Statistical Inference of Moment Structures

Alexander Shapiro

School of Industrial and Systems Engineering,
Georgia Institute of Technology,
Atlanta, Georgia 30332-0205, USA
e-mail: ashapiro@isye.gatech.edu

Abstract

This chapter is focused on statistical inference of moment structures models. Although the theory is presented in terms of general moment structures, the main emphasis is on the analysis of covariance structures. We discuss identifiability and the minimum discrepancy function (MDF) approach to statistical analysis (estimation) of such models. We address such topics of the large samples theory as consistency, asymptotic normality of MDF estimators and asymptotic chi-squaredness of MDF test statistics. Finally, we discuss asymptotic robustness of the normal theory based MDF statistical inference in the analysis of covariance structures.

Contents

1 Introduction

In this paper we discuss statistical inference of moment structures where first and/or second population moments are hypothesized to have a parametric structure. Classical examples of such models are multinomial and covariance structures models. The presented theory is sufficiently general to handle various situations, however the main focus is on covariance structures. Theory and applications of covariance structures were motivated first by the factor analysis model and its various generalizations and later by a development of the LISREL models (Jöreskog [15, 16]) (see also Browne [8] for a thorough discussion of covariance structures modeling).

2 Moment Structures Models

In this section we discuss modeling issues in the analysis of moment structures, and in particular identifiability of such models. Let $\boldsymbol{\xi} = (\xi_1, ..., \xi_m)'$ be a vector variable representing a parameter vector of some statistical population. For example, in the analysis of covariance structures,

$\boldsymbol{\xi}$ will represent the elements of a $p \times p$ covariance matrix $\boldsymbol{\Sigma}$. That is[1], $\boldsymbol{\xi} := \text{vec}(\boldsymbol{\Sigma})$, where $\text{vec}(\boldsymbol{\Sigma})$ denotes the $p^2 \times 1$ vector formed by stacking elements of $\boldsymbol{\Sigma}$. Of course, since matrix $\boldsymbol{\Sigma}$ is symmetric, vector $\text{vec}(\boldsymbol{\Sigma})$ has duplicated elements. Therefore an alternative is to consider $\boldsymbol{\xi} := \text{vecs}(\boldsymbol{\Sigma})$, where $\text{vecs}(\boldsymbol{\Sigma})$ denotes the $p(p+1)/2 \times 1$ vector formed by stacking (nonduplicated) elements of $\boldsymbol{\Sigma}$ above and including the diagonal. Note that covariance matrices $\boldsymbol{\Sigma}$ are positive semidefinite, and hence the corresponding vectors $\boldsymbol{\xi}$ are restricted to a (convex) subset of $\mathbb{R}^m$. Therefore, we assume that $\boldsymbol{\xi}$ varies in a set $\Xi \subset \mathbb{R}^m$ representing a *saturated* model for the population vector $\boldsymbol{\xi}$. In the analysis of covariance structures we have a natural question whether to use vector $\boldsymbol{\xi} = \text{vec}(\boldsymbol{\Sigma})$ or $\boldsymbol{\xi} = \text{vecs}(\boldsymbol{\Sigma})$ for the saturated model. Since in both cases the dimension of the corresponding set Ξ is $p(p+1)/2$, it seems more advantageous to use $\boldsymbol{\xi} := \text{vecs}(\boldsymbol{\Sigma})$. In that case the set Ξ has a nonempty interior[2]. However, for actual calculations it is often more convenient to use $\boldsymbol{\xi} := \text{vec}(\boldsymbol{\Sigma})$. When dealing with specific applications we will specify a choice of the corresponding vector $\boldsymbol{\xi}$.

A model for $\boldsymbol{\xi}$ is a subset Ξ_0 of Ξ. Of course, this definition is too abstract and one needs a constructive way of defining a model. There are two natural ways for constructing a model, namely either by imposing equations or by a parameterization. The parameterization approach suggests existence of an $m \times 1$ vector valued function $\boldsymbol{g}(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), ..., g_m(\boldsymbol{\theta}))$, and a parameter set $\Theta \subset \mathbb{R}^q$, which relates the parameter vector $\boldsymbol{\theta} = (\theta_1, ..., \theta_q)'$ to $\boldsymbol{\xi}$. That is,

$$\Xi_0 := \left\{ \boldsymbol{\xi} \in \Xi : \boldsymbol{\xi} = \boldsymbol{g}(\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta \right\}. \tag{2.1}$$

We refer to $\boldsymbol{g}(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, as a *structural* model for $\boldsymbol{\xi}$. We assume in the subsequent analysis that the mapping $\boldsymbol{g}(\boldsymbol{\theta})$ is sufficiently smooth. In particular, we always assume that $\boldsymbol{g}(\boldsymbol{\theta})$ is *twice continuously differentiable*. We associate with mapping $\boldsymbol{g}(\cdot)$ its $m \times q$ Jacobian matrix $\partial \boldsymbol{g}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}' = [\partial g_i(\boldsymbol{\theta})/\partial \theta_j]_{i=1,...,m, j=1,...,q}$ of partial derivatives and use notation $\boldsymbol{\Delta}(\boldsymbol{\theta}) := \partial \boldsymbol{g}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}'$.

Remark 1 It should be noted that the same set Ξ_0 could be represented by different parameterizations in the form (2.1). For example, let Ξ be the set of all $p \times p$ symmetric positive semidefinite matrices (covariance matrices) and Ξ_0 be its subset of diagonal matrices with nonnegative diagonal elements. This model can be parameterized by the set[3] $\Theta := \mathbb{R}^p_+$ and mapping $\boldsymbol{g}(\boldsymbol{\theta}) := \text{diag}(\theta_1, ..., \theta_p)$. Alternatively, it can be parameterized by $\Theta := \mathbb{R}^p$ and $\boldsymbol{g}(\boldsymbol{\theta}) := \text{diag}(\theta_1^2, ..., \theta_p^2)$. Of course, in applications the considered parameters typically have an interpretation. For instance, in the above example of diagonal covariance matrices, in the first parameterization parameters θ_i represent the corresponding variances while in the second parameterization these are the corresponding standard deviations. Note that this set Ξ_0 can be also defined by equations by setting the off-diagonal elements of $\boldsymbol{\Sigma}$ to zero. In the subsequent analysis we mainly deal with the parameterization approach.

[1] The notation " $:=$ " means "equal by definition".

[2] The interior of Ξ is the set of points $\boldsymbol{\xi} \in \Xi$ such that Ξ contains a neighborhood of $\boldsymbol{\xi}$. For example, the interior of the set of positive semidefinite matrices is formed by its subset of positive definite matrices. A singular positive semidefinite matrix can be viewed as a boundary point of this set Ξ. A neighborhood of a point $\boldsymbol{\xi} \in \mathbb{R}^m$ is a subset of $\mathbb{R}^m$ containing a ball centered at $\boldsymbol{\xi}$ of a sufficiently small positive radius.

[3] The set $\mathbb{R}^p_+$ denotes the nonnegative orthant of the space $\mathbb{R}^p$, i.e., $\mathbb{R}^p_+ := \{\boldsymbol{\theta} \in \mathbb{R}^p : \theta_i \geq 0,\ i = 1, ...p\}$.

It is said that model $\boldsymbol{g}(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, is (globally) *identified* at a point $\boldsymbol{\theta}_0 \in \Theta$ if $\boldsymbol{\theta}_0$ is a unique parameter vector corresponding to the value $\boldsymbol{\xi}_0 := \boldsymbol{g}(\boldsymbol{\theta}_0)$ of the population vector. It is said that the model is locally identified at $\boldsymbol{\theta}_0$ if such uniqueness holds in a neighborhood of $\boldsymbol{\theta}_0$. More formally we have the following definition.

Definition 2.1 *It is said that structural model $\boldsymbol{g}(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, is* identified (locally identified) *at a point $\boldsymbol{\theta}_0 \in \Theta$ if $\boldsymbol{g}(\boldsymbol{\theta}^*) = \boldsymbol{g}(\boldsymbol{\theta}_0)$ and $\boldsymbol{\theta}^* \in \Theta$ ($\boldsymbol{\theta}^*$ in a neighborhood of $\boldsymbol{\theta}_0$) implies that $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$.*

Of course, (global) identifiability implies local identifiability. A well known sufficient condition for local identifiability of $\boldsymbol{\theta}_0 \in \Theta$ is that the Jacobian matrix $\boldsymbol{\Delta}(\boldsymbol{\theta}_0)$, of $\boldsymbol{g}(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_0$, has full column rank q (e.g., [14]). In general, this condition is not necessary for local identifiability of $\boldsymbol{\theta}_0$ even if $\boldsymbol{\theta}_0$ is an interior point of Θ. Take, for example, $g(\theta) := \theta^3$, $\theta \in \mathbb{R}$. This model is locally (and globally) identified at $\theta = 0$, while $\partial g(0)/\partial\theta = 0$. This condition becomes necessary and sufficient under the following assumption of constant rank regularity which was used by several authors (e.g., [14, 23, 39]).

Definition 2.2 *We say that a point $\boldsymbol{\theta}_0 \in \Theta$ is* locally regular *if the Jacobian matrix $\boldsymbol{\Delta}(\boldsymbol{\theta})$ has the same rank as $\boldsymbol{\Delta}(\boldsymbol{\theta}_0)$ for every $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$.*

If the mapping $\boldsymbol{g}(\boldsymbol{\theta})$ is independent of, say, last s parameters $\theta_{q-s}, ..., \theta_q$, then, of course, these parameters are redundant and the model can be viewed as overparameterized. In that case the rank of the Jacobian matrix $\boldsymbol{\Delta}(\boldsymbol{\theta})$ is less than or equal to $q - s$ for any $\boldsymbol{\theta}$. In general, it is natural to view the structural model as being (locally) overparameterized, at a point $\boldsymbol{\theta}_0$ in the interior of Θ, if it can be reduced to the above case by a local transformation (reparameterization). More formally we have the following definition (cf., [31]).

Definition 2.3 *We say that structural model $\boldsymbol{g}(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, is* locally overparameterized, *at an interior point $\boldsymbol{\theta}_0$ of Θ, if the rank r of $\boldsymbol{\Delta}(\boldsymbol{\theta}_0)$ is less than q and there exists a local diffeomorphism[4] $\boldsymbol{\theta} = \boldsymbol{h}(\boldsymbol{\gamma})$ such that the composite mapping $\boldsymbol{g}(\boldsymbol{h}(\boldsymbol{\gamma}))$ is independent of, say last, $q - r$ coordinates of $\boldsymbol{\gamma}$.*

The local diffeomorphism $\boldsymbol{h}(\boldsymbol{\gamma})$, in the above definition, can be viewed as a local reparameterization of the model. We don't need to construct such a reparameterization explicitly but rather to know about its existence since it gives us an information about a (local) structure of the model. Clearly, if the model is locally overparameterized at a point $\boldsymbol{\theta}_0$, then it is not locally identified at this point. Moreover, in the case of local overparameterization the set of points $\boldsymbol{\theta}$

[4]Mapping $\boldsymbol{h}(\boldsymbol{\gamma})$ from a neighborhood of $\boldsymbol{\gamma}_0 \in \mathbb{R}^q$ to a neighborhood of $\boldsymbol{\theta}_0$, with $\boldsymbol{h}(\boldsymbol{\gamma}_0) = \boldsymbol{\theta}_0$, is called local diffeomorphism if it is continuously differentiable, locally one-to-one and its inverse is also continuously differentiable. It can be shown that $\boldsymbol{h}(\boldsymbol{\gamma})$ is a local diffeomorphism if and only if it is continuously differentiable and the Jacobian matrix $\partial\boldsymbol{h}(\boldsymbol{\gamma}_0)/\partial\boldsymbol{\gamma}$ is nonsingular. We can assume without loss of generality that $\boldsymbol{\gamma}_0 = \mathbf{0}$.

such that $\boldsymbol{g}(\boldsymbol{\theta}) = \boldsymbol{g}(\boldsymbol{\theta}_0)$ forms a smooth manifold in a neighborhood of $\boldsymbol{\theta}_0$. Note that the rank of the Jacobian matrix of the composite mapping $\boldsymbol{g}(\boldsymbol{h}(\boldsymbol{\gamma}))$, at $\boldsymbol{\gamma}_0 = \mathbf{0}$, is the same as the rank of $\boldsymbol{\Delta}(\boldsymbol{\theta}_0)$. Therefore, in the reparameterized model the remaining r coordinates of $\boldsymbol{\gamma}$ are locally identified.

A relation between the concepts of local regularity and local overparameterization is clarified by the following result known as the Rank Theorem (cf., [14]).

Proposition 2.1 *Let $\boldsymbol{\theta}_0$ be an interior point of the parameter set Θ. Then the following holds.* (i) *Suppose that $\boldsymbol{\theta}_0$ is locally regular. Then the model is locally identified at $\boldsymbol{\theta}_0$ if and only if the rank r of $\boldsymbol{\Delta}(\boldsymbol{\theta}_0)$ is equal to q, otherwise if $r < q$, then the model is locally overparameterized at $\boldsymbol{\theta}_0$.* (ii) *Conversely, if the model is locally overparameterized at $\boldsymbol{\theta}_0$, then $r < q$ and the point $\boldsymbol{\theta}_0$ is locally regular.*

The above results are not very useful for verification of (local) identifiability at an individual point $\boldsymbol{\theta}_0$. For one thing the population value of the parameter vector usually is unknown, and even if the value $\boldsymbol{\theta}_0$ is specified, it is not possible to calculate the rank of the corresponding Jacobian matrix numerically because of the round off errors. However, one can approach the identifiability problem from a generic point of view. Suppose that the mapping $\boldsymbol{g}(\cdot)$ is analytic, i.e., every coordinate function $g_i(\cdot)$ can be expended into a power series in a neighborhood of every point $\boldsymbol{\theta} \in \Theta$. Suppose also that the set Θ is connected. Let ι be an index set of rows and columns of $\boldsymbol{\Delta}(\boldsymbol{\theta})$ defining its squared submatrix and $v_\iota(\boldsymbol{\theta})$ be the determinant of that submatrix. Clearly there is only a finite number of such submatrices and hence the corresponding determinant functions. Since $\boldsymbol{g}(\cdot)$ is analytic, every such determinant function $v_\iota(\boldsymbol{\theta})$ is also analytic. Consequently, either $v_\iota(\boldsymbol{\theta})$ is identically zero for all $\boldsymbol{\theta} \in \Theta$, or $v_\iota(\boldsymbol{\theta}) \neq 0$ for almost every[5] $\boldsymbol{\theta} \in \Theta$. These arguments lead to the following result.

Proposition 2.2 *Suppose that the mapping $\boldsymbol{g}(\cdot)$ is analytic and the set Θ is connected. Then almost every point of Θ is locally regular with the same rank r of the Jacobian matrix $\boldsymbol{\Delta}(\boldsymbol{\theta})$, and, moreover, the rank of $\boldsymbol{\Delta}(\boldsymbol{\theta})$ is less than or equal to r for all $\boldsymbol{\theta} \in \Theta$.*

By the above proposition we have that with the model, defined by analytic mapping $\boldsymbol{g}(\boldsymbol{\theta})$, is associated an integer r equal to the rank of of $\boldsymbol{\Delta}(\boldsymbol{\theta})$ almost everywhere. We refer to this number r as the **characteristic rank** of the model, and whenever talking about the characteristic rank we assume that the mapping $\boldsymbol{g}(\boldsymbol{\theta})$ is analytic and the parameter set Θ is connected (the concept of characteristic rank was introduced in Shapiro [31]). It follows from the preceding discussion that either $r = q$ in which case the model is locally identified almost everywhere, or $r < q$ in which case the model is locally overparameterized almost everywhere.

[5]The "almost every" statement here can be understood in the sense that it holds for all $\boldsymbol{\theta}$ in Θ except on a subset of Θ of Lebesgue measure zero.

We say that the model is identified (locally identified) in the **generic sense** if it is identified (locally identified) at almost every $\boldsymbol{\theta} \in \Theta$ (cf., Shapiro [27]). By the above analysis we have that the model is locally identified in the generic sense if and only if its characteristic rank is equal q. Note that the characteristic rank is always less than or equal to the dimension m of the saturated model.

In situations where the model is (locally) overparameterized, the usual practice is to restrict the parameter space by imposing constraints. According to definition 2.3, if the model is locally overparameterized, at a point $\boldsymbol{\theta}_0$, then it can be reparameterized such that the reparameterized model locally does not depend on the last $q-r$ coordinates $\gamma_{q-r+1}, ..., \gamma_q$. Consequently by imposing the constraints $\gamma_i = 0$, $i = q-r+1, ..., q$, the reparameterized model becomes locally identified at $\boldsymbol{\gamma}_0 = \mathbf{0}$ while its image space Ξ_0 is not changed. For the original model this is equivalent to imposing (locally) the identifiability constraints $c_i(\boldsymbol{\theta}) = 0$, $i = q-r+1, ..., q$, where $\boldsymbol{\gamma} = \boldsymbol{c}(\boldsymbol{\theta})$ is the inverse of the mapping $\boldsymbol{\theta} = \boldsymbol{h}(\boldsymbol{\gamma})$.

Example 2.1 Consider the factor analysis model:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}, \tag{2.2}$$

which relates the $p \times p$ covariance $\boldsymbol{\Sigma}$ to the $p \times k$ matrix $\boldsymbol{\Lambda}$ of factor loadings and the $p \times p$ diagonal matrix $\boldsymbol{\Psi}$ of the residual variances. The corresponding parameter vector $\boldsymbol{\theta}$ is composed here from the elements of matrix $\boldsymbol{\Lambda}$ and diagonal elements of $\boldsymbol{\Psi}$, and hence has dimension $q = pk + p$. Note that the diagonal elements of the matrix $\boldsymbol{\Psi}$ should be nonnegative while there are no restrictions on the elements of $\boldsymbol{\Lambda}$.

By substituting $\boldsymbol{\Lambda}$ with $\boldsymbol{\Lambda T}$, where $\boldsymbol{T}$ is an arbitrary $k \times k$ orthogonal matrix, we end up with the same matrix $\boldsymbol{\Sigma}$ (this is the so-called indeterminacy of the factor analysis model). Since the dimension of the (smooth) manifold of $k \times k$ orthogonal matrices is $k(k-1)/2$ and the dimension of the space of $p \times p$ symmetric matrices is $p(p+1)/2$, it is possible to show that the characteristic rank r of the factor analysis model (2.2) is

$$r = \min\left\{pk + p - k(k-1)/2, p(p+1)/2\right\}. \tag{2.3}$$

It follows that for $k > 1$ the model (2.2) is locally overparameterized. A way of dealing with this is to reduce the number of parameters given by matrix $\boldsymbol{\Lambda}$ by setting $k(k-1)/2$ elements (say of the upper triangular part) of $\boldsymbol{\Lambda}$ to zero. Then the question of global (local) identifiability of the factor analysis model is reduced to the global (local) identifiability of the diagonal matrix $\boldsymbol{\Psi}$ (cf., [2]). We have that a necessary condition for *generic* local identifiability of $\boldsymbol{\Psi}$ is that $pk+p-k(k-1)/2$ is less than or equal to $p(p+1)/2$, which is equivalent to $(p-k)(p-k+1)/2 \geq p$, and in turn is equivalent to to $k \leq \phi(p)$, where

$$\phi(p) := \frac{2p + 1 - \sqrt{8p+1}}{2}. \tag{2.4}$$

The above function $\phi(p)$ corresponds to the so-called Ledermann bound, [19]. In the present case we have that $k \leq \phi(p)$ is a necessary and sufficient condition for local identifiability of the diagonal matrix $\boldsymbol{\Psi}$, of the factor analysis model, in the *generic* sense (cf., [27]).

It is more difficult to establish (global) identifiability of a considered model. We can also approach the (global) identifiability problem from the generic point of view. Of course, if a model is not locally identified it cannot be globally identified. Therefore, $k \leq \phi(p)$ is a necessary condition for (global) identifiability of the factor analysis model in the generic sense. It is known that matrix $\boldsymbol{\Psi}$, in the factor analysis model (2.2), is globally identified in the generic sense if and only if $k < \phi(p)$ (Bekker and ten Berge [5]).

3 Discrepancy Functions Estimation Approach

Let $\boldsymbol{\xi}_0 \in \Xi$ be a population value of the parameter vector of the saturated model. Recall that we refer to a subset Ξ_0 of Ξ as a model for $\boldsymbol{\xi}$. Unless stated otherwise it will be assumed that the model is structural, i.e., the set Ξ_0 is given in the parametric form (2.1). It is said that the model holds if $\boldsymbol{\xi}_0 \in \Xi_0$. Clearly this means that there exists $\boldsymbol{\theta}_0 \in \Theta$ such that $\boldsymbol{\xi}_0 = \boldsymbol{g}(\boldsymbol{\theta}_0)$. If the model is identified at $\boldsymbol{\theta}_0$, then this vector $\boldsymbol{\theta}_0$ is defined uniquely. In that case we refer to $\boldsymbol{\theta}_0$ as the *population* value of the parameter vector $\boldsymbol{\theta}$.

Suppose that we are given an estimator $\hat{\boldsymbol{\xi}}$ of $\boldsymbol{\xi}_0$, based on a sample of size n. We will be interested then in testing the hypothesis $H_0 : \boldsymbol{\xi}_0 \in \Xi_0$, and consequently in estimation of the population value of the parameter $\boldsymbol{\theta}$. Consider the setting of the covariance structures with $\boldsymbol{\Sigma}$ being the covariance matrix of $p \times 1$ random vector $\boldsymbol{X}$, and let $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ be an associated structural model. Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ be an iid (independent identically distributed) random sample drawn from a considered population. Then the standard estimator of the population value $\boldsymbol{\Sigma}_0$ of the covariance matrix is the sample covariance matrix

$$\boldsymbol{S} := \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})', \tag{3.1}$$

where $\bar{\boldsymbol{X}} := n^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i$. Suppose, further, that the population distribution is (multivariate) normal with mean vector $\boldsymbol{\mu}_0$ and covariance matrix $\boldsymbol{\Sigma}_0$. Then the corresponding log-likelihood function (up to a constant independent of the parameters) is[6]

$$\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\tfrac{1}{2} n \ln |\boldsymbol{\Sigma}| - \tfrac{1}{2} \mathrm{tr} \left(\boldsymbol{\Sigma}^{-1} \sum_{i=1}^{n} (\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})' \right). \tag{3.2}$$

The maximum likelihood (ML) estimator of $\boldsymbol{\mu}_0$ is $\bar{\boldsymbol{X}}$ and the ML estimator of $\boldsymbol{\Sigma}_0$, for the saturated model, is

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})' = \frac{n-1}{n} \boldsymbol{S}. \tag{3.3}$$

Of course, for reasonably large values of n, the ML estimator $\hat{\boldsymbol{\Sigma}}$ is "almost" equal to the unbiased estimator $\boldsymbol{S}$. Therefore, with some abuse of the notation, we use $\boldsymbol{S} = \hat{\boldsymbol{\Sigma}}$ as the estimator of the population covariance matrix.

[6] By $|\boldsymbol{A}|$ and $\mathrm{tr}(\boldsymbol{A})$ we denote the determinant and the trace, respectively, of a (squared) matrix $\boldsymbol{A}$.

It follows that two times log-likelihood ratio statistic for testing the null hypothesis $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$ is given by $n\hat{F}$, where

$$\hat{F} := \min_{\boldsymbol{\theta}\in\Theta} F_{ML}\left(\boldsymbol{S}, \boldsymbol{\Sigma}(\boldsymbol{\theta})\right), \tag{3.4}$$

with $F_{ML}(\cdot,\cdot)$ being a function of two (matrix valued) variables defined by

$$F_{ML}\left(\boldsymbol{S}, \boldsymbol{\Sigma}\right) := \ln|\boldsymbol{\Sigma}| - \ln|\boldsymbol{S}| + \operatorname{tr}(\boldsymbol{S}\boldsymbol{\Sigma}^{-1}) - p. \tag{3.5}$$

The corresponding ML estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$ is a minimizer of $F_{ML}\left(\boldsymbol{S}, \boldsymbol{\Sigma}(\boldsymbol{\theta})\right)$ over $\boldsymbol{\theta} \in \Theta$, i.e.[7],

$$\hat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta}\in\Theta} F_{ML}\left(\boldsymbol{S}, \boldsymbol{\Sigma}(\boldsymbol{\theta})\right). \tag{3.6}$$

Let us note at this point that the estimator $\hat{\boldsymbol{\theta}}$, defined in (3.6), can be calculated whether the model holds or not. We will discuss implications of this later.

Following Browne [8] we say that a function $F(\boldsymbol{x}, \boldsymbol{\xi})$, of two vector variables $\boldsymbol{x}, \boldsymbol{\xi} \in \Xi$, is a **discrepancy** function if it satisfies the following conditions:

(i) $F\left(\boldsymbol{x}, \boldsymbol{\xi}\right) \geq 0$ for all $\boldsymbol{x}, \boldsymbol{\xi} \in \Xi$,

(ii) $F\left(\boldsymbol{x}, \boldsymbol{\xi}\right) = 0$ if and only if $\boldsymbol{x} = \boldsymbol{\xi}$,

(iii) $F\left(\boldsymbol{x}, \boldsymbol{\xi}\right)$ is twice continuously differentiable jointly in $\boldsymbol{x}$ and $\boldsymbol{\xi}$.

Let $\boldsymbol{g}(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, be a considered structural model. Given an estimator $\hat{\boldsymbol{\xi}}$ of $\boldsymbol{\xi}_0$ and a discrepancy function $F\left(\boldsymbol{x}, \boldsymbol{\xi}\right)$, we refer to the statistic $n\hat{F}$, where

$$\hat{F} := \min_{\boldsymbol{\theta}\in\Theta} F\big(\hat{\boldsymbol{\xi}}, \boldsymbol{g}(\boldsymbol{\theta})\big), \tag{3.7}$$

as the minimum discrepancy function (MDF) test statistic, and to

$$\hat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta}\in\Theta} F\big(\hat{\boldsymbol{\xi}}, \boldsymbol{g}(\boldsymbol{\theta})\big) \tag{3.8}$$

as the MDF estimator.

The function $F_{ML}\left(\boldsymbol{S}, \boldsymbol{\Sigma}\right)$ defined in (3.5), considered as a function of $\boldsymbol{s} = \operatorname{vec}(\boldsymbol{S})$ and $\boldsymbol{\sigma} = \operatorname{vec}(\boldsymbol{\Sigma})$, is an example of a discrepancy function. It is referred to as the maximum likelihood (ML) discrepancy function. Another popular choice of the discrepancy function in the analysis of covariance structures is

$$F_{GLS}\left(\boldsymbol{S}, \boldsymbol{\Sigma}\right) := \tfrac{1}{2}\operatorname{tr}\left[(\boldsymbol{S} - \boldsymbol{\Sigma})\boldsymbol{S}^{-1}(\boldsymbol{S} - \boldsymbol{\Sigma})\boldsymbol{S}^{-1}\right]. \tag{3.9}$$

We refer to a function of the form

$$F\left(\boldsymbol{x}, \boldsymbol{\xi}\right) := (\boldsymbol{x} - \boldsymbol{\xi})'[\boldsymbol{V}(\boldsymbol{x})](\boldsymbol{x} - \boldsymbol{\xi}), \quad \boldsymbol{x}, \boldsymbol{\xi} \in \Xi, \tag{3.10}$$

[7] By $\arg\min\{f(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ we denote the set of all minimizers of $f(\boldsymbol{\theta})$ over $\boldsymbol{\theta} \in \Theta$. This set can be empty or contain more than one element.

as a generalized least squares (GLS) discrepancy function. Here $\boldsymbol{V}(\boldsymbol{x})$ is an $m \times m$ symmetric matrix valued function of $\boldsymbol{x} \in \Xi$. We assume that for any $\boldsymbol{x} \in \Xi$, the corresponding matrix $\boldsymbol{V}(\boldsymbol{x})$ is positive definite, and hence conditions (i) and (ii) hold, and that $\boldsymbol{V}(\boldsymbol{x})$ is twice continuously differentiable, and hence condition (iii) is satisfied. The $F_{GLS}(\boldsymbol{S}, \boldsymbol{\Sigma})$ function, defined in (3.9), is a particular example of GLS discrepancy functions with the weight matrix[8] $\boldsymbol{V}(\boldsymbol{s}) = \frac{1}{2}\boldsymbol{S}^{-1} \otimes \boldsymbol{S}^{-1}$.

We have the following basic result about structure of discrepancy functions (Shapiro [28]).

Proposition 3.1 *Let $F(\boldsymbol{x}, \boldsymbol{\xi})$ be a discrepancy function satisfying conditions* (i)–(iii). *Then there exists a continuous $m \times m$ symmetric matrix valued function $\boldsymbol{V}(\boldsymbol{x}, \boldsymbol{\xi})$ such that*

$$F(\boldsymbol{x}, \boldsymbol{\xi}) = (\boldsymbol{x} - \boldsymbol{\xi})'[\boldsymbol{V}(\boldsymbol{x}, \boldsymbol{\xi})](\boldsymbol{x} - \boldsymbol{\xi}) \tag{3.11}$$

for all $\boldsymbol{x}, \boldsymbol{\xi} \in \Xi$.

The above result shows that any discrepancy function can be represented in a form of an "almost" GLS function. A difference between the representation (3.11) and the general form (3.10) of GLS discrepancy functions is that the weight matrix in (3.11) can also depend on $\boldsymbol{\xi}$ as well as on $\boldsymbol{x}$.

Let $\boldsymbol{\xi}_0 \in \Xi$ be a given (say the population) value of vector $\boldsymbol{\xi}$. Consider matrix $\boldsymbol{V}_0 := \boldsymbol{V}(\boldsymbol{\xi}_0, \boldsymbol{\xi}_0)$ associated with the matrix valued function $\boldsymbol{V}(\cdot, \cdot)$ of representation (3.11). We can write then

$$F(\boldsymbol{x}, \boldsymbol{\xi}) = (\boldsymbol{x} - \boldsymbol{\xi})'\boldsymbol{V}_0(\boldsymbol{x} - \boldsymbol{\xi}) + \boldsymbol{r}(\boldsymbol{x}, \boldsymbol{\xi}), \tag{3.12}$$

where $\boldsymbol{r}(\boldsymbol{x}, \boldsymbol{\xi}) := (\boldsymbol{x} - \boldsymbol{\xi})'[\boldsymbol{V}(\boldsymbol{x}, \boldsymbol{\xi}) - \boldsymbol{V}_0](\boldsymbol{x} - \boldsymbol{\xi})$. We have that

$$|\boldsymbol{r}(\boldsymbol{x}, \boldsymbol{\xi})| \leq \|\boldsymbol{x} - \boldsymbol{\xi}\|^2 \|\boldsymbol{V}(\boldsymbol{x}, \boldsymbol{\xi}) - \boldsymbol{V}_0\|,$$

and $\boldsymbol{V}(\boldsymbol{x}, \boldsymbol{\xi})$ tends to $\boldsymbol{V}_0$ as $\boldsymbol{x} \to \boldsymbol{\xi}_0$ and $\boldsymbol{\xi} \to \boldsymbol{\xi}_0$. Consequently, for $(\boldsymbol{x}, \boldsymbol{\xi})$ near $(\boldsymbol{\xi}_0, \boldsymbol{\xi}_0)$ the remainder term $\boldsymbol{r}(\boldsymbol{x}, \boldsymbol{\xi})$ in (3.12) is of order[9]

$$\boldsymbol{r}(\boldsymbol{x}, \boldsymbol{\xi}) = o\left(\|\boldsymbol{x} - \boldsymbol{\xi}_0\|^2 + \|\boldsymbol{\xi} - \boldsymbol{\xi}_0\|^2\right).$$

This can be compared with a Taylor expansion of $F(\boldsymbol{x}, \boldsymbol{\xi})$ at $(\boldsymbol{\xi}_0, \boldsymbol{\xi}_0)$. We have that $F(\boldsymbol{\xi}_0, \boldsymbol{\xi}_0) = 0$, and since $F(\cdot, \boldsymbol{\xi}_0)$ attains its minimum (of zero) at $\boldsymbol{x} = \boldsymbol{\xi}_0$, we have that $\partial F(\boldsymbol{\xi}_0, \boldsymbol{\xi}_0)/\partial \boldsymbol{x} = 0$, and similarly $\partial F(\boldsymbol{\xi}_0, \boldsymbol{\xi}_0)/\partial \boldsymbol{\xi} = 0$. It follows that the second-order Taylor expansion of $F(\boldsymbol{x}, \boldsymbol{\xi})$ at $(\boldsymbol{\xi}_0, \boldsymbol{\xi}_0)$ can be written as follows

$$\begin{aligned} F(\boldsymbol{x}, \boldsymbol{\xi}) = \; & \tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\xi}_0)'\boldsymbol{H}_{xx}(\boldsymbol{x} - \boldsymbol{\xi}_0) + \tfrac{1}{2}(\boldsymbol{\xi} - \boldsymbol{\xi}_0)'\boldsymbol{H}_{\xi\xi}(\boldsymbol{\xi} - \boldsymbol{\xi}_0) + (\boldsymbol{x} - \boldsymbol{\xi}_0)'\boldsymbol{H}_{x\xi}(\boldsymbol{\xi} - \boldsymbol{\xi}_0) \\ & + o\left(\|\boldsymbol{x} - \boldsymbol{\xi}_0\|^2 + \|\boldsymbol{\xi} - \boldsymbol{\xi}_0\|^2\right), \end{aligned} \tag{3.13}$$

[8] By $\otimes$ we denote the Kronecker product of matrices.

[9] The notation $o(x)$ means that $o(x)$ is a function of x such that $o(x)/x$ tends to zero as $x \to 0$. For a sequence X_n of random variables the notation $X_n = o_p(a_n)$ means that X_n/a_n converges in probability to zero. In particular, $X_n = o_p(1)$ means that X_n converges in probability to zero.

where $\boldsymbol{H}_{xx} := \partial^2 F(\boldsymbol{\xi}_0, \boldsymbol{\xi}_0)/\partial \boldsymbol{x} \partial \boldsymbol{x}'$, $\boldsymbol{H}_{\xi\xi} := \partial^2 F(\boldsymbol{\xi}_0, \boldsymbol{\xi}_0)/\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}'$, $\boldsymbol{H}_{x\xi} := \partial^2 F(\boldsymbol{\xi}_0, \boldsymbol{\xi}_0)/\partial \boldsymbol{x} \partial \boldsymbol{\xi}'$ are the corresponding Hessian matrices of second order partial derivatives. By comparing (3.12) and (3.13) we obtain that

$$\frac{\partial^2 F(\boldsymbol{\xi}_0, \boldsymbol{\xi}_0)}{\partial \boldsymbol{x} \partial \boldsymbol{x}'} = \frac{\partial^2 F(\boldsymbol{\xi}_0, \boldsymbol{\xi}_0)}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}'} = -\frac{\partial^2 F(\boldsymbol{\xi}_0, \boldsymbol{\xi}_0)}{\partial \boldsymbol{x} \partial \boldsymbol{\xi}'} = 2\boldsymbol{V}_0. \tag{3.14}$$

Both discrepancy functions $F_{ML}(\boldsymbol{s}, \boldsymbol{\sigma})$ and $F_{GLS}(\boldsymbol{s}, \boldsymbol{\sigma})$, defined in (3.5) and (3.9), respectively, have the same Hessian matrix

$$\frac{\partial^2 F_{ML}(\boldsymbol{\sigma}_0, \boldsymbol{\sigma}_0)}{\partial \boldsymbol{s} \partial \boldsymbol{s}'} = \frac{\partial^2 F_{GLS}(\boldsymbol{\sigma}_0, \boldsymbol{\sigma}_0)}{\partial \boldsymbol{s} \partial \boldsymbol{s}'} = \boldsymbol{\Sigma}_0^{-1} \otimes \boldsymbol{\Sigma}_0^{-1}, \tag{3.15}$$

and hence the same second-order Taylor approximation at $(\boldsymbol{\sigma}_0, \boldsymbol{\sigma}_0)$.

Remark 2 For technical reasons we also assume the following condition for discrepancy functions.

(iv) For any (fixed) $\bar{\boldsymbol{x}} \in \Xi$, $F(\boldsymbol{x}, \boldsymbol{\xi})$ tends to infinity as $\boldsymbol{x} \to \bar{\boldsymbol{x}}$ and $\|\boldsymbol{\xi}\| \to \infty$.

It is not difficult to verify that the ML discrepancy function, defined in (3.5), and the GLS discrepancy functions satisfy this condition.

4 Consistency of MDF Estimators

Let $\hat{\boldsymbol{\xi}} = \hat{\boldsymbol{\xi}}_n$ be a given estimator[10], based on a sample of size n, of the population value $\boldsymbol{\xi}_0 \in \Xi$ of the parameter vector of the saturated model. It is said that the estimator $\hat{\boldsymbol{\xi}}_n$ is *consistent* if it converges with probability one (w.p.1) to $\boldsymbol{\xi}_0$ as $n \to \infty$. For example, by the (strong) Law of Large Numbers we have that the sample covariance matrix $\boldsymbol{S}$ converges to $\boldsymbol{\Sigma}_0$ w.p.1 as $n \to \infty$. For this to hold we only need to assume that the population distribution has finite second-order moments, and hence the covariance matrix $\boldsymbol{\Sigma}_0$ does exist, and that the corresponding random sample is iid.

Let $F(\boldsymbol{x}, \boldsymbol{\xi})$ be a chosen discrepancy function satisfying conditions (i)–(iv) specified in the previous section, and $\bar{\boldsymbol{\xi}} = \bar{\boldsymbol{\xi}}(\boldsymbol{x})$ be an optimal solution of the minimization problem:

$$\min_{\boldsymbol{\xi} \in \Xi_0} F(\boldsymbol{x}, \boldsymbol{\xi}). \tag{4.1}$$

Note that since $F(\boldsymbol{x}, \cdot)$ is continuous and because of condition (iv), such a minimizer always exists (provided that the set Ξ_0 is closed), although may be not unique. Define

$$\hat{\boldsymbol{\xi}}_n^* := \bar{\boldsymbol{\xi}}(\hat{\boldsymbol{\xi}}_n) \text{ and } \boldsymbol{\xi}^* := \bar{\boldsymbol{\xi}}(\boldsymbol{\xi}_0). \tag{4.2}$$

[10]We use the subscript n, in the notation $\hat{\boldsymbol{\xi}}_n$, to emphasize that the estimator is a function of a considered sample of size n.

That is, $\hat{\boldsymbol{\xi}}_n^*$ and $\boldsymbol{\xi}^*$ are minimizers of $F(\hat{\boldsymbol{\xi}}_n, \cdot)$ and $F(\boldsymbol{\xi}_0, \cdot)$, respectively, over Ξ_0. It could be noted that if the model holds, i.e., $\boldsymbol{\xi}_0 \in \Xi_0$, then the minimizer $\boldsymbol{\xi}^*$ coincides with $\boldsymbol{\xi}_0$ and is unique (because of the properties (i) and (ii) of the discrepancy function). It is possible to show that if the minimizer $\boldsymbol{\xi}^*$ is unique, then the function $\bar{\boldsymbol{\xi}}(\boldsymbol{x})$ is continuous at $\boldsymbol{x} = \boldsymbol{\xi}_0$, i.e., $\bar{\boldsymbol{\xi}}(\boldsymbol{x}) \to \boldsymbol{\xi}^*$ as $\boldsymbol{x} \to \boldsymbol{\xi}_0$. Together with consistency of $\hat{\boldsymbol{\xi}}_n$ this implies the following result (cf., [26]).

Proposition 4.1 *Suppose that the discrepancy function satisfies conditions* (i)–(iv) *and $\hat{\boldsymbol{\xi}}_n$ is a consistent estimator of $\boldsymbol{\xi}_0$. Then $\hat{\boldsymbol{\xi}}_n^*$ converges to $\boldsymbol{\xi}^*$ w.p.1 as $n \to \infty$, provided that the minimizer $\boldsymbol{\xi}^*$ is unique. In particular, if $\boldsymbol{\xi}_0 \in \Xi_0$, then $\hat{\boldsymbol{\xi}}_n^* \to \boldsymbol{\xi}_0$ w.p.1 as $n \to \infty$.*

Similar analysis can be applied to studying consistency of the MDF estimators of the parameter vectors in Θ. For a given $\boldsymbol{x} \in \Xi$ consider the optimization (minimization) problem:

$$\min_{\boldsymbol{\theta} \in \Theta} F(\boldsymbol{x}, \boldsymbol{g}(\boldsymbol{\theta})). \tag{4.3}$$

Recall that the MDF estimator $\hat{\boldsymbol{\theta}}_n$ is an optimal solution of problem (4.3) for $\boldsymbol{x} = \hat{\boldsymbol{\xi}}_n$. Let $\boldsymbol{\theta}^*$ be an optimal solution of (4.3) for $\boldsymbol{x} = \boldsymbol{\xi}_0$, i.e.,

$$\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta} \in \Theta} F\left(\boldsymbol{\xi}_0, \boldsymbol{g}(\boldsymbol{\theta})\right).$$

Of course, if $\boldsymbol{\xi}_0 = \boldsymbol{g}(\boldsymbol{\theta}_0)$ for some $\boldsymbol{\theta}_0 \in \Theta$ (i.e., the model holds), then $\boldsymbol{\theta}_0$ is an optimal solution of (4.3) for $\boldsymbol{x} = \boldsymbol{\xi}_0$, and we can take $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$. The optimal values of problems (4.1) and (4.3) are equal to each other and there is a one-to-one correspondence between the sets of optimal solutions of problems (4.1) and (4.3). That is, if $\bar{\boldsymbol{\theta}}$ is an optimal solution of (4.3), then $\bar{\boldsymbol{\xi}} = \boldsymbol{g}(\bar{\boldsymbol{\theta}})$ is an optimal solution of (4.1), and conversely if $\bar{\boldsymbol{\xi}}$ is an optimal solution of (4.1) and $\bar{\boldsymbol{\theta}} \in \Theta$ is a corresponding point of Θ, then $\bar{\boldsymbol{\theta}}$ is an optimal solution of (4.3). The relation between $\bar{\boldsymbol{\xi}}$ and $\bar{\boldsymbol{\theta}}$ is defined by the equation $\bar{\boldsymbol{\xi}} = \boldsymbol{g}(\bar{\boldsymbol{\theta}})$. If the model is identified at $\bar{\boldsymbol{\theta}}$, then the equation $\bar{\boldsymbol{\xi}} = \boldsymbol{g}(\bar{\boldsymbol{\theta}})$ defines the point $\bar{\boldsymbol{\theta}}$ uniquely.

It follows that, under the assumptions of Proposition 4.1, $\hat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}^*$, if the inverse of the mapping $\boldsymbol{g}(\cdot)$ is continuous at $\boldsymbol{\theta}^*$, i.e., if the following condition holds:

$$\boldsymbol{g}(\boldsymbol{\theta}_n) \to \boldsymbol{g}(\boldsymbol{\theta}^*), \text{ for some sequence } \{\boldsymbol{\theta}_n\} \subset \Theta, \text{ implies that } \boldsymbol{\theta}_n \to \boldsymbol{\theta}^*. \tag{4.4}$$

Note that the above condition (4.4) can only hold if the model is identified at $\boldsymbol{\theta}^*$. This leads to the following result (cf., [17, 26]).

Proposition 4.2 *Suppose that the discrepancy function satisfies conditions* (i)–(iv), *$\hat{\boldsymbol{\xi}}_n$ is a consistent estimator of $\boldsymbol{\xi}_0$, and for $\boldsymbol{x} = \boldsymbol{\xi}_0$ problem* (4.3) *has unique optimal solution $\boldsymbol{\theta}^*$ and condition* (4.4) *holds. Then $\hat{\boldsymbol{\theta}}_n$ converges to $\boldsymbol{\theta}^*$ w.p.1 as $n \to \infty$. In particular, if $\boldsymbol{\xi}_0 = \boldsymbol{g}(\boldsymbol{\theta}_0)$, for some $\boldsymbol{\theta}_0 \in \Theta$ (i.e., the model holds), then $\boldsymbol{\theta}_0 = \boldsymbol{\theta}^*$ and the MDF estimator $\hat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}_0$.*

Note that uniqueness of the optimal solution $\boldsymbol{\theta}^*$ implies uniqueness of the corresponding optimal solution $\boldsymbol{\xi}^*$. Converse of that also holds if the model is identified at $\boldsymbol{\theta}^*$. As it was mentioned above, identifiability of $\boldsymbol{\theta}^*$ is a necessary condition for the property (4.4) to hold. It is also sufficient if the set Θ is compact (i.e., bounded and closed). For a noncompact set Θ, condition (4.4) prevents the MDF estimator from escaping to infinity.

The above proposition shows that if the model holds, then under mild regularity conditions (in particular, identifiability of the model at $\boldsymbol{\theta}_0$) the MDF estimator $\hat{\boldsymbol{\theta}}_n$ converges w.p.1 to the true (population) value $\boldsymbol{\theta}_0$ of the parameter vector. On the other hand, if the model does not hold, then $\hat{\boldsymbol{\theta}}_n$ converges to an optimal solution $\boldsymbol{\theta}^*$ of the problem (4.3). It could be noted that if the model does not hold, then such an optimal solution depends on a particular choice of the discrepancy function.

As we can see uniqueness of the (population) minimizer $\boldsymbol{\theta}^*$ is crucial for convergence of the MDF estimator $\hat{\boldsymbol{\theta}}_n$. If the model holds, then $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$ and uniqueness of $\boldsymbol{\theta}_0$ is equivalent to identifiability of the model at $\boldsymbol{\theta}_0$. Now if a point $\bar{\boldsymbol{\theta}} = \bar{\boldsymbol{\theta}}(\boldsymbol{x})$ is an optimal solution of problem (4.3) and is an interior point of the set Θ, then it satisfies the necessary optimality condition

$$\frac{\partial F(\boldsymbol{x}, \boldsymbol{g}(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = \mathbf{0}. \tag{4.5}$$

This condition can be viewed as a system of (nonlinear) equations. Consider a point $\boldsymbol{\theta}_0$ in the *interior* of the set Θ and let $\boldsymbol{\xi}_0 := \boldsymbol{g}(\boldsymbol{\theta}_0)$. By linearizing (4.5) at $\boldsymbol{x} = \boldsymbol{\xi}_0$ and $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, we obtain the following (linear) system of equations:

$$\left[\frac{\partial^2 F(\boldsymbol{\theta}_0, \boldsymbol{g}(\boldsymbol{\theta}_0))}{\partial \boldsymbol{\theta} \partial \boldsymbol{x}'}\right](\boldsymbol{x} - \boldsymbol{\xi}_0) + \left[\frac{\partial^2 F(\boldsymbol{\theta}_0, \boldsymbol{g}(\boldsymbol{\theta}_0))}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right](\boldsymbol{\theta} - \boldsymbol{\theta}_0) = \mathbf{0}. \tag{4.6}$$

Note that by (3.14) we have that

$$\frac{\partial^2 F(\boldsymbol{\theta}_0, \boldsymbol{g}(\boldsymbol{\theta}_0))}{\partial \boldsymbol{\theta} \partial \boldsymbol{x}'} = -2\boldsymbol{\Delta}_0' \boldsymbol{V}_0 \ \text{ and } \ \frac{\partial^2 F(\boldsymbol{\theta}_0, \boldsymbol{g}(\boldsymbol{\theta}_0))}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = 2\boldsymbol{\Delta}_0' \boldsymbol{V}_0 \boldsymbol{\Delta}_0, \tag{4.7}$$

where $\boldsymbol{\Delta}_0 := \boldsymbol{\Delta}(\boldsymbol{\theta}_0)$. Since the matrix $\boldsymbol{V}_0$ is positive definite, we have that the matrix $\boldsymbol{\Delta}_0' \boldsymbol{V}_0 \boldsymbol{\Delta}_0$ is nonsingular iff the Jacobian matrix $\boldsymbol{\Delta}_0$ has full column rank q (recall that this is a sufficient condition for identifiability of $\boldsymbol{\theta}_0$). It follows then by the Implicit Function Theorem that:

> If $\boldsymbol{\Delta}_0$ has full column rank q, then for all $\boldsymbol{x}$ sufficiently close to $\boldsymbol{\xi}_0$ the system (4.5) has a unique solution $\bar{\boldsymbol{\theta}} = \bar{\boldsymbol{\theta}}(\boldsymbol{x})$ in a neighborhood of $\boldsymbol{\theta}_0$, and $\bar{\boldsymbol{\theta}}$ is the unique optimal solution of problem (4.3) in that neighborhood of $\boldsymbol{\theta}_0$.

The above is a local result. It implies that, under the specified conditions, if the estimator $\hat{\boldsymbol{\xi}}_n$ is sufficiently close to a point $\boldsymbol{\xi}_0 = \boldsymbol{g}(\boldsymbol{\theta}_0)$ satisfying the model, i.e., the fit is good enough, then the corresponding MDF estimator $\hat{\boldsymbol{\theta}}_n$ is unique, and can be obtained by solving equations (4.5) with $\boldsymbol{x} = \hat{\boldsymbol{\xi}}_n$, in a neighborhood of the point $\boldsymbol{\theta}_0 \in \Theta$. Of course, in practice it is impossible to say apriory when "sufficiently close" is close enough for the above to hold.

5 Asymptotic Analysis of the MDF Estimation Procedure

In this section we discuss a basic theory of asymptotics of the MDF estimation procedure. We assume that the considered discrepancy function satisfies conditions (i)–(iv) specified in section 3. We also assume that the estimator $\hat{\boldsymbol{\xi}}_n$ is *asymptotically normal*. That is, we assume that the sequence

$$\boldsymbol{Z}_n := n^{1/2}(\hat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_0),$$

of random vectors, with $\boldsymbol{\xi}_0$ being the population value of the parameter vector $\boldsymbol{\xi} \in \Xi$, converges in distribution[11] to multivariate normal with mean vector zero and covariance matrix $\boldsymbol{\Gamma}$, i.e., $\boldsymbol{Z}_n \Rightarrow N(\boldsymbol{0}, \boldsymbol{\Gamma})$. For example, in the analysis of covariance structures we have that vector $\boldsymbol{s} := \text{vec}(\boldsymbol{S})$, associated with the sample covariance matrix $\boldsymbol{S}$, is asymptotically normal. This follows from the Central Limit Theorem provided that the population distribution has fourth-order moments and the sample is iid. Moreover, if the population distribution is normal, then $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_N$, where

$$\boldsymbol{\Gamma}_N := 2\boldsymbol{M}_p(\boldsymbol{\Sigma}_0 \otimes \boldsymbol{\Sigma}_0) \tag{5.1}$$

with $\boldsymbol{M}_p$ being an $p^2 \times p^2$ symmetric idempotent matrix of rank $p(p+1)/2$ with element in row ij and column kl given by[12] $\boldsymbol{M}_p(ij, kl) = \frac{1}{2}(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})$ (cf., [7]). It follows that matrix $\boldsymbol{\Gamma}_N$ has also rank $p(p+1)/2$, provided that the covariance matrix $\boldsymbol{\Sigma}_0$ is nonsingular. We assume that the (asymptotic) covariance matrix $\boldsymbol{\Gamma}$, of $\boldsymbol{Z}_n$, has the *maximal rank*, which in the case of covariance structures is $p(p+1)/2$. It follows then that the linear space generated by columns of the Jacobian matrix $\boldsymbol{\Delta}(\boldsymbol{\theta})$ is contained in the linear space generated by columns of $\boldsymbol{\Gamma}$.

Denote by $\vartheta(\boldsymbol{x})$ the optimal value of problem (4.1). Recall that the optimal values of problems (4.1) and (4.3) are the same, and hence $\vartheta(\boldsymbol{x})$ is also the optimal value of problem (4.3). Denote by $\bar{\boldsymbol{\theta}}(\boldsymbol{x})$ an optimal solution of problem (4.3). By the definitions we have that $\hat{F} = \vartheta(\hat{\boldsymbol{\xi}}_n)$ and $\hat{\boldsymbol{\theta}}_n = \bar{\boldsymbol{\theta}}(\hat{\boldsymbol{\xi}}_n)$. Therefore it should be not surprising that asymptotic properties of the MDF test statistics and estimators are closely related to analytical properties of functions $\vartheta(\cdot)$ and $\bar{\boldsymbol{\theta}}(\cdot)$.

Suppose that the model holds, i.e., $\boldsymbol{\xi}_0 \in \Xi_0$ or equivalently $\boldsymbol{\xi}_0 = \boldsymbol{g}(\boldsymbol{\theta}_0)$ for some $\boldsymbol{\theta}_0 \in \Theta$. The second-order Taylor approximation of the discrepancy function, at the point $(\boldsymbol{x}, \boldsymbol{\xi}) = (\boldsymbol{\xi}_0, \boldsymbol{\xi}_0)$, can be written in the form (3.12). Suppose, further, that the set Ξ_0 can be approximated at the point $\boldsymbol{\xi}_0$ by a cone $\mathcal{T} \subset \mathbb{R}^m$ in the following sense[13]:

$$\text{dist}(\boldsymbol{\xi}_0 + \boldsymbol{z}, \Xi_0) = o(\|\boldsymbol{z}\|), \quad \boldsymbol{z} \in \mathcal{T}, \tag{5.2}$$

$$\text{dist}(\boldsymbol{\xi} - \boldsymbol{\xi}_0, \mathcal{T}) = o(\|\boldsymbol{\xi} - \boldsymbol{\xi}_0\|), \quad \boldsymbol{\xi} \in \Xi_0. \tag{5.3}$$

This definition of cone approximation is going back to Chernoff [12].

[11] By " $\Rightarrow$ " we denote convergence in distribution.

[12] Here $\delta_{ik} = 1$ if $i = k$, and $\delta_{ik} = 0$ if $i \neq k$.

[13] By $\text{dist}(\boldsymbol{x}, A) := \inf_{\boldsymbol{z} \in A} \|\boldsymbol{x} - \boldsymbol{z}\|$ we denote the distance from a point $\boldsymbol{x} \in \mathbb{R}^m$ to a set $A \subset \mathbb{R}^m$. A set $\mathcal{T} \subset \mathbb{R}^m$ is said to be a *cone* if for any $\boldsymbol{z} \in \mathcal{T}$ and $t \geq 0$ it follows that $t\boldsymbol{z} \in \mathcal{T}$.

In particular, if Ξ_0 is a smooth manifold near $\boldsymbol{\xi}_0$, then it is approximated at $\boldsymbol{\xi}_0$ by a *linear space* referred to as its tangent space at $\boldsymbol{\xi}_0$. Suppose that $\boldsymbol{\theta}_0$ is an *interior* point of Θ and $\boldsymbol{\theta}_0$ is locally regular (see Definition 2.2). Denote $\boldsymbol{\Delta}_0 := \boldsymbol{\Delta}(\boldsymbol{\theta}_0)$. Then the image $\boldsymbol{g}(\mathcal{N}) := \{\boldsymbol{\xi} : \boldsymbol{\xi} = \boldsymbol{g}(\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \mathcal{N}\}$ of the set Θ restricted to a neighborhood $\mathcal{N} \subset \Theta$ of $\boldsymbol{\theta}_0$, is a smooth manifold with the tangent space at $\boldsymbol{\xi}_0$ given by

$$\mathcal{T} = \left\{\boldsymbol{\zeta} = \boldsymbol{\Delta}_0\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^q\right\}. \tag{5.4}$$

Of course, $\boldsymbol{g}(\mathcal{N})$ is a subset of Ξ_0, restricted to a neighborhood of $\boldsymbol{\xi}_0$. The asymptotic analysis is local in nature. Therefore, there is no loss of generality here by restricting the set Θ to a neighborhood of $\boldsymbol{\theta}_0$.

Definition 5.1 *A point $\boldsymbol{\theta}_0 \in \Theta$ is said to be* regular *if $\boldsymbol{\theta}_0$ is locally regular and there exist a neighborhood $\mathcal{V}$ of $\boldsymbol{\xi}_0 = \boldsymbol{g}(\boldsymbol{\theta}_0)$ and a neighborhood $\mathcal{N} \subset \Theta$ of $\boldsymbol{\theta}_0$ such that $\Xi_0 \cap \mathcal{V} = \boldsymbol{g}(\mathcal{N})$.*

In other words, regularity of $\boldsymbol{\theta}_0$ ensures that local structure of Ξ_0 near $\boldsymbol{\xi}_0$ is provided by the mapping $\boldsymbol{g}(\boldsymbol{\theta})$ defined in a neighborhood of $\boldsymbol{\theta}_0$. Regularity of $\boldsymbol{\theta}_0$ implies that Ξ_0 is a smooth manifold near $\boldsymbol{\xi}_0$ and is approximated at $\boldsymbol{\xi}_0$ by its tangent space $\mathcal{T}$of the form (5.4).

In particular, if condition (4.4) holds and $\boldsymbol{\Delta}(\boldsymbol{\theta})$ has full column rank q for all $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$ (i.e., point $\boldsymbol{\theta}_0$ is locally regular of rank q), then Ξ_0 is a smooth manifold near $\boldsymbol{\xi}_0$ and its tangent space at $\boldsymbol{\xi}_0$ is given by (5.4). Note that $\boldsymbol{g}(\mathcal{N})$ is a smooth manifold even if the rank of the Jacobian matrix $\boldsymbol{\Delta}_0$ is less than q provided that the local regularity condition holds. Note also that if the tangent space $\mathcal{T}$ is given in the form (5.4), then its dimension, $\dim(\mathcal{T})$, is equal to the rank of the Jacobian matrix $\boldsymbol{\Delta}_0$, i.e., $\dim(\mathcal{T}) = \text{rank}(\boldsymbol{\Delta}_0)$.

Remark 3 Let us remark at this point that if the point $\boldsymbol{\theta}_0$ is a *boundary* point of Θ, then under certain regularity conditions the set Θ can be approximated at $\boldsymbol{\theta}_0$ by a cone $\mathcal{C} \subset \mathbb{R}^q$, rather than a linear space, and consequently Ξ_0 can be approximated by the cone

$$\mathcal{T} = \left\{\boldsymbol{\zeta} = \boldsymbol{\Delta}_0\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathcal{C}\right\}. \tag{5.5}$$

Later we will discuss implications of this to the asymptotics of the MDF estimators (see section 5.4).

The above discussion suggests the following approximation of the optimal value function $\vartheta(\boldsymbol{x})$ near $\boldsymbol{\xi}_0$ (cf., [29, Lemma 3.1]).

Proposition 5.1 *Suppose that the set Ξ_0 can be approximated at $\boldsymbol{\xi}_0 \in \Xi_0$ by a cone $\mathcal{T} \subset \mathbb{R}^m$. Then*

$$\vartheta(\boldsymbol{\xi}_0 + \boldsymbol{z}) = \min_{\boldsymbol{\zeta} \in \mathcal{T}} (\boldsymbol{z} - \boldsymbol{\zeta})'\boldsymbol{V}_0(\boldsymbol{z} - \boldsymbol{\zeta}) + o(\|\boldsymbol{z}\|^2). \tag{5.6}$$

Suppose, further, that the cone $\mathcal{T}$ actually is a linear space, i.e., Ξ_0 can be approximated at $\boldsymbol{\xi}_0 \in \Xi_0$ by a linear space $\mathcal{T} \subset \mathbb{R}^m$. Then the main (first) term in the right hand side of (5.6) is a quadratic function of $\boldsymbol{z}$ and can be written as $\boldsymbol{z}'\boldsymbol{Q}\boldsymbol{z}$ for some symmetric positive semidefinite matrix $\boldsymbol{Q}$. In particular, if the space $\mathcal{T}$ is given in the form (5.4), then this term can be written as

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^q} (\boldsymbol{z} - \boldsymbol{\Delta}_0\boldsymbol{\beta})'\boldsymbol{V}_0(\boldsymbol{z} - \boldsymbol{\Delta}_0\boldsymbol{\beta}) = \boldsymbol{z}'\boldsymbol{Q}\boldsymbol{z}, \tag{5.7}$$

where[14] $\boldsymbol{Q} = \boldsymbol{V}_0 - \boldsymbol{V}_0\boldsymbol{\Delta}_0(\boldsymbol{\Delta}_0'\boldsymbol{V}_0\boldsymbol{\Delta}_0)^{-}\boldsymbol{\Delta}_0'\boldsymbol{V}_0$. It is also possible to write this matrix in the form

$$\boldsymbol{Q} = \boldsymbol{\Delta}_c(\boldsymbol{\Delta}_c'\boldsymbol{V}_0^{-1}\boldsymbol{\Delta}_c)^{-1}\boldsymbol{\Delta}_c', \tag{5.8}$$

where $\boldsymbol{\Delta}_c$ is an orthogonal complement of $\boldsymbol{\Delta}_0$, i.e., $\boldsymbol{\Delta}_c$ is an $m \times (m - \text{rank}(\boldsymbol{\Delta}_0))$ matrix of full column rank such that $\boldsymbol{\Delta}_c'\boldsymbol{\Delta}_0 = 0$. This follows by the standard theory of linear models (see, e.g., [24, sections 3.6 and 3.8]). Note that $\text{rank}(\boldsymbol{Q}) = m - \text{rank}(\boldsymbol{\Delta}_0)$.

We already discussed continuity properties of $\bar{\boldsymbol{\theta}}(\cdot)$ in section 4. Suppose that the model is (globally) identified at $\boldsymbol{\theta}_0$, and hence $\bar{\boldsymbol{\theta}}(\boldsymbol{\xi}_0) = \boldsymbol{\theta}_0$ and is defined uniquely. Then under mild regularity conditions we have that $\bar{\boldsymbol{\theta}}(\cdot)$ is continuous at $\boldsymbol{\xi}_0$. Moreover, we have the following result (cf., [29, Lemma 3.1]).

Proposition 5.2 *Suppose that the set Θ can be approximated at $\boldsymbol{\theta}_0 \in \Theta$ by a convex cone $\mathcal{C} \subset \mathbb{R}^q$, the Jacobian matrix $\boldsymbol{\Delta}_0$ has full column rank q and $\bar{\boldsymbol{\theta}}(\cdot)$ is continuous at $\boldsymbol{\xi}_0 = \boldsymbol{g}(\boldsymbol{\theta}_0)$. Then*

$$\bar{\boldsymbol{\theta}}(\boldsymbol{\xi}_0 + \boldsymbol{z}) = \boldsymbol{\theta}_0 + \bar{\boldsymbol{\beta}}(\boldsymbol{z}) + o(\|\boldsymbol{z}\|), \tag{5.9}$$

where $\bar{\boldsymbol{\beta}}(\boldsymbol{z})$ is the optimal solution of the problem

$$\min_{\boldsymbol{\beta} \in \mathcal{C}} (\boldsymbol{z} - \boldsymbol{\Delta}_0\boldsymbol{\beta})'\boldsymbol{V}_0(\boldsymbol{z} - \boldsymbol{\Delta}_0\boldsymbol{\beta}). \tag{5.10}$$

Note that since the approximating cone $\mathcal{C}$ is assumed to be *convex*, $\text{rank}(\boldsymbol{\Delta}_0) = q$ and the matrix $\boldsymbol{V}_0$ is positive definite, the minimizer $\bar{\boldsymbol{\beta}}(\boldsymbol{z})$ is unique for any $\boldsymbol{z} \in \mathbb{R}^m$. If, moreover, the point $\boldsymbol{\theta}_0$ is an interior point of Θ and hence $\mathcal{C} = \mathbb{R}^q$, then $\bar{\boldsymbol{\beta}}(\cdot)$ is a linear function and can be written explicitly as

$$\bar{\boldsymbol{\beta}}(\boldsymbol{z}) = (\boldsymbol{\Delta}_0'\boldsymbol{V}_0\boldsymbol{\Delta}_0)^{-1}\boldsymbol{\Delta}_0'\boldsymbol{V}_0\boldsymbol{z}. \tag{5.11}$$

Now if $\boldsymbol{\xi}_0 \notin \Xi_0$, then the analysis becomes considerably more involved. It will be beyond the scope of this paper to give a detailed description of such theory. We refer the interested reader to Bonnans and Shapiro [6] for a thorough development of that theory. We give below some, relatively simple, results which will be relevant for the statistical inference. In the optimization literature the following result, giving a first order approximation of the optimal value function, is often referred to as Danskin Theorem [13].

[14] By $\boldsymbol{A}^{-}$ we denote a generalized inverse of matrix $\boldsymbol{A}$.

Proposition 5.3 *Let $\mathfrak{S}$ be the set of optimal solutions of problem* (4.1) *for $\boldsymbol{x} = \boldsymbol{\xi}_0$. Then*

$$\vartheta(\boldsymbol{\xi}_0 + \boldsymbol{z}) = \vartheta(\boldsymbol{\xi}_0) + \min_{\boldsymbol{\xi} \in \mathfrak{S}} \mathbf{g}'_{\boldsymbol{\xi}} \boldsymbol{z} + o(\|\boldsymbol{z}\|), \tag{5.12}$$

where $\mathbf{g}_{\boldsymbol{\xi}} := \partial F(\boldsymbol{\xi}_0, \boldsymbol{\xi})/\partial \boldsymbol{x}$. In particular, if problem (4.1) *has unique optimal solution $\boldsymbol{\xi}^*$ for $\boldsymbol{x} = \boldsymbol{\xi}_0$, then*

$$\frac{\partial \vartheta(\boldsymbol{\xi}_0)}{\partial \boldsymbol{x}} = \frac{\partial F(\boldsymbol{\xi}_0, \boldsymbol{\xi}^*)}{\partial \boldsymbol{x}}. \tag{5.13}$$

Of course, if $\boldsymbol{\xi}_0 \in \Xi_0$, then problem (4.1) has unique optimal solution $\boldsymbol{\xi}^* = \boldsymbol{\xi}_0$ and hence $\partial \vartheta(\boldsymbol{\xi}_0)/\partial \boldsymbol{x} = 0$. The following result is a consequence of the Implicit Function Theorem (cf., [25, Theorem 4.2]).

Proposition 5.4 *Suppose that:* (i) *for $\boldsymbol{x} = \boldsymbol{\xi}_0$ problem* (4.3) *has unique optimal solution $\boldsymbol{\theta}^*$,* (ii) *the point $\boldsymbol{\theta}^*$ is an interior point of Θ,* (iii) *$\bar{\boldsymbol{\theta}}(\cdot)$ is continuous at $\boldsymbol{\theta}^*$,* (iv) *the Hessian matrix $\boldsymbol{H}_{\theta\theta} := \partial^2 F(\boldsymbol{\xi}_0, \boldsymbol{g}(\boldsymbol{\theta}^*))/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'$ is nonsingular. Then $\bar{\boldsymbol{\theta}}(\cdot)$ is continuously differentiable and $\vartheta(\cdot)$ is twice continuously differentiable at $\boldsymbol{\xi}_0$, and*

$$\frac{\partial \bar{\boldsymbol{\theta}}(\boldsymbol{\xi}_0)}{\partial \boldsymbol{x}} = -\boldsymbol{H}_{\theta\theta}^{-1} \boldsymbol{H}_{\theta x}, \tag{5.14}$$

$$\frac{\partial^2 \vartheta(\boldsymbol{\xi}_0)}{\partial \boldsymbol{x} \partial \boldsymbol{x}'} = \boldsymbol{H}_{xx} - \boldsymbol{H}'_{\theta x} \boldsymbol{H}_{\theta\theta}^{-1} \boldsymbol{H}_{\theta x}, \tag{5.15}$$

where $\boldsymbol{H}_{\theta x} := \partial^2 F(\boldsymbol{\xi}_0, \boldsymbol{g}(\boldsymbol{\theta}^*))/\partial \boldsymbol{\theta} \partial \boldsymbol{x}'$ and $\boldsymbol{H}_{xx} := \partial^2 F(\boldsymbol{\xi}_0, \boldsymbol{g}(\boldsymbol{\theta}^*))/\partial \boldsymbol{x} \partial \boldsymbol{x}'$.

Remark 4 If the model holds, and hence $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$, then $\boldsymbol{H}_{\theta\theta} = 2\boldsymbol{\Delta}_0' \boldsymbol{V}_0 \boldsymbol{\Delta}_0$, $\boldsymbol{H}_{\theta x} = -2\boldsymbol{\Delta}_0' \boldsymbol{V}_0$ and $\boldsymbol{H}_{xx} = 2\boldsymbol{V}_0$ (compare with (4.7)). In that case formula (5.14) gives the same derivatives as (5.9) and (5.11), and (5.15) is equivalent to (5.7), and these formulas involve only first-order derivatives (i.e., the Jacobian matrix) of $\boldsymbol{g}(\cdot)$. On the other hand, if the model does not hold, and hence $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}_0$, then these derivatives involve *second-order* derivatives of $\boldsymbol{g}(\cdot)$. Note also that the Hessian matrix $\boldsymbol{H}_{\theta\theta}$ can be nonsingular only if the Jacobian matrix $\boldsymbol{\Delta}(\boldsymbol{\theta}^*)$ has full column rank q and hence the model is locally identified at $\boldsymbol{\theta}^*$.

5.1 Asymptotics of MDF Test Statistics

Suppose that the model holds, i.e., $\boldsymbol{\xi}_0 \in \Xi_0$. Since $\hat{F} = \vartheta(\hat{\boldsymbol{\xi}}_n)$ we obtain from the approximation (5.6) the following asymptotic expansion of the MDF test statistic (under the null hypothesis):

$$n\hat{F} = \min_{\boldsymbol{\zeta} \in \mathcal{T}} (\boldsymbol{Z}_n - \boldsymbol{\zeta})' \boldsymbol{V}_0 (\boldsymbol{Z}_n - \boldsymbol{\zeta}) + o_p(1). \tag{5.16}$$

Recall that $\boldsymbol{Z}_n \Rightarrow N(\mathbf{0}, \boldsymbol{\Gamma})$. It follows that

$$n\hat{F} \Rightarrow \min_{\boldsymbol{\zeta} \in \mathcal{T}} (\boldsymbol{Z} - \boldsymbol{\zeta})' \boldsymbol{V}_0 (\boldsymbol{Z} - \boldsymbol{\zeta}), \tag{5.17}$$

where $\boldsymbol{Z}$ is a random vector having normal distribution $N(\mathbf{0}, \boldsymbol{\Gamma})$. The optimal value of the right hand side of (5.17) is a quadratic function of $\boldsymbol{Z}$. Under certain conditions this quadratic function $\boldsymbol{Z}'\boldsymbol{Q}\boldsymbol{Z}$ has a chi-square distribution (see, e.g., [24, section 2.4]). In particular, this holds if $\boldsymbol{V}_0 = \boldsymbol{\Gamma}^{-1}$. As it was discussed earlier, nonsingularity of the covariance matrix $\boldsymbol{\Gamma}$ depends on a choice of the space where the saturated model is defined. In applications it is often convenient to take a larger space in which case $\boldsymbol{\Gamma}$ becomes singular. It is said that the discrepancy function is **correctly specified** if $\boldsymbol{V}_0$ is equal to a generalized inverse of $\boldsymbol{\Gamma}$, that is, $\boldsymbol{\Gamma}\boldsymbol{V}_0\boldsymbol{\Gamma} = \boldsymbol{\Gamma}$. Of course, if $\boldsymbol{\Gamma}$ is nonsingular, then this is the same as $\boldsymbol{V}_0 = \boldsymbol{\Gamma}^{-1}$. As it was mentioned earlier we assume that the asymptotic covariance matrix $\boldsymbol{\Gamma}$ has the maximal rank, e.g., in the case of covariance structures we assume that $\mathrm{rank}(\boldsymbol{\Gamma}) = p(p+1)/2$. It follows then that each column vector of the Jacobian matrix $\boldsymbol{\Delta}(\boldsymbol{\theta})$ is contained in the linear space generated by columns of $\boldsymbol{\Gamma}$.

We have the following result (cf., [8, 31]) giving asymptotics of the null distribution of the MDF test statistic. Recall Definition 5.1 of a regular point.

Theorem 5.1 *Suppose that the model holds, the discrepancy function is correctly specified and the point* $\boldsymbol{\theta}_0$ *is regular (and hence the set* Ξ_0 *is approximated at* $\boldsymbol{\xi}_0 = \boldsymbol{g}(\boldsymbol{\theta}_0)$ *by a linear space* $\mathcal{T}$ *of the form* (5.4)*). Then the MDF test statistic* $n\hat{F}$ *converges in distribution to a (central) chi-square with*

$$\nu = \mathrm{rank}(\boldsymbol{\Gamma}) - \dim(\mathcal{T}) = \mathrm{rank}(\boldsymbol{\Gamma}) - \mathrm{rank}(\boldsymbol{\Delta}_0) \tag{5.18}$$

degrees of freedom.

Suppose that the mapping $\boldsymbol{g}(\cdot)$ is analytic and let r be the *characteristic rank* of the model. The above results imply that if the discrepancy function is correctly specified, then under the null hypothesis, generically, the MDF test statistic $n\hat{F}$ has asymptotically a chi-square distribution, with $\nu = \mathrm{rank}(\boldsymbol{\Gamma}) - r$ degrees of freedom. Recall that "generically" means that this holds for almost every population value $\boldsymbol{\theta}_0 \in \Theta$ of the parameter vector. For example, consider the setting of covariance structures and suppose that the population distribution is normal. Then the covariance matrix $\boldsymbol{\Gamma}$ can be written in the form (5.1), has rank $p(p+1)/2$ and matrix $\boldsymbol{V}_0 := \frac{1}{2}\boldsymbol{\Sigma}_0^{-1} \otimes \boldsymbol{\Sigma}_0^{-1}$ is its generalized inverse. It follows that for a normally distributed population, both discrepancy functions F_{ML} and F_{GLS}, defined in (3.5) and (3.9), respectively, are correctly specified. Therefore, we have that:

> Under the null hypothesis, generically, the MDF test statistics, associated with F_{ML} and F_{GLS}, are asymptotically chi-square distributed, with $\nu = p(p+1)/2 - r$ degrees of freedom, provided that the population distribution is normal.

It is possible to extend this basic result in various directions. Suppose now that the model does not hold, i.e., $\boldsymbol{\xi}_0 \notin \Xi_0$. Let $\boldsymbol{\xi}^*$ be a minimizer of $F(\boldsymbol{\xi}_0, \cdot)$ over Ξ_0, i.e., $\boldsymbol{\xi}^*$ is an optimal solution of problem (4.1) for $\boldsymbol{x} = \boldsymbol{\xi}_0$. Since the model does not hold, we have here that $\boldsymbol{\xi}^* \neq \boldsymbol{\xi}_0$. Suppose, however, that the population value $\boldsymbol{\xi}_0$ is close to the model set Ξ_0, i.e., there is no big difference between $\boldsymbol{\xi}^*$ and $\boldsymbol{\xi}_0$. We can employ approximation (5.6) at the point $\boldsymbol{\xi}^*$, instead of $\boldsymbol{\xi}_0$, by taking $\boldsymbol{V}_0 := \boldsymbol{V}(\boldsymbol{\xi}^*, \boldsymbol{\xi}^*)$, i.e., making the second-order Taylot expansion of the discrepancy function at the point $(\boldsymbol{x}, \boldsymbol{\xi}) = (\boldsymbol{\xi}^*, \boldsymbol{\xi}^*)$, and using the tangent space $\mathcal{T}$ at $\boldsymbol{\xi}^*$. We obtain the following approximation of the MDF statistic:

$$n\hat{F} = \min_{\boldsymbol{\zeta} \in \mathcal{T}} (\boldsymbol{Z}_n^* - \boldsymbol{\zeta})' \boldsymbol{V}_0 (\boldsymbol{Z}_n^* - \boldsymbol{\zeta}) + o(\|\boldsymbol{Z}_n^*\|^2), \tag{5.19}$$

where

$$\boldsymbol{Z}_n^* := n^{1/2}(\hat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}^*) = \underbrace{n^{1/2}(\hat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_0)}_{\boldsymbol{Z}_n} + \underbrace{n^{1/2}(\boldsymbol{\xi}_0 - \boldsymbol{\xi}^*)}_{\boldsymbol{\mu}_n}.$$

Recall that it is assumed that $\boldsymbol{Z}_n \Rightarrow N(\boldsymbol{0}, \boldsymbol{\Gamma})$. On the other hand, as n tends to infinity, the "deterministic" part $\boldsymbol{\mu}_n := n^{1/2}(\boldsymbol{\xi}_0 - \boldsymbol{\xi}^*)$ of $\boldsymbol{Z}_n^*$ grows indefinitely. However, the quadratic approximation, given by the right hand side of (5.19), could be reasonable if the "stochastic" part $\boldsymbol{Z}_n$ is bigger than the "deterministic" part $\boldsymbol{\mu}_n$ (we will discuss this in more details later). In order to formulate this in a mathematically rigorous way, we make the so-called assumption of a *sequence of local alternatives*. That is, we assume that there is a sequence $\boldsymbol{\xi}_0 = \boldsymbol{\xi}_{0,n}$ of population values (local alternatives) converging to a point $\boldsymbol{\xi}^* \in \Xi_0$ such that $n^{1/2}(\boldsymbol{\xi}_{0,n} - \boldsymbol{\xi}^*)$ converges to a (deterministic) vector[15] $\boldsymbol{\mu}$. It follows then that $\boldsymbol{Z}_n^* \Rightarrow N(\boldsymbol{\mu}, \boldsymbol{\Gamma})$ and the remainder term $o(\|\boldsymbol{Z}_n^*\|^2)$ in (5.19) converges in probability to zero. If, moreover, $\boldsymbol{V}_0 = \boldsymbol{\Gamma}^-$, then the quadratic term in (5.19) converges in distribution to noncentral chi-square with the same degrees of freedom ν and the noncentrality parameter

$$\delta = \min_{\boldsymbol{\zeta} \in \mathcal{T}} (\boldsymbol{\mu} - \boldsymbol{\zeta})' \boldsymbol{V}_0 (\boldsymbol{\mu} - \boldsymbol{\zeta}). \tag{5.20}$$

Moreover, by (5.6) we have that

$$\delta = \lim_{n \to \infty} \left[n \min_{\boldsymbol{\xi} \in \Xi_0} F(\boldsymbol{\xi}_{0,n}, \boldsymbol{\xi}) \right]. \tag{5.21}$$

This leads to the following result (cf., [25, Theorem 5.5],[36]).

Theorem 5.2 *Suppose that the assumption of a sequence of local alternatives (Pitman drift) holds, the discrepancy function is correctly specified and the set* Ξ_0 *is approximated at* $\boldsymbol{\xi}^* = \boldsymbol{g}(\boldsymbol{\theta}^*)$ *by a linear space* $\mathcal{T}$ *generated by the columns of the matrix* $\boldsymbol{\Delta}^* = \boldsymbol{\Delta}(\boldsymbol{\theta}^*)$. *Then the MDF test statistic* $n\hat{F}$ *converges in distribution to a noncentral chi-square with* $\nu = \mathrm{rank}(\boldsymbol{\Gamma}) - \dim(\mathcal{T})$ *degrees of freedom and the noncentrality parameter* δ *given in* (5.20) *or, equivalently,* (5.21).

From the practical point of view it is important to understand when the noncentral chi-square distribution gives a reasonable approximation of the true distribution of the MDF test statistics. By the analysis of section 4 we have that $\hat{F}$ converges w.p.1 to the value

$$F^* := \min_{\boldsymbol{\xi} \in \Xi_0} F(\boldsymbol{\xi}_0, \boldsymbol{\xi}) = F(\boldsymbol{\xi}_0, \boldsymbol{\xi}^*). \tag{5.22}$$

[15] This assumption is often referred to as *Pitman drift*.

Recall that $\vartheta(\boldsymbol{x})$ denotes the optimal value of problem (4.1), and hence $\hat{F} = \vartheta(\hat{\boldsymbol{\xi}}_n)$ and $F^* = \vartheta(\boldsymbol{\xi}_0)$. Suppose that $\boldsymbol{\xi}^*$ is the *unique* minimizer of $F(\boldsymbol{\xi}_0, \cdot)$ over Ξ_0. Then we have by Danskin Theorem (see Proposition 5.3) that $\partial\vartheta(\boldsymbol{\xi}_0)/\partial\boldsymbol{x} = \partial F(\boldsymbol{\xi}_0, \boldsymbol{\xi}^*)/\partial\boldsymbol{x}$. It follows that

$$n^{1/2}\left(\hat{F} - F^*\right) = \mathbf{g}_0' \boldsymbol{Z}_n + o_p(1), \tag{5.23}$$

where $\mathbf{g}_0 := \frac{\partial F(\boldsymbol{x}, \boldsymbol{\xi}^*)}{\partial \boldsymbol{x}}\big|_{\boldsymbol{x}=\boldsymbol{\xi}_0}$, which in turn implies the following asymptotic result (Shapiro [25, Theorem 5.3]).

Theorem 5.3 *Suppose that $\boldsymbol{\xi}^*$ is the unique minimizer of $F(\boldsymbol{\xi}_0, \cdot)$ over Ξ_0. Then*

$$n^{1/2}\left(\hat{F} - F^*\right) \Rightarrow N\left(0, \mathbf{g}_0' \boldsymbol{\Gamma} \mathbf{g}_0\right). \tag{5.24}$$

If the model holds, then $F^* = 0$ and $\mathbf{g}_0 = \mathbf{0}$. In that case the asymptotic result (5.24) degenerates into the trivial statement that $n^{1/2}\hat{F}$ converges in probability to zero. And, indeed, as it was discussed above, under the null hypothesis one needs to scale $\hat{F}$ by the factor of n, instead of $n^{1/2}$, in order to get meaningful asymptotics. However, as the distance between the population value $\boldsymbol{\xi}_0$ and the model set Ξ_0 becomes larger, the noncentral chi-square distribution approximation deteriorates and the normal distribution, with mean F^* and variance $n^{-1}\mathbf{g}_0'\boldsymbol{\Gamma}\mathbf{g}_0$, could become a better approximation of the distribution of $\hat{F}$. The noncentral chi-square approximation is based on the distribution of the quadratic form

$$\min_{\zeta \in \mathcal{T}} (\boldsymbol{Z}_n^* - \zeta)' \boldsymbol{V}_0 (\boldsymbol{Z}_n^* - \zeta) = \boldsymbol{Z}_n^{*\prime} \boldsymbol{Q} \boldsymbol{Z}_n^* = \boldsymbol{Z}_n' \boldsymbol{Q} \boldsymbol{Z}_n + \boldsymbol{\mu}_n' \boldsymbol{Q} \boldsymbol{\mu}_n + 2\boldsymbol{\mu}_n' \boldsymbol{Q} \boldsymbol{Z}_n. \tag{5.25}$$

Recall that $\boldsymbol{Z}_n^* = \boldsymbol{Z}_n + \boldsymbol{\mu}_n$ and $\boldsymbol{\mu}_n = n^{1/2}(\boldsymbol{\xi}_0 - \boldsymbol{\xi}^*)$. The first term, in the right hand side of (5.25), has approximately a central chi-square distribution with $\nu = \text{rank}(\boldsymbol{\Gamma}) - \dim(\mathcal{T})$ degrees of freedom. Suppose that $\boldsymbol{\xi}_0$ is close to Ξ_0. By (5.6) the second term in the right hand side of (5.25) can be approximated as follows

$$\boldsymbol{\mu}_n' \boldsymbol{Q} \boldsymbol{\mu}_n = n \min_{\zeta \in \mathcal{T}} (\boldsymbol{\xi}_0 - \boldsymbol{\xi}^* - \zeta)' \boldsymbol{V}_0 (\boldsymbol{\xi}_0 - \boldsymbol{\xi}^* - \zeta) \approx nF^*. \tag{5.26}$$

Recall that, by (5.21), nF^* is approximately equal to the noncentrality parameter δ. We also have that

$$\mathbf{g}_0 = \frac{\partial F(\boldsymbol{x}, \boldsymbol{\xi}^*)}{\partial \boldsymbol{x}}\Big|_{\boldsymbol{x}=\boldsymbol{\xi}_0} \approx \frac{\partial (\boldsymbol{x} - \boldsymbol{\xi}^*)' \boldsymbol{V}_0 (\boldsymbol{x} - \boldsymbol{\xi}^*)}{\partial \boldsymbol{x}}\Big|_{\boldsymbol{x}=\boldsymbol{\xi}_0} = 2\boldsymbol{V}_0(\boldsymbol{\xi}_0 - \boldsymbol{\xi}^*). \tag{5.27}$$

Moreover, by the first-order optimality conditions the gradient vector $\mathbf{g}_0$ is orthogonal to the space $\mathcal{T}$, and hence $\boldsymbol{V}_0(\boldsymbol{\xi}_0 - \boldsymbol{\xi}^*) \approx \boldsymbol{Q}(\boldsymbol{\xi}_0 - \boldsymbol{\xi}^*)$.

It follows that the sum of the second and third terms in the right hand side of (5.25) has approximately a normal distribution with mean nF^* and variance $n\mathbf{g}_0'\boldsymbol{\Gamma}\mathbf{g}_0$. Therefore, for $\boldsymbol{\xi}_0$ close to Ξ_0 the difference between the noncentral chi-square and normal approximations, given in Theorems 5.2 and 5.3, respectively, is the first term in the right hand side of (5.25). The expected value and the variance of a chi-square random variable with ν degrees of freedom is equal to ν and 2ν, respectively. Therefore, when the number of degrees of freedom $\nu = \text{rank}(\mathbf{\Gamma}) - \dim(\mathcal{T})$ is bigger or comparable with the noncentrality parameter $\delta \approx nF^*$, the noncentral chi-square approximation, which is based on a second-order Taylor expansion at the point $\boldsymbol{\xi}^*$, should be better than the normal approximation, which is based on a first-order approximation at $\boldsymbol{\xi}_0$. On the other hand, if δ is significantly bigger than ν, then the first term in the right hand side of (5.25) becomes negligible and the normal approximation could be reasonable. This is in agreement with the property that a noncentral chi-square distribution, with ν degrees of freedom and noncentrality parameter δ, becomes approximately normal if δ is much bigger than ν. In such a case the normal approximation can be used to construct a confidence interval, for F^*, of the form $\hat{F} \pm \kappa\, \hat{\sigma}_F$. Here $\hat{\sigma}_F$ is an estimate of the standard deviation of $\hat{F}$ and κ is a critical value. Recall that the expected value and variance of a noncentral chi-square random variable, with ν degrees of freedom and noncentrality parameter δ, is $\nu + \delta$ and $2\nu + 4\delta$, respectively. Therefore, the normal approximation could be reasonable if

$$\frac{n\hat{F} - \nu}{\sqrt{4n\hat{F} + 2\nu}} \geq \kappa, \tag{5.28}$$

where κ is a critical value, say $\kappa = 3$.

Let us finally remark that from a theoretical point of view one can obtain a better approximation of the distribution of the MDF test statistic by using a second-order Taylor expansion of the optimal value function at the population point $\boldsymbol{\xi}_0$. The corresponding first and second order derivatives are given in (5.13) and (5.15), respectively, provided that the optimal solution $\boldsymbol{\xi}^*$ is unique. Note, however, that in practical applications this will require an accurate estimation of the corresponding first and second order derivatives which could be a problem.

5.2 Nested Models

Suppose now that we have two models $\Xi_1 \subset \Xi$ and $\Xi_2 \subset \Xi$ for the same parameter vector $\boldsymbol{\xi}$. It is said that the second model is *nested*, within the first model, if Ξ_2 is a subset of Ξ_1, i.e., $\Xi_2 \subset \Xi_1$. We refer to the models associated with the sets Ξ_1 and Ξ_2 as *full* and *restricted* models, respectively. If Ξ_1 is given in the parametric form

$$\Xi_1 := \left\{ \boldsymbol{\xi} \in \Xi : \boldsymbol{\xi} = \boldsymbol{g}(\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta_1 \right\}, \tag{5.29}$$

i.e., the full model is structural, then it is natural to define a nested model by restricting the parameter space Θ_1 to a subset Θ_2. Typically the subset $\Theta_2 \subset \Theta_1$ is defined by imposing constraints on the parameter vector $\boldsymbol{\theta}$. In this section we discuss asymptotics of the MDF test statistics $n\hat{F}_i$, $i = 1, 2$, where

$$\hat{F}_i := \min_{\boldsymbol{\xi} \in \Xi_i} F(\hat{\boldsymbol{\xi}}_n, \boldsymbol{\xi}). \tag{5.30}$$

Suppose that the population value $\boldsymbol{\xi}_0 \in \Xi_2$, i.e., the restricted model holds. Suppose, further, that the sets Ξ_1 and Ξ_2 are approximated at $\boldsymbol{\xi}_0$ by linear spaces $\mathcal{T}_1$ and $\mathcal{T}_2$, respectively. This holds if both Ξ_1 and Ξ_2 are smooth manifolds near $\boldsymbol{\xi}_0$ with respective tangent spaces $\mathcal{T}_1$ and $\mathcal{T}_2$. Note that $\mathcal{T}_2 \subset \mathcal{T}_1$ since $\Xi_2 \subset \Xi_1$. We have then (see (5.16)) that

$$n\hat{F}_i = \min_{\boldsymbol{\zeta} \in \mathcal{T}_i} (\boldsymbol{Z}_n - \boldsymbol{\zeta})' \boldsymbol{V}_0 (\boldsymbol{Z}_n - \boldsymbol{\zeta}) + o_p(1), \quad i = 1, 2. \tag{5.31}$$

Suppose, further, that the discrepancy function is *correctly specified.* Then by the analysis of the previous section we have that $n\hat{F}_i$, $i = 1, 2$, converges in distribution to a (central) chi-square with $\nu_i = \text{rank}(\boldsymbol{\Gamma}) - \dim(\mathcal{T}_i)$ degrees of freedom. Moreover, it follows from the representation (5.31) that $n\hat{F}_1$ and $n\hat{F}_2 - n\hat{F}_1$ are asymptotically independent of each other. The corresponding arguments are analogous to derivations of the statistical inference of linear constraints in the theory of linear models (e.g., [24, Section 4.5.1]). This can be extended to the setting of a sequence of local alternatives, where there is a sequence $\boldsymbol{\xi}_{0,n}$ of population values converging to a point $\boldsymbol{\xi}^* \in \Xi_2$ such that the following limits exist

$$\delta_i = \lim_{n \to \infty} \left[n \min_{\boldsymbol{\xi} \in \Xi_i} F(\boldsymbol{\xi}_{0,n}, \boldsymbol{\xi}) \right], \quad i = 1, 2. \tag{5.32}$$

Then the following asymptotic results hold (Steiger et al [36]).

Theorem 5.4 *Suppose that the assumption of a sequence of local alternatives holds, the discrepancy function is correctly specified and the sets Ξ_i, $i = 1, 2$, are approximated at $\boldsymbol{\xi}^* \in \Xi_2$ by respective linear spaces $\mathcal{T}_i$. Then the following holds:* (i) *the MDF test statistics $n\hat{F}_i$ converge in distribution to noncentral chi-square with respective degrees of freedom $\nu_i = \text{rank}(\boldsymbol{\Gamma}) - \dim(\mathcal{T}_i)$ and noncentrality parameter δ_i given in* (5.32), (ii) *the statistic $n\hat{F}_2 - n\hat{F}_1$ converges in distribution to a noncentral chi-square with $\nu_2 - \nu_1$ degrees of freedom and noncentrality parameter $\delta_2 - \delta_1$,* (iii) *the statistics $n\hat{F}_1$ and $n\hat{F}_2 - n\hat{F}_1$ are asymptotically independent of each other,* (iv) *the ratio statistic $\frac{(\hat{F}_2 - \hat{F}_1)/(\nu_2 - \nu_1)}{\hat{F}_1/\nu_1}$ converges in distribution to doubly noncentral F-distribution with noncentrality parameters $\delta_2 - \delta_1$ and δ_1 and with $\nu_2 - \nu_1$ and ν_1 degrees of freedom.*

It is straightforward to extend the above result to a sequence of nested models. Also we have that $n\hat{F}_2 = n\hat{F}_1 + (n\hat{F}_2 - n\hat{F}_1)$. Recall that the variance of a noncentral chi-square random variable with ν degrees of freedom and noncentrality parameter δ is $2\nu + 4\delta$. Therefore, under the assumptions of the above theorem, the asymptotic covariance between $n\hat{F}_1$ and $n\hat{F}_2$ is equal to the asymptotic variance of $n\hat{F}_1$, which is equal to $2\nu_1 + 4\delta_1$. Consequently, the asymptotic correlation between the MDF statistics $n\hat{F}_1$ and $n\hat{F}_2$ is equal to $\sqrt{\frac{\nu_1 + 2\delta_1}{\nu_2 + 2\delta_2}}$ (cf., [36]).

5.3 Asymptotics of MDF Estimators

In this section we discuss asymptotics of the MDF estimator $\hat{\boldsymbol{\theta}}_n$. Suppose that the model holds and $\hat{\boldsymbol{\theta}}_n$ is a consistent estimator of the population value $\boldsymbol{\theta}_0 \in \Theta$ (see section 4 and Proposition 4.2 in particular). Since $\hat{\boldsymbol{\theta}}_n = \bar{\boldsymbol{\theta}}(\hat{\boldsymbol{\xi}}_n)$, we have by (5.9) that, under the assumptions of Proposition 5.2,

$$n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \bar{\boldsymbol{\beta}}(\boldsymbol{Z}_n) + o_p(1). \tag{5.33}$$

Recall that $\bar{\boldsymbol{\beta}}(\boldsymbol{z})$ is the optimal solution of (5.10) and note that $\bar{\boldsymbol{\beta}}(\cdot)$ is positively homogeneous, i.e., $\bar{\boldsymbol{\beta}}(t\boldsymbol{z}) = t\bar{\boldsymbol{\beta}}(\boldsymbol{z})$ for any $\boldsymbol{z}$ and $t \geq 0$. This leads to the following asymptotics of the MDF estimator (cf., [7],[25]).

Theorem 5.5 *Suppose that the model holds, $\hat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}_0$, the set Θ is approximated at $\boldsymbol{\theta}_0$ by a convex cone $\mathcal{C}$ and* $\operatorname{rank}(\boldsymbol{\Delta}_0) = q$*. Then $n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \Rightarrow \bar{\boldsymbol{\beta}}(\boldsymbol{Z})$, where $\boldsymbol{Z} \sim N(0, \boldsymbol{\Gamma})$. If, furthermore, $\boldsymbol{\theta}_0$ is an interior point of Θ, and hence $\mathcal{C} = \mathbb{R}^q$, then $n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ converges in distribution to normal with mean vector zero and covariance matrix*

$$\boldsymbol{\Pi} = (\boldsymbol{\Delta}_0' \boldsymbol{V}_0 \boldsymbol{\Delta}_0)^{-1} \boldsymbol{\Delta}_0' \boldsymbol{V}_0 \boldsymbol{\Gamma} \boldsymbol{V}_0 \boldsymbol{\Delta}_0 (\boldsymbol{\Delta}_0' \boldsymbol{V}_0 \boldsymbol{\Delta}_0)^{-1}. \tag{5.34}$$

Moreover, if the discrepancy function is correctly specified, then $\boldsymbol{\Pi} = (\boldsymbol{\Delta}_0' \boldsymbol{V}_0 \boldsymbol{\Delta}_0)^{-1}$.

In particular, if in the setting of covariance structures the population distribution is normal and the employed discrepancy function is normal-theory correctly specified, then the asymptotic covariance matrix of $n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ can be written as

$$\boldsymbol{\Pi}_N = 2\left[\boldsymbol{\Delta}_0' (\boldsymbol{\Sigma}_0^{-1} \otimes \boldsymbol{\Sigma}_0^{-1}) \boldsymbol{\Delta}_0\right]^{-1}. \tag{5.35}$$

Note that since it is assumed that the asymptotic covariance matrix $\boldsymbol{\Gamma}$ has the maximal rank, and hence the linear space generated by columns of $\boldsymbol{\Delta}_0$ is included in the linear space generated by columns of $\boldsymbol{\Gamma}$, we have here that matrix $\boldsymbol{\Delta}_0' \boldsymbol{\Gamma}^- \boldsymbol{\Delta}_0$ is independent of a particular choice of the generalized inverse $\boldsymbol{\Gamma}^-$ and is positive definite. In particular, if the discrepancy function is correctly specified, then $\boldsymbol{\Delta}_0' \boldsymbol{\Gamma}^- \boldsymbol{\Delta}_0 = \boldsymbol{\Delta}_0' \boldsymbol{V}_0 \boldsymbol{\Delta}_0$. It is possible to show (cf., [7, Proposition 3]) that the inequality[16]

$$\boldsymbol{\Pi} \succeq (\boldsymbol{\Delta}_0' \boldsymbol{\Gamma}^- \boldsymbol{\Delta}_0)^{-1} \tag{5.36}$$

always holds. Basically this is the Gauss-Markov Theorem. That is, for a correctly specified discrepancy function, the asymptotic covariance matrix of the corresponding MDF estimator attains its lower bound given by the right hand side of (5.36). Therefore, for a correctly specified discrepancy function the corresponding MDF estimator is *asymptotically efficient* within the class of MDF estimators.

The above asymptotics of MDF estimators were derived under the assumption of identifiability of the model. If the model is overparameterized, then it does not make sense to talk about distribution of the MDF estimators since these estimators are not uniquely defined. However, even in the case of overparameterization some of the parameters could be defined uniquely. Therefore it makes sense to consider the following concept of *estimable functions* borrowed from the theory of linear models (e.g., [24, section 3.8.2]).

[16] For $q \times q$ symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ the inequality $\boldsymbol{A} \succeq \boldsymbol{B}$ is understood in the Loewner sense, i.e., that matrix $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite.

Definition 5.2 *Consider a continuously differentiable function* $a(\boldsymbol{\theta})$*. We say that* $\alpha = a(\boldsymbol{\theta})$ *is an* estimable *parameter if* $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$ *and* $\boldsymbol{g}(\boldsymbol{\theta}_1) = \boldsymbol{g}(\boldsymbol{\theta}_2)$ *imply that* $a(\boldsymbol{\theta}_1) = a(\boldsymbol{\theta}_2)$*. If this holds in a neighborhood of a point* $\boldsymbol{\theta}_0 \in \Theta$*, we say* α *is* locally estimable*, near* $\boldsymbol{\theta}_0$*.*

In the analysis of covariance structures the above concept of estimable parameters was discussed in Shapiro [31, p.146]. By using local reparameterization (see Proposition 2.1) it is possible to show the following.

> If $\boldsymbol{\theta}_0$ is a locally regular interior point of Θ, then a parameter $\alpha = a(\boldsymbol{\theta})$ is locally estimable, near $\boldsymbol{\theta}_0$, iff vector $\partial a(\boldsymbol{\theta})/\partial \boldsymbol{\theta}'$ belongs to the linear space generated by rows of $\boldsymbol{\Delta}(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$.

Consider now vector $\boldsymbol{a}(\boldsymbol{\theta}) = (a_1(\boldsymbol{\theta}), ..., a_s(\boldsymbol{\theta}))$ of locally estimable, near a population point $\boldsymbol{\theta}_0$, parameters. Suppose that $\boldsymbol{\theta}_0$ is locally regular and let $\boldsymbol{\alpha}_0 := \boldsymbol{a}(\boldsymbol{\theta}_0)$ and $\hat{\boldsymbol{\alpha}}_n := \boldsymbol{a}(\hat{\boldsymbol{\theta}}_n)$. Note that, by local estimability of $\boldsymbol{\alpha}$, the estimator $\hat{\boldsymbol{\alpha}}_n$ is defined uniquely for $\hat{\boldsymbol{\theta}}_n$ sufficiently close to $\boldsymbol{\theta}_0$. We have then that $n^{1/2}(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0)$ converges in distribution to normal with mean vector zero and covariance matrix

$$\boldsymbol{A}_0(\boldsymbol{\Delta}_0'\boldsymbol{V}_0\boldsymbol{\Delta}_0)^{-}\boldsymbol{\Delta}_0'\boldsymbol{V}_0\boldsymbol{\Gamma}\boldsymbol{V}_0\boldsymbol{\Delta}_0(\boldsymbol{\Delta}_0'\boldsymbol{V}_0\boldsymbol{\Delta}_0)^{-}\boldsymbol{A}_0', \tag{5.37}$$

where $\boldsymbol{A}_0 := \partial \boldsymbol{a}(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}'$ is $s \times q$ Jacobian matrix. Note that because of the local estimability of $\boldsymbol{\alpha}$, we have that row vectors of $\boldsymbol{A}_0$ belong to the linear space generated by rows of $\boldsymbol{\Delta}_0$, and hence the expression in (5.37) does not depend on a particular choice of the generalized inverse of $\boldsymbol{\Delta}_0'\boldsymbol{V}_0\boldsymbol{\Delta}_0$. In particular, for correctly specified discrepancy function this expression becomes $\boldsymbol{A}_0(\boldsymbol{\Delta}_0'\boldsymbol{V}_0\boldsymbol{\Delta}_0)^{-}\boldsymbol{A}_0'$.

We can also consider a situation when the model does not hold. Under the assumptions of Proposition 5.4, in particular that $\boldsymbol{\theta}^*$ is the unique optimal solution of problem (4.3), we have that $\hat{\boldsymbol{\theta}}_n$ converges w.p.1 to $\boldsymbol{\theta}^*$ and, by (5.14), that (cf., [25, Theorem 5.4]):

$$n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \Rightarrow N\left(0, \boldsymbol{H}_{\theta\theta}^{-1}\boldsymbol{H}_{\theta x}\boldsymbol{\Gamma}\boldsymbol{H}_{\theta x}'\boldsymbol{H}_{\theta\theta}^{-1}\right). \tag{5.38}$$

5.4 The Case of Boundary Population Value

In the previous sections we discussed asymptotics of MDF test statistics and estimators under the assumption that the set Ξ_0 can be approximated, at a considered point, by a *linear* space $\mathcal{T}$. In this section we consider a situation when $\boldsymbol{\theta}_0$ is a boundary point of the set Θ, and as a consequence the set Ξ_0 should be approximated at $\boldsymbol{\xi}_0 = \boldsymbol{g}(\boldsymbol{\theta}_0)$ by a (convex) cone rather than a linear space. This may happen if the set Θ is defined by inequality constraints and some of these inequality constraints are active at the population point. For instance, in example 2.1 (Factor Analysis model) the diagonal entries of the matrix $\boldsymbol{\Psi}$ should be nonnegative. If the population value $\boldsymbol{\Psi}_0$ has zero diagonal entries (i.e., some residual variances are zeros), then the corresponding value of the parameter vector can be viewed as lying on the boundary of the feasible region. One can think about more sophisticated examples where, for instance, it is hypothesized that some of the residual variances (i.e., diagonal entries of $\boldsymbol{\Psi}$) are bigger than the others, or that elements of matrix $\boldsymbol{\Lambda}$ are nonnegative. Statistical theory of parameters estimation under inequality type constraints is often referred to as order restricted statistical inference. The interested reader can be referred to the recent comprehensive monograph [35] for a thorough treatment of that theory. We give below a few basic results which are relevant for our discussion.

Suppose that the model holds, the Jacobian matrix $\boldsymbol{\Delta}_0$ has full column rank q and the set Θ is approximated at $\boldsymbol{\theta}_0$ by convex cone $\mathcal{C}$. We have then that

$$n\hat{F} \Rightarrow \min_{\boldsymbol{\beta}\in\mathcal{C}} (\boldsymbol{Z} - \boldsymbol{\Delta}_0\boldsymbol{\beta})'\boldsymbol{V}_0(\boldsymbol{Z} - \boldsymbol{\Delta}_0\boldsymbol{\beta}), \tag{5.39}$$

where $\boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{\Gamma})$. Let us look at the optimal value of the minimization problem in the right hand side of (5.39). Suppose, for the sake of simplicity, that $\boldsymbol{\Delta}_0$ has full column rank q. Then we can decompose this optimal value into a sum of two terms as follows

$$\min_{\boldsymbol{\beta}\in\mathcal{C}} (\boldsymbol{Z} - \boldsymbol{\Delta}_0\boldsymbol{\beta})'\boldsymbol{V}_0(\boldsymbol{Z} - \boldsymbol{\Delta}_0\boldsymbol{\beta}) = (\boldsymbol{Z} - \boldsymbol{\Delta}_0\tilde{\boldsymbol{\beta}})'\boldsymbol{V}_0(\boldsymbol{Z} - \boldsymbol{\Delta}_0\tilde{\boldsymbol{\beta}}) + \min_{\boldsymbol{\beta}\in\mathcal{C}} (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{\Delta}_0'\boldsymbol{V}_0\boldsymbol{\Delta}_0(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}), \tag{5.40}$$

where $\tilde{\boldsymbol{\beta}} = (\boldsymbol{\Delta}_0'\boldsymbol{V}_0\boldsymbol{\Delta}_0)^{-1}\boldsymbol{\Delta}_0'\boldsymbol{V}_0\boldsymbol{Z}$ is the corresponding unconstrained minimizer (compare with (5.11)). The first term in the right hand side of (5.40) is the corresponding unconstrained minimum over $\boldsymbol{\beta} \in \mathbb{R}^q$, and in a sense the above decomposition is just the Pythagorus Theorem. Suppose, further, that the discrepancy function is correctly specified. Recall that it is assumed that the asymptotic covariance matrix $\boldsymbol{\Gamma}$ has maximal rank. By reducing the saturated space, if necessary, we can assume here that $\boldsymbol{\Gamma}$ is nonsingular, and hence "correctly specified" means that $\boldsymbol{V}_0 = \boldsymbol{\Gamma}^{-1}$. It follows then that $\tilde{\boldsymbol{\beta}} \sim N\left(\boldsymbol{0}, (\boldsymbol{\Delta}_0'\boldsymbol{V}_0\boldsymbol{\Delta}_0)^{-1}\right)$.

Assuming that the model holds and $\boldsymbol{V}_0 = \boldsymbol{\Gamma}^{-1}$, we have that the first term in the right hand side of (5.40) has chi-square distribution, with $\nu = m - q$ degrees of freedom, and is distributed independently of the second term. The second term in the right hand side of (5.40) is a *mixture* of chi-square distributions (such distributions are called *chi-bar-squared* distributions). With various degrees of generality this result was derived in [4, 18, 22], it was shown in [30] that this holds for any *convex* cone $\mathcal{C}$. Denote by $n\tilde{F}$ the corresponding *unconstrained* MDF test statistic, i.e.,

$$\tilde{F} := \min_{\boldsymbol{\theta}\in\mathbb{R}^q} F(\hat{\boldsymbol{\xi}}_n, \boldsymbol{g}(\boldsymbol{\theta})).$$

We have that $n\tilde{F}$ is asymptotically equivalent to the first term in the right hand side of (5.40). Under the above assumptions we obtain the following results:

(i) The unrestricted MDF test statistic $n\tilde{F}$ converges in distribution to chi-square with $\nu = m - q$ degrees of freedom and is asymptotically independent of the difference statistic $n\hat{F} - n\tilde{F}$.

(ii) The difference statistic $n\hat{F} - n\tilde{F}$ is asymptotically equivalent to the second term in the right hand side of (5.40) and converges in distribution to a mixture of chi-square distributions, that is,

$$\lim_{n\to\infty} \text{Prob}\left\{n\hat{F} - n\tilde{F} \geq c\right\} = \sum_{i=0}^{q} w_i \,\text{Prob}\left\{\chi_i^2 \geq c\right\}, \tag{5.41}$$

where χ_i^2 is a chi-square random variable with i degrees of freedom, $\chi_0^2 \equiv 0$ and w_i are nonnegative weights such that $w_0 + ... + w_q = 1$.

Of course, the asymptotic distribution, given by the right hand side of (5.41), depends on the weights w_i, which in turn depend on the covariance matrix[17] of $\tilde{\boldsymbol{\beta}}$ and cone $\mathcal{C}$. A general property of these weights is that $\sum_{i=0}^{q}(-1)^i w_i = 0$ if at least two of these weights are nonzeros. If $\boldsymbol{\theta}_0$ is an interior point of Θ, and hence $\mathcal{C} = \mathbb{R}^q$, then $w_0 = 1$ and all other weights are zeros. In that case we have the same asymptotics of the MDF statistic $n\hat{F}$ as given in Theorem 5.1. Often the set $\Theta \subset \mathbb{R}^q$ is defined by inequality constraints. Then, under mild regularity conditions (called *constraint qualifications*), the approximating cone $\mathcal{C}$ is obtained by linearizing the active at $\boldsymbol{\theta}_0$ inequality constraints. In particular, if only one inequality constraint is active at $\boldsymbol{\theta}_0$, then $\mathcal{C}$ is defined by one linear inequality constraint and hence is a half space of $\mathbb{R}^q$. In that case $w_0 = w_1 = 1/2$ and all other weights are zeros. If two inequality constraints are active at $\boldsymbol{\theta}_0$, then only weights w_0, w_1 and w_2 can be nonzeros, with $w_1 = 1/2$, etc. For a general discussion of how to calculate these weights we can refer, e.g., to [34, 35].

6 Asymptotic Robustness of the MDF Statistical Inference

An important condition in the analysis of the previous section was the assumption of correct specification of the discrepancy function. In particular, the discrepancy functions F_{ML} and F_{GLS}, defined in (3.5) and (3.9), respectively, are motivated by the assumption that the underline population has a normal distribution and are correctly specified in that case. Nevertheless, these discrepancy functions are often applied in situations where the normality assumption has no justification or even can be clearly wrong. It turns out, however, that the asymptotic chi-square distribution of MDF test statistics, discussed in Theorems 5.1 and 5.2, can hold under considerably more general conditions than correct specification of the discrepancy function. This was discovered in Amemiya and Anderson [1] and Anderson and Amemiya [3] for a class of factor analysis models, and in Browne [10], Browne and Shapiro [11] and Shapiro [32] for general linear models, by using approaches based on different techniques. In this section we are going to discuss this theory following the Browne-Shapiro approach, which in a sense is more general although uses a slightly stronger assumption of existence of fourth-order moments. As in the previous section we assume that $n^{1/2}(\hat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_0) \Rightarrow N(\mathbf{0}, \boldsymbol{\Gamma})$.

[17] Recall that, for correctly specified discrepancy function, the covariance matrix of $\tilde{\boldsymbol{\beta}}$ is $(\boldsymbol{\Delta}_0' \boldsymbol{V}_0 \boldsymbol{\Delta}_0)^{-1}$.

Let us start with the following algebraic result (Shapiro [32, Theorem 3.1]). It is based on a verification that the corresponding quadratic form has a chi-square distribution.

Proposition 6.1 *Suppose that the assumption of a sequence of local alternatives holds and the set* Ξ_0 *is approximated at the point* $\boldsymbol{\xi}^* = \boldsymbol{g}(\boldsymbol{\theta}^*)$ *by a linear space* $\mathcal{T}$ *generated by the columns of the matrix*[18] $\boldsymbol{\Delta} = \boldsymbol{\Delta}(\boldsymbol{\theta}^*)$ *(e.g., the point* $\boldsymbol{\theta}^*$ *is regular). Suppose, further, that the discrepancy function is correctly specified with respect to an* $m \times m$ *positive semidefinite matrix* $\boldsymbol{\Gamma}_0$ *of maximal rank. Then* $n\hat{F}$ *is asymptotically chi-squared, with degrees of freedom* ν *and the noncentrality parameter* δ *given in* (5.18) *and* (5.21)*, respectively, if and only if* $\boldsymbol{\Gamma}$ *is representable in the form*

$$\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_0 + \boldsymbol{\Delta}\boldsymbol{C}' + \boldsymbol{C}\boldsymbol{\Delta}', \tag{6.1}$$

where $\boldsymbol{C}$ *is an arbitrary* $m \times q$ *matrix.*

In particular, we have that for the normal-theory discrepancy functions F_{ML} and F_{GLS} (in the analysis of covariance structures) the MDF test statistics are asymptotically chi-squared if and only if the corresponding $p^2 \times p^2$ asymptotic covariance matrix $\boldsymbol{\Gamma}$ can be represented in the form

$$\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_N + \boldsymbol{\Delta}\boldsymbol{C}' + \boldsymbol{C}\boldsymbol{\Delta}', \tag{6.2}$$

where matrix $\boldsymbol{\Gamma}_N$ is defined in (5.1).

The representation (6.1) is slightly more general than the following representation

$$\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_0 + \boldsymbol{\Delta}\boldsymbol{D}\boldsymbol{\Delta}', \tag{6.3}$$

where $\boldsymbol{D}$ is an arbitrary $q \times q$ symmetric matrix. Clearly, (6.3) is a particular form of the representation (6.1) with $\boldsymbol{C} := \frac{1}{2}\boldsymbol{\Delta}\boldsymbol{D}$. It turns out that under various structural assumptions, in the analysis of covariance structures, it is possible to show that the corresponding asymptotic covariance matrix $\boldsymbol{\Gamma}$ is of the form

$$\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_N + \boldsymbol{\Delta}\boldsymbol{D}\boldsymbol{\Delta}', \tag{6.4}$$

and hence to verify that the normal-theory MDF test statistics have asymptotically chi-square distributions. We also have the following result about asymptotic robustness of MDF estimators (Shapiro [32, Corollary 5.4]).

Proposition 6.2 *Suppose that the model holds, the set* Ξ_0 *is approximated at the point* $\boldsymbol{\xi}_0$ *by a linear space* $\mathcal{T}$ *generated by the columns of the matrix* $\boldsymbol{\Delta} = \boldsymbol{\Delta}(\boldsymbol{\theta}_0)$*, the Jacobian matrix* $\boldsymbol{\Delta}$ *has full column rank* q*, the MDF estimator* $\hat{\boldsymbol{\theta}}_n$ *is a consistent estimator of* $\boldsymbol{\theta}_0$*, the discrepancy function is correctly specified with respect to an* $m \times m$ *positive semidefinite matrix* $\boldsymbol{\Gamma}_0$ *of maximal rank, and the representation* (6.3) *holds. Then* $n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ *converges in distribution to normal*

[18] For the sake of notational convenience we drop here the superscript of the Jacobian matrix $\boldsymbol{\Delta}^*$.

$N(\mathbf{0}, \mathbf{\Pi})$*, the MDF estimator* $\hat{\boldsymbol{\theta}}_n$ *is asymptotically efficient within the class of MDF estimators, and*

$$\mathbf{\Pi} = \mathbf{\Pi}_0 + \boldsymbol{D}, \tag{6.5}$$

where $\mathbf{\Pi}_0 := (\mathbf{\Delta}'\boldsymbol{V}_0\mathbf{\Delta})^{-1}$.

We have that if the representation (6.3) holds, then the MDF test statistics designed for the asymptotic covariance matrix $\mathbf{\Gamma}_0$ still have asymptotically a chi-square distribution (under a sequence of local alternatives) and the asymptotic covariance matrix of the corresponding MDF estimators needs a simple correction given by formula (6.5). In the remainder of this section we discuss situations in the analysis of covariance structures which lead to the representation (6.4).

We assume below, in the remainder of this section, the setting of the analysis of covariance structures, with structural model $\mathbf{\Sigma} = \mathbf{\Sigma}(\boldsymbol{\theta})$ and with $\mathbf{\Gamma}$ being the $p^2 \times p^2$ asymptotic covariance matrix of $n^{1/2}(\boldsymbol{s} - \boldsymbol{\sigma}_0)$, where $\boldsymbol{s} := \mathrm{vec}(\boldsymbol{S})$ and $\boldsymbol{\sigma}_0 := \mathrm{vec}(\mathbf{\Sigma}_0)$. We assume that the underlying population has finite fourth-order moments, and hence the asymptotic covariance matrix $\mathbf{\Gamma}$ is well defined. As before, we denote by $\mathbf{\Gamma}_N$ and $\mathbf{\Pi}_N$ the normal-theory asymptotic covariance matrices given in (5.1) and (5.35), respectively. We also assume that the employed discrepancy function is *correctly specified with respect to a normal distribution* of the data, i.e., $\boldsymbol{V}_0$ is a generalized inverse of $\mathbf{\Gamma}_N$. Recall that the normal-theory discrepancy functions F_{ML} and F_{GLS} satisfy this property.

6.1 Elliptical distributions

In this section we assume that the underlying population has an *elliptical* distribution. We may refer to [20] for a thorough discussion of elliptical distributions. In the case of elliptical distributions the asymptotic covariance matrix $\mathbf{\Gamma}$ has the following structure:

$$\mathbf{\Gamma} = \alpha\mathbf{\Gamma}_N + \beta\boldsymbol{\sigma}_0\boldsymbol{\sigma}_0', \tag{6.6}$$

where $\alpha = 1 + \kappa$, $\beta = \kappa$, and κ is the kurtosis parameter of a considered elliptical distribution. This basic asymptotic result was employed in the studies of Muirhead and Waternaux [21], Tyler [37, 38] and Browne [8, 9].

It can be seen that the corrected covariance matrix $\alpha^{-1}\mathbf{\Gamma}$ has the structure specified in equation (6.4), provided that $\boldsymbol{\sigma}_0$ can be represented as a linear combination of columns of the Jacobian matrix $\mathbf{\Delta} = \mathbf{\Delta}(\boldsymbol{\theta}_0)$. If the point $\boldsymbol{\theta}_0$ is regular, and hence Ξ_0 is a smooth manifold near $\boldsymbol{\sigma}_0 = \boldsymbol{\sigma}(\boldsymbol{\theta}_0)$ and the tangent space $\mathcal{T}$, to Ξ_0 at $\boldsymbol{\sigma}_0$, is the linear space generated by the columns of $\mathbf{\Delta}$ (i.e., can be written in the form (5.4)), then this condition is equivalent to the condition that $\boldsymbol{\sigma}_0 \in \mathcal{T}$. This, in turn, holds if the set Ξ_0 is *positively homogeneous*, i.e., it satisfies the property that if $\boldsymbol{\sigma} \in \Xi_0$ and $t > 0$, then $t\boldsymbol{\sigma} \in \Xi_0$. For structural models, positive homogeneity of Ξ_0 can be formulated in the following form (this condition was introduced in [33] and models satisfying this condition were called *invariant under a constant scaling factor* in [8]):

(C) For every $t > 0$ and $\boldsymbol{\theta} \in \Theta$ there exists $\boldsymbol{\theta}^* \in \Theta$ such that $t\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}(\boldsymbol{\theta}^*)$.

The above condition (C) is easy to verify and it holds for many models used in applications. For example, it holds for the factor analysis model (2.2). By the above discussion we have the following results, which in somewhat different forms were obtained in Tyler [38] and Browne [8, 9], and in the present form in Shapiro and Browne [33].

Theorem 6.1 *Suppose that the assumption of a sequence of local alternatives holds, the set* Ξ_0 *is approximated at the point* $\boldsymbol{\sigma}^* = \boldsymbol{\sigma}(\boldsymbol{\theta}^*)$ *by a linear space* $\mathcal{T}$*, the representation* (6.6) *holds with* $\alpha > 0$ *and that* $\boldsymbol{\sigma}_0 \in \mathcal{T}$*. Let* $\hat{\alpha}$ *be a consistent estimator of the parameter* α*. Then* $\hat{\alpha}^{-1} n\hat{F}$ *has asymptotically chi-squared distribution with* ν *degrees of freedom and the noncentrality parameter* $\alpha^{-1}\delta$*, where* ν *and* δ *are defined in* (5.18) *and* (5.21)*, respectively.*

Recall that the condition "$\boldsymbol{\sigma}_0 \in \mathcal{T}$", used in the above theorem, holds if the set Ξ_0 is positively homogeneous, which in turn is implied by condition (C) (invariance under a constant scaling factor). We also have the following result about asymptotic robustness of the MDF estimators (Shapiro and Browne [33]).

Theorem 6.2 *Suppose that the model holds, the set* Ξ_0 *is approximated at the point* $\boldsymbol{\sigma}_0 \in \Xi_0$ *by a linear space* $\mathcal{T}$ *generated by the columns of the matrix* $\boldsymbol{\Delta} = \boldsymbol{\Delta}(\boldsymbol{\theta}_0)$*, the MDF estimator* $\hat{\boldsymbol{\theta}}_n$ *is a consistent estimator of* $\boldsymbol{\theta}_0$*, the representation* (6.6) *holds with* $\alpha > 0$ *and that* $\boldsymbol{\sigma}_0 = \boldsymbol{\Delta}\boldsymbol{\zeta}$ *for some* $\boldsymbol{\zeta} \in \mathbb{R}^q$*. Then* $n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ *converges in distribution to normal* $N(\mathbf{0}, \boldsymbol{\Pi})$*, the MDF estimator* $\hat{\boldsymbol{\theta}}_n$ *is asymptotically efficient within the class of MDF estimators, and*

$$\boldsymbol{\Pi} = \alpha\boldsymbol{\Pi}_N + \beta\boldsymbol{\zeta}\boldsymbol{\zeta}'. \tag{6.7}$$

The vector $\boldsymbol{\zeta}$ can be obtained by solving the system of linear equations $\boldsymbol{\Delta}\boldsymbol{\zeta} = \boldsymbol{\sigma}_0$, which is consistent if $\boldsymbol{\sigma}_0 \in \mathcal{T}$, and in particular if the model is invariant under a constant scaling factor. In some cases $\boldsymbol{\zeta}$ is readily available. For example, if the model is linear in $\boldsymbol{\theta}$, then $\boldsymbol{\zeta} = \boldsymbol{\theta}_0$. Of course, in practice the (unknown) population value $\boldsymbol{\theta}_0$ should be replaced by its estimator $\hat{\boldsymbol{\theta}}$.

6.2 Linear latent variate models

Presentation of this section is based on Browne and Shapiro [11]. We assume here that the observed $p \times 1$ vector variate $\boldsymbol{X}$ can be written in the form

$$\boldsymbol{X} = \boldsymbol{\mu} + \sum_{i=1}^{s} \boldsymbol{A}_i \boldsymbol{z}_i, \tag{6.8}$$

where $\boldsymbol{\mu}$ is a $p \times 1$ mean vector, $\boldsymbol{z}_i$ is an (unobserved) $m_i \times 1$ vector variate and $\boldsymbol{A}_i$ is a (deterministic) $p \times m_i$ matrix of regression weights of $\boldsymbol{X}$ onto $\boldsymbol{z}_i$, $i = 1, ..., s$. We assume that random vectors $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ are independently distributed for all $i \neq j$ and have finite fourth-order moments. The above model implies the following structure of the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{X}$:

$$\boldsymbol{\Sigma} = \sum_{i=1}^{s} \boldsymbol{A}_i \boldsymbol{\Phi}_i \boldsymbol{A}_i', \tag{6.9}$$

where $\boldsymbol{\Phi}_i$ is the $m_i \times m_i$ covariance matrix of $\boldsymbol{z}_i$, $i = 1, ..., s$.

For example, consider the factor analysis model:

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\Lambda} \boldsymbol{f} + \boldsymbol{u}, \tag{6.10}$$

where $\boldsymbol{\Lambda}$ is a $p \times k$ matrix of factor loadings, $\boldsymbol{f} = (f_1, ..., f_k)'$ is a $k \times 1$ common vector variate, and $\boldsymbol{u} = (u_1, ..., u_p)'$ is a $p \times 1$ unique factor vector variate. It is assumed that random variables $f_1, ..., f_k, u_1, ..., u_p$, are mutually independently distributed, and hence random vectors $\boldsymbol{f}$ and $\boldsymbol{u}$ are independent. This implies that the covariance matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$, of $\boldsymbol{f}$ and $\boldsymbol{u}$, respectively, are diagonal. Of course, we can write model (6.10) in the form

$$\boldsymbol{X} = \boldsymbol{\mu} + \sum_{i=1}^{k} \boldsymbol{\Lambda}_i f_i + \sum_{j=1}^{p} \boldsymbol{E}_j u_j, \tag{6.11}$$

where $\boldsymbol{\Lambda}_i$ is the i-th column vector of $\boldsymbol{\Lambda}$ and $\boldsymbol{E}_j$ is the j-th coordinate vector. This shows that this model is a particular case of model (6.8) with $s = k + p$ and $m_i = 1$, $i = 1, ..., s$. Now model (6.10) (or, equivalently, model (6.11)) generates the following covariance structures model:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda} \boldsymbol{\Phi} \boldsymbol{\Lambda}' + \boldsymbol{\Psi}. \tag{6.12}$$

The only difference between the above model (6.12) and the model (2.2) is that in (2.2) the covariance matrix $\boldsymbol{\Phi}$ is assumed to be the identity matrix.

The linear model (6.8) implies the following structure of the asymptotic covariance matrix $\boldsymbol{\Gamma}$ (Browne and Shapiro [11, Theorem 2.1]):

$$\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_N + \sum_{i=1}^{s} (\boldsymbol{A}_i \otimes \boldsymbol{A}_i) \boldsymbol{C}_i (\boldsymbol{A}_i' \otimes \boldsymbol{A}_i'), \tag{6.13}$$

where $\boldsymbol{C}_i$ is the $m_i^2 \times m_i^2$ forth-order cumulant matrix of $\boldsymbol{z}_i$, $i = 1, ..., s$.

Suppose now that the weight matrices have parametric structures $\boldsymbol{A}_i = \boldsymbol{A}_i(\boldsymbol{\upsilon})$, $i = 1, ..., s$, where $\boldsymbol{\upsilon} \in \Upsilon$ is a parameter vector varying in space $\Upsilon \subset \mathbb{R}^\ell$. Then (6.9) becomes the following covariance structures model

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \sum_{i=1}^{s} \boldsymbol{A}_i(\boldsymbol{\upsilon}) \boldsymbol{\Phi}_i \boldsymbol{A}_i(\boldsymbol{\upsilon})', \tag{6.14}$$

with the parameter vector $\boldsymbol{\theta} := (\boldsymbol{\upsilon}', \boldsymbol{\varphi}_1', ..., \boldsymbol{\varphi}_s')'$, where $\boldsymbol{\varphi}_i := \text{vec}(\boldsymbol{\Phi}_i)$, $i = 1, ..., s$. Note that the only restriction on the $m_i^2 \times 1$ parameter vectors $\boldsymbol{\varphi}_i$ imposed here is that the corresponding covariance matrix $\boldsymbol{\Phi}_i$ should be positive semidefinite. Note also that if $m_i > 1$ for at least one i, then such choice of the parameter vector results in overparameterization of model (6.14) since $\boldsymbol{\varphi}_i$ will have duplicated elements. It is possible to include in the parameter vector $\boldsymbol{\theta}$ only nonduplicated elements of matrices $\boldsymbol{\Phi}_i$. This, however, is not essential at the moment.

By applying the 'vec' operator to both sides of the equation (6.14) we can write this model in the form

$$\boldsymbol{\sigma}(\boldsymbol{\theta}) = \sum_{i=1}^{s} \big(\boldsymbol{A}_i(\boldsymbol{v}) \otimes \boldsymbol{A}_i(\boldsymbol{v})\big)\boldsymbol{\varphi}_i. \tag{6.15}$$

It can be seen that the model is linear in parameters $\boldsymbol{\varphi}_i$, $i = 1, ..., s$, and

$$\frac{\partial \boldsymbol{\sigma}(\boldsymbol{\theta})}{\partial \boldsymbol{\varphi}_i'} = \boldsymbol{A}_i(\boldsymbol{v}) \otimes \boldsymbol{A}_i(\boldsymbol{v}). \tag{6.16}$$

That is, the corresponding Jacobian matrix can be written as

$$\boldsymbol{\Delta}(\boldsymbol{\theta}) = \big[\boldsymbol{\Delta}(\boldsymbol{v}), \boldsymbol{A}_1(\boldsymbol{v}) \otimes \boldsymbol{A}_1(\boldsymbol{v}), ..., \boldsymbol{A}_s(\boldsymbol{v}) \otimes \boldsymbol{A}_s(\boldsymbol{v})\big]. \tag{6.17}$$

Together with (6.13) this implies that the equation (6.4) holds with matrix

$$\boldsymbol{D} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{C}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{C}_2 & \cdots & \mathbf{0} \\ & \cdots & \cdots & \cdots & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{C}_s \end{bmatrix}. \tag{6.18}$$

We obtain the following result (Browne and Shapiro [11, Proposition 3.3]).

Theorem 6.3 *Consider the linear latent variate model* (6.8) *and the corresponding covariance structures model* (6.14). *Suppose that random vectors* $\boldsymbol{z}_i$, $i = 1, .., s$, *are mutually independently distributed, the assumption of a sequence of local alternatives (for the covariance structures model) holds, and* Ξ_0 *is a smooth manifold near the point* $\boldsymbol{\sigma}^*$. *Then the MDF test statistic* $n\hat{F}$ *has asymptotically noncentral chi-squared distribution with* $\nu = p(p+1)/2 - \operatorname{rank}(\boldsymbol{\Delta}_0)$ *degrees of freedom and the noncentrality parameter* δ.

In particular, the above theorem can be applied to the factor analysis model (6.10). Note that the MDF test statistics for the model (6.12), with the covariance term $\boldsymbol{\Phi}$, and model (2.2), without this term, are the same since $\boldsymbol{\Phi}$ can be absorbed into $\boldsymbol{\Lambda}$ and hence the corresponding set Ξ_0 is the same. Note also that in order to derive the asymptotic chi-squaredness of the MDF test statistics we only used the corresponding independence condition, no other assumptions about distributions of $\boldsymbol{f}$ and $\boldsymbol{u}$ were made (accept existence of fourth-order moments). For the factor analysis model this result was first obtained by Amemiya and Anderson [1] by employing different techniques and without the assumption of finite forth-order moments.

It is also possible to give corrections for the asymptotic covariance matrix $\boldsymbol{\Pi}$ of $n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ (compare with formula (6.5) of Proposition 6.2). In order to do that we need to verify identifiability of the parameter vector $\boldsymbol{\theta}$. Let us replace now $\boldsymbol{\varphi}_i$ with parameter vector $\boldsymbol{\varphi}_i^* :=$ vecs$(\boldsymbol{\Phi}_i)$, $i = 1, ..., s$, i.e., $\boldsymbol{\varphi}_i^*$ is $m_i(m_i+1)/2 \times 1$ vector formed from the nonduplicated elements of $\boldsymbol{\Phi}_i$. The equation (6.4) still holds with the matrix $\boldsymbol{D}$ reduced to a smaller matrix $\boldsymbol{D}^*$ by replacing each $m_i^2 \times m_i^2$ matrix $\boldsymbol{C}_i$ by the corresponding $m_i(m_i+1)/2 \times m_i(m_i+1)/2$ matrix $\boldsymbol{C}_i^*$ formed by the nonduplicated rows and columns of $\boldsymbol{C}_i$. We have then that, under the above assumptions,

$$\boldsymbol{\Pi} = \boldsymbol{\Pi}_N + \boldsymbol{D}^*. \tag{6.19}$$

It follows that the asymptotic covariance matrix of the MDF estimator $\hat{\boldsymbol{\upsilon}}$ is independent of the particular distribution of the $\boldsymbol{z}_i$, $i = 1, ..., s$, while the asymptotic covariance matrix of the MDF estimator $\hat{\varphi}_i^*$ needs the correction term $\boldsymbol{C}_i^*$ as compared with the normal case. Asymptotic covariances between $\hat{\boldsymbol{\upsilon}}$ and $\hat{\boldsymbol{\varphi}}_i^*$, and between $\hat{\boldsymbol{\varphi}}_i^*$ and $\hat{\boldsymbol{\varphi}}_j^*$, for $i \neq j$, are the same as in the normal case.

Remark 5 Suppose, furthermore, that the population value $\boldsymbol{\upsilon}_0$, of the parameter vector $\boldsymbol{\upsilon}$, lies on the *boundary* of the parameter space Υ, and that Υ is approximated at $\boldsymbol{\upsilon}_0$ by convex cone $\mathcal{C}$ (recall that the parameter vectors $\hat{\boldsymbol{\varphi}}_i^*$, $i = 1, ..., s$, are assumed to be unconstrained). Let $\tilde{\boldsymbol{\upsilon}}$ be the *unconstrained* MDF estimator of $\boldsymbol{\upsilon}_0$ (compare with the derivations of section 5.4). Then, under the above assumptions, $n^{1/2}(\tilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}_0) \Rightarrow N(\boldsymbol{0}, \boldsymbol{U})$, where the asymptotic covariance matrix $\boldsymbol{U}$ is independent of the particular distribution of the $\boldsymbol{z}_i$, $i = 1, ..., s$. We also have then that the MDF test statistic $n\hat{F}$ converges in distribution to the sum of two stochastically independent terms (compare with equations (5.39) and (5.40)), one term having the usual chi-square distribution and the other term given by $\min_{\boldsymbol{\upsilon} \in \mathcal{C}} (\tilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon})' \boldsymbol{U}^{-1} (\tilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon})$. It follows that the asymptotic distribution of the MDF test statistic $n\hat{F}$ is chi-bar-squared and is independent of the particular distribution of the $\boldsymbol{z}_i$, $i = 1, ..., s$. That is, under these assumptions, distribution of the MDF test statistic is again asymptotically robust.

References

[1] Amemiya, Y. and Anderson, T.W. (1990), Asymptotic chi-square tests for a large class of factor analysis models, *Annals of Statistics*, 18, 1453–1463.

[2] Anderson, T.W. and Rubin, H. (1956), Statistical inference in factor analysis, in *Proceedings of the Third Berkley Symposium on Mathematical Statistics and Probability*, vol. 5, J. Neyman (ed.), Berkeley, CA: University of California Press, pp. 111-150.

[3] Anderson, T. W. and Amemiya, Y. (1988), The Asymptotic normal distribution of estimators in Factor Analysis under general conditions, *Annals of Statistics*, 16, 759-771.

[4] Bartholomew, D.J. (1961), Ordered tests in the analysis of variance, *Biometrika*, 48, 325-332.

[5] Bekker, P.A. and ten Berge J.M.F. (1997), Generic global identification in factor analysis, *Linear Algebra and its Applications*, 264, 255-263.

[6] Bonnans, J.F. and Shapiro, A. (2000). *Perturbation Analysis of Optimization Problems.* New York: Springer-Verlag.

[7] Browne, M. W. (1974), Generalized least squares estimators in the analysis of covariance structures, *South African Statistical Journal*, 8, 1-24.

[8] Browne, M.W. (1982), Covariance structures, in: D.M. Hawkins (ed.), *Topics in Applied Multivariate Analysis*, Cambridge University Press.

[9] Browne, M.W. (1984), Asymptotically distribution-free methods in the analysis of covariance structures, *British Journal of Mathematical and Statistical Psychology*, 37, 62-83.

[10] Browne, M.W. (1987), Robustness in statistical inference in factor analysis and related models, *Biometrika*, 74, 375-384.

[11] Browne, M.W. and Shapiro, A. (1988), Robustness of normal theory methods in the analysis of linear latent variate models, *British Journal of Mathematical and Statistical Psychology*, 41, 193-208.

[12] Chernoff, H. (1954), On the distribution of the likelihood ratio, *Ann. Math. Statist.*, 25, 573-578.

[13] Danskin, J.M. (1967), *The Theory of Max-Min and its Applications to Weapon Allocation Problems*, Econometrics and Operations Research **5**, Berlin and New York, Springer-Verlag.

[14] Fisher, F.M. (1966), *The Identification Problem in Econometrics*, McGraw-Hill, New York.

[15] Jöreskog, K.G. (1977), Structural equation models in the social sciences: specification, estimation and testing. In P.R. Krishnaiah (ed.), *Applications of Statistics*, pp. 265-287, Amsterdam, North Holland.

[16] Jöreskog, K.G. (1981), Analysis of covariance structures (with discussion), *Scandinavian Journal of Statistics*, 8, 65-92.

[17] Kano, Y. (1986), Conditions on consistency of estimators in covariance structure model, *Journal of the Japan Statistical Society*, 16, 75-80.

[18] Kudô, A. (1963), A multivariate analogue of the one-sided test, *Biometrika*, 50, 403-418.

[19] Ledermann, W. (1937), On the rank of reduced correlation matrices in multiple factor analysis. *Psychometrika*, 2, 85-93.

[20] Muirhead, R.J. (1982), *Aspects of Multivariate Statistical Theory*, Wiley, New York.

[21] Muirhead, R.J. and Waternaux, C.M. (1980), Asymptotic distribution in canonical correlation analysis and other multivariate procedures for nonnormal populations, *Biometrika*, 67, 31-43.

[22] Nüesch, P.E. (1966), On the problem of testing location in multivariate populations for restricted alternatives, *Ann. Math. Stat.*, 37, 113-119.

[23] Rothenberg, T.J. (1971), Identification in parametric models, *Econometrica*, 39, 577-591.

[24] Seber, G.A.F. (1977), *Linear Regression Analysis*, Wiley, New York.

[25] Shapiro, A. (1983), Asymptotic distribution theory in the analysis of covariance structures (a unified approach), *South African Statistical Journal*, 17, 33-81.

[26] Shapiro, A. (1984), A note on consistency of estimators in the analysis of moment structures, *British Journal of Mathematical and Statistical Psychology*, 37, 84-88.

[27] Shapiro, A. (1985), Identifiability of factor analysis: some results and open problems, *Linear Algebra and its Applications*, 70, 1-7.

[28] Shapiro, A. (1985), Asymptotic equivalence of minimum discrepancy function estimators to GLS estimators, *South African Statistical Journal*, 19, 73-81.

[29] Shapiro, A. (1985), Second-order sensitivity analysis and asymptotic theory of parametrized nonlinear programs, *Mathematical Programming*, 33, 280-299.

[30] Shapiro, A. (1985), Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints, *Biometrika*, 72, 133-144.

[31] Shapiro, A. (1986), Asymptotic theory of overparametrized structural models, *Journal of the American Statistical Association*, 81, 142-149.

[32] Shapiro, A. (1987), Robustness properties of the MDF analysis of moment structures, *South African Statistical Journal*, 21, 39-62.

[33] Shapiro, A. and Browne, M.W. (1987), Analysis of covariance structures under elliptical distributions, *Journal of the American Statistical Association*, 82, 1092-1097.

[34] Shapiro, A. (1988), Towards a unified theory of inequality constrained testing in multivariate analysis, *International Statistical Review*, 56, 49-62.

[35] Silvapulle, M.J. and Sen, P. K. (2005), *Constrained Statistical Inference: order, inequality, and shape constraints*, Wiley, New York.

[36] Steiger, J.H., Shapiro, A. and Browne, M.W. (1985), On the multivariate asymptotic distribution of sequential chi-square statistics, *Psychometrika*, 50, 253-264.

[37] Tyler, E.D. (1982), Radial estimates and the test for sphericity, *Biometrika*, 69, 429-436.

[38] Tyler, E.D. (1983), Robustness and efficiency properties of scatter matrices, *Biometrika*, 70, 411-420.

[39] Wald, A. (1950), Note on the identification of economic relations. In T.C. Koopmans (ed.), *Statistical Inference in Dynamic Economic Models*, New York, John Wiley.