Chapter 5

Higher order asymptotic theory for independent replications

1 Introduction

In this chapter we review some basic higher order asymptotic theory for analytic models. As mentioned in the introduction to Chapter 4 we confine ourselves to the case of independent replications to avoid the technical complications that arise in connection with the Cramér condition for characteristic functions, cf. Lemma 2.2, for more general cases. The moment conditions required to derive higher order asymptotic expansions would be fulfilled for any sequence of models for which the index tends to zero.

The purpose of the chapter is not to develop the higher order asymptotic theory as such, but to justify the claim made earlier, cf. the preface and the introduction to Chapter 2, that the analytic models are 'regular' in the sense that they fulfil the regularity conditions employed to prove standard asymptotic results. Therefore it is not attempted to give a review of the vast amount of literature on higher order asymptotic results. For likelihood based inference in parametric models some references giving accounts of, at least, parts of this theory, are McCullagh (1987), Amari (1985), Pfanzagl (1982) and Pfanzagl (1985). Many important results, such as the existence of the Bartlett correction factor to the log likelihood ratio test, are outside the scope of the present chapter, although they are exactly the kind of results that were the motivation for the development of the present theory.

Certain modifications may have to be made to adapt theorems to our framework; for example, Bhattacharya and Ghosh (1978) in their proof of the validity of the Edgeworth expansions for the local maximum likelihood estimator, assume that the observations belong to a finite-dimensional real vector space. While this may be convenient for applications it does not fit in very well with our framework, where it has been completely avoided to specify the type of the sample space, because the importance is the structure of the parameter space. The difference is unimportant, because the log-likelihood differentials, which are sufficient as shown in Section 2.8, will be elements of finite-dimensional real vector spaces. However, the different setup requires a minor change in the assumptions. Apart from such minor differences only two further assumptions are needed for an analytic model to derive the higher order expansions, namely the Cramér condition mentioned above, or another continuity condition on the distribution, and the condition that the Fisher information is positive definite.

At the end of this section we describe the basic setup and the notation to be used throughout the chapter. In Section 2 we briefly review the definition of the Edgeworth expansions and then demonstrate the validity of these expansions for any finite sequence of log-likelihood differentials. The expansions for these statistics are, not surprisingly considering their sufficiency or approximate sufficiency as discussed in Sections 2.7 and 2.8, the basis of most higher order expansions of relevance to likelihood based statistical inference.

The conditions provided by Bhattacharya and Ghosh (1978) for the validity of the Edgeworth expansions for the local maximum likelihood estimator are verified in Section 3 as the main example of regularity conditions satisfied by the analytic models. Given the theory developed so far, and the theory of Edgeworth expansions, it would not be hard to prove the validity of these expansions directly, but, as argued above, the purpose here is to demonstrate that these models satisfy commonly used conditions. Of course, many other results and variations of the conditions could equally well have been chosen, but to a large extent the basic conditions are the same, namely conditions of uniform integrability of certain derivatives of the log-likelihood function.

In Section 4 we turn to the problem of the consistency of the maximum likelihood estimator for a compact parameter space, although this is not an asymptotic result of higher order. The proof, essentially due to Wald (1949), although various minor improvements have been given later, does, however, require the observations to be independent and identically distributed and therefore we include the verifications of the conditions for this result in the present chapter.

Throughout the chapter we will be working with sequences of models as described in the remainder of the present section.

Consider a model

$$\{(E, \nu); f_1(y;\beta); \beta \in B \subseteq V\} \tag{1.1}$$

parametrized by β taking values in the finite-dimensional real vector space V. Let the point $\beta_0 \in \text{int}(B)$ be fixed and assume that the model is analytic at this point. Let $Y_1, \ldots, Y_n$ be independent and identically distributed random variables on E, each with density $f_1(y;\beta)$, where β is still the (unknown) parameter. Thus, the model for $Y^{(n)} = (Y_1, \ldots, Y_n)$ is

$$\{f^{(n)}(y^{(n)};\beta); \beta \in B \subseteq V\}, \tag{1.2}$$

where

$$f^{(n)}(y^{(n)};\beta) = \prod_{i=1}^{n} f_1(y_i;\beta) \tag{1.3}$$

is the density on the product space at the point $y^{(n)} = (y_1, \ldots, y_n)$. It follows from Theorem 2.5.2 that the model (1.2) for $Y^{(n)}$ is analytic at the point β_0.

The log-likelihood differentials from (2.2.1) for the model (1.1) for the ith observation Y_i, are denoted

$$D_{k,i}(\beta) = D_\beta^k \log f_1(Y_i;\beta), \tag{1.4}$$

while the corresponding differentials for the model (1.2) for $Y^{(n)}$ are denoted

$$D_k^{(n)}(\beta) = D_\beta^k \log f^{(n)}(Y^{(n)}; \beta). \tag{1.5}$$

For the model (1.1) for a single observation we denote the Fisher information from (2.3.16), the cumulants of the log-likelihood differentials from (2.3.15), and the index from Definition 2.5.1, by $I(\beta)$, $\chi_{k_1 \cdots k_m}(\beta)$, and $\lambda(\beta)$, respectively, while the corresponding quantities are denoted

$$I^{(n)}(\beta) = nI(\beta), \tag{1.6}$$

$$\chi^{(n)}_{k_1 \cdots k_m}(\beta) = n\chi_{k_1 \cdots k_m}(\beta), \tag{1.7}$$

$$\lambda^{(n)}(\beta) = \lambda(\beta)/\sqrt{n} \tag{1.8}$$

for the model (1.2) for $Y^{(n)}$, cf. (2.5.11).

We let $\langle \cdot, \cdot \rangle$ denote an arbitrary pre-given inner product on V, and $\|\cdot\|$ the corresponding norm. The same notation is used for the dual inner product and the corresponding norm on the dual space V^*. By $\|\cdot\|_n$ we denote the Fisher information semi-norm on V, i.e.,

$$\|v\|_n^2 = I^{(n)}(\beta_0)(v^2), \qquad v \in V, \tag{1.9}$$

and we use the same notation for the corresponding norm induced by the inner product $I^{(n)}(\beta_0)^{-1}$ on V^*, whenever $I^{(n)}(\beta_0)$ is positive definite.

We adopt the previously used standard that if the argument β is omitted, evaluation at β_0 is understood.

Notice that with the only reservation that the Fisher information is positive definite, all the results from Sections 4.3 and 4.4 hold for the models considered here. This is easily seen by verification of Conditions 4.2.1, and is mainly due to the result (1.8) taken from Theorem 2.5.2, which shows that the index tends to zero as n tends to infinity. Thus, there is no need to consider first order asymptotic results here, except the consistency of the maximum likelihood estimator verified in Section 4.

2 Edgeworth expansions for the log-likelihood differentials

In this section we give conditions for the validity of an Edgeworth expansion to any order of the joint distribution of $\left(D_1^{(n)}, \ldots, D_K^{(n)}\right)$ for any $K \in \mathbf{N}$. The only extra condition needed, beside the assumption that the model is analytic at β_0, is a continuity condition on the distribution, or an assumption that the distribution is a lattice distribution. Since the statistic considered is a sum of independent and identically distributed random variables the result follows immediately from the 'classical' theory of Edgeworth expansions, an account of which may be found in Bhattacharya and Rao (1976), and proofs are therefore not given here. Because the model is analytic, the statistic considered has moments of all orders. Hence it follows from the theory of Edgeworth expansions that such are valid for the density if an integrability condition is imposed on the characteristic function, for the probabilities of 'regular sets', see (2.6) below, if Cramér's condition is required instead, and for the point probabilities if the distribution is a lattice distribution.

Below we briefly review the construction of the Edgeworth approximations. A more comprehensive account, using a notation similar to the one used here, may be found in Skovgaard (1986b), and a fairly short review, using a more traditional notation, in Bhattacharya and Ghosh (1978).

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables on a finite-dimensional real vector space W, and assume that X_j has moments of all orders. Denote the cumulants of X_j by

$$\kappa_k = \mathrm{cum}_k(X_j), \qquad k \in \mathbf{N}, \tag{2.1}$$

and suppose that the variance $\Gamma = \kappa_2$ is positive definite. Recall the definition of the characteristic function of X_j, as the function

$$\xi_1(t) = \mathrm{E}\exp\{it(X_j)\}, \qquad t \in W^*. \tag{2.2}$$

We want an expansion of the distribution of

$$U_n = (X_1 + \cdots + X_n - n\kappa_1)/\sqrt{n}$$

the leading term of which is, of course, the normal distribution with mean zero and variance Γ. The logarithm of the characteristic function $\xi^{(n)}$ of U_n has the Taylor series expansion

$$\log \xi^{(n)}(t) \sim -\frac{1}{2}\Gamma(t^2) + \sum_{k=3}^{\infty} \tau^{k-2} \frac{i^k}{k!} \kappa_k(t^k), \tag{2.3}$$

as $\tau \to 0$, where $\tau = 1/\sqrt{n}$. On expansion of the exponential of the sum on the right hand side in powers of τ, we obtain an expansion of the form

$$\xi^{(n)}(t) = \exp\left\{-\frac{1}{2}\Gamma(t^2)\right\}\left(1 + \sum_{k=1}^{M-1} \tau^k \tilde{P}_k(it : \{\kappa_j\})\right) + O(\tau^M) \tag{2.4}$$

as $\tau \to 0$, where $M \in \mathbf{N}$, and $\tilde{P}_k(it : \{\kappa_j\})$ is the kth Cramér-Edgeworth polynomial in (it), cf. Bhattacharya and Rao, Section 7. The polynomial $\tilde{P}_k$ has coefficients depending on $\kappa_3, \ldots, \kappa_{k+2}$, and it is of degree $3k$.

By Fourier inversion of the right hand side of (2.4), disregarding the error term, we get the approximate density

$$g_{M,n}(u : \{\kappa_j\}) = \left\{1 + \sum_{k=1}^{M-1} \tau^k P_k(-D : \{\kappa_j\})\right\} \phi_{0,\Gamma}(u), \tag{2.5}$$

which will be referred to in the sequel as the Edgeworth approximation of order M to the density of U_n. Here, D denotes the differential operator and $P_k(-D : \{\kappa_j\})$ is the same polynomial as $\tilde{P}_k$ above, except that the argument is $(-D)$ instead of (it). Thus, any power of $(-D)$ is supposed to operate on the normal density function $\phi_{0,\Gamma}$, with mean 0 and variance Γ, such that the result becomes a differential of a certain order of this density. The resulting density therefore becomes a linear combination of such differentials, i.e., a polynomial multiplied by the normal density function.

The approximation $g_{M,n}(u : \{\kappa_j\})$ involves the cumulants to order $M + 1$, i.e., including κ_{M+1}. For example, the second order Edgeworth approximation is based on the first three cumulants.

It is important to notice that given the first $M + 1$ cumulants of any statistic we can always write down the Edgeworth approximation of order M; just take $n = 1$ in (2.5) and insert the cumulants in the formula. In the case of a standardized sum of independent and identically distributed random variables, as U_n above, this approach leads to the same expression as (2.5), because the powers of τ will be 'inherent' in the cumulants of U_n. The quality of the approximation (2.5), or even the existence of the exact density, is a quite different matter which is the theme of the theorems of Edgeworth expansions.

The measure with the density (2.5) will be referred to as the Edgeworth measure (of order M) based on the first $M + 1$ cumulants of U_n, and it is denoted $Q_{M,\tau}(\cdot; \{\kappa_j\})$. It is a finite signed measure assigning measure one to the entire set W. It is not in general a probability measure, since it may assign negative measure to some sets. An important fact in connection to the applications of the Edgeworth approximations is that when W is one-dimensional it is a trivial matter to integrate the density in (2.5) explicitly.

In many cases the 'cumulants' on which an Edgeworth approximation is based are not cumulants of any probability distribution, but only approximations to such cumulants. For these cases we may still use the expression from (2.5), and we still speak of the Mth order Edgeworth approximation based on these 'cumulants'.

Whenever the mean κ_1 is non-zero, the Edgeworth approximation is assumed to be the approximation obtained as above (with $\kappa_1 = 0$) translated by the amount κ_1.

Excluding the lattice case and assuming the existence of all moments, there are two main versions of the Mth order Edgeworth expansion of a probability distribution. One gives an expansion for the density with an additive error of order

$O(\tau^M)$ uniformly in $u \in W$ as $n \to \infty$. This implies that the expansion of the distribution by the Edgeworth measure also gives a uniform additive error of the same order of magnitude uniformly over all Borel sets, cf. Skovgaard (1986b). The other version holds under a weaker continuity condition and requires the technique of smoothing to be used, thereby leading to an expansion of the measure, giving an error which is again $O(\tau^M)$, but this time uniformly only over any class $\mathbf{A}$ of Borel sets satisfying the condition

$$\sup\{\Phi_{0,\Gamma}([\partial A]^\epsilon) : A \in \mathbf{A}\} = O(\epsilon) \quad \text{as} \quad \epsilon \to 0, \tag{2.6}$$

where $\Phi_{0,\Gamma}$ is the normal distribution with mean 0 and variance Γ, and $[\partial A]^\epsilon$ is the ϵ-boundary of A, see Bhattacharya and Rao (1976), Section 1, and Bhattacharya and Ghosh (1978), Theorem 1. It may be noted that the class of convex sets satisfies (2.6), cf. Bhattacharya and Rao (1976), Section 2.

Let us return now to the distribution of $(D_1^{(n)}, \ldots, D_K^{(n)})$ from the analytic model (1.2). Let W denote the vector space spanned by the distribution of

$$X_i = (D_{1,i} - \chi_1, \ldots, D_{K,i} - \chi_K),$$

the dependence on K being notationally subpressed. Let $\kappa_1, \kappa_2, \ldots$, denote the cumulants of $X_i \in W$. By construction the mean κ_1 is zero and the variance $\Gamma = \kappa_2$ is positive definite. We want an expansion of the distribution of $X^{(n)} = \sum X_i$, or equivalently of the standardized version $U^{(n)} = X^{(n)}/\sqrt{n}$.

Lemma 2.1. *Assume that some power of the characteristic function of X_i is absolutely integrable. Then, for sufficiently large n, the density $g^{(n)}(u)$ of $U^{(n)}$ exists, and for any $M \in \mathbf{N}$ we have*

$$|g^{(n)}(u) - g_{M,n}(u : \{\kappa_j\})| = O(\tau^M) \tag{2.7}$$

as $n \to \infty$ uniformly in $u \in W$, where $\tau = 1/\sqrt{n}$. Also, we have the distributional expansion

$$|\mathrm{P}\{U^{(n)} \in A\} - Q_{M,n}(A : \{\kappa_j\})| = O(\tau^M) \tag{2.8}$$

as $n \to \infty$ uniformly over all Borel sets $A \subseteq W$.

Proof. From the fact that the model is analytic we know that all moments of X_i are finite. Hence the assumptions of Theorem 19.2 in Bhattacharya and Rao (1976) are fulfilled, and consequently the Edgeworth expansion to any order is valid as stated in the lemma. The fact that we obtain the error $O(\tau^M)$ instead of just $o(\tau^{M-1})$ is a simple consequence of the fact that also the Edgeworth to next order, i.e., the expansion of order $M+1$, is valid. From that expansion the error is $o(\tau^M)$ and since the difference between the two expansions is of order $O(\tau^M)$ the result follows.

∎

Lemma 2.2. *Assume that Cramér's condition is satisfied for the characteristic function*

$$\xi(t) = \mathrm{E}\exp\{it(X_i)\}, \qquad t \in W^*, \tag{2.9}$$

i.e., that

$$\sup\{\,|\xi(t)| : \|t\| > c\,\} < 1 \tag{2.10}$$

for some, and hence for any, $c > 0$. Then, for any $M \in \mathbf{N}$, we have the distributional expansion

$$\left|\mathrm{P}\{U^{(n)} \in A\} - Q_{M,n}(A : \{\kappa_j\})\right| = O(\tau^M) \tag{2.11}$$

as $n \to \infty$ uniformly over A in any class $\mathbf{A}$ satisfying (2.6).

Proof. Here we use Corollary 20.4 from Bhattacharya and Rao (1976) and note again that the error becomes $O(\tau^M)$ because the Edgeworth expansion to the next order is also valid. ∎

We omit the corresponding result for the case of lattice random variables. The result is an expansion for the point probabilities similar to the one for densities obtained in Lemma 2.1, except for the appearance of a normalizing factor. This result is, however, less useful for inferential purposes because a (non-linear) transformation of a lattice distribution generally leads to a discrete, but non-lattice, distribution. For such transformations there are great technical difficulties in obtaining expansions beyond the first order results and therefore the expansions for the log-likelihood differentials are not easily transformed to expansions for other statistics, such as the local maximum likelihood estimator as considered in the next section for the continuous case.

3 Edgeworth expansions for the local maximum likelihood estimator

We now turn to Edgeworth expansions of the distribution of the local maximum likelihood estimator (LMLE). We use the results from Bhattacharya and Ghosh (1978) and verify these by proving that their conditions hold for the models considered here, except for minor modifications discussed in connection with the theorem given below. First we give a very brief heuristic summary of the method of proof used, since the same scheme may be employed to derive expansions for other estimators or statistics.

The first step is to derive an Edgeworth expansion for a certain number of the log-likelihood differentials, as we did in the previous section. Since these are approximately sufficient, cf. Section 2.7, it follows that to the order of approximation considered, the statistic, here the LMLE, is a function of such a finite number of differentials. Hence the next step is to transform the Edgeworth expansion of these differentials to an expansion of some smooth function of these. In Bhattacharya and Ghosh (1978) a general theorem of such transformations was given

as Theorem 2. This transformation result was generalized to the non i.i.d. case in Skovgaard (1981). For the case of the LMLE we know from Section 4.3 that for sufficiently large n the LMLE is a unique solution to the likelihood equation

$$0 = D_1^{(n)} + D_2^{(n)}(\hat{\beta}_n - \beta_0) + \frac{1}{2}D_3^{(n)}(\hat{\beta}_n - \beta_0)^2 + \cdots \tag{3.1}$$

within some neighbourhood of β_0. This equation gives $\hat{\beta}_n$ implicitly as a function of the log-likelihood differentials $(D_1^{(n)}, D_2^{(n)}, \ldots)$. A Taylor series expansion around the point given by $\hat{\beta}_n = \beta_0$ and $D_k^{(n)} = \chi_k^{(n)}$ for all k then leads to an expansion of $\hat{\beta}_n - \beta_0$ as a polynomial in the log-likelihood differentials. From this expansion we may then formally calculate approximations to the cumulants of $\hat{\beta}_n - \beta_0$, by the so-called delta method, keeping only terms of the desired order of magnitude. On the basis of these 'formal cumulants' we may then write down an Edgeworth expansion for $\hat{\beta}_n - \beta_0$. To avoid specifying the approximate cumulants, we denote the Mth order Edgeworth expansion for the distribution of $\hat{\beta}_n$ obtained in this way by $\hat{Q}_M$.

The theorem stating the validity of the resulting expansion for the LMLE is stated below. As a, somewhat brief, proof we verify the conditions of Theorem 3 in Bhattacharya and Ghosh (1978), in a slightly modified form. We first list these conditions and then, for each of these, comment on any changes made and verify their validity in our case. It should be noted that Bhattacharya and Ghosh (1978) state their result in the more general framework of minimum contrast estimators, and that their formulation of the 'locality' of the estimator is slightly different from our Definition 4.3.2 of the LMLE. However their result is easily seen to apply to the LMLE as defined here, regardless of the value of δ chosen in the definition of the LMLE. Furthermore, they proved the uniformity of the resulting error with respect to the parameter point β_0 varying in a compact subset. While it would not be difficult to do the same in our setting, by assuming the model to be analytic throughout this compact set, we stick to the case of a single parameter point, partly to avoid notational complications, and partly to demonstrate the result of the assumption that the model is analytic at one particular point.

Theorem 3.1. *Assume that the model* $\{f_1(\beta); \beta \in B \subseteq V\}$ *is analytic at* β_0 *and that the Fisher information* $I(\beta_0)$ *is positive definite. Let* W_M *be the finite-dimensional vector space spanned by the distribution of the statistic*

$$X_{M,i} = \big(D_{1,i}(\beta_0) - \chi_1(\beta_0), \ldots, D_{M,i}(\beta_0) - \chi_M(\beta_0)\big) \tag{3.2}$$

and assume that for any $M \in \mathbf{N}$, *some power of the characteristic function of this distribution on* W_M *is absolutely integrable.*

Then, for any $M \in \mathbf{N}$, *the Edgeworth expansion* $\hat{Q}_M$ *of the distribution of the LMLE* $\hat{\beta}_n(\delta)$ *satisfies*

$$\left|P_{\beta_0}\{\hat{\beta}_n(\delta) \in A\} - \hat{Q}_M(A)\right| = O(n^{-M/2}) \tag{3.3}$$

as $n \to \infty$, uniformly over all sets A within any class of Borel sets satisfying (2.6) with Γ equal to the inverse Fisher information.

Proof. The conditions that we need to verify are the following ones, the numbers referring to the corresponding number in Bhattacharya and Ghosh (1978). The 'missing' condition (A5) is not needed for our case.

(A1) There is a measurable subset $E_1 \subseteq E$ and a neighbourhood $U(\beta_0) \subseteq B$ of β_0, such that $P_\beta(E_1) = 1$ for all $\beta \in U(\beta_0)$, and such that $\log f_1(y;\beta)$ is infinitely often differentiable with respect to β on $E_1 \times U(\beta_0)$.

(A2) For any $k \in \mathbf{N}$ we have

$$\mathrm{E}_{\beta_0}(\|D_{k,i}(\beta_0)\|) < \infty,$$

and for any $k \in \mathbf{N}$ an $\epsilon > 0$ exists, such that

$$\mathrm{E}_{\beta_0}(\sup\{\|D_{k,i}(\beta)\| : \|\beta - \beta_0\| \leq \epsilon\})^{k-1} < \infty.$$

(A3) The mean $\chi_1(\beta_0)$ of $D_{1,i}(\beta_0)$ is zero, and the matrices $-\chi_2(\beta_0)$ and $\chi_{11}(\beta_0)$ are both positive definite.

(A4) The functions $I(\beta)$, $\chi_k(\beta)$, and $\chi_{kk}(\beta)$ are all continuous in a neighbourhood of β_0 which may not depend on k.

(A6) For any M, some power of the characteristic function of the distribution of $X_{M,i} \in W_M$, from the statement of the theorem above, is absolutely integrable.

We now go through this list of conditions, one by one, and comment on the changes compared to the ones from Bhattacharya and Ghosh (1978), in the sequel referred to as BG, and demonstrate that they are satisfied for any model fulfilling the conditions of the theorem above.

The condition (A1) has been restricted to a neighbourhood $U(\beta_0)$ of β_0 instead of any compact set of parameters as in BG. This is entirely due to the restriction of our theorem to the single parameter point β_0, instead of proving the uniformity of the error over any compact subset of β_0-parameters. That we require infinitely many derivatives to exist is because we have stated the validity of (3.3) for any $M \in \mathbf{N}$. Finally, we do not require the set E_1 to be open, because this condition was required only for the use of Lemma 2.2 in BG to demonstrate the existence of an absolutely continuous component of some convolution power of the distribution of $X_{M,i}$. We have instead imposed a continuity condition directly on the distribution of $X_{M,i}$ as the condition (A6). The requirements in (A1) are immediate consequences of the conditions for the model to be analytic at β_0, except for the positivity of the density, required for the differentiability of its logarithm, which is a consequence of Lemma 2.3.4.

In the condition (A2) in BG the suprema of the expectations over any compact set of β_0's were considered, but again, since we are proving the result only at a fixed point β_0 we need only the weaker form stated above. The condition follows trivially from Lemma 2.3.4 together with condition (iv) in Definition 2.2.1 of an analytic model.

The only change in condition (A3) is that we only require the condition to hold at the single point β_0 because we only prove the result for this parameter value. The two matrices are identical in our case and the condition is seen to follow from Lemma 2.3.5, cf. (2.3.14) and (2.3.15), and from the assumption that the Fisher information is positive definite at β_0.

In condition (A4) we have, for brevity, imposed the continuity condition on the first two cumulants of the log-likelihood differentials, instead of on the second moment. For reasons mentioned under the discussion of (A1) we only need it to hold in a neighbourhood of β_0, instead of throughout the parameter space. This condition follows trivially from Lemma 2.3.5.

The condition (A5) in BG states that the mapping $\beta \mapsto P_\beta$ is continuous with respect to the total variation norm on the space of finite signed measures. While it follows from Section 2.7 that this is true at any point where the model is analytic, this condition is not required for our case since we are not claiming the uniformity of the error bound with respect to a set of β_0-points.

The condition (A6) is the one which differs most from the corresponding one in BG. Bhattacharya and Ghosh impose a continuity condition directly on the distribution of Y_i, which they assume takes values in a vector space, and then prove the corresponding continuity condition for some convolution power of the distribution of $X_{M,i}$ by use of their Lemma 2.2. This condition of the existence of an absolutely continuous component of the convolution power of the distribution of $X_{M,i}$ follows trivially from the assumption we have made. Thus, we have essentially picked out this assumption from a later stage of their proof. There is no verification of this condition in our case and we have stated the condition directly as a condition in the theorem. ∎

Notice that it is condition (A6) that rules out the lattice case from the result.

4 Consistency of the maximum likelihood estimator

Although the consistency of the maximum likelihood estimator is an asymptotic property of first order we include this result in the present chapter because it is confined to the framework of independent replications as mentioned in the introduction, and generalizations of the result outside this framework become much more complicated. We base the theorem on the classical result from the paper by Wald (1949), refined and developed in several papers, cf. Kiefer and Wolfowitz (1956), Perlman (1972) and references therein. We only treat the case of consistency within a compact parameter space, first of all because the conditions for a model to be analytic are local conditions which do not suffice to prove the result for a non-compact parameter space. To deal with that case various kinds of global uniformity conditions have been proposed, either in the form of a compactification of the parameter space, cf. Assumption 2 in Kiefer and Wolfowitz (1956), or in the form of some uniform bound over a non-compact space, cf. Wald (1949) and the discussion in Perlman (1972). While such an assumption can certainly be added to the conditions of the theorem below, we take the viewpoint that this is a somewhat separate matter that does not have much to do with the present

(local) theory of analytic models. Thus the point of the present section is to show that analytic models satisfy the local regularity conditions required to prove the consistency of the maximum likelihood estimator.

The theorem below is based mainly on that in Wald (1949) with generalizations as described in Kiefer and Wolfowitz (1956). In the latter paper the result covered the more general setting with an infinite sequence of nuisance parameters, but their conditions are trivially simplified to the present case. In the numbering of the conditions discussed in the proof of the theorem below we refer to the assumptions listed in the paper by Kiefer and Wolfowitz.

Notations and setup from the introduction to this chapter is used in the sequel. The probability measure with density $f_1(y;\beta)$ from (1.1) will be denoted P_β, and the maximum likelihood estimator (MLE) $\hat{\beta}_n(K)$, where $K \subseteq B$, is any function of $Y^{(n)}$ which maximizes the likelihood function

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

on the set K. From now on the notation $\hat{\beta}_n(K)$ is used exclusively to denote the MLE, although it formally is the same as the notation for the local maximum likelihood estimator which is not considered in the present section. Notice that while the maximum may not be unique, its existence is guaranteed in the theorem below by the compactness of K and the requirement that the model is analytic on K and hence that the likelihood function is continuous.

Theorem 4.1. *Assume that the model (1.1) is analytic in the open connected set $B \subseteq V$ and let $K \subseteq B$ be any compact subset of B. Furthermore assume that the identifiability condition*

$$P_{\beta_1} = P_{\beta_2} \quad \Rightarrow \quad \beta_1 = \beta_2 \tag{4.1}$$

holds for $\beta_1, \beta_2 \in B$. Then any maximum likelihood estimator $\hat{\beta}_n$ of the model restricted to the parameter space K is strongly consistent, i.e.,

$$\hat{\beta}_n(K) \to \beta_0 \qquad \text{almost surely } (P_{\beta_0}),$$

for any $\beta_0 \in K$.

Proof. As in the proof of the theorem in the previous section we go through the list of assumptions required to prove the theorem, comment on any changes for our case, and verify that they are satisfied by any model fulfilling the requirements stated in the theorem. Here we take the five assumptions from the paper by Kiefer and Wolfowitz (1956), in the sequel referred to as KW, as the starting point, and the numbers below correspond to their numbers. Occasionally we refer to the similar assumptions from Wald (1949).

(1) As the first assumption in KW, the density is required to be a density with respect to a σ-finite measure on a Euclidean space. As argued in the previous section we do not wish to impose such a restriction on the sample

space, and the importance for the proof is the log likelihood function and its distribution. Since we have the local representation

$$\log f_1(y;\beta) - \log f_1(y;\beta_1) = \sum_{k=1}^{\infty} \frac{1}{k!} D_k(\beta_1)(\beta - \beta_1)^k \tag{4.2}$$

in a neighbourhood of any fixed point $\beta_1 \in K$, measurability of the likelihood function follows from the measurability of the D_k's. In fact, since the differentials of the log likelihood are elements of real vector spaces, the modification of the assumption here is quite trivial and unimportant.

(2) On a set of P_{β_0}-probability one, the density function $f_1(y;\beta)$ is continuous with respect to β in K.

This condition is stronger as stated here than in KW. First of all KW allow (as Wald did) that the set of probability one may depend on the point at which continuity is considered. We have no need of this generality since the set E_1 of probability one from Definition 2.2.1 of an analytic model, may be chosen to be independent of the parameter point β_0 at which the model is analytic. To see this, notice that the set E_1 may be chosen independently of the point β in some neighbourhood of any point $\beta_0 \in B$, cf. Lemma 2.3.4, and hence it follows by a usual compactness argument that the set E_1 may be chosen as a fixed set throughout K. In fact, only σ-compactness is used here, because it is sufficient to show that the 'global' set of probability one may be represented as an intersection of a countable number of sets of probability one. An important simplification of the condition as stated above compared to the one in KW is that we have assumed the parameter space considered to be compact whereas a compactification is considered in KW.

(3) For any $\beta_1 \in K$ and any $\delta > 0$, the function

$$w(y;\beta_1,\delta) = \sup\{\log f_1(y;\beta) : \|\beta - \beta_1\| < \delta\} - \log f_1(y;\beta_0) \tag{4.3}$$

is measurable as a function of $y \in E_1$.

This measurability follows from the representation (4.2) and the continuity of the function of which the supremum is taken, cf. the remark following Condition 8 in Wald (1949). Compared to KW we have, for convenience, used the log-density, instead of the density, and subtracted the log-density at the point β_0 in the definition of the function w.

(4) For $\beta_1, \beta_2 \in B$, the identifiability condition

$$P_{\beta_1} = P_{\beta_2} \quad \Rightarrow \quad \beta_1 = \beta_2$$

holds.

This is the same identifiability condition as stated in KW, although the formulation is slightly different. The condition is stated directly in the

theorem above. It must be imposed as a separate condition, as it does not follow from the fact that the model is analytic.

(5) For any $\beta_1 \in K$ we have

$$\lim_{\delta \to 0} \mathrm{E}\{w(Y; \beta_1, \delta\} < \infty.$$

Due to the representation (4.2), or more directly because the model is analytic, infinite values of the log-densities or of the expectations of differences between log-densities, as in (4.2), do not occur in our case. Therefore we have not stated the condition in its weaker form as in KW, where only the positive part of the function w is included. From the representation (4.2) and the condition (iv) from Definition 2.2.1 we see that the inequality

$$\begin{aligned} &|\sup\{\log f_1(y;\beta) : \|\beta - \beta_1\| < \delta\} - \log f_1(y;\beta_1)| \\ &\quad \leq M(y;\beta_1) \sum_{k=1}^{\infty} \rho(\beta_1)^{k-1} \delta^k \end{aligned}$$

holds on a set of probability one, where $M(y;\beta_1)$ and $\rho(\beta_1)$ are the functions from Definition 2.2.1. For δ sufficiently small this bound is seen to have finite expectation if we can prove that $M(Y;\beta_1)$ may be chosen such that it has finite expectation with respect to P_{β_0} for any $\beta_0 \in K$. Before we go on to the proof of this claim, notice that as a general result we know that also the function

$$\log f_1(Y;\beta_1) - \log f_1(Y;\beta_0)$$

has finite P_{β_0}-expectation, and consequently it follows that the condition (5) holds. In fact, this latter function has negative expectation because of the identifiability condition, cf., e.g., the proof in KW.

Let us turn now to the proof that $E_{\beta_0} M(Y;\beta_1)$ is finite. In fact, we show that $M(Y;\beta_1)$ may be chosen such that it has finite exponential moments with respect to any P_{β_0}, $\beta_0 \in B$. Consider the mixed cumulant condition (v) from Theorem 2.4.2, and let $c(\beta)$ and $\lambda(\beta)$, $\beta \in B$, denote functions satisfying this condition. As the first step we want to prove that it is possible to choose these two functions such that they are both bounded on K.

Let $c(\beta)$ and $\lambda(\beta)$ be given as any functions satisfying the condition (v) in Theorem 2.4.2. From Lemma 2.5.5 it is then seen that for any $\beta_1 \in K$, and any $\beta \in B$ satisfying

$$\|\beta - \beta_1\| \leq \{4\lambda(\beta_1)\}^{-1},$$

the condition (v) holds at β with $c(\beta) \leq 16c(\beta_1)$ and $\lambda(\beta) \leq 4\lambda(\beta_1)$, because

$$\{1 - 2\lambda(\beta_1)\|\beta - \beta_1\|\}^{-1} \leq 2$$

and the same bound applies to the left hand side if the factor 2 is omitted. For any $\beta_1 \in K$ consider the open ball centered at β_1 and with radius

$\{4\lambda(\beta_1)\}^{-1}$. Since the compact set K may be covered by a finite number of these balls, the argument above shows that $\lambda(\beta)$ and $c(\beta)$ may be chosen to be bounded on K, still satisfying the mixed cumulant condition (v) in Theorem 2.4.2. In the sequel we assume these two functions to be chosen in this way.

The next step is to notice, by use of Lemma 2.4.1 and Theorem 2.4.2, that the condition (iv) from Definition 2.2.1 is satisfied for

$$\rho(\beta) = 1 + (2e\sqrt{p})\lambda(\beta)$$

and

$$M(y;\beta) = c(\beta)H(y;\beta) + c(\beta)^2, \tag{4.4}$$

where 1 has been added to the expression for $\rho(\beta)$ from Lemma 2.4.1 to avoid problems related to the case $\lambda(\beta) = 0$. Here $H(y;\beta)$ is known from Lemma 2.4.1 to satisfy the moment condition

$$\mathrm{E}_\beta \exp\{sc(\beta)H(Y;\beta)\} < \infty$$

for any $s < \rho(\beta)^{-1}$. Hence, with $M(y;\beta)$ chosen as in (4.4), it follows that for any $\beta_1 \in K$ and any $\beta \in B$ satisfying

$$\|\beta - \beta_1\| \le a < \{3\rho(\beta_1)\}^{-1},$$

we have

$$\begin{aligned}
\mathrm{E}_{\beta_1}\left(\{f(Y;\beta)/f(Y;\beta_1)\}^2\right) &= \mathrm{E}_{\beta_1} \exp\left\{2\sum_{k=1}^{\infty}\frac{1}{k!}D_k(\beta_1)(\beta-\beta_1)^k\right\} \\
&\le \mathrm{E}_{\beta_1} \exp\left\{2M(Y;\beta_1)\sum_{k=1}^{\infty}a^k\rho(\beta_1)^{k-1}\right\} \\
&= \mathrm{E}_{\beta_1} \exp\left\{M(Y;\beta_1)\,2a/\left(1-a\rho(\beta_1)\right)\right\} \\
&< \infty.
\end{aligned}$$

Now, let $\beta_1 \in K$ be an arbitrary fixed point and consider the set of β_0's for which $M(Y;\beta_1)$, defined in (4.4), has finite exponential moments, i.e., the set

$$A = \{\beta_0 \in K : \mathrm{E}_{\beta_0} \exp\{sM(Y;\beta_1)\} < \infty \text{ for some } s > 0\}. \tag{4.5}$$

Since $\beta_1 \in A$, this set is not empty. Consider a point $\beta_0 \in A$ and let $s_0 > 0$ be such that

$$\mathrm{E}_{\beta_0} \exp\{s_0 M(Y;\beta_1)\} < \infty.$$

Notice that for any $\beta \in B$ satisfying $\|\beta - \beta_0\| < \{3\rho(\beta_0)\}^{-1}$, Cauchy-Schwartz' inequality gives

$$\begin{aligned}
&\mathrm{E}_\beta \exp\left\{\frac{1}{2}s_0 M(Y;\beta_1)\right\} \\
&\quad = \int \exp\left\{\frac{1}{2}s_0 M(y;\beta_1)\right\}\{f(y;\beta)/f(y;\beta_0)\}\, dP_{\beta_0}(y) \\
&\quad \leq \mathrm{E}_{\beta_0}\{s_0 M(Y;\beta_1)\}\,\mathrm{E}_{\beta_0}\left(\{f(Y;\beta)/f(Y;\beta_0)\}^2\right) \\
&\quad < \infty.
\end{aligned}$$

Hence it follows that for any $\beta_0 \in A$, the set of β's within a distance $\{3\rho(\beta_0)\}^{-1}$ from β_0 is also contained in A. Since this distance is bounded away from zero, it follows by standard arguments that $A = K$. In fact, since B is σ-compact, the result implies that $M(Y;\beta_1)$ from Definition 2.2.1 may be chosen to have finite exponential moments with respect to any P_{β_0}, $\beta_0 \in B$.

This completes the proof of the fifth condition from KW.

Thus, it follows that all the assumptions from the paper by Kiefer and Wolfowitz are fulfilled, and hence that the result holds as claimed in the theorem. ∎

That the strong consistency of the estimator is the result in the theorem should not be taken to imply that this is more relevant than the weak consistency, i.e., the convergence in probability of the estimator towards the fixed point β_0. In fact, we take the point of view that the weak consistency is more likely to be of statistical relevance, but it follows from the strong consistency which happens to be a consequence of the assumptions.