6}$$

while all the mixed differentials are noticed to be non-random. Since all these differentials are affine functions of the statistic $(U, \log U)$, where $U = Y/\psi$ follows a Γ-distribution with scale parameter $\psi = 1$, we need the joint cumulant generating function for this two-dimensional statistic.

Lemma 9.1. *The cumulant generating function of* $(U, \log U)$ *is*

$$\log \mathrm{E} \exp\{sU + t\log U\} = \log\Gamma(\phi + t) - \log\Gamma(\phi) - (\phi + t)\log(1-s), \tag{9.7}$$

for $s < 1$ *and* $t \in \mathbf{R}$, *where* U *follows a* Γ*-distribution with scale parameter* $\psi = 1$ *and shape parameter* ϕ, *cf. (9.1).*

Proof. Trivial.

By differentiation of the cumulant generating function we obtain the following expressions for the cumulants of $(U, \log U)$:

$$\mathrm{cum}_m\{U\} = (m-1)!\,\phi, \tag{9.8}$$

$$\mathrm{cum}_m\{\log U\} = D^m \log\Gamma(\phi), \tag{9.9}$$

$$\mathrm{cum}\{U,\ldots,U,\log U\} = (m-2)!\,, \tag{9.10}$$

for $m \geq 2$, where the number of U's that appear in the last cumulant expression is $m-1$, and all mixed cumulants with two or more $\log U$'s are zero. On combination with (9.3) and (9.4) we see that the Fisher information is given by

$$I(\psi,\phi)(a,b)^2 = a^2\phi/\psi^2 + 2ab/\psi + b^2 D^2 \log\Gamma(\phi), \tag{9.11}$$

for $(a,b) \in \mathbf{R}^2$. In matrix notation this Fisher information becomes

$$\begin{pmatrix} \phi/\psi^2 & 1/\psi \\ 1/\psi & D^2\log\Gamma(\phi) \end{pmatrix}. \tag{9.12}$$

In the computation of cumulants of higher order, non-random terms may be disregarded. Thus, we consider terms of the form

$$\begin{aligned} &\left\{D_{(\psi,\phi)}^k \log g(Y;\psi,\phi)\right\}(a,b)^k \\ &\quad = a^k D_\psi^k \log g(Y;\psi,\phi) + b^k D_\phi^k \log g(Y;\psi,\phi) \quad + \quad \text{non-random terms} \\ &\quad = -k!\,(-a/\psi)^k U + b^k (\log U) I_{\{k=1\}} \quad + \quad \text{non-random terms.} \end{aligned} \tag{9.13}$$

Now consider a fixed sequence $(k_1, \dots, k_m) \in \mathbf{N}^m$ with $m \geq 2$ and $\sum k_j \geq 3$ and consider the joint cumulant from (2.3.15) of the log-likelihood differentials, still working in the reference model. From (9.13) and (9.8)–(9.10) we see that if we let $J = \{\, j : k_j = 1 \,\}$, then

$$\begin{aligned}
&\{\chi_{k_1 \cdots k_m}(\psi, \phi)\} \left\{(a_1, b_1)^{k_1}, \dots, (a_m, b_m)^{k_m}\right\} \\
&\quad = \operatorname{cum}\{-k_1!\,(-a_1/\psi)^{k_1} U, \dots, -k_m!\,(-a_m/\psi)^{k_m} U\} \\
&\qquad + I_{\{\Sigma k_j = m\}} \operatorname{cum}\{b_1 \log U, \dots, b_m \log U\} \\
&\qquad + \sum_{j \in J} \left[\prod_{i \neq j} \{-k_i!\,(-a_i/\psi)^{k_i}\}\right] \operatorname{cum}\{b_j \log U, U, \dots, U\} \\
&\quad = (-1)^{m + \Sigma k_j}(m-1)!\, \left\{\prod \left(k_j!\,(a_j/\psi)^{k_j}\right)\right\} \phi \\
&\qquad + I_{\{\Sigma k_j = m\}} \left(\prod b_j\right) D^m \log \Gamma(\phi) \\
&\qquad + (-1)^{m + \Sigma k_j - 1}(m-2)! \sum_{j \in J} b_j \prod_{i \neq j} \{k_i!\,(a_i/\psi)^{k_i}\}\,. \qquad (9.14)
\end{aligned}$$

Notice that this expression depends on ψ and on the vectors (a_j, b_j) only through the vectors $(a_j/\psi, b_j)$, and that the same is true for the Fisher information in (9.11). Hence, it follows that the index $\lambda(\psi, \phi)$ is independent of ψ, because each of the expressions of the two sides of the inequality in (2.5.3), for a particular ψ and particular vectors (a_j, b_j), is matched by the corresponding expressions with $\psi = 1$ and each a_j replaced by a_j/ψ. That the index is finite follows from the fact that no linear combination of U and $\log U$ has a one-point distribution, implying that the Fisher information in (9.11) and (9.12) is positive definite.

Now that we know that the index $\lambda(\psi, \phi)$ is constant, and hence bounded, as a function of ψ, the verification of the conditions (1) and (3) in Theorem 5.1 will suffice to prove that the index $\lambda^{(n)}$ for the generalized linear model tends to zero, and hence that the asymptotic results from Sections 3 and 4, relating to the local maximum likelihood estimator, hold. These two conditions can be simplified somewhat more. In fact, we show below that they are equivalent to the same two conditions for the model with known shape parameter ϕ. Thus, the asymptotic results for the generalized linear models with unknown shape parameter are derived under exactly the same conditions as for the corresponding model with this parameter known.

To see this, notice first that we may rewrite the Fisher information (9.11) as

$$I(\psi, \phi)(a, b)^2 = A_\phi \left(\frac{a}{\psi}, b\right)^2, \qquad (9.15)$$

where $(a, b) \in \mathbf{R}^2$ and $A_\phi = I(1, \phi)$ is the inner product on $\mathbf{R}^2$ given by

$$A_\phi(s, t)^2 = s^2 \phi + 2st + t^2 D^2 \log \Gamma(\phi), \qquad (s, t) \in \mathbf{R}^2, \qquad (9.16)$$

with the matrix representation

$$\begin{pmatrix} \phi & 1 \\ 1 & D^2 \log \Gamma(\phi) \end{pmatrix} . \tag{9.17}$$

From the equivalence of any two norms on $\mathbf{R}^2$ we know that there exist two constants $c > 0$ and $C > 0$ such that

$$c \, \|(s,t)\|^2 \leq A_\phi(s,t)^2 \leq C \, \|(s,t)\|^2, \tag{9.18}$$

where $\|(s,t)\|^2 = s^2 + t^2$ is the square of the usual Euclidean norm on $\mathbf{R}^2$. For the generalized linear model we see from (5.10) that for any $v_1 \in V_1$ and $v_2 \in V_2 = \mathbf{R}$ we have

$$\begin{aligned} I^{(n)}(\theta,\phi)(v_1,v_2)^2 &= \sum_{i=1}^{n} I(\psi_i,\phi)(a_i(v_1),v_2)^2 \\ &= \sum_{i=1}^{n} A_\phi(a_i(v_1)/a_i(\theta),v_2)^2 \end{aligned} \tag{9.19}$$

and hence

$$\begin{aligned} c\left\{\left[\sum_{i=1}^{n}\left(\frac{a_i(v_1)}{a_i(\theta)}\right)^2\right] + nv_2^2\right\} &\leq I^{(n)}(\theta,\phi)(v_1,v_2)^2 \\ &\leq C\left\{\left[\sum_{i=1}^{n}\left(\frac{a_i(v_1)}{a_i(\theta)}\right)^2\right] + nv_2^2\right\} . \end{aligned}$$

Thus, $I^{(n)}(\theta,\phi)(v_1,v_2)^2$ obviously tends to infinity as n tends to infinity, if $v_2 \neq 0$. Therefore condition (1) in Theorem 5.1 reduces to the condition that the information on θ tends to infinity, ϕ being regarded as fixed. This condition may be expressed

$$\sum_{i=1}^{n}\left\{\frac{[a_i]_1[v]_1 + \ldots + [a_i]_r[v]_r}{[a_i]_1[\theta]_1 + \ldots + [a_i]_r[\theta]_r}\right\}^2 \to \infty \quad \text{as} \quad n \to \infty \tag{9.20}$$

for all $v = ([v]_1, \ldots, [v]_r) \in V$ with $v \neq 0$, cf. (5.4).

For the special case of a simple linear regression, i.e., with $r = 2$, $[a_i]_1 = 1$, and one covariate $[a_i]_2 = x_i$, say, this reduces to the condition

$$\sum_{i=1}^{n}\left\{\frac{[v]_1 + x_i[v]_2}{[\theta]_1 + x_i[\theta]_2}\right\}^2 \to \infty \quad \text{as} \quad n \to \infty \tag{9.21}$$

for all $([v]_1, [v]_2) \neq (0,0)$. Notice that the denominators in (9.20) and (9.21) equal ψ_i and must be positive by assumption.

Concerning condition (3) in Theorem 5.1 we see in a similar way that this holds if and only if the corresponding condition holds for the model with known ϕ, i.e., if

$$\left\{\frac{a_n(v)}{a_n(\theta)}\right\}^2 \Big/ \sum_{i=1}^{n}\left\{\frac{a_i(v)}{a_i(\theta)}\right\}^2 \to 0 \quad \text{as} \quad n \to \infty \tag{9.22}$$

uniformly in $v \in V$, $v \neq 0$. In the general case (9.20) may hold but (9.22) fail, heuristically speaking, e.g., if the length of the vector $([a_n]_1, \ldots, [a_n]_r)$ increases with n while the angle between this vector and the vector $([\theta]_1, \ldots, [\theta]_r)$ decreases in such a way that the inner product of these two vectors, i.e., the denominator in the nth term in (9.20), is constant. In such a case the numerator may "blow up" such that the nth term of the sum constitutes a non-decreasing fraction of the sum in (9.20).

However, for the simple linear regression, corresponding to (9.21), this cannot happen if we impose the natural conditions $[\theta]_1 > 0$, $[\theta]_2 > 0$, and $x_i > 0$ for all i. In that case we have

$$\left|\frac{[v]_1 + x_n[v]_2}{[\theta]_1 + x_n[\theta]_2}\right| \leq \frac{[v]_1}{[\theta]_1} + \frac{[v]_2}{[\theta]_2} \tag{9.23}$$

which is bounded in n. In conclusion, for the simple linear regression with positivity conditions on the θ-parameters and on the covariates, the condition (9.21) suffices to establish the asymptotic results for the local maximum estimator from Sections 3 and 4. For the general case the conditions (9.20) and (9.22) must both be required to hold for the same conclusions to be valid.

10 Log-linear Poisson models

Consider the Poisson distribution with mean $\mu > 0$, given by the point probabilities

$$g(y;\psi) = \mathrm{P}(Y = y) = \frac{\mu^y}{y!}e^{-\mu}, \qquad \mu = e^{\psi}, y \in \mathbf{N}, \tag{10.1}$$

where $\psi = \log\mu$ denotes the canonical parameter in this exponential family of distributions. The log-linear Poisson models are the generalized linear models as described in Section 5, based on this reference distribution with no ϕ-parameter. Thus, ψ_i is a known linear function of the unknown parameter $\theta \in \Theta \subseteq V_1$. Since this generalized linear model is a linear exponential family it follows that the results obtained for the local maximum likelihood estimator in Sections 3 and 4 apply to the (global) maximum likelihood estimator whenever the conditions for these results hold.

Direct computations give

$$D_\psi \log g(Y;\psi) = Y - e^{\psi} \tag{10.2}$$

while higher-order derivatives are non-random. It follows that all cumulants of log-likelihood differentials of the form (2.3.15) with $m \geq 2$ vanish if any differential of order k_j greater than 1 is involved, and consequently we need only consider the cumulants of the score function in (10.2) to compute the index $\lambda(\psi)$ in the reference model.

The cumulants of Y are all equal to $\mu = \exp(\psi)$ and therefore we have

$$\mathrm{cum}_m\{D_\psi \log g(Y;\psi)\} = e^\psi, \qquad m \geq 2. \tag{10.3}$$

In particular, the Fisher information is the variance

$$I(\psi) = e^\psi. \tag{10.4}$$

From Definition 2.5.1 we then see that the index of the reference model is

$$\lambda(\psi) = \exp\{-\psi/2\} \sup_{m\geq 3} \left\{(m-1)!^{\,-1/(m-2)}\right\} = \frac{1}{2}\exp\{-\psi/2\}. \tag{10.5}$$

Returning to the generalized linear model with $\psi_i = a_i(\theta)$ as in (5.2) we obtain the Fisher information

$$I^{(n)}(\theta)(v^2) = \sum_{i=1}^{n} a_i(v)^2 \exp\{a_i(\theta)\}, \tag{10.6}$$

where $v \in V = V_1$. Theorem 2.5.3 now gives the following bound for the index $\lambda^{(n)}(\theta)$:

$$\begin{aligned}\lambda^{(n)}(\theta)^2 &\leq \sup\Big\{\, \lambda(\psi_i)^2 \left[I_i(\theta)(v^2)\right] \,/\, I^{(n)}(\theta)(v^2) : 1 \leq i \leq n, v \in V, v \neq 0 \Big\}\\ &= \frac{1}{4}\sup\Big\{\, a_i(v)^2 \,/\, \sum_{j=1}^{n} a_j(v)^2 \exp\{a_j(\theta)\} : 1 \leq i \leq n, v \in V, v \neq 0 \Big\}.\end{aligned} \tag{10.7}$$

The validity of the asymptotic results for the maximum likelihood estimator now follows under the single condition that this quantity tends to zero as n tends to infinity. The supremum in (10.7) is exactly the quantity d_t^2 from Haberman (1977), Section 2, and his Condition 2 for this type of model states accordingly that d_t tends to zero, on the basis of which he proved similar results. The ease with which the condition was derived above cannot be compared directly with the proof in Haberman's paper since his results were adapted to models for which the dimension of the parameter space may change with the sample size.

Just as it was shown in the proof of Theorem 5.1 the supremum in (10.7) tends to zero as n tends to infinity if and only if

$$\frac{a_n(v)^2}{\sum_{i=1}^{n} a_i(v)^2 \exp\{a_i(\theta)\}} \to 0 \tag{10.8}$$

uniformly in $v \neq 0$ as $n \to \infty$. For this to be the case we must assume, as usual, that the denominator in (10.8), i.e., the Fisher information from (10.6), tends to infinity for any fixed $v \neq 0$. This assumption is, however, not sufficient to imply (10.8). For example, (10.8) may be violated if $a_i(\theta)$ stays bounded while $a_i(v)$ tends rapidly to infinity.

Instead of basing the results concerning the index on the inequality (10.7), derived by use of Theorem 2.5.3, we may in this case derive an expression for the index $\lambda^{(n)}$ directly from its definition. Proceeding from (10.3) and (10.4) we see that for the generalized linear model we have

$$\mathrm{cum}_m\{D_1^{(n)}(\psi)\} = \sum_{i=1}^{n} a_i(v)^m \exp\{\psi_i\}, \tag{10.9}$$

for any $m \geq 2$. Hence, the definition in (2.5.3) of the index of this model becomes

$$\lambda^{(n)}(\theta) = \sup\left\{ \Lambda_m(v;\theta)^{1/(m-2)} : m \geq 3, v \in V, v \neq 0 \right\}, \tag{10.10}$$

where

$$\Lambda_m(v;\theta) = \frac{1}{(m-1)!} \left| \sum_{i=1}^{n} a_i(v)^m \exp\{a_i(\theta)\} \right| \Bigg/ \left\{ \sum_{i=1}^{n} a_i(v)^2 \exp\{a_i(\theta)\} \right\}^{m/2}. \tag{10.11}$$

From this expression it is fairly easy to derive sufficient conditions on the sequence of covariates for the index to tend to zero. It is a delicate matter, however, to determine precisely for which sequences the index tends to zero, even in the case of a simple linear regression. In this case, with only one covariate, it takes some effort to provide a sequence of covariates such that the Fisher information tends to infinity but the index does not tend to zero.

11 Logistic regression

As a final example of a generalized linear model in the framework of Section 5, consider the reference model with point probabilities

$$g(y;\psi) = \mathrm{P}(Y = y) = \begin{cases} 1-p & \text{for } y = 0, \\ p & \text{for } y = 1, \end{cases} \tag{11.1}$$

where $\psi = \log\{p/(1-p)\}$, $0 < p < 1$. Again the model for $\psi_i = a_i(\theta)$ is linear in $\theta \in \Theta \subseteq V_1$ and there is no ϕ-parameter. The results and methods for this type of model are quite similar to those for the log-linear Poisson model and we include it here mainly because it is of some independent interest. The basic scheme of development is exactly the same as for the Poisson case and is therefore described less detailed here.

The first differential of the log-density in the reference model is

$$D_\psi \log g(Y;\psi) = Y - e^\psi/(1+e^\psi), \tag{11.2}$$

which has the cumulants

$$\operatorname{cum}_m\{D_\psi \log g(Y;\psi)\} = \operatorname{cum}_m(Y) \tag{11.3}$$

for $m \geq 2$. The Fisher information is

$$I(\psi) = \operatorname{var}(Y) = p(1-p) = e^\psi/(1+e^\psi)^2. \tag{11.4}$$

For the higher order cumulants we show below that

$$|\operatorname{cum}_m(Y)| < e(e-1)(m-1)!\operatorname{var}(Y) = e(e-1)(m-1)!p(1-p), \tag{11.5}$$

for $m \geq 3$, from which we see that the index $\lambda(\psi)$ satisfies the inequality

$$\begin{aligned}\lambda(\psi) &\leq e(e-1)\{p(1-p)\}^{-1/2}\\ &= e(e-1)\left\{e^\psi/(1+e^\psi)^2\right\}^{-1/2}.\end{aligned} \tag{11.6}$$

It is not hard to improve on the constant $e(e-1)$ but that is not of importance here.

A proof of (11.6) starts from the cumulant generating function for $Y-p$,

$$\kappa(z) = \log \mathrm{E}\exp\{z(Y-p)\} = \log\{pe^z + (1-p)\} - zp, \qquad z \in \mathbb{C}, \tag{11.7}$$

the first derivative of which is

$$D\kappa(z) = p(1-p)\frac{e^z-1}{1+p(e^z-1)}. \tag{11.8}$$

This is an analytic function without singularities on the disc $|z| \leq 1$. Therefore Cauchy's inequalities tell us that

$$|D^m\kappa(0)| \leq (m-1)!M_1,$$

where

$$\begin{aligned}M_1 = \sup_{|z|=1}\{|D\kappa(z)|\} &\leq p(1-p)\frac{e-1}{1+p(e^{-1}-1)}\\ &< p(1-p)e(e-1)\end{aligned}$$

justifying the claim from (11.5).

For the generalized linear model the Fisher information is

$$I^{(n)}(\theta)(v^2) = \sum_{i=1}^{n} a_i(v)^2 e^{\psi_i}/(1+e^{\psi_i})^2, \tag{11.9}$$

for $v \in V = V_1$, and from Theorem 2.5.3 it then follows that

$$\lambda^{(n)}(\theta)^2 \leq e(e-1)\sup\left\{ a_i(v)^2 \Bigg/ \left(\sum_{j=1}^{n} a_j(v)^2 e^{\psi_j}/(1+e^{\psi_j})^2\right) \right. \\ \left. : 1 \leq i \leq n, v \in V, v \neq 0 \right\}, \tag{11.10}$$

where $\psi_i = a_i(\theta)$. The condition that $\lambda^{(n)}(\theta)$ tends to zero as n tends to infinity implies again that the asymptotic results for the maximum likelihood estimator from Sections 3 and 4 hold, and this is the case if

$$a_n(v)^2 \Big/ \sum a_j(v)^2 e^{a_j(\theta)} \Big/ \left(1+e^{a_j(\theta)}\right)^2 \to 0$$

uniformly in $v \neq 0$ as $n \to \infty$.

As for the Poisson model we may compare with Condition 2 in Haberman (1977), Section 2, which for the present model reduces to the condition that $d_t \to 0$, where d_t in our notation equals

$$\sup\left\{ \max\{1-p_i, p_i\}\, |a_i(v)| \Big/ \{I^{(n)}(\theta)(v^2)\}^{1/2} : 1 \leq i \leq n, v \in V, v \neq 0 \right\}. \tag{11.11}$$

The inequality $\frac{1}{2} \leq \max\{1-p_i, p_i\} < 1$ immediately shows the equivalence of Haberman's condition to the condition that the right hand side of (11.10) tends to zero as n tends to infinity.

Finally, we may also for this model improve the result from (11.10) by working directly from the definition of the index of the generalized linear model. Then, analogously to the Poisson case, we arrive at the result that

$$\lambda^{(n)}(\theta) = e(e-1)\sup\left\{ \Lambda_m(v;\theta)^{1/(m-2)} : m \geq 3, v \in V, v \neq 0 \right\}, \tag{11.12}$$

where

$$\Lambda_m(v;\theta) = \left|\sum_{i=1}^{n} a_i(v)^m e^{\psi_i}/(1+e^{\psi_i})^2\right| \Bigg/ \left\{\sum_{i=1}^{n} a_i(v)^2 e^{\psi_i}/(1+e^{\psi_i})^2\right\}^{m/2}. \tag{11.13}$$

The considerations of which sequences of covariates that make the index tend to zero are almost the same as for the log-linear Poisson models. The fact that the

Fisher information from a single observation is bounded for the logistic models does not lead to any simplification or change of importance. Thus, beside the condition that the Fisher information tends to infinity, we need the condition that the right hand side of (11.12) tends to zero for the asymptotic results to hold.

12 Generalized non-linear models

In analogy with the generalized linear models from Section 5 we consider here the generalized non-linear models, which differ from the linear ones only by the relaxation of the requirement that the mappings a_i in (5.2) are linear to the requirement that they are analytic. Despite the term 'non-linear' we do not exclude the special case of linear a_i's here. In view of the generality of the models covered by this framework it is somewhat surprising that the complexity of proofs of asymptotic results for sequences of such models may not be much greater than for the generalized linear models, indicated by the examples in Sections 8-11. The basic technique is again to use Theorem 2.5.3 and Theorem 2.6.1 to derive conditions for the index of the model to tend to zero as n tends to infinity. Although the conditions on the mappings (a_i) for this to be the case, are usually weaker for the special case of a normal non-linear regression model than for other generalized non-linear models, the considerations involved in the proof are mainly related to the computation of the constant R from (2.6.8) corresponding to each of the mappings a_i, and this computation is not more complicated for general models of this type than for the case of a normal reference distribution.

Let us first quickly review the basic setup for a generalized non-linear model in our context. The *reference model* is a model for $Y \in E$ parametrized by two (vector) parameters $\psi \in \Psi \subseteq W$ and $\phi \in \Phi \subseteq V_2$, where W and V_2, like V_1 below, are finite-dimensional real vector spaces. This reference model has densities

$$g(y; \psi, \phi), \qquad (\psi, \phi) \in \Psi \times \Phi \subseteq W \times V_2, \tag{12.1}$$

with respect to some underlying measure ν on E. The reference model is assumed to be analytic throughout its parameter space. A generalized non-linear model is a model for a sequence $Y^{(n)} = (Y_1, \dots, Y_n)$ of independent random variables in E. The distribution of Y_i is assumed to have density

$$f(y_i; \theta, \phi) = g(y_i; \psi_i, \phi), \tag{12.2}$$

with respect to ν, where

$$\psi_i = a_i(\theta), \tag{12.3}$$

and $a_i : \Theta \to \Psi$ is a known *analytic* function of the unknown parameter $\theta \in \Theta \subseteq V_1$. Thus the parameter space for the generalized non-linear model is

$$B = \Theta \times \Phi \subseteq V = V_1 \times V_2$$

and it follows immediately from Theorem 2.6.1 and Theorem 2.5.3 that the model is analytic. In some cases the model contains no ϕ-parameter, and consequently some trivial notational changes are required.

Often the functions a_i are given as some fixed function of the parameter θ and some covariates for the ith observation, but we have no need for that assumption here. In specific cases, however, it may be a useful as well as natural assumption.

The Fisher information and the index of the reference model are denoted $I(\psi, \phi)$ and $\lambda(\psi, \phi)$, respectively. The corresponding quantities in the model for Y_i are denoted $I_i(\theta, \phi)$ and $\lambda_i(\theta, \phi)$. For the entire generalized non-linear model the index is denoted $\lambda^{(n)}(\theta, \phi)$, and the Fisher information is given by

$$\begin{aligned} I^{(n)}(\theta,\phi)(v_1,v_2)^2 &= \sum_{i=1}^{n} I_i(\theta,\phi)(v_1,v_2)^2 \\ &= \sum_{i=1}^{n} I(\psi_i,\phi)\big(\{Da_i(\theta)\}(v_1),v_2\big)^2, \end{aligned} \tag{12.4}$$

for $(v_1, v_2) \in V_1 \times V_2$. This information differs from that for the generalized linear models in (5.10), only by the appearance of the first differential, $Da_i(\theta) : V_1 \to W$, instead of a_i. In the linear case the two expressions agree.

To investigate whether the index of a sequence of generalized non-linear models tends to zero as n tends to infinity we need a generalization of Theorem 5.1. In the proof of that theorem we used the fact, proved in Theorem 2.6.1, that the linear reparametrization a_i does not increase the index. For the case of a non-linear mapping a_i we must instead use the bound in (2.6.9) for the index of the model for Y_i. Otherwise the theorem and its proof are the same.

Let $R_i(\theta)$ denote a constant satisfying the inequality (2.6.8) for the model for Y_i, where the reparametrization in this connection is the mapping a_i. Thus, the inequality (2.6.8) becomes

$$||D^k a_i(\theta)(v_1)^k||_{I(\psi_i,\phi)} \leq k!\, R_i(\theta)^{k-1} ||Da_i(\theta)(v_1)||^k_{I(\psi_i,\phi)}, \tag{12.5}$$

where $v_1 \in V_1$ and this inequality is required to hold for all v_1 and all $k \in \mathbb{N}$. This quantity measures the 'degree of non-linearity' of the mapping a_i in relation to the model, and bounds the amount by which the index of the model for Y_i can be inflated by this non-linear mapping, as is seen from (2.6.9).

Theorem 12.1. *Assume that the reference model from (12.1) is analytic with a finite index $\lambda(\psi, \phi)$ on $\Psi \times \Phi$. Then the index $\lambda^{(n)}(\theta, \phi)$ of the model for $Y_1, \ldots, Y_n$ tends to zero as n tends to infinity if both of the following conditions hold:*

(1) The smallest eigenvalue of the Fisher information $I^{(n)}(\theta, \phi)$, relative to a fixed inner product on $V_1 \times V_2$, tends to infinity as n tends to infinity.

(2) Uniformly in $(v_1, v_2) \in V_1 \times V_2$ we have

$$\{\lambda(\psi_n,\phi) + R_n(\theta)\}^2 \{I_n(\theta,\phi)(v_1,v_2)^2\} / I^{(n)}(\theta,\phi)(v_1,v_2)^2 \to 0 \tag{12.6}$$

as $n \to \infty$, where $\psi_n = a_n(\theta)$, cf. (12.3).

Proof. The only change compared to the proof of Theorem 5.1 is that the index for the model for Y_i is not $\lambda(\psi_i, \phi)$, but is bounded instead by $\lambda(\psi_i, \phi) + R_i(\theta)$, as was shown in Theorem 2.6.1. ▮

A difficulty in the application of the theorem is that the quantity $R_i(\theta)$ from (12.5) may well be infinite. This is particularly likely to occur when the dimension of the θ-parameter is greater than the dimension of the ψ-parameter, as is often the case. In that case the model for a single observation is over-parametrized and its index may therefore be infinite. The reason that this problem did not occur with the generalized linear models is that although the model for a single observation in that case also might be over-parametrized, this would be due to the linear reparametrization a_i, and a linear reparametrization does not increase the index. Stated in another way, the Fisher information for the model for a single observation may be zero in certain directions in both cases, but in the linear case this would imply that the model was constant in those directions, and according to Corollary 2.5.9 this does not make the index infinite.

A remedy for this difficulty for the present, non-linear, models is to group the observations such that we let the reference model be a model for several independent observations. Usually we would need to group m observations, say, where m is such that $m \dim \Psi \geq \dim \Theta$ in which case the Fisher information is likely to be positive definite and the index therefore finite. The drawback is that in this way the dimensionality, and hence the complexity, of the computations are increased. The last example in Section 14 illustrates this problem.

13 One-parameter exponential regression function

Consider a generalized non-linear model in the framework of Section 12, where θ is one-dimensional and

$$\psi_i = a_i(\theta) = e^{\theta x_i}, \qquad \theta \in \mathbf{R}, \tag{13.1}$$

are the expressions for the ψ_i's in terms of θ and the (known) covariates (x_i). For the moment we leave the reference model unspecified, but we assume throughout the present section that the model contains no ϕ-parameter. This is done purely for the convenience of not having to repeat the arguments needed to handle this extra parameter for reference models such as the normal distribution with unknown variance (Section 7), location models with unknown scale (Section 8), and the Gamma distribution with unknown shape (Section 9). The results given below all generalize to these models.

The Fisher information for the reference model is $I(\psi)$ and that for the model for Y_i is

$$I_i(\theta) = I(\psi_i) x_i^2 e^{2\theta x_i}. \tag{13.2}$$

Since the kth derivative of a_i is $x_i^k \exp(\theta x_i)$, the inequality (12.5), after cancellation of the factor v_1^k from the two sides of the equation, becomes

$$|x_i|^k e^{\theta x_i} I(\psi_i)^{1/2} \leq k!\, R_i(\theta)^{k-1} \left\{ |x_i| e^{\theta x_i} I(\psi_i)^{1/2} \right\}^k ,$$

from which it is seen that

$$R_i(\theta) = \frac{1}{2} I(\psi_i)^{-1/2} e^{-\theta x_i} \tag{13.3}$$

is the smallest possible value of $R_i(\theta)$. Hence it follows from (2.6.9) in Theorem 2.6.1 that the index $\lambda_i(\theta)$ of the model for Y_i satisfies the inequality

$$\lambda_i(\theta) \leq \lambda(\psi_i) + \frac{1}{2} I(\psi_i)^{-1/2} e^{-\theta x_i}. \tag{13.4}$$

We now consider the case where the reference model consists of normal distributions with fixed known variance σ^2 and mean ψ, as in Section 6. For that case we know that $\lambda(\psi) = 0$, and hence the condition (2) in Theorem 12.1 reduces to the condition that

$$\sigma^2 x_n^2 \Big/ \sum_{i=1}^n x_i^2 \exp(\theta x_i) \to 0 \tag{13.5}$$

as $n \to \infty$. The condition (1), that the Fisher information tends to infinity, demands the denominator of this expression to tend to infinity as n tends to infinity. Only two types of violation of condition (13.5) can occur if the Fisher information tends to infinity; either a convergence of the x_i's to zero at an appropriate rate, or the existence of a sub-sequence of the x_i's of opposite sign of θ (or of either sign if $\theta = 0$), for which $|x_i|$ tends to infinity sufficiently rapidly.

As in Sections 10 and 11 it is not hard to write down an expression for the index $\lambda^{(n)}(\theta)$ directly for this model. To do that we only have to consider cumulants of the form $\chi_{kk}^{(n)}(\theta)$, because cumulants of order three or higher are zero for the normal distribution. From this directly derived expression for the index it is seen that $\lambda^{(n)}(\theta) \to 0$ as $n \to \infty$ if and only if

$$\sup_{k \geq 3} \left\{ \frac{\sigma}{k} \left[\sum_{i=1}^n x_i^{2k} e^{2\theta x_i} \Big/ \left(\sum_{i=1}^n x_i^2 e^{2\theta x_i} \right)^k \right]^{1/(2k-2)} \right\} \to 0 \tag{13.6}$$

as $n \to \infty$. For sequences of x_i's for which the information tends to infinity this condition is strictly weaker than (13.5).

Notice from the condition in (13.6) that the index obviously tends to zero under the limiting operation $\sigma \to 0$ with n fixed, unless all the covariates are zero. This is the kind of asymptotics that suggests the adequacy of the first order asymptotic theory for non-linear normal regression models when the standard deviation is small.

Let us turn now to the case where the reference model is an arbitrary (analytic) generalized non-linear model without any ϕ-parameter. From (13.2) and (13.3) we then see that the condition (2) in Theorem 12.1 becomes

$$x_n^2 \left(\frac{1}{2} + \lambda(\psi_n) I(\psi_n)^{1/2} e^{\theta x_n} \right)^2 \Big/ \sum_{i=1}^{n} I(\psi_n) x_i^2 e^{2\theta x_i} \to 0 \tag{13.7}$$

as $n \to \infty$, while condition (1) still requires the denominator to tend to infinity. A direct computation of the index becomes more complicated for this general case. Notice that although the condition in (13.7) is, of course, more restrictive than the corresponding condition in (13.5) for the normal case, its derivation was equally easy and the final result is usually of similar complexity.

For the special case of a Gamma reference model with fixed known shape parameter α, cf. Section 9, we get

$$I_i(\theta) = \alpha x_i^2, \tag{13.8}$$

and

$$R_i(\theta) = 1/(2\sqrt{\alpha}), \tag{13.9}$$

from which it is seen that the condition (2) in Theorem 12.1 becomes

$$x_n^2 \Big/ \sum_{i=1}^{n} x_i^2 \to 0 \tag{13.10}$$

as $n \to \infty$, the simplicity of which is striking. Condition (1) is here the condition that $\sum x_i^2 \to \infty$ as $n \to \infty$.

Also for the case where the reference model is an analytic location model, as considered in Section 3.4, the condition (13.7) reduces somewhat because $I(\psi)$ and $\lambda(\psi)$ are both constant.

14 Two-parameter exponential regression function

We now extend the model from the previous section by adding another parameter. This will be done in two different ways; first we consider the regression function

$$\psi_i = a_i(\alpha, \gamma) = \alpha e^{\gamma x_i}, \qquad \alpha > 0, \gamma \in \mathbf{R}, \tag{14.1}$$

where (α, γ) plays the role of θ from (13.1). Otherwise we still consider the same setting as in Section 13, i.e., the model is a generalized non-linear model of the form described in Section 12, without any ϕ-parameter.

Although the model for a single observation, Y_i say, is over-parametrized by (α, γ) instead of just ψ_i, we may derive conditions for the index to tend to zero by the same method as in the previous section, based on Theorem 12.1. The reason is that the function in (14.1) allows a reparametrization from (α, γ) to (η, γ), where $\eta = \log \alpha$, such that

$$\psi_i = \tilde{a}_i(\eta, \gamma) = e^{\eta + \gamma x_i}, \qquad \eta \in \mathbf{R}, \gamma \in \mathbf{R}. \tag{14.2}$$

The point is that in this new parametrization the parameters enter only through a one-dimensional linear function. Therefore the model for Y_i is constant in any direction in the parameter space in which the Fisher information is zero, and the index is therefore finite, provided that this is so in the reference model, cf. Corollary 2.5.9.

We proceed now with the derivation in the parametrization (14.2) and use a tilde, as in $\tilde{\lambda}(\eta, \gamma)$, to indicate that the parametrization has been changed.

Let $\lambda(\psi)$ denote the index of the reference model, and $\tilde{\lambda}_i(\eta, \gamma)$ that of the model for Y_i. This latter index is known from Theorem 2.6.1 to be bounded by the index of the model parametrized by the single parameter $\eta + \gamma x_i$, which in turn was shown in the previous section to be bounded by the inequality (13.4), because the choice of parametrization by θ or by θx_i of the one parameter model (13.1) for Y_i does not affect the index. Hence we have

$$\tilde{\lambda}_i(\eta, \gamma) \leq \lambda(\psi_i) + \frac{1}{2} I(\psi_i)^{-1/2} e^{-(\eta + \gamma x_i)}. \tag{14.3}$$

The Fisher information $\tilde{I}_i(\eta, \gamma)$ in the model for Y_i is given by

$$\tilde{I}_i(\eta, \gamma)(s, t)^2 = (s^2 + 2stx_i + t^2 x_i^2) I(\psi_i) e^{\eta + \gamma x_i}, \tag{14.4}$$

where $(s, t) \in \mathbf{R}^2$, with the matrix representation

$$\begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix} I(\psi_i) e^{(\eta + \gamma x_i)}. \tag{14.5}$$

It is now a matter of insertion to see that the condition (2) in Theorem 12.1

states that

$$\frac{(s^2+2stx_n+t^2x_n^2)\left(\frac{1}{2}+\lambda(\psi_n)I(\psi_n)^{1/2}\exp\{\eta+\gamma x_n\}\right)}{\sum_{i=1}^n(s^2+2stx_i+t^2x_i^2)I(\psi_i)^{1/2}\exp\{\eta+\gamma x_i\}}\to 0 \tag{14.6}$$

uniformly in $(s,t)\neq(0,0)$ as $n\to\infty$. The condition (1) requires the denominator to tend to infinity. For the special case of normal distributions with fixed known variances the second factor in the numerator reduces to a constant. For the Gamma distribution with a fixed known shape parameter, (14.6) reduces to the condition that

$$(s^2+2stx_n+t^2x_n^2)\Big/\sum_{i=1}^n(s^2+2stx_i+t^2x_i^2)\to 0 \tag{14.7}$$

uniformly in $(s,t)\neq(0,0)$ as $n\to\infty$, cf. (13.10).

These two conditions for the index $\tilde{\lambda}^{(n)}(\eta,\gamma)$ to tend to zero imply the same conclusion to hold for the index $\lambda^{(n)}(\alpha,\gamma)$ in the original parametrization, as noted in the comments following Conditions 2.1. Thus, the change of parametrization has reduced the complexity of the problem, in effect by reducing the dimensionality.

Consider now, instead of (14.1), another two parameter regression function, namely

$$\psi_i=a_i(\mu,\gamma)=\mu+e^{\gamma x_i},\qquad \mu\in\mathbf{R},\gamma\in\mathbf{R}, \tag{14.8}$$

otherwise in the same setting as above. Here the method of reparametrization provides no simplification because ψ_i cannot be written as a function of any linear combination of two parameters that do not depend on i. Hence we are facing the problem that the index $\lambda_i(\mu,\gamma)$ of the model Y_i, is infinite. As noted below Theorem 12.1, a remedy is to group the observations, in the present case in pairs. This means that we modify the reference model to consist of two independent observations, such as (Y_1,Y_2), and consequently extend the ψ-parameter to be two-dimensional corresponding to the pair (ψ_1,ψ_2). Provided that the corresponding two covariates x_1 and x_2 are different, the Fisher information for (μ,γ) is positive definite and we can therefore proceed as before, using Theorem 12.1 to obtain conditions for the index to tend to zero. The price paid is an increased complexity of the problem, because of the higher dimensions of the parameter spaces involved, and perhaps more importantly a less 'streamlined' derivation of the conditions. For example, there are many ways in which the observations can be grouped in pairs and some may be less effective than others. The model in (14.8) approaches a complexity where additional assumptions on the covariates should be introduced to simplify the problem. Thus, it might be quite natural to assume that the covariates are independent identically distributed random variables. Then the approach sketched above, with the pairwise grouping, might be used to derive conditions on the distribution of covariates, since all groupings in pairs lead to the same result. We shall, however, not pursue the computations for the present example since it is included only to illustrate the theory. Note, however, that the more natural extension of the model to a three parameter regression function with both of the α and μ parameters from (14.1) and (14.8), is not much more complicated than the two parameter function in (14.8), because we can still use the reparametrization from (14.2) to collapse two of the parameters into a single one at the intermediate stage of the computations.