Likelihood Asymptotics in Nonregular Settings: A Review and Annotated Bibliography with Emphasis on the Likelihood Ratio

Submitted to the Statistical Science

# Likelihood Asymptotics in Nonregular Settings: A Review and Annotated Bibliography with Emphasis on the Likelihood Ratio

**Alessandra R. Brazzale** * and **Valentina Mameli** †

*Abstract.* This paper reviews the most common situations where one or more regularity conditions which underlie likelihood-based parametric inference fail. We identify three main classes of problems: boundary problems, indeterminate parameter problems—which include non-identifiable parameters and singular information matrices—and change-point problems. The review focuses on the large-sample properties of the likelihood ratio statistic, though other approaches to hypothesis testing and connections to estimation will be mentioned in passing. We emphasize analytical solutions and mention software implementations where available. Some summary insights about the possible tools to derivate the key results are given.

*Key words and phrases:* boundary point, change-point, finite mixture, first order theory, identifiability, large-sample inference, singular information.

## 1. INTRODUCTION

It is commonly believed that under the null hypothesis the three classical tests of likelihood-based inference—that is, those based on the Wald, score and likelihood ratio statistics—are asymptotically equivalent and, to the first order of approximation, follow a chi-squared distribution. However, in order to hold true this statement requires a number of regularity conditions. These conditions, which are typically of Cramér type (Cramér, 1946, §33.3), require, among others, differentiability of the underlying joint probability or density function up to a suitable order and finiteness of the Fisher information matrix. Models which satisfy these requirements are said to be 'regular' and cover a wide range of applications. However, there are many important cases where one or more conditions break down. A classical example, which is traditionally used to demonstrate the failure of parametric likelihood theory, is Neyman and Scott's (1948) paradox.

Example 1.1 (Growing number of parameters). *Let $(X_1, Y_1), \dots, (X_n, Y_n)$ denote $n$ independent pairs of mutually independent and normally distributed random variables such that for each $i = 1, \dots, n$, $X_i$ and $Y_i$ have mean $\mu_i$ and common variance $\sigma^2$. The maximum likelihood estimator of $\sigma^2$ is*

$$\hat{\sigma}_n^2 = \frac{1}{2n}\sum_{i=1}^{n}\{(X_i - \hat{\mu}_i)^2 + (Y_i - \hat{\mu}_i)^2\},$$

*with $\hat{\mu}_i = (X_i + Y_i)/2$. Straightforward calculation shows that, for $n \to \infty$, $\hat{\sigma}_n^2$ converges in probability to $\sigma^2/2$ instead of the true value $\sigma^2$. The reason is that only a finite number of observations, in fact*

*Alessandra R. Brazzale is Associate Professor of Statistics, Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121 Padova, Italy (e-mail: alessandra.brazzale@unipd.it; url: www.stat.unipd.it/~brazzale).

†Valentina Mameli is tenure track Assistant Professor at the Department of Economics and Statistics, University of Udine, Via Tomadini 30/A, 33100 Udine, Italy (e-mail: valentina.mameli@uniud.it; url: https://people.uniud.it/page/valentina.mameli).

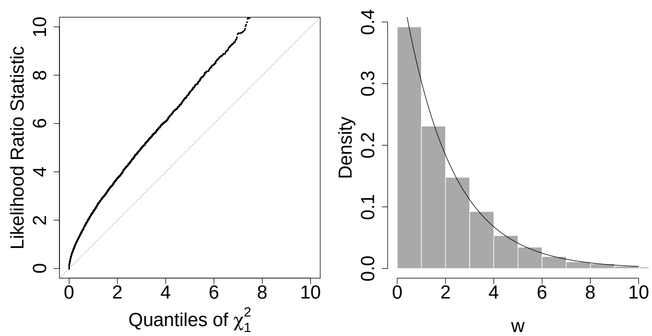


FIG 1. *Example 1.2: Translated exponential distribution. Values of the likelihood ratio $W(3)$ observed in 10,000 exponential samples of size $n = 50$ generated with rate equal to 1 and translated by $\theta_0 = 3$. Left: $\chi^2_1$ quantile plot. The diagonal dotted line is the theoretical $\chi^2_1$ approximation. Right: histogram and superimposed $\chi^2_2$ density (solid line).*

*two, is available for estimating the unknown sample means $\mu_i$. This violates a major requirement which underlies the consistency of the maximum likelihood estimator, namely that the uncertainty of all parameter estimates goes to zero.*

Example 1.1 is an early formulation of an incidental parameters problem. Other examples of this type are reviewed in Lancaster (2000), who also discusses the relevance of the Neyman–Scott paradox in statistics and economics. A recent contribution is Feng et al. (2012). Non-regularity may also arise when the parameter space is constrained and the null hypothesis lies on its boundary, or when some of the parameters disappear under the null hypothesis. The following simple example shows what may happen when the support of the distribution depends on the parameter $\theta$ so that the likelihood function cannot be differentiated over the entire parameter space.

EXAMPLE 1.2 (Translated exponential distribution). *Let $X_1, \ldots, X_n$ be an independent and identically distributed sample from an exponential distribution with rate equal to 1. Consider the translation $Y_i = X_i + \theta$, with $\theta > 0$ unknown. Given the minimum observed value $Y_{(1)}$, the likelihood ratio statistic for testing the hypothesis that $\theta = \theta_0$ is $W(\theta_0) = 2n(Y_{(1)} - \theta_0)$. Straightforward calculation proves that under the null hypothesis $W(\theta_0)$ has a $\chi^2_2$ distribution, not the classical $\chi^2_1$ limiting distribution. Furthermore, the maximum likelihood estimator of $\theta$ is no longer asymptotically normal. Indeed, it is easy to show that $Y_{(1)} - \theta$ follows exactly an exponential distribution with rate $n$. The left panel of Figure 1 shows the $\chi^2_1$ quantile plot of the likelihood ratio statistic observed in 10,000 exponential samples of size $n = 50$ generated with rate equal to 1 and translated by $\theta_0 = 3$. The finite-sample distribution of $W(3)$ is visibly far from the theoretical $\chi^2_1$ approximation represented by the dotted diagonal line. The right panel reports the empirical distribution of the likelihood ratio statistics with superimposed the $\chi^2_2$ density (solid line).*

These situations are not mere mathematical artifacts, but include many models of practical interest, such as mixture distributions and change-point problems, in genetics, reliability, econometrics, and many other fields. Especially practitioners may be less familiar with the resulting limiting distributions. As will be shown in Section 3, the distribution of the likelihood ratio statistic may converge to a mixture of chi-squared distributions, such as when the true value of the parameter belongs to the boundary of its parameter space, with mixing proportions which are awkward to determine. Or, its asymptotic behaviour may be characterised by the distribution of the supremum of a squared truncated Gaussian process, which is the common case for the finite mixture models reviewed in Section 5.

Asymptotic theory is an essential part of statistical methodology. It provides first thing approximate answers where exact ones are unavailable. Beyond this, it serves to check if a proposed inferential solution provides a sensible answer when the amount of information in the data increases without limit. Given the tremendous advances in computer age statistical inference (Efron and Hastie, 2016) one could be tempted to by-pass the often rather demanding algebraic derivations of asymptotic approximation. Gaining insight in what happens to the limiting distribution of likelihood-based test statistics when one or more regularity conditions fail is a central issue to decide whether and to which extent to rely upon simulation. The following simple example tries and makes the point.

Example 1.3 (Testing for homogeneity in a von Mises mixture). *Suppose we observe a random sample $y_1, \ldots, y_n$ from the mixture model*

$$(1.1) \qquad (1-p)f(y_i; 0, \kappa) + pf(y_i; \mu, \kappa),$$

*where $0 \leq p \leq 1$ is the mixing proportion. Furthermore, $f(y_i; \mu, \kappa)$ denotes the von Mises distribution with mean direction $|\mu| \leq \pi$ and concentration parameter $\kappa \geq 0$. Fu et al. (2008) prove that the asymptotic null distribution of the likelihood ratio statistic for testing the hypothesis $p = 0$ is the squared supremum of a truncated Gaussian process. The quantiles of the process can in principle be approximated to desirable precision by simulation, this way overcoming the algebraic difficulties of the exact solution. However, the same authors also show that if a suitable penalisation term is used, the distribution of the corresponding modified likelihood ratio statistic converges to the simple $\chi^2_1$ distribution for $n \to \infty$. This is wholly different from what happens in the Gaussian case. If the component densities $f(y_i; \mu, \kappa)$ in (1.1) represent normal distributions with unknown mean $\mu \in \mathbb{R}$ and variance $\kappa > 0$, then the distribution of the likelihood ratio statistic for testing model homogeneity diverges to infinity unless suitable constraints are imposed (Chen and Chen, 2003). This is because normal mixtures with unknown variance are not identifiable unlike the von Mises mixture model (1.1); see Section 5.4. Trying and simulating the limiting distribution would lead to totally misleading results.*

The purpose of this paper is to present the most common situations where one or more regularity conditions fail. A highly cited review of nonregular problems is Smith (1989); see also the discussion paper by Cheng and Traylor (1995). Further examples can be found in Barndorff-Nielsen and Cox (1994, §3.8), Davison (2003, §4.6) and Cox (2006, Chapter 7). The majority of existing results consider the failure of one condition at a time, but failure of two assumptions simultaneously has also received attention. Indeed, there is a rich literature on this topic. Since it is nearly impossible to cover all aspects of the subject, here, we will focus on the large-sample properties of likelihood-based parametric test statistics derived under non-standard conditions, that is, when the likelihood function is nonregular. Special attention will be paid to the likelihood ratio and its limiting distribution, although analogies with alternative test statistics and/or nonparametric and semiparametric models may be mentioned in passing. This is justified by the widespread use of Wilks' statistic, and its chi-squared limiting distribution, for hypothesis testing, model selection and other related uses. We furthermore restrict our attention to the key results and the corresponding prototype derivations; further contributions are mentioned in the annotated bibliography.

The paper is organised as follows. First order parametric inference based on the likelihood function of a regular model is reviewed in Section 2, together with the conditions upon which it is based. However, when these are not fulfilled, deriving the finite and/or asymptotic properties of the likelihood ratio statistic can be very challenging. In the absence of a unifying theory, most of the individual problems have been treated on their own. After careful consideration, we decided to group them into three broad classes. The first considers the case where the parameter space is bounded and embraces, in particular, testing for a value of the parameter which lies on its boundary; see Section 3. Section 4 concerns models where one part of the parameter vanishes when the remaining one is set to a particular value. The best-studied indeterminate parameter problem are finite mixture models. Given their widespread use in statistical practice, and their closeness to boundary problems, we will consider them separately in Section 5. Change-point problems are the third broad class of nonregular models, which we review in Section 6. Most articles investigate the consequences of the failure of one regularity condition at a time. Mixture distributions and change-point problems deserve special attention as they represent situations where two conditions fail simultaneously. Section 7 reviews cases which do not fit into the above three broad model classes, but still fall under the big umbrella of nonregular problems. These include, among others, shape constrained inference, a genre of nonparametric problem which leads to highly nonregular models.

Despite the many remarkable theoretical developments in likelihood-based asymptotic theory for nonregular parametric models, one may wonder why the corresponding results are little known especially among practitioners. We believe there are at least two reasons. The first is that the results are highly scattered, in time and scope, which makes it difficult to get the general picture. The second reason is that the limiting distributions are often fairly complex in their derivation and implementation. Section 8 reviews the few software implementation we are aware of.

The paper closes with the short summary discussion of Section 9.

## 2. LIKELIHOOD ASYMPTOTICS

### 2.1 First order theory

*2.1.1 General notation.* Consider a parametric statistical model with probability density or mass function $f(y;\theta)$, where the parameter $\theta$ takes values in a subset $\Theta \subseteq \mathbb{R}^p$, $p \geq 1$, and $y = (y_1, \ldots, y_n)$ are $n$ observations from $Y = (Y_1, \ldots, Y_n)$. Throughout the paper we will consider these an independent and identically distributed random sample unless stated differently. Let $L(\theta) = L(\theta; y) \propto f(y;\theta)$ and $l(\theta) = \log L(\theta)$ denote the likelihood and the log-likelihood functions, respectively. The maximum likelihood estimate (MLE) $\hat\theta$ of $\theta$ is the value of $\theta$ which maximises $L(\theta)$ or equivalently $l(\theta)$. Under mild regularity conditions on the log-likelihood function to be discussed in Section 2.2, $\hat\theta$ solves the score equation $u(\theta) = 0$, where $u(\theta) = \partial l(\theta)/\partial\theta$ is the score function. We furthermore define the observed information function $j(\theta) = -\partial^2 l(\theta)/\partial\theta\partial\theta^\top$ and the expected or Fisher information $i(\theta) = E\,[j(\theta;Y)]$, where $\theta^\top$ denotes transposition of $\theta$.

*2.1.2 No nuisance parameter.* The three classical likelihood-based statistics for testing $\theta = \theta_0$ are the

| | |
|---|---|
| standardized MLE, | $(\hat\theta-\theta_0)^\top j(\hat\theta)(\hat\theta-\theta_0)$, |
| score statistic, | $u(\theta_0)^\top j(\hat\theta)^{-1} u(\theta_0)$, |
| likelihood ratio | $W(\theta_0) = 2\{l(\hat\theta) - l(\theta_0)\}$, |

where the observed information $j(\hat\theta)$ is at times replaced by the Fisher information $i(\theta)$. These statistics are also known under the names of Wald's, Rao's and Wilks' tests, respectively. If the parametric model is regular, the finite-sample null distribution of the above three statistics converges to a $\chi^2_p$ distribution to the order $O(n^{-1})$ as $n \to \infty$. For $\theta$ scalar, inference may be based on the corresponding signed versions, that is, on the signed Wald statistic, $(\hat\theta-\theta_0)j(\hat\theta)^{1/2}$, score statistic, $u(\theta_0)j(\theta_0)^{-1/2}$, and likelihood root,

$$r(\theta_0) = \operatorname{sign}(\hat\theta-\theta_0)[2\{l(\hat\theta) - l(\theta_0)\}]^{1/2},$$

whose finite-sample distributions converge to the standard normal distribution to the order $O(n^{-1/2})$.

*2.1.3 Nuisance parameters.* Suppose now that the parameter $\theta = (\psi, \lambda) \in \Psi \times \Lambda$ is partitioned into a $p_0$-dimensional parameter of interest, $\psi \in \Psi \subseteq \mathbb{R}^{p_0}$, and a vector of nuisance parameters $\lambda \in \Lambda \subseteq \mathbb{R}^{p-p_0}$ of dimension $p - p_0$. Large-sample inference for $\psi$ is commonly based on the profile log-likelihood function

$$l_{\mathrm{p}}(\psi) = \sup_{\lambda\in\Lambda} l(\psi,\lambda),$$

which maximises the log-likelihood $l(\psi,\lambda)$ with respect to $\lambda$ for fixed $\psi$. The profile likelihood ratio statistic for testing $\psi \in \Psi_0$ is

$$W_{\mathrm{p}}(\psi_0) = 2\{\sup_{\psi\in\Psi} l_{\mathrm{p}}(\psi) - \sup_{\psi\in\Psi_0} l_{\mathrm{p}}(\psi)\},$$

where $\Psi_0 \subset \Psi$ is the parameter space specified under the null hypothesis. If the null hypothesis is $\psi = \psi_0$, the finite-sample distribution of $W_{\mathrm{p}}(\psi_0)$ converges to the $\chi^2_{p_0}$ distribution to the order $O(n^{-1})$ for $n \to \infty$.

If there exists a closed form expression for the constrained maximum likelihood estimate $\hat\lambda_\psi$ of $\lambda$ for given $\psi$, the profile log-likelihood function may be written as

$$l_{\mathrm{p}}(\psi) = \sup_{\lambda\in\Lambda} l(\psi,\lambda) = l(\psi,\hat\lambda_\psi). \tag{2.1}$$

A typical situation where $\hat\lambda_\psi$ is not available in closed form is when the nuisance parameter $\lambda$ vanishes under the null hypothesis, as will be addressed in Section 4.1. If (2.1) holds, we may define the profile Wald, score and likelihood ratio statistics for testing $\psi = \psi_0$ as in Section 2.1.2, but now in terms of the profile log-likelihood $l_{\mathrm{p}}(\psi)$, with $u_{\mathrm{p}}(\psi) = \partial l_{\mathrm{p}}(\psi)/\partial\psi$ and $j_{\mathrm{p}}(\psi) = \partial l_{\mathrm{p}}(\psi)/\partial\psi\partial\psi^\top$ being the profile score and profile observed information functions. The asymptotic null distribution of these statistics is a $\chi^2_{p_0}$ distribution up to the order $O(n^{-1})$. If $\psi$ is scalar, the distributions of the corresponding signed versions, $(\hat{\psi}-\psi_0)j_{\mathrm{p}}(\hat{\psi})^{1/2}$, $u_{\mathrm{p}}(\psi_0)j_{\mathrm{p}}(\psi_0)^{-1/2}$, and

$$r_{\mathrm{p}}(\psi_0) = \mathrm{sign}(\hat{\psi}-\psi_0)[2\{l_{\mathrm{p}}(\hat{\psi}) - l_{\mathrm{p}}(\psi_0)\}]^{1/2}, \tag{2.2}$$

may be approximated by standard normal distributions up to the order $O(n^{-1/2})$.

## 2.2 Regularity conditions

The first step in the derivation of the large-sample approximations and statistics of Sections 2.1 is typically Taylor series expansion of the log-likelihood function $l(\theta)$, or quantities derived thereof, in $\hat{\theta}$ around $\theta$. We illustrate this by considering the expansion to the order $O_p(n^{-1/2})$ of the likelihood ratio $W(\theta) = 2\{l(\hat{\theta}) - l(\theta)\}$ for the scalar parameter case.

EXAMPLE 2.1 (Asymptotic expansion of likelihood ratio). *Let $p = 1$ and $l_m = l_m(\theta) = d^m l(\theta)/d\theta^m$ be the derivative of order $m = 2, 3, \ldots$ of $l(\theta)$, the log-likelihood function for $\theta$ in a regular parametric model. Recall that $u = u(\theta) = dl(\theta)/d\theta$ represents the score function, while $i = i(\theta) = E[-l_2(\theta; Y)]$ is the Fisher information. Taylor series expansion of $l(\hat{\theta})$ around $\theta$ yields*

$$\begin{aligned} l(\hat{\theta}) - l(\theta) &= (\hat{\theta}-\theta)u + \frac{1}{2}(\hat{\theta}-\theta)^2 l_2 \\ &+ \frac{1}{6}(\hat{\theta}-\theta)^3 l_3 + \frac{1}{24}(\hat{\theta}-\theta)^4 l_4 + \cdots . \end{aligned} \tag{2.3}$$

*Rewriting (2.3) using notation (A.3) and replacing $(\hat{\theta}-\theta)$ with expansion (A.5) yields, after suitable rearrangement of the terms,*

$$\begin{aligned} l(\hat{\theta}) - l(\theta) &= \frac{1}{2}i^{-1}u^2 + \frac{1}{6}i^{-2}\left(i^{-1}u\nu_3 + 3H_2\right) l_2 \\ &+ O_p(n^{-1}). \end{aligned} \tag{2.4}$$

*Here, $H_2 = l_2 - \nu_2$, with $\nu_m = E[l_m(\theta; Y)]$, for $m = 2, 3$. The leading term, $i^{-1}u^2$, in (2.4) converges asymptotically to the $\chi^2_1$ distribution, while the second addend is of order $n^{-1/2}$. This leads to the well known result for Wilks' statistic. See Pace and Salvan (1997, §9.4.4) for the details.*

The derivation of Example 2.1 requires that the model under consideration is regular. This implies first of all that the log-likelihood function can be differentiated to whatever order is required, but also that the asymptotic order of expected values of log-likelihood derivatives is proportional to the sample size. Wald (1949)—who is generally acknowledged for having provided the earliest proof of consistency of the maximum likelihood estimator which is mathematically correct—furthermore emphasized the importance of the compactness of the parameter space $\Theta$ and that the maximum likelihood estimator be unique. Indeed, the former condition was missing in Cramér's (1946) and Huzurbazar's (1948) proofs.

The required regularity conditions may be formulated in several ways; see e.g. Cox and Hinkley (1974, p. 281), Barndorff-Nielsen and Cox (1994, §3.8), Azzalini (1996, §3.2.3), Severini (2000, §4.7), van der Vaart (2000, Chap. 5), Davison (2003, §4.6), Hogg, McKean and Craig (2019, §6.1, §6.2 and A.1). Here, we will assume that the following five conditions on the model function $f(y; \theta)$ hold.

*Condition 1* All components of $\theta$ are identifiable. That is, two model functions $f(y; \theta^1)$ and $f(y; \theta^2)$ defined by any two different values $\theta^1 \neq \theta^2$ of $\theta$ are distinct almost surely.

*Condition 2* The support of $f(y; \theta)$ does not depend on $\theta$.

*Condition 3* The parameter space $\Theta$ is a compact subset of $\mathbb{R}^p$, for a fixed positive integer $p$, and the true value $\theta^0$ of $\theta$ is an interior point of $\Theta$.

*Condition 4* The partial derivatives of the log-likelihood function $l(\theta; y)$ with respect to $\theta$ up to the order three exist in a neighbourhood of the true parameter value $\theta^0$ almost surely. Furthermore, in such a neighbourhood, $n^{-1}$ times the absolute value of the log-likelihood derivatives of order three are bounded above by a function of $Y$ whose expectation is finite.

*Condition 5* The first two Bartlett identities hold, which imply that

$$E[u(\theta; Y)] = 0, \quad i(\theta) = \mathrm{Var}[u(\theta; Y)].$$

Conditions 1–5 are relevant in many important models of practical interest, and can fail in as many ways. For instance, from the perspective of significance testing, Condition 1 fails when under the null hypothesis parameters defined for the whole model become undefined and therefore inestimable. We already mentioned this situation when introducing the profile log-likelihood function and will come back to it in Section 4.1. Further examples are treated in Sections 4.2 and 5. Condition 2 typically does not hold in change-point problems, which will be treated in Section 6. Failure of Condition 2 is furthermore addressed in Hirano and Porter (2003) and Severini (2004). Failure of Condition 3 characterises the first and most extensively explored nonregular setting, that is, boundary problems; see Section 3. The compactness condition, in particular, can be omitted, provided it is replaced by some other requirements; see, for instance, Pfanzagl (2017, Page 119). This will be also the case for a number of the large-sample results derived for nonregular models; see, for instance, Section 5. A prominent example where Condition 4 is not satisfied, is the double exponential, or Laplace, distribution, which arises in quantile regression. For a book-length review of this topic we refer the Reader to Koenker et al. (2017). Condition 5 is guaranteed if standard results on the interchanging of integration and differentiation hold, Condition 2 is satisfied, and the log-likelihood derivatives are continuous functions of $\theta$. A typical situation where this condition fails is when the data under analysis are derived from a probability density which does not belong to the model $f(y;\theta)$, a topic of much investigation in robustness (Huber and Ronchetti, 2009). A remedy is provided by Godambe's theory of estimating equations (Godambe, 1991).

Conditions 4 and 5, as used by Cramér (1946), Wald (1949) and others, imply the existence of at least three derivatives of the log-likelihood function together with some uniform integrability restrictions. However, these conditions do not have by themselves any direct statistical interpretation. LeCam (1970) presents a different type of regularity assumption—differentiability in quadratic mean of the log-likelihood function—which involves only one differentiation step and may be justified from a statistical point of view. As shown in his 1970 paper, the regularity conditions of Cramér type imply differentiability in quadratic mean, while the opposite does not hold true. This way, LeCam gives rise to a radically different type of asymptotic inference called *local asymptotics*, which is based upon the concept of a 'contiguity neighbourhood'. Under Conditions 1–5, this translates into a sequence of alternative hypotheses of the form $\theta_n = \theta_0 + \eta/\sqrt{n}$, where $\eta$ is any given real number. The properties of the likelihood-based procedures are hence studied in the Euclidean $n^{-1/2}$–neighbourhood of the fixed parameter $\theta_0$ defined by $\eta$. In particular, the log-likelihood function is said to be 'locally asymptotically quadratic' if there exist two random sequences $U_n(\theta_0)$ and $I_n(\theta_0)$ which do not depend on $\eta$ such that

$$l_n\left(\theta_0 + \frac{\eta}{\sqrt{n}}\right) - l_n(\theta_0) = \eta U_n(\theta_0) - \frac{\eta^2 I_n(\theta_0)}{2} + R_n(\eta,\theta_0).$$

Here, the sequence $I_n(\theta_0)$ is positive and bounded in probability away from zero, while the residual term $R_n(\eta,\theta_0)$ converges in probability to zero for $n \to \infty$. Note how this definition mimics Taylor series expansion in classical likelihood-based asymptotics, where $U_n(\theta_0)$ and $I_n(\theta_0)$ replace the score and expected information functions.

In the remainder of the paper, we review the most common situations where one or some of Conditions 1–5 fail. We will also provide some summary insight into the main prototype derivations of the corresponding asymptotic results. The vast majority of the proofs require conditions of Cramér type; in some occasions, as for instance in Section 4.1, LeCam's local asymptotic theory will be used.

## 3. BOUNDARY PROBLEMS

Boundary problems represent the first and most extensively explored nonregular setting. Furthermore, small-sample solutions seem to have been addressed only for this case. A boundary problem arises when the value $\theta_0$ specified by the null hypothesis, or parts of it, fall on the boundary of the parameter space. Informally, the methodological difficulties in likelihood-based inference occur because the maximum likelihood estimate can only fall 'on the side' of $\theta_0$ that belongs to the parameter space $\Theta$. This implies that if the maximum occurs on the boundary, the score function need not be zero and the distributions of the related likelihood statistics will not converge to the typical normal or chi-squared distributions. Because of the difficulties inherent the derivation of the limiting distribution of the likelihood ratio statistic, especially practitioners tend to ignore the boundary problem and to proceed as if all parameters where interior points of $\Theta$. This is commonly called the naïve approach. An alternative approach is to suitably enlarge the parameter space so as to guarantee that the likelihood ratio maintains the common limiting distribution; see, for instance, Feng and McCulloch (1992). The literature on boundary problems is very rich and includes, among others, solutions for random effects and frailty models, and for times series analysis. The following example gives a flavour of the statistical issues.

Example 3.1 (Bivariate normal). *Consider a single observation $y = (y_1, y_2)$ from the bivariate normal random variable $Y = (Y_1, Y_2) \sim N_2(\theta, I_2)$, where $\theta = (\theta_1, \theta_2)$, with $\theta_1 \geq 0$ and $\theta_2 \geq 0$, and $I_2$ is the $2 \times 2$ identity matrix. Straightforward calculation shows that the null distribution of the likelihood ratio statistic for $\theta_0 = (0, 0)$ versus the alternative hypothesis that at least one equality does not hold, converges to a mixture of a point mass $\chi^2_0$ at 0 and two chi-squared distributions, $\chi^2_1$ and $\chi^2_2$ (DasGupta, 2008, Example 21.3). Figure 2 provides a graphical representation of the problem. Because of the boundedness of the parameter space, we have that $\hat\theta_1 = \max(y_1, 0)$ and $\hat\theta_2 = \max(y_2, 0)$. The grey shaded area is the parameter space into which the MLE is bound to fall. However, the random observation $Y = (Y_1, Y_2)$ can fall into any of the 4 quadrants of $\mathbb{R}^2$ with equal probability $1/4$. When $Y$ falls into the first quadrant, that is, when $y_1, y_2 > 0$, the likelihood ratio statistic is $W(\theta_0) = Y_1^2 + Y_2^2$ and follows the common $\chi^2_2$ distribution. However, if $y_1 > 0$ and $y_2 < 0$ or when $y_1 < 0$ and $y_2 > 0$, we have that $W(\theta_0) = Y_1^2 \sim \chi^2_1$ and $W(\theta_0) = Y_2^2 \sim \chi^2_1$, respectively. Lastly, when $Y$ lies in the third quadrant, $W(\theta_0) = 0$ and its distribution is a point mass in 0. Summing up, we can informally write*

$$W(\theta_0) \sim \frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2. \tag{3.1}$$

Distribution (3.1) is a special case of the so-called chi-bar squared distribution (Kudô, 1963), denoted by $\bar\chi^2(\omega, N)$, with cumulative distribution function

$$\Pr(\bar\chi^2 \leq c) = \sum_{\nu=0}^{N} \omega_\nu \Pr(\chi^2_\nu \leq c),$$

which corresponds to a mixture of chi-squared distributions with degrees of freedom $\nu$ from 0 to $N$.

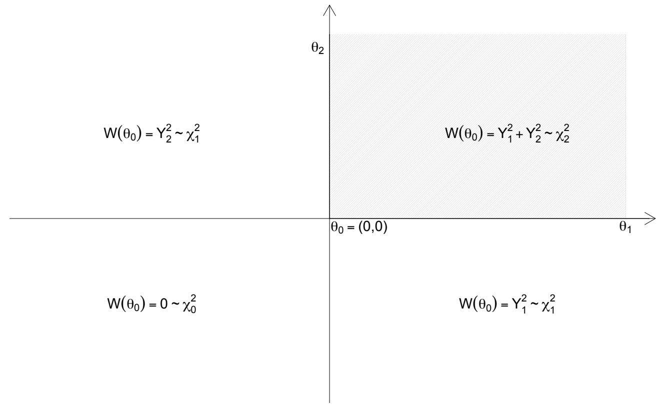


Fig 2. *Example 3.1: Bivariate normal. The grey shaded area represents the parameter space $\Theta$. Under the null hypothesis $\theta_0 = (0, 0)$, the parameter space collapses with the origin. The asymptotic distribution of the corresponding likelihood ratio statistics is a mixture of $\chi^2_0$, $\chi^2_1$ and $\chi^2_2$ distributions with weights $(0.25, 0.5, 0.25)$.*

In some cases, explicit and computationally feasible formulae are available for the weights $\omega = (\omega_0, \ldots, \omega_N)$. Extensive discussion on their computation and use, with special emphasis on inequality constrained testing, is given in Robertson et al. (1988, Chapters 2 and 3), Wolak (1987), Shapiro (1985, 1988) and Sun (1988).

## 3.1 General results

The research on boundary problems was initiated by Chernoff (1954) who derives the asymptotic null distribution of the likelihood ratio statistic for testing whether $\theta$ lies on one or the other side of a smooth $(p-1)$-dimensional surface in a $p$-dimensional space when the true parameter value lies on the surface. Using a geometrical argument, Chernoff establishes that this distribution is equivalent to the distribution of the likelihood ratio statistic for testing suitable restrictions on the mean of a multivariate normal distribution with covariance matrix given by the inverse of the Fisher information matrix using a single observation. In particular, Chernoff proves that the limiting distribution is a $\bar\chi^2(\omega, 1)$ distribution, with $\omega = (0.5, 0.5)$, that is, a mixture of a point mass at zero and a $\chi^2_1$, with equal weights. This generalizes Wilks (1938) result when the parameter space under the null hypothesis is not a hyperplane.

In Chernoff (1954), the parameter spaces $\Theta_0$ and $\Theta_1$, specified by the null and the alternative hypotheses, are assumed to have the same dimension. Furthermore, the true parameter value falls on the boundary of both, $\Theta_0$ and $\Theta_1$, while it is still an interior point of the global parameter space $\Theta = \Theta_0 \cup \Theta_1$. The no doubt cornerstone contribution which inspired many researchers and fuelled an enormous literature, is the highly-cited article by Self and Liang (1987). Using geometrical arguments similar to those of Chernoff (1954), Self and Liang (1987) study the asymptotic null distribution of the likelihood ratio statistic for testing the null hypothesis $\theta \in \Theta_0$ against the alternative $\theta \in \Theta_1 = \Theta \setminus \Theta_0$. This time, the true parameter value $\theta^0$ no longer needs be an interior point, but can fall onto the boundary of $\Theta$. The two sets $\Theta$ and $\Theta_0$ must be regular enough to be approximated by two cones, $C_\Theta$ and $C_{\Theta_0}$, with vertex at $\theta_0$ (Chernoff, 1954, Definition 2). Under this scenario and provided their Assumptions 1–4 hold—which translate into our Conditions 1–2 and 4–5, with likelihood derivatives taken from the appropriate side—Self and Liang (1987, Theorem 3) show that the distribution of the likelihood ratio converges to the distribution of

$$\sup_{\theta \in C_{\Theta-\theta^0}} \left\{ -(\tilde{Z}-\theta)^\top i_1(\theta^0)(\tilde{Z}-\theta) \right\} - \sup_{\theta \in C_{\Theta_0-\theta^0}} \left\{ -(\tilde{Z}-\theta)^\top i_1(\theta^0)(\tilde{Z}-\theta) \right\}. \tag{3.2}$$

Here, $C_{\Theta-\theta^0}$ and $C_{\Theta_0-\theta^0}$ are the translations of the cones $C_\Theta$ and $C_{\Theta_0}$, such that their vertices are at the origin, and $\tilde{Z}$ is a multivariate Gaussian variable with mean 0 and covariance matrix given by $i_1(\theta^0)^{-1}$, which is the Fisher information matrix for a single observation. If we transform the random variable $\tilde{Z}$ so that it follows a multivariate standard Gaussian distribution $Z$, we can re-express Equation (3.2) as

$$\inf_{\theta \in \tilde{C}_0} ||Z-\theta||^2 - \inf_{\theta \in \tilde{C}} ||Z-\theta||^2 = ||Z - \mathcal{P}_{\tilde{C}_0}(Z)||^2 - ||Z - \mathcal{P}_{\tilde{C}}(Z)||^2, \tag{3.3}$$

where $\tilde{C}$ and $\tilde{C}_0$ are the corresponding transformations of the cones $C_{\Theta-\theta^0}$ and $C_{\Theta_0-\theta^0}$ and $||\cdot||$ is the Euclidean norm. Finding the null distribution requires to work out the two projections $\mathcal{P}_{\tilde{C}}(Z)$ and $\mathcal{P}_{\tilde{C}_0}(Z)$ of $Z$ onto the cones $\tilde{C}$ and $\tilde{C}_0$. This must be done on a case by case basis as shown by the following revisitation of Example 3.1.

Example 3.2 (Bivariate normal revisited). *In Example 3.1 we faced a typical non-standard situation where both components of the parameter $\theta$ are of interest and both lie on the boundary of the parameter space. Here, the Fisher information matrix is the identity matrix which is why $\tilde{Z} = Z = Y$ and the original two set $\Theta$ and $\Theta_0$ agree with the approximating cones. That is, the grey shaded region $[0,\infty) \times [0,\infty)$ in Figure 2 represents the sets $\Theta = C_\Theta = C_{\Theta-\theta_0} = \tilde{C}$, while the origin $\{0\}$ corresponds to the sets $\Theta_0 = C_{\Theta_0} = C_{\Theta_0-\theta_0} = \tilde{C}_0$. The derivation of the second term of (3.3) depends on the projection of $Z$ onto $\tilde{C}$, which is*

$$\mathcal{P}_{\tilde{C}}(Z) = \begin{cases} Z = (Z_1, Z_2) & \text{if} \quad Z_1, Z_2 > 0 \\ Z_2 & \text{if} \quad Z_1 < 0, Z_1 > 0 \\ 0 & \text{if} \quad Z_1, Z_2 < 0 \\ Z_1 & \text{if} \quad Z_1 > 0, Z_2 < 0, \end{cases}$$

*while $\mathcal{P}_{\tilde{C}_0}(Z) = 0$. As shown in Example 3.1, $\mathcal{P}_{\tilde{C}}(Z)$ takes on the four possible values with equal probability 1/4. By simple algebra, we can prove that the distribution of the likelihood ratio statistics is given by the mixture of Equation (3.1).*

A sketch of the derivation of Equation (3.2) is given in Example A.1. The proof consists of two steps. We first consider a quadratic Taylor series expansion of the log-likelihood $l(\theta)$ around $\theta^0$, the true value of the parameter. The asymptotic distribution of the likelihood ratio statistic is then derived as in Chernoff (1954) by approximating the sets $\Theta$ and $\Theta_0$ using the cones $C_\Theta$ and $C_{\Theta_0}$. Self and Liang (1987) present a number of special cases in which the representations (3.2) and (3.3) are used to derive the asymptotic null distribution of the likelihood ratio statistic. In most cases, this results in a chi-bar squared distribution whose weights depend, at times in a rather tricky way, on the partition of the parameter space induced by the geometry of the cones.

A further major step forward in likelihood asymptotics for boundary problems was marked by Kopylev and Sinha (2011) and Sinha et al. (2012). Now, the null distribution of the likelihood ratio statistic is derived by using algebraic arguments. A first simple case considers the scalar hypothesis $\theta_1 = \theta_{10}$ against the alternative $\theta_1 > \theta_{10}$ on the first component of the $p$-dimensional parameter $\theta$ under the assumption that the remaining components of $\theta$ are interior points. The corresponding asymptotic null distribution of the likelihood ratio statistic is a fifty-fifty mixture of a $\chi^2_0$ and a $\chi^2_1$ distribution, in agreement with Case 5 of Self and Liang (1987). From the technical point of view, the derivation of a closed form expression for the limiting distribution of the likelihood ratio becomes the more difficult the more nuisance parameter lie on the boundary of the parameter space. In particular, the derivation of the limiting distribution becomes awkward when there are more than four boundary points and/or the Fisher information matrix is not diagonal. Sinha et al. (2012) furthermore show that when one or more nuisance parameters are on the boundary, following the naïve approach can result in inferences which are anti-conservative. In general, the asymptotic distribution turns out to be a chi-bar squared distribution with weights that depend on the number of parameters of interest and of nuisance parameters, and on where these lie in $\Theta$. However, limiting distributions other than the $\bar{\chi}^2$ distribution are found as well; see, for instance, Theorem 2.1 of Sinha et al. (2012).

Susko (2013) proposes a data-dependent solution to Self and Liang's (1987) problem which avoids the calculation of the mixing weights of the chi-bar squared limit distribution and performs well in terms of power and type I error provided all nuisance parameters are interior points of $\Theta$. In particular, Susko (2013) shows that the likelihood ratio $W$ conditioned on the number of parameters $\nu$ which are estimated to fall within the parameter space, converges under the null hypothesis weakly to a simple $\chi^2_\nu$ distribution with $\nu$ degrees of freedom. Further recent alternatives, which avoid the calculation of the mixing weights of the $\bar{\chi}^2$ distribution and/or lead to the classical $\chi^2$ limiting distribution, are mentioned in the annotated bibliography.

A concise review of the cases considered in Self and Liang (1987), Kopylev and Sinha (2011) and Sinha et al. (2012), with some interesting examples and an account of the areas of interest in genetics and biology, is given by Kopylev (2012). The following sections treat three special cases, namely testing for a zero variance component, constrained one-sided tests and the few treatments of a nonregular problem in higher order asymptotics we are aware of. We mention the mainstream contributions while further related work can be found in the annotated bibliography.

### 3.2 Null variance components

In linear and generalized linear mixed models a boundary problem arises as soon as we want to assess the significance of one or more variance components. The two reference papers are Crainiceanu and Ruppert (2004) and Stram and Lee (1994). Both consider a linear mixed effects model and test for a zero scalar variance component. However, Stram and Lee (1994) assume that the data vector can be partitioned into a large number of independent and identically distributed sub-vectors, which needs not hold for Crainiceanu and Ruppert (2004). The limiting distributions are derived from the spectral decomposition of the likelihood ratio statistic.

More precisely, assume the following model holds,

$$Y = X\beta + Zb + \varepsilon,$$

where $Y$ is a vector of observations of dimension $n$, $X$ is a $n \times p$ fixed effects design matrix and $\beta$ is a $p$-dimensional vector of fixed effects. In addition, $Z$ is a $n \times k$ random effects design matrix and $b$ is a $k$-dimensional vector of random effects which are assumed to follow a multivariate Gaussian distribution with mean 0 and covariance matrix $\sigma^2_b\Sigma$ of order $k \times k$. The error term $\varepsilon$ is assumed to be independent of $b$ and distributed as a normal random vector with zero mean and covariance matrix $\sigma^2_\varepsilon I_n$, where $I_n$ is the identity matrix. Suppose we are interested in testing

$$H_0 : \beta_{p+1-q} = \beta^0_{p+1-q}, \ldots, \beta_p = \beta^0_p, \qquad \sigma^2_b = 0$$

against

$$H_1 : \beta_{p+1-q} \neq \beta^0_{p+1-q}, \ldots, \beta_p \neq \beta^0_p, \quad \text{or} \quad \sigma^2_b > 0$$

for some positive value of $q \in \{1, \ldots, p\}$. Nonregularity arises as under the null hypothesis $\sigma^2_b = 0$ falls on the boundary of the parameter space. Furthermore, the alternative hypothesis that $\sigma^2_b > 0$ induces dependence among the observations $Y$. Crainiceanu and Ruppert (2004, Theorem 1) show that the finite-sample distribution of the likelihood ratio statistic agrees with the distribution of

$$n\left(1 + \frac{\sum_{s=1}^{q} u_s^2}{\sum_{s=1}^{n-p} w_s^2}\right) + \sup_{\lambda \geq 0} f_n(\lambda), \tag{3.4}$$

where $u_s$ for $s = 1, \ldots, k$ and $w_s$ for $s = 1, \ldots, n-p$ are independent standard normal variables, $\lambda = \sigma_b^2/\sigma_\varepsilon^2$, and

$$f_n(\lambda) = n \log\left\{1 + \frac{N_n(\lambda)}{D_n(\lambda)}\right\} - \sum_{s=1}^{k} \log(1 + \lambda\xi_{s,n}),$$

where

$$N_n(\lambda) = \sum_{s=1}^{k} \frac{\lambda\mu_{s,n}}{1+\lambda\mu_{s,n}} w_s^2,$$

and

$$D_n(\lambda) = \sum_{s=1}^{k} \frac{w_s^2}{1+\lambda\mu_{s,n}} + \sum_{s=k+1}^{n-p} w_s^2.$$

Here, $\mu_{s,n}$ and $\xi_{s,n}$ are the $k$ eigenvalues of the matrices $\Sigma^{\frac{1}{2}} Z^T P_0 Z \Sigma^{\frac{1}{2}}$ and $\Sigma^{\frac{1}{2}} Z^T Z \Sigma^{\frac{1}{2}}$, respectively. The matrix $P_0 = I_n - X(X^T X)^{-1} X^T$ is the matrix which projects onto the orthogonal complement to the subspace spanned by the columns of the design matrix $X$. Theorem 2 of Crainiceanu and Ruppert (2004) shows that the asymptotic null distribution of the likelihood ratio statistic depends on the asymptotic behaviour of the eigenvalues $\mu_{s,n}$ and $\xi_{s,n}$. The limiting distribution, in general, differs from the chi-bar squared distribution which often holds for independent and identically distributed data.

Formula (3.4) represents the spectral decomposition of the likelihood ratio statistic. A similar result is also derived for the restricted likelihood ratio (Crainiceanu and Ruppert, 2004, Formula 9). The unquestioned advantage of these two results is that they allow us to simulate the finite-sample null distribution of the two test statistics once the eigenvalues are calculated. Furthermore, this simulation is more efficient than bootstrap resampling, as the speed of the algorithm only depends on the number of random effects $k$, and not on the number of observations $n$. Applications of Crainiceanu and Ruppert's (2004) results include testing for level- or subject-specific effects in a balanced one-way ANOVA, testing for polynomial regression versus a general alternative described by P-splines and testing for a fixed smoothing parameter in a P-spline regression.

### 3.3 Constrained one-sided tests

Multistage dose-response models are a further example of boundary problem. A $K$-stage model is characterised by a dose-response function of the form

$$g(d;\beta) = g(\beta_0 + \beta_1 d + \beta_2 d^2 + \cdots + \beta_K d^K),$$

where $d$ is the tested dose and $g(\cdot)$ is a function of interest such as, for instance, the probability of developing a disease. The coefficients $\beta_k \geq 0$, for $k = 1, \ldots, K$, are often constrained to be non-negative so that the dose-response function will be non-decreasing. There is no limit on the number of stages $K$, though in practice this is usually specified to be no larger than the number of non-zero doses. Testing whether $\beta_k = 0$ results in a boundary problem and requires the application of a so-called constrained one-sided test. Apart from clinical trials, constrained one-sided tests are common in a number of other areas, where the constraints on the parameter space are often natural such as testing for over-dispersion, for the presence of clusters and for homogeneity in stratified analyses. All these instances amount to having the parameter value lying on the boundary of the parameter space under the null hypothesis. Despite their importance in statistical practice, few contributions are available on the asymptotic behaviour of the most commonly used test statistics, and of the likelihood ratio in particular.

A first contribution which evaluates the asymptotic properties of constrained one-sided tests is Andrews (2001), who establishes the limiting distributions of the Wald, score, quasi-likelihood and rescaled quasi-likelihood ratio statistics under the null and the alternative hypotheses. The results are used to test for no conditional heteroscedasticity in a GARCH(1,1) regression model and zero variances in random coefficient models. Sen and Silvapulle (2002) review refinements of likelihood-based inferential procedures for a number of parametric, semiparametric, and nonparametric models when the parameters are subject to inequality constraints. Special emphasis is placed on their applicability, validity, computational flexibility and efficiency. Again, the chi-bar squared distribution plays a central role in characterising the limiting null distribution of the test statistics, while the corresponding proof requires tools of convex analysis, such projections onto cones. See Silvapulle and Sen (2005) for a book-length account of constrained statistical inference.

### 3.4 Small-sample results

In addition to Crainiceanu and Ruppert (2004) we found two further contributions which explore the higher order properties of likelihood-based test statistics in a nonregular setting.

del Castillo and Lopez-Ratera (2006) consider testing for a boundary point in a scalar exponential family. In particular, they consider the family $\mathcal{F}$ of real valued random variables with probability density function

$$f(y;\theta) = e^{\theta y - \kappa(\theta)} f(y), \quad \theta \in \Theta \in \mathbb{R}, \tag{3.5}$$

where $\Theta$ is the set of parameters for which the function $\kappa(\theta) < +\infty$. The family $\mathcal{F}$ is said to be the coniugate family of $f(y)$, obtained from its cumulant generator function $\kappa(\theta)$. If $\Theta$ is an open convex set, model (3.5) is a regular exponential family. Otherwise, if $\Theta$ includes some of its boundary points, $\mathcal{F}$ is called a nonregular exponential model. del Castillo and Lopez-Ratera (2006) characterise the asymptotic null distribution of the likelihood ratio for testing the hypothesis $\theta = 0$, where $\Theta = \{c < \theta \leq 0\}$, when the variance of $Y$ is finite. The resulting distribution is a fifty-fifty mixture of a $\chi^2_1$ and a $\chi^2_0$, similar to the findings by Self and Liang (1987, Case 5) where one component of the parameter vector lies on the boundary of its parameter space. The approach is illustrated for testing exponentiality in reliability theory and survival analysis.

Sørensen (2008) examines the small-sample distribution of the likelihood ratio statistic in the random effects model which is often recommended for meta-analyses, and in a related over-dispersion model. For small sample sizes the distribution of the likelihood ratio for the overall treatment effect is not $\chi^2$ distributed and depends on the true value of the heterogeneity parameter (or between-study variance) of the model. Sørensen (2008) suggests a simulation-based method to investigate how strong this dependence is.

## 4. INDETERMINATE PARAMETER PROBLEMS

An 'indeterminate parameter' problem occurs when setting one of the components of the parameter $\theta = (\theta_1, \theta_2)$ to a particular value, say $\theta_1 = \theta_{10}$, leads to the disappearance of some or all components of $\theta_2$. The model is no longer identifiable, as all model functions $f(y;\theta)$ with $\theta_1 = \theta_{10}$ and arbitrary $\theta_2$ identify the same distribution. The following simple example illustrates this point.

Example 4.1 (Loss of identifiability in jump regression). *Consider the model*

$$Y = \theta_{11} + \theta_{12}\mathbb{1}(X > \theta_2) + \varepsilon, \quad \varepsilon \sim f(\varepsilon),$$

*where $Y$ is a continuous response, $X$ a corresponding covariate and $\mathbb{1}(X > \theta_2)$ represents the indicator function which assumes value 1 if $X > \theta_2$ and zero otherwise. Furthermore, $\theta_1 = (\theta_{11}, \theta_{12})$ is a real valued vector of regression coefficients, while $\theta_2 \in \mathbb{R}$ defines the point at which the jump occurs. Assume that $\varepsilon$ is a zero-mean error term with density function $f(\varepsilon)$. The mean of the variable $Y$ is $\theta_{11}$ for values of $X$ less or equal to $\theta_2$ and is equal to $\theta_{11} + \theta_{12}$ for values of $X$ larger than $\theta_2$. Under the null hypothesis of no jump, $\theta_{10} = (\theta_{11}, 0)$ with arbitrary $\theta_{11}$, the parameter $\theta_2$ disappears and the model is no longer identifiable. Arbitrary values of $\theta_2$ identify the same distribution for the variable $Y$.*

When the parameter which indexes the true distribution is not unique, the classical likelihood theory of Section 2 no longer applies. Various difficulties accompany the derivation of the asymptotic properties of likelihood-based statistics. For instance, the maximum likelihood estimator may not converge to any point in the parameter space specified by the null hypothesis. Or, the Fisher information matrix degenerates. Typically, the limiting distribution of the likelihood ratio statistics will not be chi-squared. Loss of identifiability occurs in areas as diverse as econometrics, reliability theory and survival analysis (Prakasa Rao, 1992), and has been the subject of intensive research. Rothenberg (1971) studied the conditions under which a general stochastic model whose probability law is determined by a finite number of parameters is identifiable. Paulino and Pereira (1994) present a systematic and unified description of the aspects of the theory of identifiability.

In the remainder of the section we will consider two special cases: non-identifiable parameters and singular information matrix. We will report the main research strains; related contributions can be looked up in the annotated bibliography.

### 4.1 Non-identifiable parameters

The general framework for deriving the asymptotic null distribution of the likelihood ratio statistic was developed by Liu and Shao (2003). They address the common hypothesis testing problem $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta \setminus \Theta_0$, where $\Theta_0 = \{\theta \in \Theta : F_\theta = F^0\}$ with $F_\theta$ the distribution function indexed by $\theta$ and $F^0$ the true distribution. The true distribution is hence unique and $H_0$ is a simple null hypothesis. However, the set $\Theta_0$ may contain more than one value. When the true parameter value $\theta^0$ is not unique, the classical quadratic approximation of the likelihood ratio function in an Euclidean neighbourhood of $\theta^0$ no longer holds. Liu and Shao (2003) bypass this problem by establishing a general quadratic approximation of the likelihood ratio function

$$lr(\theta) = \sum_{i=1}^{n} \log\{\lambda_i(\theta)\},$$

this time in a so-called Hellinger neighbourhood of the true model, which is valid with or without loss of identifiability of the true distribution $F^0$. Here, $\lambda_i(\theta) = \lambda(Y_i; \theta)$ denotes the Radon-Nikodym derivative, $\lambda(\theta) = dF_\theta/dF^0$, evaluated at $Y_i$, for $i = 1, \ldots, n$. The Hellinger neighbourhood of $F^0$ is defined as

$$\Theta_\epsilon = \{\theta \in \Theta \mid 0 < H(\theta) \leq \epsilon\},$$

where

$$H^2(\theta) = \frac{1}{2} E_{F^0}\left[\left\{\sqrt{\lambda_i(\theta)} - 1\right\}^2\right]$$

is the squared Hellinger distance between $F_\theta$ and $F^0$. Under suitable regularity conditions, which assure Hellinger consistency of the maximum likelihood estimator despite loss of identifiability, the distribution of the likelihood ratio statistic

$$W(H_0) = 2 \sup_{\theta \in \Theta \setminus \Theta_0} \{lr(\theta) \vee 0\},$$

with $\{a \vee b\} = \max(a, b)$, converges to the distribution of the square of a left-truncated centered Gaussian process with uniformly continuous sample paths (Liu and Shao, 2003, Theorem 2.3). The proof, which is detailed in Appendix A.2, involves two steps. We first derive a generalized differentiable quadratic in mean (GDQM) expansion of the likelihood ratio function

$$\begin{aligned} lr(\theta) &= 2\sqrt{n} H(\theta) \nu_n(S_i(\theta)) \\ &- nH^2(\theta)\left\{2 + F_n(S_i^2(\theta))\right\} + o_p(1), \end{aligned} \tag{4.1}$$

where $S_i(\theta)$ is such that $E_{F^0}[S_i(\theta)] = 0$, $F_n(\cdot)$ indicates the empirical distribution function and $\nu_n(g) = \sqrt{n}(E_{F_n} - E_{F^0})[g]$ is a random process defined for any integrable function $g$. Expansion of $lr(\theta)$ is valid in a Hellinger neighborhood $\Theta_\epsilon$ of $F^0$ and is not unique. As $lr(\theta)$ can diverge to $-\infty$ for some $\theta \in \Theta_\epsilon$, it is not always easy to find a general approximation with uniform residual terms on $\Theta_\epsilon$. We then have to maximise $\{lr(\theta) \vee 0\}$ which has a general quadratic expansion. This expansion is then used to prove that the distribution of the likelihood ratio function converges to the distribution of the supremum of a squared left-truncated centered Gaussian process with uniformly continuous sample paths. In principle, the distribution of the Gaussian process can be approximated by simulation, since its covariance kernel is known. The most crucial aspect, however, is the derivation of the set which contains the $\mathcal{L}^2$ limits of the generalized score function

$$\frac{S_i(\theta)}{\sqrt{1 + E_{F^0}[S_i^2(\theta)]/2}}$$

over which the supremum is to be taken. This needs be worked out on a case by case basis.

Liu and Shao (2003, Section 3) also consider square-integrable likelihood ratios, for which they derive a quadratic approximation to the likelihood ratio based on the Pearson-type $\mathcal{L}^2$ distance

$$E_{F^0}\left[\{lr(\theta) - 1\}^2\right]$$

using arguments similar to the ones contained in the prototype proof of Appendix A.2. As as prominent example, they work out the results for finite mixture models whose component distributions belong to an exponential family.

An alternative, and less general, contribution is Ritz and Skovgaard (2005). These authors derive the asymptotic distribution of the likelihood ratio and of the related score statistic for a general curved exponential family for which some nuisance parameters vanish under the null hypothesis. Their results are illustrated using the multivariate normal model whose covariance matrix can be written as

$$(\varphi - \varphi_0)\Sigma(\rho) + \gamma_1\Sigma_1 + \cdots + \gamma_k\Sigma_k \tag{4.2}$$

where $\varphi, \rho, \gamma_1, \ldots, \gamma_1$ are unknown variance parameters and $\Sigma(\rho), \Sigma_1, \ldots, \Sigma_k$ are suitable matrices. The null hypothesis $\varphi = \varphi_0$ reduces the model to a random coefficients model, while making the parameter $\rho$ non-identifiable. The results are derived without the need to assume compactness of the parameter space, a condition which, as we will see in Section 5, is generally required when some parameters vanish under the null hypothesis. Again, the proof evolves along two steps and uses argument similar to those provided in Appendix A.3 which we will discuss in Section 5. The likelihood ratio function is first approximated by a quadratic expansion with respect to the identifiable parameter. Under the null hypothesis, this expansion converges to the square of a Gaussian random process indexed by the non-identifiable parameter $\rho$. The supremum of this process with respect to $\rho$ is then taken. The Gaussian process has a covariance function that can be estimated consistently, which allows us to simulate the limiting process. The numerical investigation of Ritz and Skovgaard (2005) shows that the limiting distribution for the motivating example (4.2) lies between a $\bar{\chi}^2(\omega, 1)$ with $\omega = (0.5, 0.5)$ and a $\chi^2_1$ distribution. The authors furthermore show that their approximation performs well also in small or moderate samples, and remains stable over a wide range of parameter values.

## 4.2 Singular information matrix

A further case of indeterminate parameter problem is when Fisher's information matrix is singular at the true value $\theta^0$ of the parameter. Singularity of the information matrix is linked to non-identifiability as shown by the following example.

Example 4.2 (Singularity and non-identifiability). *Consider a normal random variable $Y$ with mean $\theta^q$, for a given even integer $q$, and variance 1. Globally, the parameter $\theta$ is identifiable for $\theta_0 = 0$, although this value results to be a singular point for the information function $i(\theta) = q^2\theta^{2(q-1)}$. Moreover, locally the parameter is identifiable for any $\theta_0 \neq 0$ in an open neighbourhood of $\theta_0$ with non singular information function at that point. Remember that for scalar $\theta$, zero information implies a null score statistic with probability 1, while for multidimensional $\theta$, a singular information matrix implies linear dependence among the different components of the score vector.*

Singularity of $i(\theta)$ can lead to multiple maxima of the log-likelihood function $l(\theta)$ in a neighbourhood of $\theta^0$ and to inconsistency of the maximum likelihood estimator $\hat{\theta}$. Moreover, the limiting distribution of the likelihood ratio statistic may not be chi-squared. The, to our knowledge, earliest contribution who addresses the problem of singular information matrix is Silvey (1959). The author proposes to modify the curvature of the quadratic approximation of the likelihood ratio by replacing the inverse of the Fisher information matrix with a generalized inverse matrix obtained by imposing suitable constraints on the model parameters. The cornerstone contribution to the development of the theory of singular information matrices is Rotnitzky et al. (2000) who derive the asymptotic null distribution of the likelihood ratio statistic for testing the null hypothesis $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, when $\theta$ is a $p$-dimensional parameter of an identifiable parametric model and the information matrix is singular at $\theta_0$ and has rank $p-1$. The theory is developed only for independent and identically distributed random variables, though the authors point out that the same theory may straightforwardly be extended to non-identically distributed observations. When $\theta$ is scalar, the asymptotic properties of the maximum likelihood estimator and of the likelihood ratio statistic depend on the integer $m_0$, which represents the order of the first partial derivative of the log-likelihood function which does not vanish at $\theta = \theta_0$; see Theorems 1 and 2 of Rotnitzky et al. (2000). If $m_0$ is odd, the distribution of the likelihood ratio converges under the null hypothesis to a $\chi^2_1$ distribution, while for even $m_0$ it converges to a $\bar{\chi}^2(\omega, 1)$ with $\omega = (0.5, 0.5)$. Extensions of these results when the parameter $\theta$ is $p$-dimensional are also provided. These are generally based on suitable reparametrizations of the model which remove the specific causes of the singularity, but are difficult to generalize as they are ad-hoc solutions.

## 5. FINITE MIXTURE MODELS

Finite mixture models deserve special attention, because of their widespread use in statistical practice, but also because of the methodological challenges posed by the derivation of their asymptotic properties. They probably represent the best-studied indeterminate parameter problem, though we may also treat them as a boundary case. Indeed, testing a hypothesis such as model homogeneity against the alternative that the model is a finite mixture of two or more components will most likely lead to the failure of two regularity conditions. As we shall see in Section 5.1, this occurs because while under the null hypothesis the mixing proportions fall on the boundary of their parameter space, some of the parameters of the corresponding component distributions become indeterminate. Under this scenario, the asymptotic distribution of the likelihood ratio statistic does not follow the commonly believed chi-squared distribution, and its limiting distribution has for long been unknown.

The remainder of the section outlines the many mainstream contributions for this class of models, with special emphasis on hypothesis testing using the likelihood ratio. Further related work is listed in the annotated bibliography. General reference for mixture distributions are Lindsay (1995) and McLachlan and Peel (2000).

### 5.1 Testing for homogeneity

Consider the two-component mixture model

$$(1-\pi)f_1(y;\theta_1)+\pi f_2(y;\theta_2), \tag{5.1}$$

where the probability density or mass functions $f_1(y;\theta_1)$ and $f_2(y;\theta_2)$, with $\theta_1 \in \Theta_1 \subseteq \mathbb{R}^{p_1}$ and $\theta_2 \in \Theta_2 \subseteq \mathbb{R}^{p_2}$, represent the mixture components and $0 \leq \pi \leq 1$ is the mixing probability. The null hypothesis of homogeneity can be written in different ways. We may set $\pi = 0$, which corresponds to $H_0 : f^0 = f_1(y;\theta_1)$, where $f^0$ represents the true unknown distribution, or alternatively, $\pi = 1$ and $H_0 : f^0 = f_2(y;\theta_2)$. If the two components, $f_1(y;\theta_1)$ and $f_2(y;\theta_2)$, are known, then the limiting distribution is a $\bar{\chi}^2(\omega, 1)$ with $\omega = (0.5, 0.5)$ (Lindsay, 1995, p. 75). Otherwise, for $f_1(y;\theta) = f_2(y;\theta)$ a third possibility arises: in this case homogeneity assumes that $H_0 : \theta_1 = \theta_2$. Whatever choice is made, some model parameters, that is, $\theta_2$ and $\theta_1$, respectively, in the first two cases and $\pi$ in the third, vanish under the null hypothesis. This contradicts classical likelihood theory, where the parameter which characterises the true distribution is typically assumed to be a unique point $\theta^0$ in the open subset $\Theta \subseteq \mathbb{R}^p$. As we have seen in Section 3, the failure of Condition 3 generally implies that the limiting distribution is truncated on its left to account for the fact that the maximum likelihood estimate can only fall on one side of the true parameter value. The failure of Condition 1 in addition implies that there is no value to which the maximum likelihood estimator of the indeterminate parameters can converge.

*5.1.1 General results* The first discussion of asymptotic theory for testing homogeneity of model (5.1) when all parameters are unknown was provided by Ghosh and Sen (1985). As the two authors point out, there is an additional major difficulty in dealing with finite mixture models: though the mixture itself may be identifiable, the parameters $\pi$, $\theta_1$ and $\theta_2$ may not be. For instance, for the simple mixture where $f_1(y;\theta) = f_2(y;\theta) = f(y;\theta)$, the equality

$$\begin{aligned}(1-\pi)f(y;\theta_1)+\pi f(y;\theta_2)\\ =(1-\pi')f(y;\theta_1')+\pi' f(y;\theta_2')\end{aligned}$$

holds for $\pi = \pi'$, $\theta_1 = \theta_1'$, $\theta_2 = \theta_2'$, but also for $1-\pi = \pi'$, $\theta_1 = \theta_2'$, $\theta_2 = \theta_1'$. That is, if the alternative hypothesis is true, there is a second set of parameters which gives rise to the same distribution. Furthermore, under the null hypothesis of homogeneity the model is represented by the three curves $\pi = 1$, $\pi = 0$ and $\theta_1 = \theta_2$. As illustrated by Ghosh and Sen (1985), choosing an identifiable parametrisation doesn't bring any improvement as the density is then no longer differentiable.

The first result derived by Ghosh and Sen (1985) characterises the limiting distribution of the likelihood ratio statistic for strongly identifiable continuous mixtures. Write $f(y;\theta) = (1-\pi)f_1(y;\theta_1) + \pi f_2(y;\theta_2)$ with the convention that $\theta = (\pi, \theta_1, \theta_2)$. Strong identifiability holds if $f(y;\theta) = f(y;\theta')$ implies that $\pi = \pi'$, $\theta_1 = \theta_1'$ and $\theta_2 = \theta_2'$. Ghosh and Sen (1985) furthermore assume that $\Theta_2$ is a closed bounded interval of $\mathbb{R}$, while $\Theta_1 \subseteq \mathbb{R}^{p_1}$, $p_1 \geq 1$. The distribution of the likelihood ratio statistic for testing $H_0 : \pi = 0$ then converges to the distribution of

$T^2 I_{\{T>0\}}$, where $T = \sup_{\theta_2}\{Z(\theta_2)\}$ and $Z(\theta_2)$ is a zero-mean Gaussian process on $\Theta_2$ whose covariance function depends on the true value of the parameters under the null hypothesis (Ghosh and Sen, 1985, Theorem 2.1). This results from proceeding in two steps. We first approximate the log-likelihood function by a quadratic expansion with respect to $\pi$ and $\theta_1$ which, under the null hypothesis, converges to the square of a Gaussian random process indexed by the non-identifiable parameter $\theta_2$. The supremum of this process with respect to $\theta_2$ is then taken. The sketch of this proof is given in Appendix A.3

A similar result holds if the finite mixture is not strongly identifiable, such as when $f_1(y;\theta) = f_2(y;\theta)$ in (5.1). In this case, a separation condition between $\theta_1$ and $\theta_2$ of the form $||\theta_1 - \theta_2|| \geq \epsilon$ for a fixed quantity $\epsilon > 0$ needs be imposed, so that $H_0$ is described by either $\pi = 0$ or $\pi = 1$ (Ghosh and Sen, 1985, §5). The two authors furthermore restrict the parameter space of $\pi$ to $[0, 0.5]$ and impose again that $\Theta_1$ be an open set containing the true value $\theta_1^0$ and $\Theta_2$ be a closed set such that $\Theta_1 \cap \Theta_2 = \emptyset$. These additional conditions guarantee that the maximum likelihood estimate $(\hat{\pi}, \hat{\theta}_1)$ will fall with high probability into the $n^{-1/2}$–neighbourhood of $(0, \theta_1^0)$. The proof outlined in Appendix A.3 still applies with the exception that now the non-identifiable parameter $\theta_2$ varies in a subset of $\Theta_2$ which depends on the given $\epsilon$. Ghosh and Sen (1985, §4) also discuss the link to Bayesian testing and develop asymptotically locally minimax tests for some special cases.

Removing the above separation condition without imposing further constraints is challenging. Several authors have addressed this issue. As we will see in the following section, some require reparametrization of the model function, other penalise the log-likelihood function or rely on simulation.

## 5.2 Alternative approaches

*5.2.1 Reparametrization* The first contribution which, to our knowledge, uses ad hoc reparametrization in place of a separation condition between the parameters $\theta_1$ and $\theta_2$ to derive the limiting distribution of the likelihood ratio statistic for testing model homogeneity, is Chernoff and Lander (1995). The two authors study several versions of the two-component binomial mixture model, which is typically used in linkage analysis. They heuristically prove that the finite-sample null distribution of the likelihood ratio statistic again converges to the supremum of the square of a left-truncated zero-mean unit-variance Gaussian process with well-behaved covariance function. The formal proof is given in Lemdani and Pons (1997) for several classical models. Later, Lemdani and Pons (1999) study the limiting distribution of the likelihood ratio statistic to test whether a known density $f(y;\theta_0)$ is contaminated by another density of the same parametric family. In particular, the null hypothesis corresponds to assuming $f^0 = f(y;\theta_0)$ while under the alternative hypothesis the model becomes $(1-\pi)f(y;\theta_0) + \pi f(y;\theta)$. By reparametrizing to $\mu = \pi||\theta - \theta_0||$, they express the null hypothesis as $H_0 : \mu = 0$, that is, as a function of the single parameter $\mu$, and avoid any separation condition on the parameters $\theta_0$ and $\theta$. The likelihood ratio statistic is again shown to converge to the distribution of the supremum of a squared left-truncated Gaussian process. The result is extended to the case where a mixture of $K_0$ known densities is contaminated by additional $K_1$ ones of the same family. We will come back to this scenario in Section 5.3.

Testing for homogeneity of the two-component mixture model (5.1) is furthermore considered in Ciuperca (2002) who assumes that $f_1(y;\theta)$ belongs to an exponential family and $f_2(y;\theta,\tau) = f_1(y-\tau;\theta)$ is a translation of the same by an unknown amount $\tau \in \mathbb{R}$. Here, the limiting distribution of the likelihood ratio statistic is shown to converge to a fifty-fifty mixture of a point mass at zero and of a distribution which diverges in probability to $+\infty$, and this despite the fact that all parameters are assumed to belong to a compact set. This shows that Condition 3 of Section 2.2 is necessary, but not sufficient.

Dacunha-Castelle and Gassiat (1997, 1999) introduce a reparametrization of the model which they call 'locally conic'. Roughly speaking, the novel parametrization is represented by two parameters, $\alpha$ and $\beta$, in which the Fisher information is normalized to be uniformly equal to one. The first parameter, $\alpha$, represents the 'distance' to the true model and is entirely identifiable under the null hypothesis. It is the point around which now it is possible to have an asymptotic expansions of the log-likelihood function. The second parameter, $\beta$, represents the 'direction' of the perturbation of the model and includes all non-identifiable parts. The key assumption is that the closure of the set of derivatives of the log-likelihood function with respect to $\alpha$ for any $\beta$ at the true value $\alpha^0$ is a Donsker class (van der Vaart and Wellner, 1996). The unboundedness behaviour of the likelihood ratio of Ciuperca (2002) is because their model does not satisfy this latter condition.

*5.2.2 Penalisation* A rather different route is taken in Chen et al. (2001). To overcome the two difficulties of asymptotic theory for mixture models—the boundary problem and non-identifiability under the null hypothesis—they suggest to penalise the log-likelihood function

$$l(\pi, \theta; y) + c \log\{4\pi(1-\pi)\}, \tag{5.2}$$

where the degree of penalisation is controlled by the constant term $c$. As the authors point out, the penalisation term can be justified from the Bayesian perspective. It furthermore guarantees that the maximum likelihood estimate of the mixing proportion $0 < \hat{\pi} < 1$ will not fall on the boundary of the parameter space and that the maximum likelihood estimators of all parameters are consistent under the null hypothesis $\pi = 0$. Provided Conditions 1–5 of their paper hold, the distribution of the modified likelihood ratio statistic derived from (5.2) converges to a $\bar{\chi}^2(\omega, 1)$ distribution with $\omega = (0.5, 0.5)$ instead of the unquestioned supremum of a squared truncated Gaussian random process. Numerical assessment for Poisson and Gaussian mixtures reveals that their proposal competes well with alternative solutions especially with respect to power.

Chen et al. (2008) derive the asymptotic null distribution of the modified likelihood ratio test introduced in Chen et al. (2001) and of a further modification, called the iterative modified likelihood ratio test, for testing model homogeneity against the alternative that the model is a two-component von Mises mixture with unknown mean directions without and with nuisance parameters. A further example of penalisation for von Mises mixtures is Fu et al. (2008); see Example 1.3. Both papers outline how to improve the accuracy of the asymptotic approximation in finite samples.

*5.2.3 Simulation* A third route to investigate the asymptotic null distribution of the likelihood ratio statistic for finite mixture models is by simulation. Thode et al. (1988) consider testing the hypothesis that the sample comes from a normal random variable with unknown mean and unknown variance against the alternative that the sample comes from the two-component Gaussian mixture (5.6) with $\mu_1 \neq \mu_2$ and common variance $\sigma_1^2 = \sigma_2^2 = \sigma^2$. All model parameters are assumed to be unknown. Their extensive numerical investigation shows that the distribution of the likelihood ratio statistic converges very slowly to a limiting distribution, if any exists, and is rather unstable even for sample sizes as large as $n = 1,000$. For very large sample sizes, the empirical distributions rather closely agree with the commonly assumed $\chi^2_2$, though this may be too liberal for small to moderate $n$. This gives little support to Hartigan's (1977) conjecture that the asymptotic distribution may lie between a $\chi^2_1$ and a $\chi^2_2$. An example of application to a study of population genetics is given, motivated by the fact that these studies are typically of small to moderate sample sizes, which justifies the use of empirical approximations. The distribution of the likelihood ratio under the alternative hypothesis (5.6) is investigated numerically in Mendell et al. (1991) for a wide range of mixing proportions $\pi$. The authors conjecture that the limiting distribution is a non-central $\chi^2_2$ distribution.

Böhning et al. (1994) investigate numerically the asymptotic properties of the likelihood ratio statistic for testing homogeneity in the two-component mixture model (5.1) when the component distributions $f_k(y; \theta_k)$, $k = 1, 2$ are binomial, Poisson, exponential or Gaussian with known common variance. They establish that, for sufficiently large sample sizes, the null distribution is well approximated by a $\bar{\chi}^2(\omega, 1)$ which remains stable across the possible range of values for the parameters $\theta_1$ and $\theta_2$, but is model-specific in the sense that the weights $\omega$ depend on the model under consideration. Chen and Chen (2001b) consider the same setting as Böhning et al. (1994), though the component distributions are allowed to belong to a generic parametric family. They show that under suitable conditions which guarantee identifiability of the mixture and regularity of the component distributions $f_k(y; \theta_k)$, the limiting distribution of the likelihood ratio is the distribution of the squared supremum of a left-truncated standard Gaussian process, whose autocorrelation function is explicitly presented; see Sections 2 and 3 of their paper. Chen and Chen (2001b) recommend using resampling to calculate the desired tail probabilities. The procedure is illustrated for normal, binomial and Poisson models.

Lo (2008) shows that the commonly used $\chi^2$ approximation for testing the null hypothesis of a homoscedastic normal mixture against the alternative that the data arise from a heteroscedastic model is reasonable only for samples as large as $n = 2,000$ and component distributions that are well separated under the alternative. Furthermore, the restrictions of Hathaway (1985) need be imposed to ensure that the likelihood is bounded and to rule out spurious maxima under the alternative. Otherwise, the author suggests use of parametric resampling.

### 5.3 Assessing the number of components

Consider now the general $K$-component mixture model

$$\sum_{k=1}^{K} \pi_k f_k(y;\theta_k), \quad K \geq 2, \tag{5.3}$$

where $f_k(y;\theta_k)$ are probability density or mass functions indexed by $\theta_k \in \Theta_k \subseteq \mathbb{R}^{p_k}$ and $0 \leq \pi_k \leq 1$, $k = 1,\dots,K$, with $\sum_{k=1}^{K} \pi_k = 1$. Developing a formal test for the null hypothesis $H_0 : K = K_0$ against the alternative that the mixture includes $K > K_0$ components is a difficult task. Many routes have been taken, including Wald-type statistics derived from moment or alternative estimators, adaptation of model selection techniques and the use of simulation. For instance, using the findings of Vuong (1989), who develop likelihood ratio tests for non-nested models, Lo et al. (2001) claim that in the Gaussian case the distribution of the likelihood ratio statistic based on the Kullback-Leibler information criterion converges under the null hypothesis to a weighted sum of $\chi^2_1$ distributions. Jeffries (2003) disproves this result based on the fact that it requires conditions on the structure of the parameter space that are generally not met when the null hypothesis of a $K_0$-component model holds. Oliveira-Brochado and Martins (2005) give a partial review of these techniques. In the remainder of the section, we focus on the proper likelihood ratio test and its asymptotic distribution.

Using the inequalities on likelihood ratios developed in Gassiat (2002), Azaïs et al. (2006) provide the asymptotic distribution of the likelihood ratio statistic under the null hypothesis of a $K_0$-component model and under contiguous alternatives for a general mixture of parametric populations for a bounded parameter space. More precisely, if we define $\mathbb{K} = [-K, K]$ and $\mathcal{F} = \{f_k, k \in \mathbb{K}\}$ is a parametric set of probability densities on $\mathbb{R}$, they consider testing

$$H_0 : f^0 = f_0 \quad \text{against} \quad H_1 : f^0 : (1-\pi)f_0 + \pi f_k,$$

with $k \in \mathbb{K}$ and $0 \leq \pi \leq 1$. In the particular case of Gaussian components, they prove that if the parameter space is unbounded, the likelihood ratio statistic cannot distinguish the null hypothesis from any contiguous alternative. A by-product of their paper is the characterisation of the asymptotic properties of the likelihood ratio statistic for testing homogeneity of the means in the two-component normal mean mixture model of Section 5.4. Azaïs et al. (2009) consider likelihood ratio testing homogeneity in the general $K$-component model (5.3), with application to Gaussian, Poisson and binomial distributions, and testing for the number of components of a finite mixture with or without a nuisance parameter. A number of conditions need be imposed to avoid divergence of the limiting distribution of the likelihood ratio test.

### 5.4 Gaussian mixtures

Theoretical results are particularly generous if the two-component model is a normal mixture. Goffinet et al. (1992) consider an i.i.d. sample from a $d$-dimensional random variable with density function

$$(1-\pi)\phi_d(y;\mu_1,\Sigma) + \pi\phi_d(y;\mu_2,\Sigma),$$

with $0 \leq \pi \leq 1$ and $\phi_d(y;\mu,\Sigma)$ the $d$-dimensional normal density with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma$. They derive the asymptotic distribution of the likelihood ratio statistic for testing the null hypothesis of homogeneity of the means, that is, $H_0 : \mu_1 = \mu_2$, with known mixing proportion $\pi$. Their Theorem 1 treats the univariate case, while its bivariate extension is given in their Theorem 2. For $d = 1$ the null distribution of the likelihood ratio converges to a $\chi^2_1$ distribution if $\Sigma$ is unknown and $\pi \neq 0.5$. In all other scenarios, it converges to a $\bar{\chi}^2(\omega, 1)$ distribution with $\omega = (0.5, 0.5)$. The convergence rate depends on the mixing proportion $\pi$ and is particularly slow if $\pi$ is close to 0.5.

If $d = 2$ the limiting distribution of the likelihood ratio for known $\Sigma$ is the distribution of

$$\frac{1}{2}\{\sup(0, T)\}^2, \qquad T = Z + \sqrt{W},$$

where $Z$ is the standard normal and $W$ is an independent $\chi^2_2$ random variable. This corresponds to a fifty-fifty mixture of a point mass at zero and the squared sum of a standard normal plus the square root of an independent $\chi^2_2$. No result is given for $d = 2$ and $\Sigma$ unknown.

Chen and Chen (2001a) consider the slightly different univariate setting

$$(1 - \pi)\phi(y; \mu_1, 1) + \pi\phi(y; \mu_2, 1), \tag{5.4}$$

where $\phi(y; \mu, 1)$ is the univariate normal density with unit variance and mean $\mu \in \mathbb{R}$. The mixing proportion $\pi$ is unknown and the two means lie in an interval $[-M, M]$ for $M$ finite. Chen and Chen (2001a) consider two cases: where only $\mu_2$ is unknown and $\mu_1 = 0$, or where both location parameters are unknown. In both cases the asymptotic null distribution of the likelihood ratio statistics for testing homogeneity involves the distribution of the supremum of a squared Gaussian random process. If both means are unknown, $\pi \leq 0.5$ to ensure identifiability and we want to test $\mu_1 = \mu_2 = 0$, the limiting distribution agrees with the distribution of

$$\Big\{\sup_{|t| \leq M} Z(t)\Big\}^2 + W, \tag{5.5}$$

where $Z(t)$, $t \in [-M, M]$, is a Gaussian process and $W$ is an independent chi-squared random variable with one degree of freedom. The Gaussian process $Z(t)$ has zero mean and covariance function (Chen and Chen, 2001a, Theorem 3)

$$Cov\{Z(s), Z(t)\} = \frac{e^{st} - 1 - st}{\sqrt{(e^{s^2} - 1 - s^2)(e^{t^2} - 1 - t^2)}},$$

for $st \neq 0$, and $Cov\{Z(s), Z(t)\} = 0$ when $st = 0$. If instead we want to test the composite hypothesis $\mu_1 = \mu_2$ or the simple hypothesis $\mu_2 = 0$ with the assumption that $\mu_1 = 0$, (5.5) stil holds but the chi-squared term is absent and the expression of the covariance is slightly different; see Chen and Chen (2001a, Theorems 2 and 4).

As mentioned in Section 2.2, the compactness of the parameter space is a necessary condition to avoid that the distribution of the likelihood ratio statistic diverges to infinity. This was already proved by Hartigan (1985) and is an immediate implication of Theorem 2 by Chen and Chen (2001a) as $\{\sup_{|t| \leq M} Z(t)\}^2$ tends in probability to infinity if $M \to \infty$. For the latter proof, see Chernoff and Lander (1995, Section 5.6 and Appendix D).

The generalization to the two-component mixture model

$$(1 - \pi)\phi(y; \mu_1, \sigma^2) + \pi\phi(y; \mu_2, \sigma^2), \tag{5.6}$$

which now includes an unknown variance parameter $\sigma^2 > 0$, can be found in Chen and Chen (2003). They prove that the asymptotic distribution of the likelihood ratio for testing model homogeneity is the distribution of the sum of a $\chi^2_2$ variable and the supremum of the square of a left-truncated Gaussian process with zero mean and unit variance. Again, the correlation structure of the process involved in the limiting distribution is presented explicitly; see their Theorem 2.

The proofs of the Theorems in Chen and Chen (2001a, 2003) essentially are suitable adaptations of the prototype derivation for finite mixture models reported in Appendix A.3. All passages are detailed in the original contributions to which we refer the interested Reader. As in most cases the asymptotic distribution of the likelihood ratio is related to a Gaussian random field, the computation of percentile points becomes tricky or impossible. That is why other tests or methods have been proposed. Reviewing all these would go beyond the scope of the paper. Let us mention, here, the most fruitful research strained initiated by Li et al. (2009) who propose an EM-test for homogeneity, which Chen and Li (2009) decline in the case of a two-component Gaussian mixture. A most recent treatment is Chauveau et al. (2018).

## 6. CHANGE-POINT PROBLEMS

A change-point problem arises when we seek to identify a possible change in the probability distribution of a univariate or multivariate random sequence, in a series of time-dependent observations or in a sample of responses whose regime may suddenly change. A modification in the data generating process generally affects the support of the random variable and/or implies that the log-likelihood function is no longer differentiable with respect to some values of the parameter. This typically leads to the concurrent failure of Conditions 2 and 4 of Section 2.2. Furthermore, setting one of the components of the model to a particular value, can make other components, or parts of it, disappear, as in Example 4.1, which links change-point problems to indeterminate parameter problems.

Change-point problems have been the subject of intensive research owing to their widespread use whenever the constancy over time of random events is questioned. The theory has evolved over the past five decades to the extent that summarizing all contributions would fill in book-length accounts. A first annotated bibliography of change-point problems is Shaban (1980). Krishnaiah and Miao (1988) give an overview of change-point estimation up to their time of writing; Csörgö and Horváth (1997) focus their review monograph on limit theorems for change-point analysis. Khodadadi and Asgharian (2008) is a more than 200 pages length annotated bibliography of change-point problems in regression. Lee (2010) summarizes the most recent literature and gives a comprehensive bibliography for the five types of change-point problems characterised by a shift in the mean, a change in the variance, a switch in the regression slope, a change in the hazard rate or a change in the distribution. A recent book-length account of change-point problems with examples from medicine, genetics and finance is Chen and Gupta (2012). Niu et al. (2016) provide a selected overview of multiple change point detection. The proposed inferential solutions range from parametric to non-parametric techniques and include frequentist and Bayesian approaches. In the remainder of this section we will again focus on the parametric likelihood ratio statistic and its asymptotic distribution.

The most basic change-point problem tries and identifies patterns in a random sequence. Among the earliest contributions is Page (1957). Given $n$ independent observations $y_1, \ldots, y_n$, listed in the order they occurred, Page (1957) considers the problem of verifying whether these were generated by a random variable with distribution function $F(y;\theta)$ against the alternative that only the first $\tau$, $0 \leq \tau < n$, observations are generated from $F(y;\theta)$ while the remaining $n-\tau$ come from $F(y;\theta')$ with $\theta \neq \theta'$ and $\tau$ unknown. We will come back to this problem in Section 6.4. Generally speaking, two questions are of interest in change-point analysis: identifying the unknown number of changes and estimating where these occur, together with further quantities of interest such as the size of the change. As highlighted by Chen and Gupta (2012), the majority of reference models which have been proposed for change-point detection assume normality of the observations. These will be treated extensively in Sections 6.1–6.3 with special emphasis on regression type problems. In particular, Section 6.1 addresses the issue of detecting possible shifts in the location and/or the scale of the distribution. Sections 6.2 and 6.3 extend the treatment to linear regression and piecewise linear models. Section 6.4 resumes the original problem of Page (1957) and discusses change-point detection in a random sequence of discrete or continuous observations. Given the breadth of the available solutions, each section contains a selection of contributions which illustrate the main currents of research. Further related work is listed in the annotated bibliography.

### 6.1 Shifts in location and scale

The reference model for testing a change in the mean value of a random variable can generally be written as

$$y_i = \eta_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{6.1}$$

where the $\varepsilon_i$'s are independent zero-mean random errors. Again, all observations are considered in the order they appear, an assumption which will hold for the whole section. The function $\eta_i$ may change $K$ times,

$$\begin{aligned} \eta_i &= \mu_1, \quad 0 < i \leq \tau_1, \\ &= \mu_2, \quad \tau_1 < i \leq \tau_2, \\ &\vdots \\ &= \mu_{K+1}, \quad \tau_K < i \leq n, \end{aligned} \tag{6.2}$$

where the change-points $\tau_k$ can only assume integer values. Both the $K+1$ different mean values $\mu_k$ and the $K$ change-points $\tau_k$ are generally supposed to be unknown, although the very early contributions focus on the simpler setting where one or both pieces of information are given. The pioneering paper by Page (1955) assumes $K = 1$, a known mean value $\mu$, but unknown change-point $\tau$. The proposed test statistic records the largest difference between the partial deviation $\bar{D}_\tau = \sum_{i=1}^{\tau}(Y_i - \mu)$, for $\tau = 1, \ldots, n$, and its least value, that is,

$$\max_{0<\tau\leq n} (\bar{D}_\tau - \min_{0\leq i<\tau} \bar{D}_i), \quad \text{where } \bar{D}_0 = 0.$$

Large values support the hypothesis that the mean has changed to $\mu'$, with $\mu \neq \mu'$. Table 1 of Page (1955) gives some critial values for the binomial case, and is supplemented by the power calculations of Table 2. The same setting is considered in Hinkley (1970) with the additional assumption that the errors $\varepsilon_i \sim N(0, \sigma^2)$ are centered normal variables with constant variance $\sigma^2 > 0$. Using results from the theory on random walks, Hinkley (1970) determines the asymptotic distribution of the maximum likelihood estimator of $\tau$ and of the likelihood ratio statistic for testing the null hypothesis $H_0 : \tau = \tau_0$, that is, that the change occurred at a given time point $\tau_0$. The former distribution is tabulated in Table 3.3 of the paper, while critical values of the latter are given in Table 4.1 of the same. Numerical investigation shows that the validity of the asymptotic approximations depends on how large the location shift is.

Hawkins (1977) considers the same model than Hinkley (1970) though this time the null hypothesis is of no mean change, that is,

$$H_0 : Y_i \sim N(\mu, \sigma^2), \quad i = 1, \ldots, n,$$

against the alternative that there exists a $0 < \tau < n$ at which the unknown mean switches from $\mu$ to $\mu' \neq \mu$. The variance $\sigma^2$ is assumed to be known and we set it to one without loss of generality. The corresponding likelihood ratio statistic is a function of

$$U^2 = V_{\tau^*} = \max_{1\leq\tau<n} V_\tau,$$

where

$$V_\tau = \tau(\bar{Y}_\tau - \bar{Y})^2 + (n - \tau)(\bar{Y}_{n-\tau} - \bar{Y})^2$$

with

$$\bar{Y}_\tau = \frac{1}{\tau}\sum_{i=1}^{\tau} Y_i, \quad \bar{Y}_{n-\tau} = \frac{1}{(n-\tau)}\sum_{i=\tau+1}^{n} Y_i. \tag{6.3}$$

To derive the exact null distribution of the likelihood ratio statistic, Hawkins (1977) re-expresses $V_\tau$ as

$$V_\tau = T_\tau^2,$$

where

$$T_\tau = \sqrt{\frac{n}{\tau(n-\tau)}}\sum_{i=1}^{\tau}(Y_i - \bar{Y})$$

has standard normal distribution. It follows that the finite-sample distribution of

$$U = \sqrt{V_{\tau^*}} = \max_{1\leq\tau<n} |T_\tau| \tag{6.4}$$

agrees with the distribution of the maximum absolute value attained by a Gaussian process in discrete time having zero mean, unit variance and autocorrelation function given by Expression (3.2) of Hawkins (1977). In particular, the null distribution of $U$ has density function

$$f_U(u) = 2\phi(u)\sum_{\tau=1}^{n-1} g_\tau(u) g_{n-\tau}(u), \tag{6.5}$$

where $\phi(u)$ is the density of the standard normal, $g_1(u) = 1$ for $u \geq 0$ and $g_\tau(u)$ is a recursive function such that

$$g_\tau(u) = \Pr(|T_i| < u, i = 1, \ldots, \tau - 1 \mid |T_\tau| = u).$$

The sketch of the proof of (6.5) is given in Appendix A.4.

The finite-sample null distribution of the likelihood ratio statistic for $\sigma^2$ unknown is worked out in Worsley (1979). The likelihood ratio statistic is now expressed as a function of

$$U = \max_{1\leq\tau<n} (n-2)^{\frac{1}{2}} \frac{|T_\tau|}{S_\tau}, \tag{6.6}$$

where $S_\tau$ is the square root of

$$S_\tau^2 = \sum_{i=1}^{\tau}(Y_i - \bar{Y}_\tau)^2 + \sum_{i=\tau+1}^{n}(Y_i - \bar{Y}_{n-\tau})^2,$$

that is, of the within-group sum of squares of the observations split at $\tau$. Now, $T_\tau \sim N(0, \sigma^2)$ under the null hypothesis of no change and $S_\tau^2/\sigma^2$ follows a $\chi^2$-distribution with $n - 2$ degrees of freedom independently of $T_\tau$. It follows that

$$(n-2)^{\frac{1}{2}} \frac{T_\tau}{S_\tau}$$

distributes like a $t$ distribution with $n-2$ degrees of freedom under $H_0$. Tail probabilities for (6.6) are calculated by numerical integration for sample sizes $n \leq 10$ and using simulation if $10 < n \leq 50$. An approximation to the asymptotic null distribution of (6.6) is provided using Bonferroni-type inequalities. For large $n$, percentage points can be calculated also by using Hawkins's (1977) recursion rule.

To avoid the cumbersome calculation of the exact distribution, Yao and Davis (1986) derive the asymptotic null distribution of $U$ using results from the theory of Brownian processes. Equation (A.4) is rewritten as

$$U = \max_{1\leq\tau<n} \frac{\left|\frac{\tau}{\sqrt{n}}\left(\bar{Y}_\tau - \bar{Y}_n\right)\right|}{\sqrt{\frac{\tau}{n}\left(1-\frac{\tau}{n}\right)}}.$$

Let $\{B(t); 0 \leq t \leq \infty\}$ be a standard Brownian motion. Under $H_0$ the process

$$\left\{\frac{\tau(\bar{Y}_\tau - \mu)}{\sqrt{n}}; 1 \leq \tau \leq n\right\}$$

distributes like $\{B(\tau/n); 1 \leq \tau \leq n\}$. We can hence rewrite $U$ as

$$U = \max_{1\leq t<n} \frac{|B_0(t)|}{\sqrt{t(1-t)}},$$

where $B_0(t) = B(t) - tB(1)$ is a Brownian bridge. A suitably normalized version of $U$ converges then under $H_0$ to the double exponential, or Gumbel, distribution (Yao and Davis, 1986, Theorem 2.1). The same result was derived independently by Irvine (1986).

The theory developed so far has been generalized to the multivariate case and/or to account for a possible change in the scale of the distribution; see Chen and Gupta (2012, §§2.2–2.3 and 3.2–3.3) and the selection of references given in the annotated bibliography. Nonparametric methods for change-point analysis are discussed in Brodsky and Darkhovsky (1993).

### 6.2 Change-point detection in regression

A further extension of Model (6.2) with respect to location,

$$\begin{aligned}(6.7)\quad \eta_i &= \alpha_1 + \beta_1 x_i, \quad 0 < i \leq \tau_1,\\ &= \alpha_2 + \beta_2 x_i, \quad \tau_1 < i \leq \tau_2,\\ &\;\vdots\\ &= \alpha_{K+1} + \beta_{K+1} x_i, \quad \tau_K < i \leq n,\end{aligned}$$

is used for change-point detection in simple linear regression. The early contributions by Quandt (1958, 1960) derive the likelihood ratio statistic under the null hypothesis of no switch against the alternative that the model possibly obeys two separate regimes under the assumption of independent and zero-mean normal error terms $\varepsilon_i$. Under the alternative hypothesis, the variance is furthermore allowed to switch from $\sigma_1^2$ to $\sigma_2^2$ at instant $\tau$, when the linear predictor $\eta_i$ undergoes a structural change. The likelihood ratio statistic

$$W = \max_{3\leq\tau\leq n-3} W(\tau), \tag{6.8}$$

with

$$W = -2\log\left(\frac{\hat{\sigma}_1^{2\tau}\;\hat{\sigma}_2^{2(n-\tau)}}{\hat{\sigma}^{2n}}\right),$$

is a function of the least squares estimators $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ of $\sigma_1^2$ and $\sigma_2^2$, respectively, computed using the corresponding subsets of observations, and of the MLE $\hat{\sigma}^2$ of the common variance $\sigma^2 = \sigma_1^2 = \sigma_2^2$ based upon the entire sample. Quandt (1958) initially conjectured that the asymptotic distribution of $W$ may be $\chi_4^2$ under the null hypothesis of no change. However, the numerical investigation he reported in a later publication for the three sample sizes $n = 20, 40, 60$ (Quandt, 1960, Table 3) revealed that the limiting distribution depends on the number of observations $n$. Quandt (1960) furthermore derives three alternative small-sample test statistics, which he obtains again by splitting the observations $\tau$ into two groups as done for the calculation of (6.8).

Change-point detection in simple linear regression using the likelihood ratio is also the subject of Kim and Siegmund (1989). These authors consider two situations: where only the intercept is allowed to change and where both, the intercept and the slope change. The variance remains constant. Under the first scenario, we reject the null hypothesis of no change for large values of $\max_\tau |U(\tau)|/\hat{\sigma}$, where $\hat{\sigma}^2$ is again the maximum likelihood estimator of the common variance $\sigma^2$ and

$$U(\tau) = \left(\frac{n\tau}{n-\tau}\right)^{1/2} \left[\frac{\bar{Y}_\tau - \hat{\alpha} - \hat{\beta}\bar{x}_\tau}{\sqrt{1 - \frac{\tau}{n-\tau}\frac{(\bar{x}_\tau - \bar{x})^2}{\hat{\sigma}_x^2}}}\right].$$

Here, $\hat{\sigma}_x^2$ is the sample variance of $(x_1, \ldots, x_n)$ and $(\hat{\alpha}, \hat{\beta})$ are the maximum likelihood estimators of $(\alpha, \beta)$. A similar result is derived for the second scenario. The null distribution of the likelihood ratio statistics is shown to depend on the independent variable $x$. Again, the Brownian Bridge process is central to the derivation of the corresponding limiting distributions as in Yao and Davis (1986). Approximations for the corresponding tail probabilities are given by Kim and Siegmund (1989) under reasonably general assumptions.

### 6.3 Piecewise linear models

The piecewise linear or multi-phase regression model with $K$ possibly a priori known change-points is a further extension of model (6.2). Broken-line regression is a particular case, where

$$\begin{aligned}(6.9)\quad \eta_i &= \alpha_1 + \beta_1 x_i, \quad x_i \le \tau_1,\\ &= \alpha_2 + \beta_2 x_i, \quad \tau_1 < x_i \le \tau_2,\\ &\vdots\\ &= \alpha_{K+1} + \beta_{K+1} x_i, \quad \tau_K < x_i \le n,\end{aligned}$$

and, in analogy to Section 6.1, we assume that $x_1 \le x_2 \le \cdots \le x_n$. Note that while in Model (6.7) the changes were in time, that is, occurred as the observed sequence $y_i$ moved from the earlier to its later part, now the changes depend on the covariate $x_i$ as it assumes values from the smallest to the largest. The $\tau_k$'s represent the $x$ values at which the changes occur, while the corresponding time points are identified by the values $i_k$ such that $x_{i_k} \le \tau_k < x_{i_k+1}$. Piecewise linear regression is very popular in a large number of disciplines which include environmental sciences (Piegorsch and Bailer, 1997, Section 2.2; Muggeo, 2008a), medical sciences (Smith and Cook, 1980; Muggeo et al., 2014), epidemiology (Ulm, 1991) and econometrics (Zeileis, 2006). The first contributions date back to the early 60's. A review of likelihood ratio testing for piecewise linear regression up to his time of writing is Bhattacharya (1994). The same author treats also the time-varying situation represented by model (6.7) and the simpler situation of identifying a shift in location considered in Section 6.1.

For a known change-point $\tau$, Sprent (1961) uses the likelihood ratio to test a number of hypotheses on the relationship between the two straight lines which form the broken-line regression model (6.9) with $K = 1$. Successive work by Hinkley (1969, 1971) specifically focuses on making inference on

$$\gamma = \frac{\alpha_1 - \alpha_2}{\beta_2 - \beta_1},$$

which identifies the $x$ value at which the two straight lines cross each other. In particular, Hinkley (1969) focuses on testing whether $\gamma = \gamma_0$ when the variance $\sigma^2$ of the error term $\varepsilon_i$ in model (6.1) is known. He shows that the finite-sample distribution of the likelihood ratio statistic

$$(6.10)\qquad W = \frac{1}{\sigma^2}\left\{D_{i^*}^2(\hat{\gamma}) - D_{i_0}^2(\gamma_0)\right\},$$

where $x_{i_0} \le \gamma_0 < x_{i_0+1}$ and $x_{i^*} \le \hat{\gamma} < x_{i^*+1}$, converges to a $\chi_1^2$ distribution. Here,

$$D_i^2(\gamma) = S_0^2 - S_i^2(\gamma)$$

is the difference between the residual sum of squares $S_0^2$ for a single regression line and the residual sum of squares $S_i^2(\gamma)$ for the two regression lines which are constrained to meet at $x = \gamma$. The maximum likelihood estimate $\hat{\gamma}$ is the value of $\gamma$ which maximises $D_i^2(\gamma)$ over $x_i \le \gamma < x_{i+1}$ and $i = 2, \ldots, n-2$. Numerical investigation suggests that the $\chi_1^2$ approximation works well, especially in the upper tail of the distribution, provided the sample size is sufficiently large. For small $n$, instead, the finite-sample distribution of the likelihood ratio has slightly heavier tails. When $\sigma^2$ is unknown, it is replaced in (6.10) by the residual sum of squares $S_0^2$; in this case the limiting distribution is better approximated by an $F_{1,n-4}$ distribution. An interesting by-product is that the finite-sample distribution of the likelihood ratio statistic for testing $\beta_1 = \beta_2$ when $\sigma^2$ is known is very close to a chi-squared distribution with 3 degrees of freedom. The reasons are unknown. For sure, the problem is ill-defined under the null hypothesis as reflected by the distribution of the maximum likelihood estimator of $\beta = (\beta_2 - \beta_1)/\sigma$, which is clearly non normal and heavily biased.

Hinkley (1971) deepens the investigation of (6.10) by deriving the corresponding limiting distributions for the two cases where: (i) there is no change in $\eta_i$ ($\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$), and (ii) the response $y$ is constant until the change-point $\tau$ ($\beta_1 = 0$). On empirical grounds, Hinkley (1971) suggests as limiting distribution an $F_{3,n-4}$ and an $F_{1,n-4}$, respectively. He furthermore develops confidence intervals for the change-point $\tau$ and joint confidence regions for the change-point and the model parameters. Lund and Reeves (2002) revise Hinkley's (1971) first distributional claim, that is, that the distribution of the likelihood ratio converges to an $F_{3,n-4}$ distribution under the null hypothesis of no change. Their Table 1 gives the critical values of (6.10), which result to be much larger than expected. An approximation for the 95% percentile is furthermore given which holds for $n \geq 100$. Lund and Reeves (2002) conjecture that the asymptotic approximation of the finite-sample distribution of (6.10) may involve the Gumbel distribution, as the likelihood ratio statistic seems to behave under the null hypothesis as the maximum of a sequence of positively correlated $F$ distributions. However, the classic extreme value results would have to be adapted to account for the rather strong dependence structure. Or, the limiting distribution of the likelihood ratio (6.10) may involve the distribution of the supremum of a Brownian Bridge process.

A contribution related to Hinkley (1969) is Feder (1975b) who studies the asymptotic distribution of the likelihood ratio statistic in segmented regression, which are models where the analytical form of $\eta_i$ changes according to the values the covariate $x$ takes on. In particular, he proves that under suitable identifiability conditions the limiting distributions of Wilks and Chernoff still apply. However, if the model is not identified and contains less segments than initially assumed, the likelihood ratio statistics is no longer chi-squared. In case of the broken-line regression model considered so far, that is, with $K = 1$, the limiting null distribution for testing equality of the slopes is rather given by the maximum of a large number of correlated $\chi^2_1$ and $\chi^2_2$ distributions, where their number increases with the sample size. The correlation structure furthermore depends on the spacings of the observations $y_i$ and approaches 1 as $n$ tends to infinity.

All results mentioned so far assume that the model is continuous at the change-point $\tau$. Indeed, Hawkins (1980) points out that the asymptotic behaviour of the likelihood ratio statistic depends strongly on whether this assumptions holds. In case of model (6.9) the condition $\alpha_k + \beta_k\tau_k = \alpha_{k+1} + \beta_{k+1}\tau_k$ needs be satisfied for every change-point $\tau_k$. Otherwise, the model is discontinuous. If so, the distribution of the likelihood ratio for verifying the presence of two segments diverges to infinity.

### 6.4 Changes in random sequences

Several authors applied the likelihood ratio statistic to test for sudden changes in a random sequence. Most results consider continuous probability models which belong to exponential families in the first place. The only results we came across for discrete outcomes consider the binomial and Poisson cases. The asymptotic distribution of the likelihood ratio statistic is generally found, as in Section 6.1, by splitting the observations before and after the change point $\tau$. The remainder of the section illustrates some revealing examples where the test statistics can be unbounded. Further related work is listed in the annotated bibliography.

Worsley (1983) derives the exact null and alternative distributions of the likelihood ratio statistic and of the cumulative sum (cumsum) statistic to detect a possible change in the probability of sequence of independent binomial random variables. These distributions are obtained by conditioning on the total number of successes and using an iterative procedure similar to the one developed by Hawkins (1977). Numerical investigation indicates that the likelihood ratio test is more powerful than the cumsum test if the change occurs at the beginning or towards the end of the sequence, while it is slightly less powerful if the change occurs in the middle of the same. However, the likelihood ratio statistic is not bounded in probability.

Worsley (1986) extends his previous results to test for a change in the mean value of independent observations from an exponential family, with particular emphasis on the exponential distribution. The exact null and alternative distributions of the likelihood-based statistics are found, and their power is compared with a test based on a linear trend statistic. The likelihood ratio is a function of both, of the sample sum $\bar{Y} = \sum_{i=1}^{n} Y_i/n$ and of the partial sums, $\bar{Y}_\tau$ and $\bar{Y}_{n-\tau}$, given at (6.3). These represent the sufficient statistics for the natural parameter $\theta$ which indexes the exponential family under the null hypothesis of no change and of the the natural parameters $\theta_1$ and $\theta_2$, which index the two distributions under the hypothesis that a change occurred at $\tau$. An exact confidence region for the change-point $\tau$ is also derived.

Worsley (1988) considers survival data, in particular testing for a change in the hazard function. The likelihood ratio statistic is shown to be unbounded, but the exact null distribution of a suitably modified likelihood ratio is provided. Modified likelihood ratio statistics for the same setting are furthermore considered by Henderson (1990). Recently, Robbins et al. (2011, 2016) addressed the problem of change-point detection in time series. The former considers the mean-shift model of Section 6.1, while the latter assumes the linear regression model of Section 6.2 and, in the supplementary material, the extension to multi-phase regression of Section 6.3. A present-day example of application is the identification of a possible shift in mean temperature values (Reeves et al., 2007).

Gombay and Horváth (1994) derive the limiting distribution of the likelihood ratio type statistic for testing whether there is a change in the parameter $\theta$ which indexes a general distribution $f(y;\theta)$; this can be seen as the continuation of Page (1957). Given $f(y;\theta)$, the likelihood ratio statistic agrees again with the absolute maximum of the $U$ statistic

$$U_\tau = \max_{1\le\tau\le n-1} \left[-2\log\left\{\sup_{\theta\in\Theta_0}\prod_{i=1}^{n} f(y_i;\theta)\right\} + 2\log\left\{\sup_{\theta\in\Theta_1}\prod_{i=1}^{\tau} f(y_i;\theta)\prod_{i=\tau+1}^{n} f(y_i;\theta)\right\}\right].$$

Using results of extreme value theory, the authors prove that the limit distribution of $U_\tau$, suitably centered and rescaled, converges to a Gumbel distribution under the null hypothesis of no change.

## 7. BEYOND PARAMETRIC INFERENCE

This section reviews cases of interest which do not fit into the previously mentioned three broad model classes, but still fall under the big umbrella of non-standard problems. In particular we will focus on shape constrained inference, a genre of nonparametric problem which leads to highly nonregular models.

As brought to our attention by an anonymous Referee, the asymptotic theory of semiparametric and nonparametric inference has interesting analogues to the classical parametric likelihood theory reviewed in Section 2. Indeed, the parameter space of a semiparametric model is an infinite-dimensional metric space. This makes the model non-standard as we typically consider a real parameter of interest in the presence of an infinitely large nuisance parameter. Despite this departure from regularity, the likelihood ratio statistic still behaves as we would expect it. Murphy and van der Vaart (1997, 2000), for instance, show that the corresponding limiting distribution is chi-squared also when we profile out the infinite-dimensional nuisance parameter. The classical approximations of Section 2 also hold for the asymptotic theory of empirical likelihood (Owen, 1990, 1991); see Chen and Van Keilegom (2009) for a review. These results are quite remarkable given that the underlying distributional assumptions are much less strict.

An area of research which has received much attention in the last decade is nonparametric inference under shape constraints (Samworth and Bodhisattva, 2018). Shape constraints originate as a natural modelling assumption and lead to highly non-regular models. As highlighted by Groeneboom and Jongbloed (2018), the probability/density functions of many of the widely used parametric models satisfy shape constraints. For example, the exponential density is decreasing, the Gaussian density is unimodal, while the Gamma density can be both, depending on whether its shape parameter is smaller or larger than one. Estimation under shape constraints leads to an M-estimation problem where the parameter vector typically has the same length as the sample size and is constrained to lie in a convex cone. Non-regularity arises since the M-estimator typically falls on the face of the cone. As for boundary problems, convex geometry is an essential tool to treat shape constrained problems.

The field of shape constraint problems originated from 'monotone' estimation problems, where functions are estimated under the condition that they are monotone. The maximum likelihood estimator converges typically at the rate $n^{-1/3}$ if reasonable conditions hold, that is, at a slower pace than the $n^{-1/2}$ rate attained by regular problems. Moreover, the MLE has a non-standard limiting distribution known as Chernoff's distribution (Groeneboom and Wellner, 2001). A considerable body of work has studied the asymptotic properties of the nonparametric likelihood ratio statistic under monotonicity. In particular, Banerjee and Wellner (2001) initiated the research strain of testing whether a monotone function $\psi$ assumes the particular value $\psi(t_0) = \psi_0$ at a fixed point $t_0$. An extension to regression is given by Banerjee (2007), who assumes that the conditional distribution $p(y, \theta(x))$, of the response variable $Y$ given the covariate $X = x$, belongs to a regular parametric model, where the parameter $\theta$, or part of it, is specified by a monotone function $\theta(x) \in \Theta$ of $x$.

Other types of shape constraint problems have emerged in the meantime which entail concavity or convexity and uni-modality of the functions to be estimated; see the annotated bibliography. Many high-dimensional problems fall in this framework, which opens frontiers for research in nonregular settings; see for example Bellec (2018). Most recently, Doss and Wellner (2019) showed that the likelihood ratio statistic is asymptotically pivotal if the density is log-concave. The class of log-concave densities has many attractive properties from a statistical viewpoint; an account of the key aspects is given in Samworth (2018). Non-standard limiting distributions characterize shape constrained inferential problems. Generally, the likelihood ratio statistic converges to a limiting distribution which can be described by a functional of a standard Brownian motion plus a quadratic drift. In addition, the limiting distribution is asymptotically pivotal, that is, it doesn't depend on the nuisance parameters, as happens for the common $\chi^2$ distribution of regular parametric problems.

## 8. COMPUTATIONAL ASPECTS AND SOFTWARE

Deriving the asymptotic distribution of the likelihood ratio statistic under non standard conditions is generally a cumbersome task. In some cases the limiting distribution is well defined and usable, as for instance when it boils down to a chi or chi-bar squared distribution. Quite often, however, the analytical approximation is intractable, so as when we have to determine the percentiles of a Gaussian random field. This fact has motivated the development of alternative test statistics whose null distribution presents itself in a more manageable form; see, for instance, the contributions mentioned in Section 5.2.2. Or, we may rely upon simulation, as mentioned in passing in Sections 3.2, 4.1, 5.2.3 and 6.1. A compromise between analytical approximation and simulation is the hybrid approach described in Brazzale et al. (2007, Section 7.7) where parts of the analytical approximation are obtained by simulation. However, simulation becomes useless if the limiting distribution diverges to infinity; a non exhaustive list of examples is provided in Section 6.4 and in paragraphs 5.2–5.4 of the annotated bibliography. Substantive applications in which the approximations have been found useful and details of how to implement the methods in standard computing packages are generally missing.

Reviewing all software contributions which implement likelihood ratio based inference for nonregular problems in a more or less formalized way is beyond the scope of this paper. In the following we try and give a selected list of packages for the numerical computing environment R (R Core Team, 2020). We will again group them into the three broad classes reviewed in the previous Sections 3–6, that is, boundary problems, mixture models and change point problems.

Crainiceanu and Ruppert's (2004) proposal, which tests for a null variance component, is implemented in the `RLRsim` package by Scheipl et al. (2008). We furthermore mention the `varTestnlme` package by Baey and Kuhn (2019) and the `lmeVarComp` package by Zhang (2018). The first agains tests for null variance components in linear and non linear mixed effects model, while the second implements the method proposed by Zhang et al. (2016) for testing additivity in nonparametric regression models.

An account of some early software implementations to handle mixture models can be found in Haughton (1997), in the Appendix of McLachlan and Peel (2000) and also in the Software section of the recent review paper by McLachlan et al. (2019). A most recent implementation for use in astrostatistics is the `TOHM` package by Algeri and van Dyk (2020) which implements a computationally efficient approximation of the likelihood ratio statistic for a multimensional two-component finite-mixture model. The package is also available for the Python programming language. The code provided by Chauveau et al. (2018) for testing a two-component Gaussian mixture versus the null hypothesis of homogeneity using the EM test is available through the `MixtureInf` package by Li et al. (2016). Maximum likelihood estimation in finite mixture models based on the EM algorithm is furthermore addressed in the `mixR` package by Yu (2018), which also considers different information criteria and bootstrap resampling. The `clustBootstrapLRT` function of the `mclust` package by Scrucca et al. (2016) also implements bootstrap inference for the likelihood ratio to test the number of mixture components. A further implementation of the likelihood ratio test for mixture models is the `mixtools` package by Benaglia et al. (2009). All R packages linked to finite mixture models are listed on the CRAN Task View webpage for Cluster Analysis & Finite Mixture Models[1].

The `changepoint` package by Killick and Eckley (2014) considers a variety of test statistics for detecting change points among which the likelihood ratio. The `strucchange` package by Zeileis et al. (2002) provides methods for detecting changes in linear regression models. We may furthermore mention the `segmented` package by Muggeo (2008b) for change point detection in piecewise linear models, the `bcp` package by Erdman and Emerson (2007) for Bayesian analysis of a single change in univariate time series and the `CPsurv` package by Brazzale et al. (2019) for nonparametric change point estimation in survival data.

## 9. DISCUSSION

Non-regularity can arise in many different ways, though all entail the failure of one, at times even two, regularity conditions. Many problems can be dealt with straightforwardly; other require sophisticated tools including limit theorems and extreme value theory for random fields. The wealth of contributions, which has been produced during the last 70 years, testifies that the interest in this type of problems has not faded since they made their entrance back in the early 50's. Most solutions, however, are freestanding and scattered in time and scope. We grouped them into boundary, indeterminate parameter and change-point problems, according to which conditions fail and the type of asymptotic arguments used.

The best-studied nonregular case are boundary problems. Common examples of application are testing for a zero variance component in mixed effect models and constrained one-sided tests. The limiting distribution of the likelihood ratio is generally a chi-bar squared distribution with a number of components and mixing weights that depend on the number of parameters which fall on the boundary, and on the design matrices in regression problems. This is also the only type of problem for which higher order results are available.

Indeterminate parameter problems are far more heterogenous. Apart from finite mixtures, the remaining cases can be put under the two umbrellas of non-identifiable parameters and singular information matrix. The methodological difficulties increase as the limiting distributions depend on the parametric family and on the unknown parameters. If $\theta$ is scalar and we want to test homogeneity against a two-component mixture, the distribution of the likelihood ratio converges to the distribution of the supremum of a Gaussian process. For a larger number of mixture components and/or multidimensional $\theta$, this becomes the distribution of the supremum of a Gaussian random field. In these cases, simulation-based approaches are often needed to obtain the required tail probabilities. Moreover, constraints must be imposed to guarantee identifiability of the mixture parameters. As outlined by Garel (2007), these may act on the parameter space, by bounding it or imposing suitable separation conditions among the parameters, or on the alternative hypotheses which must be contiguous. A further possibility is to penalize the likelihood function so that the limiting distribution of the corresponding modified likelihood ratio statistic is chi-squared or well approximated by a chi-bar squared distribution.

Change-point problems range from the simple situation of detecting an alteration in the regime of a random sequence to identifying a structural break in multiple linear regression with possibly correlated errors. Although in the second case the change point can assume any value, in the first situation it must lie in a discrete set. The behaviour of the likelihood

[1] http://cran.r-project.org/web/views/Cluster.html

ratio heavily depends on whether the model is identifiable and/or the regression function is continuous. In some situations the likelihood ratio statistic for the unknown change-point is unbounded. Limit theorems for processes based on $U$-statistics and extreme value theory for random processes play a central role.

From the more practical point of view, use of the asymptotic distribution of the likelihood ratio statistic loses its appeal once it goes beyond the common $\chi^2$ distribution. As a result, simulation-based tests that circumvent the asymptotic theory are often used. Indeed, simulation may nowadays be used to establish the desired empirical distributions of the estimators and to compute approximations for $p$-values obtained from Wald-type statistics. For the most intricate situations, the authors suggest to use resampling-based techniques, such as parametric and nonparametric bootstrapping, to explore the finite-sample properties of likelihood-based statistics. Methodological difficulties, such as the possibile divergence of the likelihood ratio statistic, and prohibitive computational costs limit, however, this possibility to specific applications.

The review has focused on frequentist hypothesis testing using the likelihood ratio statistic. Maximum likelihood estimation for a class of nonregular cases, which include the three-parameter Weibull, the gamma, log-gamma and beta distributions, is considered in Smith (1985). A significant literature has grown since then, parts of which culminated in the book-length account of techniques for parameter estimation in non-standard settings by Cheng (2017). Most of the difficulties encountered in nonregular settings vanish if the model is analysed using Bayes' rule, though one has always to be cautious. Bayesian and nonparametric contributions were mentioned in passing throughout the paper with suitable links to their frequentist counterparts.

## APPENDIX A: APPENDIX

### A.1 Asymptotic expansion of $(\hat\theta - \theta)$

Let $p = 1$ and $l(\theta)$ be the log-likelihood function for a regular parametric model. Write $l_m = l_m(\theta) = d^m l(\theta)/d\theta^m$ for the derivative of order $m = 2, 3, \ldots$, of $l(\theta)$, while $u = u(\theta) = dl(\theta)/d\theta$ represents the score function. We start by expanding the likelihood equation around $\theta$ to give

$$\begin{aligned} 0 = u(\hat\theta) &= u + (\hat\theta-\theta)l_2 + \frac{1}{2}(\hat\theta-\theta)^2 l_3 + \\ &+ \frac{1}{6}(\hat\theta-\theta)^3 l_4 + \cdots , \end{aligned}$$

where $\hat\theta$ indicates the maximum likelihood estimate. Reordered, this expression gives an asymptotic expansion for $(\hat\theta - \theta)$ of the form

$$\begin{aligned} \hat\theta-\theta &= j^{-1}u + \frac{1}{2}j^{-1}(\hat\theta-\theta)^2 l_3 + \\ &+ \frac{1}{6}j^{-1}(\hat\theta-\theta)^4 l_4 + \cdots , \end{aligned} \tag{A.1}$$

where $j^{-1}$ is the inverse of the observed information $j = -l_2$. Next, iteratively substitute in the right-hand part of (A.1) $\hat\theta - \theta$ with its expansion and rearrange terms; this leads to

$$\begin{aligned} \hat\theta-\theta &= j^{-1}u + \frac{1}{2}j^{-3}u^2 l_3 + \\ &+ \frac{1}{6}j^{-4}(l_4 + 3j^{-1}l_3^2)u^3 + \cdots \end{aligned} \tag{A.2}$$

To reorder the terms in (A.2) according to their asymptotic order, we need to introduce the general notation

$$H_m = l_m - \nu_m, \quad \nu_m = E[l_m(\theta;Y)], \tag{A.3}$$

$m \geq 2$. The score function $u(\theta)$ and $H_m$ are of order $n^{1/2}$ under repeated sampling, while $\nu_m$ is of order $n$. We further write $j = i\{1 - i^{-1}(i-j)\}$ and expand $j^{-1}$ as

$$\begin{aligned} j^{-1} &= i^{-1} + i^{-2}(i-j) + \\ &+ i^{-3}(i-j)^2 + \cdots , \end{aligned} \tag{A.4}$$

where $i = E[j(\theta;Y)]$ is the expected information. Now, inserting (A.4) into (A.2) and using notation (A.3), we may rewrite the asymptotic expansion of $(\hat\theta - \theta)$ to obtain

$$\begin{aligned} \hat\theta-\theta &= i^{-1}u + i^{-2}H_2 u + \\ &+ \frac{1}{2}i^{-3}u^2 v_3 + O_p(n^{-3/2}). \end{aligned} \tag{A.5}$$

See Pace and Salvan (1997, Chapter 9) and Barndorff-Nielsen and Cox (1994, Chapter 5) for a detailed treatment.

### A.2 Prototype demonstrations

PROOF SKETCH A.1. *Boundary problem (Self and Liang, 1987, Theorem 3)* Let $y_1, \ldots, y_n$ be $n$ independent observations on the random variable $Y$, and let $l(\theta)$ denote the associated log-likelihood function, where $\theta$ takes values in the parameter space $\Theta$, a subset of $\mathbb{R}^p$. We want to test whether the true value of $\theta$ lies in the subset of $\Theta$ denoted by $\Theta_0$ versus the alternative that it falls in the complement of $\Theta_0$ in $\Theta$, denoted by $\Theta_1$. Let $\theta^0$ be the true value of $\theta$, which may fall on the boundary of $\Theta$. First, expand $2\left\{l(\theta) - l(\theta^0)\right\}$ around $\theta^0$,

$$\begin{aligned} 2\left\{l(\theta) - l(\theta^0)\right\} &= 2(\theta - \theta^0)^\top u(\theta^0) \\ &\quad - (\theta - \theta^0)^\top i(\theta^0)(\theta - \theta^0) \\ &\quad + o_p(||\theta - \theta^0||^3), \end{aligned}$$

where $u(\theta)$ is the score function, $i(\theta)$ the Fisher information matrix and $||\cdot||$ represents the Euclidean norm. Rewrite this expansion as a function of the variable $\tilde{Z}_n = n^{-1} i_1(\theta^0)^{-1} u(\theta^0)$, where $i(\theta^0) = n i_1(\theta^0)$ and $i_1(\theta^0)$ is the Fisher information matrix associated with a single observation. This yields

$$\begin{aligned} 2\left\{l(\theta) - l(\theta^0)\right\} &= \\ &- \{\sqrt{n}\tilde{Z}_n - \sqrt{n}(\theta - \theta^0)\}^\top i_1(\theta^0) \\ &\qquad \{\sqrt{n}\tilde{Z}_n - \sqrt{n}(\theta - \theta^0)\} \\ &+ u(\theta^0)^\top i(\theta^0)^{-1} u(\theta^0) \\ &+ o_p(||\theta - \theta^0||^3). \end{aligned}$$

Consider now the likelihood ratio statistic

$$\begin{aligned} W &= 2\left\{\sup_{\theta\in\Theta} l(\theta) - \sup_{\theta\in\Theta_0} l(\theta)\right\} \\ &= \sup_{\theta\in\Theta}\Big[-\{\sqrt{n}\tilde{Z}_n - \sqrt{n}(\theta - \theta^0)\}^\top i_1(\theta^0) \\ &\qquad \{\sqrt{n}\tilde{Z}_n - \sqrt{n}(\theta - \theta^0)\}\Big] \\ &- \sup_{\theta\in\Theta_0}\Big[-\{\sqrt{n}\tilde{Z}_n - \sqrt{n}(\theta - \theta^0)\}^\top i_1(\theta^0) \\ &\qquad \{\sqrt{n}\tilde{Z}_n - \sqrt{n}(\theta - \theta^0)\}\Big] \\ &+ o_p(||\theta - \theta^0||^3). \end{aligned}$$

Approximate the two sets $\Theta$ and $\Theta_0$ by the cones $C_{\Theta-\theta^0}$ and $C_{\Theta_0-\theta^0}$ centered at $\theta^0$, respectively, and rewrite the likelihood ratio statistic as

$$\begin{aligned} W = &\sup_{\theta\in C_{\Theta-\theta^0}}\left\{-(\tilde{Z}_n - \theta)^\top i_1(\theta^0)(\tilde{Z}_n - \theta)\right\} \\ &- \sup_{\theta\in C_{\Theta_0-\theta^0}}\left\{-(\tilde{Z}_n - \theta)^\top i_1(\theta^0)(\tilde{Z}_n - \theta)\right\} \\ &+ o_p(||\theta||^3). \end{aligned}$$

Now, $\sqrt{n}\tilde{Z}_n$ converges in distribution to a multivariate normal distribution with mean zero and covariance matrix $i_1(\theta^0)^{-1}$. It follows that for all $\theta$ such that $\theta - \theta^0 = O_p(n^{-1/2})$, the limiting distribution of $W$ becomes

$$\begin{aligned} &\sup_{\theta\in\tilde{C}}\left\{-(Z - \theta)^\top(Z - \theta)\right\} - \\ &\sup_{\theta\in\tilde{C}_0}\left\{-(Z - \theta)^\top(Z - \theta)\right\}, \end{aligned}$$

or equivalently as in Expression (3.3), where $\tilde{C}$ and $\tilde{C}_0$ are the corresponding transformations of the cones $C_{\Theta-\theta^0}$ and $C_{\Theta_0-\theta^0}$, respectively, and $Z$ is multivariate standard normal.

PROOF SKETCH A.2. *Non-identifiable parameter (Liu and Shao, 2003, Theorem 2.3)* Let $Y_1, \ldots, Y_n$ be $n$ independent and identically distributed random observations from the true distribution function $F^0$. Suppose that we want to test $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta \setminus \Theta_0$, where $\Theta_0 = \{\theta \in \Theta : F_\theta = F^0\}$ with $F_\theta$ the distribution indexed by $\theta$. Let

$$lr(\theta) = \sum_{i=1}^{n} \log\{\lambda_i(\theta)\}$$

be the log-likelihood ratio function, where $\lambda_i(\theta) = \lambda(Y_i; \theta)$ denotes the Radon-Nikodym derivative, $\lambda(\theta) = dF_\theta/dF^0$, evaluated at $Y_i$. Define the likelihood ratio statistic as

$$W(H_0) = 2\sup_{\theta\in\Theta\setminus\Theta_0}\{lr(\theta) \vee 0\}, \tag{A.6}$$

where $\{a \vee b\} = \max(a, b)$. Assume that there exists a trio $\{S_i(\theta), H(\theta), R_i(\theta)\}$ which satisfies the generalized differentiable in quadratic mean expansion (GDQM)

$$h_i(\theta) = H(\theta)S_i(\theta) - H^2(\theta) + H^2(\theta)R_i(\theta),$$

with $h_i(\theta) = \sqrt{\lambda_i(\theta)} - 1$. $H(\theta)$ is the Hellinger distance between $F_\theta$ and $F^0$ defined as

$$H^2(\theta) = E_{F^0}\left[\left\{\sqrt{\lambda_i(\theta)} - 1\right\}^2\right]/2$$

and $S_i(\theta)$ and $R_i(\theta)$ are such that $E_{F^0}\{S_i(\theta)\} = E_{F^0}\{R_i(\theta)\} = 0$. Furthermore assume that

$$\sup_{\theta\in\Theta_{c/\sqrt{n}}} |\nu_n\left(S_i(\theta)\right)| = O_p(1)$$

and

$$\sup_{\theta\in\Theta_{c/\sqrt{n}}} |E_{F_n}\left[R_i(\theta)\right]| = o_p(1),$$

for all $c > 0$, where $F_n(\cdot)$ indicates the empirical distribution function and $\nu_n(g) = \sqrt{n}(E_{F_n} - E_{F^0})[g]$ is a random process defined for any integrable function $g$. Here, $\Theta_\epsilon = \{\theta \in \Theta \mid 0 < H(\theta) \le \epsilon\}$ defines the Hellinger neighbourhood of $F^0$. Now, using the GDQM expansion and a Taylor series expansion of $2\log\{1 + h_i(\theta)\}$, the log-likelihood ratio function $lr(\theta)$ can be expressed as

$$\begin{aligned} lr(\theta) &= 2\sum_{i=1}^n \log\{1 + h_i(\theta)\} \\ &= 2\sqrt{n}H(\theta)\nu_n(S_i(\theta)) \\ &- nH^2(\theta)[2 + F_n(S_i^2(\theta))] + o_p(1), \end{aligned} \tag{A.7}$$

in $\Theta_{c/\sqrt{n}}$ for all $c > 0$. Under some general conditions on the trio $\{S_i(\theta), H(\theta), R_i(\theta)\}$ (Liu and Shao, 2003, Theorem 2.2), the quadratic expansion in (A.7) holds uniformly in $\theta \in \Theta_\epsilon$ for some small enough $\epsilon > 0$. Direct maximization of (A.6) by $\sqrt{n}H(\theta)$ allows us to approximate the likelihood ratio statistic by the quadratic form

$$\frac{\{\nu_n(S_i(\theta)) \vee 0\}^2}{1 + E_{F_n}[S_i^2(\theta)]/2} \approx \{\nu_n(S_i^*(\theta)) \vee 0\}^2$$

Let $\mathcal{F}$ be the se of all $\mathcal{L}^2$ limits of the standardized score function

$$S_i^*(\theta) = \frac{S_i(\theta)}{\sqrt{1 + E_{F^0}[S_i^2(\theta)]/2}}$$

as $H(\theta) \to 0$. To complete the proof we assume there exists a centered Gaussian process $\{G_S : S \in \mathcal{F}\}$ on the same probability space of the empirical process $\nu_n$ with uniformly continuous sample paths and covariance kernel $E_{F^0}[G_{S_1}G_{S_2}] = E_{F^0}[S_1S_2]$, for all $S_1, S_2$ belonging to $\mathcal{F}$. Using results from statistical limit theory, it is possibile to prove the following two inequalities

$$W(H_0) \le \sup_{S\in\mathcal{F}}\{G_S \vee 0\}^2 + o_p(1),$$

$$W(H_0) \ge \sup_{S\in\mathcal{F}}\{G_S \vee 0\}^2 + o_p(1),$$

which imply that

$$\lim_{n\to\infty} W(H_0) = \sup_{S\in\mathcal{F}}\{G_S \vee 0\}^2.$$

PROOF SKETCH A.3. *Finite mixture model (Ghosh and Sen, 1985, Theorem 2.1)* Let $y_1, \dots, y_n$ be a sample of $n$ i.i.d. observations from the strongly identifiable mixture model (5.1) and

$$l(\theta) = \sum_{i=1}^n \log\left\{(1-\pi)f_1(y_i;\theta_1) + \pi f_2(y_i;\theta_2)\right\}$$

be the corresponding log-likelihood function. Suppose that $H_0 : \pi = 0$ is true, so the true model density is $f_1(y;\theta_1^0)$, where $\theta_1^0$ is the true value of $\theta_1$. Unless differently stated, all functions and expectations will be evaluated under this assumption, that is, for $\theta^0 = (0, \theta_1^0, \theta_2)$, with arbitrary $\theta_2$. Let $W(H_0)$ be the likelihood ratio statistic

$$\begin{aligned} W(H_0) &= 2\{ \sup_{\substack{\pi\in[0,1]\\ \theta_1\in\Theta_1\\ \theta_2\in\Theta_2}} l(\theta) - \sup_{\substack{\pi=0\\ \theta_1\in\Theta_1\\ \theta_2\in\Theta_2}} l(\theta) \} \\ &= \sup_{\theta_2\in\Theta_2} 2\{ \sup_{\substack{\pi\in[0,1]\\ \theta_1\in\Theta_1}} l(\theta) - \sup_{\substack{\pi=0\\ \theta_1\in\Theta_1}} l(\theta)\}. \end{aligned} \tag{A.8}$$

Expand $l(\theta)$ with respect to the first two components of $\theta = (\pi, \theta_1, \theta_2)$ around $\pi = 0$ and $\theta_1 = \theta_1^0$. This yields

$$l(\theta) = l_1(\theta_1^0) + A_n(\theta) + o_p(1), \tag{A.9}$$

where $l_1(\theta_1) = \sum_{i=1}^n \log f_1(y_i;\theta_1)$ and

$$\begin{aligned} A_n(\theta) &= \pi l_\pi + (\theta_1 - \theta_1^0)^\top l_{\theta_1} + \frac{1}{2}\left\{\pi^2 l_{\pi\pi}\right. \\ &+ 2\pi(\theta_1 - \theta_1^0)^\top l_{\pi\theta_1} \\ &+ \left.(\theta_1 - \theta_1^0)^\top l_{\theta_1\theta_1}(\theta_1 - \theta_1^0)\right\}. \end{aligned}$$

Here, the two indexes $\pi$ and $\theta_1$ denote differentiation with respect to the corresponding parameter components. As shown in Ghosh and Sen (1985), in virtue of the Kuhn-Tucker-Lagrange theorem, the unconstrained supremum of $A_n(\theta)$ becomes

$$\sup_{\substack{\pi\in[0,1]\\ \theta_1\in\Theta_1}} A_n(\theta) = \frac{1}{2}\left\{u_0(\theta_2), u_1^\top\right\} i(\theta_2)^{-1}\left\{u_0(\theta_2), u_1^\top\right\}^\top$$

if $Z_n(\theta_2) \geq 0$ and

$$\sup_{\substack{\pi\in[0,1]\\ \theta_1\in\Theta_1}} A_n(\theta) = \frac{1}{2}u_1^\top i_{11}^{-1} u_1$$

if $Z_n(\theta_2) < 0$, where we define

$$Z_n(\theta_2) = \frac{\left\{u_0(\theta_2) i^{00} + u_1(\theta_2)^\top i^{01}(\theta_2)\right\}}{\{i^{00}(\theta_2)\}^{1/2}}.$$

In the previous three expressions, $u_0(\theta_2) = l_\pi(\theta^0)$, $u_1 = l_{\theta_1}(\theta^0)$, $i$ represents the expected information matrix with respect to $\pi$ and $\theta_1$, $i_{jk}(\theta_2)$ denotes the $(jk)$-th component of $i$, for $j = 0, 1$ and $k = 0, 1$, while $i^{jk}(\theta_2)$ denotes the $(jk)$-th component of $i^{-1}$. Similarly, the constrained supremum of $A_n(\theta)$ is

$$\sup_{\substack{\pi=0\\ \theta_1\in\Theta_1}} A_n(\theta) = \frac{1}{2}u_1^\top i_{11}^{-1} u_1.$$

Using known results on the inversion of block matrices, the likelihood ratio statistic (A.8) reduces to

$$W(H_0) = \sup_{\theta_2\in\Theta_2} Z_n^2(\theta_2)\ I_{\{Z_n\geq 0\}} + o_p(1).$$

To ensure the convergence of $Z_n(\theta_2)$ to the zero-mean Gaussian processes $Z(\theta_2)$, the set $\Theta_2$ needs be bounded and a Lipschitz condition has to hold for the $u_0$ component of the score vector which, in turn, implies tightness of $u_0$. These conditions furthermore guarantee that the remainder term in expansion (A.9) is $o_p(1)$ over the two bounded sets of $\pi$ and $\theta_1$ and uniformly in $\theta_2$.

PROOF SKETCH A.4. *Shift in location for Gaussian model (Hawkins, 1977, Theorem 1)* Given $n$ independent Gaussian observations, we want to test whether

$$Y_i \sim N(\mu, \sigma^2), \quad i = 1, \ldots, n,$$

against the alternative that there exists a $0 < \tau < n$ at which the unknown mean $\mu$ switches to $\mu' \neq \mu$. The variance $\sigma^2$ is assumed to be known; we set it to one without loss of generality. Recall from Section 6.1 that the likelihood ratio statistic can be re-expressed as a function of

$$U = \max_{1\leq\tau<n} |T_\tau|,$$

where

$$T_\tau = \sqrt{\frac{n}{\tau(n-\tau)}}\sum_{i=1}^{\tau}(Y_i - \bar{Y}).$$

The null distribution of $U$ is given at (6.5). The proof considers the following events

$$A_\tau = \{|T_\tau| \in (u, u+du)\},$$

$$B_\tau = \{|T_i| < |T_\tau|, \forall i \in (1, \ldots, \tau-1)\},$$

and

$$C_\tau = \{|T_i| < |T_\tau|, \forall i \in (\tau+1, \ldots, n)\}.$$

Define

$$\begin{aligned} F_U(u+du) - F_U(u) &= \Pr\{U \in (u, u+du)\} \\ &= \Pr\left(\bigcup_{\tau=1}^{n-1}\Big[\{|T_\tau| \in (u, u+du)\} \cap \{|T_\tau| > |T_i|, i \neq \tau\}\Big]\right) \\ &= \sum_{\tau=1}^{n-1}\Pr(A_\tau \cap B_\tau \cap C_\tau) \\ &= \sum_{\tau=1}^{n-1}\Pr(A_\tau)\Pr(B_\tau|A_\tau)\Pr(C_\tau|A_\tau\cap B_\tau). \end{aligned}$$

Since $T_\tau \sim N(0,1)$, we have that

$$\Pr(A_\tau) = 2\phi(u) + o(du).$$

Moreover,

$$\begin{aligned} \Pr(B_\tau|A_\tau) &= \Pr(|T_i| < u, \forall i \in (1, \ldots, \tau-1) \mid |T_\tau| = u) \\ &= g_\tau(u) + o(du), \end{aligned} \tag{A.10}$$

where $g_1(u) = 1$ for $u \geq 0$. Since the series $\{T_1, T_2, \ldots, T_{n-1}\}$ is Markovian, $\{T_1, T_2, \ldots, T_{\tau-1}\}$

and $\{T_{\tau+1}, T_{\tau+2}, \ldots, T_{n-1}\}$ are independent. It follows that the events $B_\tau$ and $C_\tau$ are independent given $T_\tau = u$, that is,

$$\Pr(C_\tau | A_\tau \cap B_\tau) = \mathrm{P}(C_\tau | A_\tau).$$

According to the probability symmetry between $B_\tau$ and $C_\tau$ (Chen and Gupta, 2012, §2.1.1), similar to $\Pr(B_\tau | A_\tau)$, it follows that

$$\Pr(C_\tau | A_\tau) = g_{n-\tau}(u) + o(du). \tag{A.11}$$

Combining (A.10) and (A.11), we obtain

$$\Pr\{U \in (u, u+du)\} = 2\phi(u) \sum_{\tau=1}^{n-1} g_\tau(u) g_{n-\tau}(u) + o(du),$$

which corresponds to Expression (6.5).

## ACKNOWLEDGEMENTS

We thank the Editor, the Associate Editor and three anonymous Referees for bringing to our attention a number of relevant contributions which we hadn't included in the previous version of the manuscript. Their valuable suggestions greatly helped us improve many aspects of the paper. It's furthermore a pleasure to acknowledge discussion with Prof. Ruggero Bellio, Prof. Anthony C. Davison and Prof. Nancy Reid. This research was supported by University of Padova grant no. CPDA101912 *Large- and small-sample inference under non-standard conditions* (Progetto di Ricerca di Ateneo 2010).

## ANNOTATED BIBLIOGRAPHY

*The following section-wise list of references supplements the work cited in the text in an attempt to provide a comprehensive overview of the asymptotic properties of the likelihood ratio statistic in nonregular problems.*

CHANT, D. (1974). On asymptotic tests of composite hypotheses in non standard conditions. *Biometrika*, **61**, 291–298.
[§3.1] *Investigates the limiting distribution of the maximum likelihood estimator when θ lies on the boundary of a closed parameter space.*

CHEN, Y., HUANG, J., NING, Y., LIANG, K.-Y., and LINDSAY, B.G. (2015). A conditional composite likelihood ratio test with boundary constraints. *Biometrika*, **105**, 225–232.
[§3.1] *Generalize Susko's (2013) results to composite likelihoods.*

CHEN, Y., NING, J., NING, Y., LIANG, K.-Y., and BANDEEN-ROCHE, K. (2017). On pseudolikelihood inference for semiparametric models with boundary problems. *Biometrika*, **104**, 165–179.
[§3.1] *Establish the asymptotic behaviour of the pseudo likelihood ratio statistic under semiparametric models when testing the hypothesis that the parameter of interest lies on the boundary of its parameter space.*

DRTON, M. (2009). Likelihood ratio test and singularities. *The Annals of Statistics*, **37**, 979–1012.
[§3.1] *Uses tools from algebraic geometry to study the asymptotic distribution of the likelihood ratio statistic when the true parameter value is a singularity, as for instance a cusp.*

FEDER, P. I. (1968). On the distribution of the log likelihood ratio test statistic when the true parameter is near the boundaries of the hypothesis regions. *The Annals of Mathematical Statistics*, **39**, 2044–2055.
[§3.1] *Extends Chernoff's (1954) results to the case where the true parameter value is 'near' the boundary of $\Theta_0$ and $\Theta_1$. Assumes that the true parameter value $\theta_n^0 = \theta^0 + o(1)$ is a sequence of points, not necessarily in $\Theta_0$ or $\Theta_1$, such that $\theta_n^0$ approaches the boundary given by the intersection $\bar\Theta_0 \cap \bar\Theta_1$ of their complementary subsets $\bar\Theta_0$ and $\bar\Theta_1$.*

MORAN, P. A. P. (1971). Maximum likelihood estimation in non standard conditions. *Mathematical Proceedings of the Cambridge Philosophical Society*, **70**, 441–450.
[§3.1] *Investigates the limiting distribution of the maximum likelihood estimator when θ lies on the boundary of a closed parameter space.*

VU, H. T. V. AND ZHOU, S. (1997). Generalization of likelihood ratio tests under nonstandard conditions. *The Annals of Statitics*, **25**, 897–916.
[§3.1] *Derive the large-sample distribution of estimators obtained from estimating functions for models involving covariates. The non-standard asymptotic distribution of the likelihood ratio statistic for the two-way nested variance components model is derived as an example.*

WOOD, S. N. (2013). A simple test for random effects in regression models. *Biometrika*, **100**, 1005–1010.
[§3.2] *Extends Crainiceanu and Ruppert (2004) to generalized linear mixed models with multiple variance components. Exploits the link between random effects and penalized regression to develop a simple simulation-free test for a null variance component based on the restricted likelihood ratio, which under the null hypothesis follows a weighted sum of squared independent standard normal random variables.*

ZHANG, D. and LIN, X. (2008). Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics. In *Random Effect and Latent Variable Model Selection* (D. B. Dunson Editor), p. 19-36, Lecture Notes in Statistics, Springer-Verlag, New York.
[§3.2] *Extend Stram and Lee (1994) ideas to generalized linear mixed effects models to test if between-subject variation is absent.*

CLAESKENS, G., NGUTI, R. and JANSSEN, P. One-sided tests in shared frailty models. *Test*, **17**, 69–82.

[§3.3] *Propose likelihood ratio tests in frailty models. Extend Maller and Zhou (2003) to allow for the presence of covariates. Consider also the limiting null distribution of the score statistic.*

MALLER, R.A. and ZHOU, X. (2003). Testing for individual heterogeneity in parametric models for event-history data. *Mathematical Methods of Statistics*, **12**, 276–304.

[§3.3] *Show that under minimal conditions on the censoring mechanism, the likelihood ratio statistic for homogeneity testing asymptotically distributes as a* $\bar{\chi}^2(\omega, 1)$ *with* $\omega = (0.5, 0.5)$.

MOLENBERGHS, G. and VERBEKE, G. (2007). Likelihood ratio, score and Wald tests in a constrained parameter set. *The American statistician*, **61**, 22–27.

[§3.3] *Compare the performance of the Wald, score and likelihood ratio statistics in multivariate one-sided testing. Suggest to consider the likelihood ratio as the default choice.*

ANDREWS, D. W. K. and PLOBERGER W. (1994). Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica*, **62**, 1383–1414.

[§4.1]. *Discuss asymptotically optimal tests for the linear model with unknown error variance. Use their results to test for a one-time structural change with unknown change-point and discuss several other examples.*

DAVIES, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, **64**, 247–254.

[§4.1]. *Investigates the construction of optimal likelihood-based tests under loss of identifiability for a two-parameter model when the test statistic follows a normal distribution.*

DAVIES, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, **64**, 33–43.

[§4.1]. *Extends Davies (1977) to the case when the model distribution is chi-squared.*

DAVIES, R. B. (2002). Hypothesis testing when a nuisance parameter is present only under the alternative: Linear model case. *Biometrika*, **89**, 484–489.

[§4.1]. *Extends Davies (1977) and Davies (1987) to the linear model with unknown error variance.*

FORTUNATI, S., GINI, F., GRECO, M., FARINA, A., GRAZIANO, A. and GIOMPAPA, S. (2012). An identifiability criterion in the presence of random nuisance parameters. *Proceedings of the 20th European Signal Processing Conference* (EUSIPCO 2012). Bucharest, August 27–31, 2012.

[§4.1]. *Extend Bowden's (1973) result which connects parameter identifiability to non-singularity of the information matrix to the nuisance parameter case.*

SONG, R., KOSOROK, M. R. and FINE, J. P. (2009). On asymptotically optimal tests under loss of identifiability in semiparametrics models. *The Annals of Statistics*, **37**, 2409–2444.

[§4.1]. *Consider tests of hypothesis when the parameters are not identifiable under the null hypothesis in the context of semiparametric models.*

AITCHISON, J. and SILVEY, S. D. (1960). Maximum likelihood procedures and associated tests of significance. *Journal of the Royal Statistical Society, Series B (Methodological)*, **22**, 154-171.

[§4.2] *Address the problem of singular information matrix when the null hypothesis is specified by constraints on the parameters and the outcome of the test dictates whether it is necessary to provide estimates of these parameters.*

BARNABANI, M. (2002). Wald-based approach with singular information matrix. *Workind Paper*, Department of Statistics "P. Fortunati", University of Florence.

[§4.2] *Proposes to maximise a suitably penalized log-likelihood function which guarantees that the corresponding estimator of the parameter is consistent and asymptotically normal. Allows one to construct a Wald type test statistic which has a limiting chi-squared distribution both under null and alternative hypotheses.*

EL-HELBAWY, A. T. and HASSAN, T. (1994). On the Wald, Lagrangian multiplier and likelihood ratio tests when the information matrix is singular. *Journal of the Italian Statistical Society*, **1**, 51–60.

[§4.2] *Build upon Silvey (1959) and develop modified formulae for the Wald, score and likelihood ratio statistics which, under standard regularity conditions, asymptotically follow a chi-squared distribution with degrees of freedom specified by the number of constraints.*

JIN, F. and LEE, L.-F. (2018). Lasso maximum likelihood estimation of parametric models with singular information matrices. *Econometrics*, **6**, 8.

[§4.2] *Propose to fit the parameters of models with singular information matrix by adaptive lasso while allowing the true parameter vector to lie on the boundary of the parameter space.*

LIU, X., PASARICA, C. and SHAO, Y. (2003). Testing homogeneity in gamma mixture models. *Scandinavian Journal of Statistics*, **30**, 227–239.

[§5.2] *Characterise the asymptotic behaviour of the likelihood ratio for testing model homogeneity against a two-component gamma mixture with known shape and unknown rate parameters. Show that under the null hypothesis the asymptotic distribution agrees with the distribution of the square of Davies's (1977) statistic. Further show that if the unknown rate parameter belongs to an unbounded set, the likelihood ratio diverges to infinity in probability at rate* $O\{\log(\log n)\}$, *in accordance with Hartigan (1985).*

CHEN, J. and KALBFLEISCH, J. D. (2005). Modified likelihood ratio test in finite mixture models with a structural parameter. *Journal of Statistical Planning and Inference*, **129**, 93–107.

[§5.3] *Study a modification of the likelihood ratio statistic similar to that proposed by Chen et al. (2001) to verify the hypothesis of a homogeneous model against the alternative of a Gaussian mixture of two or more components with a common and unknown variance. Show, in particular, that the* $\chi^2_2$ *distribution represents a stochastic upper bound to the limiting null distribution of the test statistic.*

KASAHARA, H. and SHIMOTSU, K. (2015). Testing the number of components in finite mixture models. *Global COE Hi-Stat Discussion Paper Series*, **gd12-259**, Institute of Economic Research, Hitotsubashi University.

[§5.3] *Derive the asymptotic distribution of the likelihood* *ratio statistic for the simpler case of testing whether the mixture contains $K_0$ components or $K_0 + 1$ components. Propose in addition a likelihood-based procedure for generally identifying the number of components.*

GAREL, B. (2001). Likelihood ratio test for univariate Gaussian mixture. *Journal of Statistical Planning and Inference*, **96**, 325–350.
[§5.4] *Discusses seven distinct cases of homogeneity testing using the likelihood ratio for the general two-component Gaussian mixture model* (5.6) *by imposing different restrictions on the means and the variances.*

LEI, Q. L. and QIN, Y. S. (2009). A modified likelihood ratio test for homogeneity in bivariate normal mixtures of two samples. *Journal of Systems Science and Complexity*, **22**, 460–468.
[§5.4] *Extend Qin and Smith (2006) to the two-sample problem by using the modified likelihood statistic.*

LIU, X. and SHAO, Y. (2004). Asymptotics for likelihood ratio test in a two-component normal mixture model. *Journal of Statistical Planning and Inference*, **123**, 61–81.
[§5.4] *Show that the asymptotic distribution of the likelihood ratio for model* (5.6) *is asymptotically equivalent to the distribution of the square of the supremum of the stochastic process studied in Theorem 1 of Bickel and Chernoff (1993). Further show that the likelihood ratio diverges to infinity in probability at rate $O\{\log(\log n)\}$ if the mean parameters are unbounded, in accordance with Hartigan (1985) and Chen and Chen (2001a, Theorem 2).*

POLYMENIS, A. and TITTERINGTON, D. M. (1999). A note on the distribution of the likelihood ratio statistic for normal mixture with known proportions. *Journal of Statistical Computation and Simulation*, **64**, 167–175.
[§5.4] *Analyse empirically the $d = 1$ scenarios treated by Goffinet et al. (1992) and give an heuristic explanation for the slow convergence. Propose to refer to the $\bar{\chi}^2(\tilde{\omega}, 1)$ distribution with suitably defined mixing proportions $\tilde{\omega}$ instead of the theoretical value $\omega = (0.5, 0.5)$ to improve the approximation in finite samples.*

QIN, Y. S. and SMITH, B. (2004). Likelihood ratio test for homogeneity in normal mixtures in the presence of a structural parameter. *Statistica Sinica*, **14**, 1165–1177.
§5.4] *Considers model* (5.6) *of Chen and Chen (2003) with the restrictions on the mean parameters given by Chen and Chen (2001a). Identifiability is guaranteed by setting $\pi \leq 0.5$. In addition the mixing proportion need to satisfy $\min(\pi, 1-\pi) \geq \epsilon$ for some positive $\epsilon < 1/2$. The likelihood ratio asymptotically follows a fifty-fifty mixture of a $\chi^2_1$ and a $\chi^2_2$ distribution under the hypothesis of homogeneity.*

QIN, Y. S. and SMITH, B. (2006). The likelihood ratio test for homogeneity in bivariate normal mixtures. *Journal of Multivariate Analysis*, **97**, 474–491.
[§5.4] *Generalize Qin and Smith (2004) to a bivariate normal mixture model with known covariance matrix under the condition $\min(\pi, 1-\pi) \geq \epsilon$ for some positive $\epsilon < 1/2$ (Theorem 1). In practice, the limiting distribution must be found numerically, though an approximation is provided in their Section 4.*

CHEN, J. and GUPTA, A. K. (1997). Testing and locating variance change points with application to stock prices. *Journal of American Statistical Association*, **92**, 739–747.
[§6.1]. *Test and locate multiple change-points in the variance of a series of independent normal observations with known mean using the Schwarz information criterion.*

CHEN, J. and GUPTA, A. K. (2004). Test and locate multiple change-points in the variance of a series of independent normal observations with known mean using the Schwarz information criterion. *Statistics*, **38**, 17–28.
[§6.1]. *Generalize Chen and Gupta (1997) to the multivariate case.*

HAWKINS, D. M. (1992). Detecting shifts in functions of multivariate location and covariance parameters. *Journal of Statistical Planning and Inference*, **33**, 233–244.
[§6.1]. *Generalizes his 1977 paper to study eight procedures—which, however, do not include the likelihood ratio—for monitoring possible shifts in the mean vector or covariance matrix of an arbitrary multivariate random variable.*

HORVÁTH, L. (1993). The maximum likelihood method for testing changes in the parameters of normal observations. *The Annals of Statistics*, **21**, 671–680.
[§6.1] *Derives the asymptotic distribution of the likelihood ratio statistic for testing whether the mean and/or the variance of a sequence of normal observations changed over time at an unknown point $\tau$.*

INCLÀN, C. (1993). Detection of multiple changes of variance using posterior odds. *Journal of Business and Economic Statistics*, **11**, 289–300.
[§6.1]. *Detects a single possible change-point in the variance of a sequence of independent Gaussian random variables with known common mean using a Bayesian approach.*

JAMES, B., JAMES, K. L. and SIEGMUND, D. (1987). Tests for a change-point. *Biometrika*, **74**, 71–83.
[§6.1]. *Compare various test statistics for detecting mean shifts in univariate normal distributions, which also include the likelihood ratio.*

JAMES, B., JAMES, K. L. and SIEGMUND, D. (1992). Asymptotic approximation for likelihood ratio test and confidence regions for a change point in mean of a multivariate normal distribution. *Statistica Sinica*, **2**, 69–90.
[§6.1]. *Compare various test statistics for detecting mean shifts as in James et al. (1987), but this time for the multivariate case.*

SEN, A. and SRIVASTAVA, M. S. (1975). Some one-sided tests for change in level. *Technometrics*, **17**, 61–64.
[§6.1]. *Provide a Bayesian solution to the problem considered in Hawkins (1977).*

SRIVASTAVA, M.S. and WORSLEY, K.J. (1986). Likelihood ratio tests for a change in the multivariate mean. *Journal of the American Statistical Association*, **81**, 199–204.
[§6.1]. *Consider change-point detection in location for the multivariate normal distribution. The likelihood ratio statistic is shown to be equivalent to the maximum of Hotteling's two sample statistic and that the same statistic can be used to test for extra-multinomial variation in a contingency table.*

TANG, J. and GUPTA, A. K.(1988). On testing homogeneity of variances for Gaussian models. *Journal of Statistical*

*Computation and Simulation*, **27(2)**, 155–173.
[§6.1]. *Use Bartlett's statistic to detect a single possible change-point in the variance of a sequence of independent Gaussian random variables with known common mean.*

Andrews, D. W. K.(1993). Tests for parameter instability and structural change with unknown change point. *Econometrica*, **61**, 821–856.
[§6.2]. *Considers Wald, score and likelihood ratio type tests based on the generalized method of moments to detect a possible structural change in multiple regression with unknown change-point. Tables of critical values are provided. If the change-point is fixed, the test statistics follow a chi-squared distribution with p degrees of freedom.*

Andrews, D. W. K., Lee, I. and Ploberger, W. (1996). Optimal changepoint tests for normal linear regression. *Journal of Econometrics*, **70**, 9–38.
[§6.2]. *Extend Andrews (1993) and determine a class of finite-sample optimal tests for the existence of a single or multiple changes at unknown time points in multiple linear regression with normal errors and known variance. Exact critical values are obtained straightforwardly on a case by case basis using simulation. Simulation is furthermore used to compare the power of the proposed test statistics.*

Aue, A., Horváth, L., Hušková, M. and Kokoszka, P. (2008). Testing for changes in polynomial regression. *Bernoulli*, **14**, 637–660.
[§6.2]. *Test the null hypothesis of no change in the polynomial regression parameters against the alternative of a break at an unknown time point. Derive the extreme value distribution of a maximum type test statistic which is asymptotically equivalent to the likelihood ratio. The approximation proves to work well even for small sample sizes. Previous ground breaking work for likelihood ratio testing in polynomial regression is reviewed in their Section 1.*

Deutsch, J. (1992). Linear regression under two separate regimes: An empirical distribution for Quandt's log-likelihood ratio. *Applied Economics*, **24**, 123–127.
[§6.2]. *Provides an empirical distribution for Quandt's (1960) statistic to verify the existence of a change point when assessing the stability of regression relationships over time.*

Kelly, G. E. (2015). Approximations to the p-values of tests for a change-point under non-standard conditions. *Journal of Statistical Computation and Simulation*, **86**, 1430–1449.
[§6.2]. *Considers three variants of the likelihood-based statistics studied by Andrews and Ploberger (1994) for the general regression setting with time trend regressors. Critical values are obtained via simulation.*

Kim, H.-J. (1994). Tests for a change-point in linear regression. In *Change-point Problems. IMS Lecture Notes - Monograph Series* (edited by E. Carlstein, H.-G. Möller and D. Siegmud), **23**, pp. 170–176. IMS, Hayward.
[§6.2]. *Extend Kim and Siegmund (1989) to multiple linear regression.*

Kim, H. J. and Cai, L. (1993). Robustness of the likelihood ratio test for a change in simple linear regression. *Journal of the American Statistical Association*, **88**, 864–871.
[§6.2] *Study the robustness of the likelihood ratio test in simple linear regression.*

Luo, X., Turnbull, B. W. and Clark, L. C. (1997). Likelihood ratio tests for a changepoint with survival data. *Biometrika*, **84**, 555-565. [§6.2]. *Derive the asymptotic distribution of the likelihood ratio statistic to test for a possible time-lag effect in covariates in the presence of right-censored observations.*

Feder, P. I. (1975a). On asymptotic distribution theory in segmented regression problems – identified case. *Annals of Statistics*, **3**, 49–83.
[§6.3]. *Puts the bases for Feder (1975b) by developing the asymptotic distribution theory of least squares estimators in broken line regression.*

Knowles, M., Siegmund, D. and Zhang, H. P. (1991). Confidence semilinear regression, *Biometrika*, **78** 15–31.
[§6.3]. *Provide confidence intervals and joint confidence regions based on the likelihood ratio statistic for the change-point in a broken-line regression model with $K = 1$.*

Koul, H. L. and Qian, L. (2002). Asymptotics of maximum likelihood estimator in a two-phase linear regression model. *Journal of Statistical Planning and Inference*, **108**, 99–119.
[§6.3]. *Consider two-phase linear regression with arbitrary error distribution and fixed jump in the linear predictor at the true change-point $\tau$. The maximum likelihood estimator $\hat{\tau}$ is shown to be consistent and the finite-sample distribution of the standardized maximum likelihood estimator to converge weakly to the distribution of a compound Poisson process.*

Robison, D. E. (1964). Estimates for the points of intersection of two polynomial regressions. *Journal of the American Statistical Association*, **59**, 214–224.rob
[§6.3]. *Generalize Sprent (1961) to polynomial regression.*

Siegmund, D. O. and Zhang, H. (1994). Confidence regions in broken-line regression. In *Change-point Problems. IMS Lecture Notes - Monograph Series* (edited by E. Carlstein, H.-G. Möller and D. Siegmud), **23**, pp. 292–316. IMS, Hayward. int Problems.
[§6.3]. *Provide as Knowles et al. (1991) confidence intervals and joint confidence regions based on the likelihood ratio statistic for the change-point in a broken-line regression model with $K = 1$.*

Haccou, P., Meelis, E. and van de Geer, S. (1987). The likelihood ratio test for the change point problem for exponentially distributed random variables. *Stochastic Processes and their Applications*, **27**, 121–139.
[§6.4]. *Show that under the null hypothesis of no change in the rate parameter of an exponential distribution, the distribution of the likelihood ratio statistic converges to an extreme value distribution. The limiting distribution of the likelihood ratio statistic is obtained by using the theory of uniform quantile process.*

Horváth, L. (1989). The limit distributions of the likelihood ratio and cumulative sum tests for a change in binomial probability. *Journal of Multivariate Analysis*, **31**, 148–159.
[§6.4]. *Derive limit theorems for the likelihood ratio for a change in a sequence of binomial distributions.*

Loader, C. R. (1992). A log-linear Model for a Poisson process change point. *The Annals of Statistics*, **20**, 1391–1411.
[§6.4]. *Tests for the presence of a change point in a non-*

*homogeneous Poisson process. Large deviation techniques are used to approximate the significance level, and approximations for the power function are provided. A coal mining accident data set is used to illustrate the methodology.*

Sadooghi-Alvandi, S. M., Nematollahi, A. R. and Habibi, R. (2011). Test procedures for change point in a general class of distributions, *Journal of Data Science*, **9**, 111–126.

[§6.4]. *Consider change point detection for a general class of distributions. Derive the exact and asymptotic null distributions of the quasi-Bayes and likelihood ratio statistics using results from the theory of Brownian motion and bridge processes. Compare the performance of the two test statistics. Tabulate the significance levels and powers of the two procedures for a number of selected values of the parameters.*

Siegmund, D. (1988). Confidence sets in change-point problems. *International Statistical Review*, **56**, 31–48.

[§6.4]. *Discusses several methods, based on the likelihood ratio, for the construction of a confidence interval for the change-point in a sequence of independent observations from completely specified distributions. The results are generalised to the construction of confidence regions for the change-point and the parameters which index the exponential family from which the independent observations are drawn.*

Visek, T. (2003). The likelihood ratio method for testing changes in the parameters of double exponential observations. textitJournal of Statistical Planning and Inference, **113**, 79–111.

[§6.4]. *Constructs procedures for testing a change in the distribution of a sequence of independent and identically distributed random variables which follow a double exponential law. The change can occur either in the location of the distribution, in the scale or in both.*

Banerjee, M. (2008). Estimating monotone, unimodal and U-shaped failure rates using asymptotic pivots. *Statistica Sinica*, **18(2)**, 467–492.

[§7]. *Proposes a method, based on asymptotic pivots, for constructing nonparametric confidence sets for a monotone failure rate, and for unimodal or U-shaped hazards.*

Groeneboom, P. and Jongbloed, G. (2015). Nonparametric confidence intervals for monotone functions. *The Annals of Statistics*, **43(5)**, 2019–2054.

[§7]. *Obtain confidence intervals for distribution functions and monotone densities by inverting the acceptance region of the nonparametric likelihood ratio test.*

## REFERENCES

*The following section-wise list of references supplements the work cited in the text in an attempt to provide a comprehensive overview of the asymptotic properties of the likelihood ratio statistic in nonregular problems.*


Algeri, S. and van Dyk, D. A. (2020). Testing one hypothesis multiple times: the multidimensional case. *Journal of Computational and Graphical Statistics*, **29**, 358–371.

Andrews, D. W. K. (2001). Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica*, **69**, 683–734.

Azaïs, J.-M., Gassiat, É. and Mercadier, C. (2006). Asymptotic distribution and local power of the log-likelihood ratio test for mixtures: bounded and unbounded cases. *Bernoulli*, **12**, 775–799.

Azaïs, J.-M., Gassiat, É. and Mercadier, C. (2009). The likelihood ratio test for general mixture models with or without a structural parameter. *ESAIM: Probability and Statistics*, **13**, 301–327.

Azzalini, A. (1996). *Statistical Inference Based on the likelihood*. Chapman & Hall, London.

Baey, C. and Kuhn, E. (2019). varTestnlme: variance components testing in mixed-effect models. https://github.com/baeyc/varTestnlme

Banerjee, M. (2007). Likelihood based inference for monotone response models. *The Annals of Statistics*, **35(3)**, 931–956.

Banerjee, M. and Wellner, J. A. (2001). Likelihood Ratio Tests for Monotone Functions. *The Annals of Statistics*. **29(6)**, 1699–1731.

Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and Asymptotics*. Chapman & Hall, London.

Bellec, P. C. (2018). Sharp oracle inequalities for Least Squares estimators in shape restricted regression. *Annals of Statistics*, **46(2)**, 745–780.

Benaglia, T., Chauveau, D., Hunter, D. R., Derek, Y. (2009). mixtools: An R Package for Analyzing Finite Mixture Models. *Journal of Statistical Software*, **32(6)**, 1–29.

Bhattacharya, P. K. (1994). Some aspects of change-point analysis. In *Change-point Problems. IMS Lecture Notes - Monograph Series* (edited by E. Carlstein, H.-G. Möller and D. Siegmud), **23**, pp. 28–56. IMS, Hayward.

Bickel, P. and Chernoff, H. (1993). Asymptotic distribution of the likelihood ratio statistic in a prototypical non regular problem. In *Statistics and Probability: A Raghu Raj Bahadur Festschrift* (edited by J. K. Ghosh, S. K. Mitra, K. R. Parthasarathy and B. L. S. Prakasa), pp. 83–96. Wiley Eastern, New Dehli.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P. and Lindsay, B. G. (1994). The distribution of the likelihood ratio for mixture of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, **46**, 373–388.

Bowden, R. (1973). The Theory of Parametric Identification. *Econometrica*, **41**, 1069–1074.

Brazzale, A. R., Davison, A. C. and Reid, N. (2007). *Applied Asymptotics: Case Studies in Small Sample Statistics*. Cambridge University Press, Cambridge.

Brazzale, A. R., Küchenhoff, H., Krügel, S., Schiergens, T. S., Trentzsch, H. and Hartl, W. (2019). Nonparametric change point estimation for survival distributions with a partially constant hazard rate. *Lifetime Data Analysis*, **25**, 301–321.

Brodsky, B.E. and Darkhovsky, B.S. (1993). *Nonparametric Methods in Change Point Problems*. Kluwer Academic, Dordrecht.

Chauveau, D., Garel, B. and Mercier, S. (2018). Testing for univariate Gaussian mixture in practice. URL: hal.archives-ouvertes.fr/hal-01659771/, Version 2.

Chen, H. and Chen, J. (2001a). Large sample distribution of the likelihood ratio test for normal mixtures. *Statistics & Probability Letters*, **52**, 125–133.

Chen, H. and Chen, J. (2001b). The likelihood ratio test for homogeneity in finite mixture models. *The Canadian Journal of Statistics*, **29**, 201–215.

Chen, H. and Chen, J. (2003). Tests for homogeneity in normal mixtures in the presence of a structural parameter. *Statistica Sinica*, **13**, 351–365.

Chen, H., Chen, J. and Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society, Series B (Methodological)*, **63**, 19–29.

Chen, J. and Gupta, A. K. (2012). *Parametric Statistical Change Point Analysis with Application to Genetics, Medicine and Finance* (2nd ed.). Birkhäuser, Boston.

Chen, J. and Li, P. (2009). Hypothesis test for normal mixture models: the EM approach. *The Annals of Statistics*, **37**, 2523–2542.

Chen, J., Li, P. and Fu, Y. (2008). Testing homogeneity in a mixture of von Mises distributions with a structural parameter. *The Canadian Journal of Statistics*, **36**, 129–142.

Chen, S. X. and Van Keilegom, I. (2009). A review on empirical likelihood methods for regression. *TEST*, **18**, 415–447.

Cheng, R. (2017). *Non-Standard Parametric Statistical Inference*. Oxford University Press, New York.

Cheng, R. C. H. and Traylor, L. (1995). Non-regular maximum likelihood problems (with Discussion). *Journal of the Royal Statistical Society, Series B (Methodological)*, **57**, 3–44.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics*, **54**, 573–578.

Chernoff, H. and Lander, E. (1995). Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial. *Journal of Statistical Planning and Inference*, **43**, 19–40.

Ciuperca, G. (2002). Likelihood ratio statistic for exponential mixtures. *The Annals of the Institute of Statistical Mathematics*, **54**, 585–594.

Cox, D. R. (2006). *Principles of Statistical Inference*. Cambridge University Press, New York.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman & Hall, London.

Crainiceanu, C. M. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society, Series B (Methodological)*, **66**, 165–185.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

Csörgö, M. and Horváth, L. (1997). *Limit Theorems in Change-point Analysis*. Wiley, New York.

Dacunha-Castelle, D. and Gassiat, É. (1997). Testing in locally conic models, and application to mixture models. *ESAIM: Probability and Statistics*, **1**, 285–317.

Dacunha-Castelle, D. and Gassiat, É. (1999). Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes. *The Annals of Statistics*, **27**, 1178–1209.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer-Verlag, New York.

Davison, A. C. (2003). *Statistical Models*. Cambridge University Press, Cambridge.

del Castillo, J. and Lopez-Ratera, A. (2006). Saddlepoint approximation in exponential models with boundary points. *Biometrika*, **12**, 491–500.

Doss, C. R. and Wellner, J. A. (2019). Inference for the mode of a log-concave density. *The Annals of Statistics*, **47**, 2950–2976.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press, USA.

Erdman, C. and Emerson, J. W. (2007). bcp: An R Package for Performing a Bayesian Analysis of Change Point Problems. *Journal of Statistical Software*, **23**, 1–13.

Feder, P. I. (1975b). The log likelihood ratio in segmented regression. *The Annals of Statistics*, **3**, 84–97.

Feng, Z. and McCulloch, C. E. (1992). Statistical inference using maximum likelihood estimation and the generalized likelihood ratio when the true parameter is on the boundary of the parameter space. *Statistics & Probability Letters*, **13**, 325–332.

Feng, C., Wang, H. and Tu, X. M. (2012). The asymptotic distribution of a likelihood ratio test statistic for the homogeneity of poisson distribution. *Sankhyā A*, **74**, 263–268.

Fu, Y., Chen, J. and Li, P. (2008). Modified likelihood ratio test for homogeneity in a mixture of von Mises distributions. *Journal of Statistical Planning and Inference*, **138**, 667–681.

Garel, B. (2007). Recent asymptotic results in testing for mixtures. *Computational Statistics & Data analysis*, **51**, 5295–5304.

Gassiat, É. (2002). Likelihood ratio inequalities with applications to various mixtures. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques*, **6**, 897–906.

Ghosh, J. K. and Sen, K. P. (1985). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (edited by L. LeCam, R. A. Olshen and C.-S. Cheng), Vol. II, pp. 789–806. Wadsworth Advanced Books & Software, Monterey.

Godambe, V. P. (1991). *Estimating Functions*. Oxford University Press, Oxford.

Goffinet, B., Loisel, P. and Laurent, B. (1992). Testing in normal mixture models when the proportions are known. *Biometrika*, **79**, 842–846.

Gombay, E. and Horváth, L. (1994). An application of the maximum likelihood test to the change-point problem. *Stochastic Processes and their Applications*, **50**, 161–171.

Gombay, E. and Horváth, L. (1999). Change-points and bootstrap. *Environmetrics*, **10**, 725–736.

Groeneboom, P. and Jongbloed, G. (2018). Some Developments in the Theory of Shape Constrained Inference. *Statistical Science*, **33**, 473–492.

Groeneboom, P. and Wellner, J. A. (2001). Computing Chernoff's distribution. *Journal of Computational and Graphical Statistics*, **10**, 388–400.

Hartigan, J. A. (1977). Distribution problems in clustering. In *Classification and Clustering* (edited by J. van Ryzin), pp. 45-72. Academic Press, New York.

Hartigan, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (edited by L. LeCam, R. A. Olshen and C.-S. Cheng), Vol. II, pp. 807–810. Wadsworth Advanced Books & Software, Monterey.

Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, **13**, 795–800.

Haughton, D. (1997). Packages for estimating finite mixtures: a review. *The American Statistician*, **51**, 194–205.

Hawkins, D. M. (1977). Testing a sequence of observations for a shift in location. *Journal of the American Statistical Association*, **72**, 180–186.

Hawkins, D. M. (1980). A note on continuous and discontinuous segmented regressions. *Technometrics*, **22**, 443–444.

Henderson, R. (1990). A problem with the likelihood ratio test for a change-point hazard rate model. *Biometrika*, **77**, 835–843.

Hinkley, D. V. (1969). Inference about the intersection in two-phase regression. *Biometrika*, **56**, 495–504.

Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, **57**, 1–17.

Hinkley, D. V. (1971). Inference in two-phase regression. *Journal of the American Statistical Association*, **66**, 736–743.

Hirano, K. and Porter, J. R. (2003). Asymptotic efficiency in parametric structural models with parameter-dependent support. *Econometrica*, **71**, 1307–1338.

Hogg, R. V., McKean, J. W. and Craig, A. T. (2019). *Introduction to Mathematical Statistics* (8th ed.). Pearson, Boston.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics* (2nd ed.). John Wiley & Sons, Hoboken, New Jersey.

Huzurbazar, V. S. (1948). The likelihood equation, consistency and the maxima of the likelihood function. *Annals of Eugenics*, **14**, 185–200.

Irvine, J. M. (1986). The asymptotic distribution of the likelihood ratio test for a change in the mean. *Statistical Research Division Report Series*, **CENSUS/SRD/RR-86/10**, Bureau of the Census, Washington, D.C. 20233.

Jeffries, N. O. (2003). A note on "Testing the number of components in a normal mixture", *Biometrika*, **90**, 991–994.

Khodadadi, A. and Asgharian, M. (2008). Change-point problem and regression: an annotated bibliography. *COBRA Preprint Series*, Working Paper **44**. https://biostats.bepress.com/cobra/art44

Killick, R. and Eckley, I.A. (2014). changepoint: An R Package for Changepoint Analysis. *Journal of Statistical Software*, **58**, 1–19.

Kim, H. J. and Siegmund, D. (1989). The likelihood ratio test for a change-point in simple linear regression. *Biometrika*, **76**, 409–423.

Koenker, R., Chernozhukov, V., He, X. and Peng, L. (2017). *Handbook of Quantile Regression.* Chapman and Hall/CRC, Boca Raton, FL.

Kopylev, L. (2012). Constrained parameters in applications: Review of issues and approaches. *International Scholarly Research Network ISRN Biomathematics.*

Kopylev, L. and Sinha, B. (2011). On the asymptotic distribution of likelihood ratio test when parameters lie on the boundary. *Sankhyā B*, **73**, 20–41.

Krishnaiah, P. R. and Miao, B. Q. (1988). Review about estimation of change points. In *Handbook of Statistics* (edited by P. R. Krishnaiah and C. R. Rao), **Vol. 7**, pp. 375–402. Elsevier, Amsterdam.

Kudô, A. (1963). A multivariate analogue of the one-sided test. *Biometrika*, **50**, 404–418.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics*, **95**, 391–413.

LeCam, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *The Annals of Mathematical Statistics*, **41**, 802–828.

Lee, T.-S. (2010). Change-point problems: bibliography and review. *Journal of Statistical Theory and Practice*, **4**, 643–662.

Lemdani, M. and Pons, O. (1997). Likelihood ratio tests for genetic linkage. *Statistics and Probability Letters*, **33**, 15–22.

Lemdani, M. and Pons, O. (1999). Likelihood ratio tests in contamination models. *Bernoulli*, **5**, 705–719.

Li, S., Chen, J. and Li, P. (2016). MixtureInf: Inference for Finite Mixture Models. R package version 1.1. https://CRAN.R-project.org/package=MixtureInf

Li, P., Chen, J. and Marriot, P. (2009). Non-finite Fisher information and homogeneity: an EM approach. *Biometrika*, **96**, 411–426.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry and Applications.* Institute of Mathematical Statistics, Hayward.

Liu, X. and Shao, Y. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *The Annals of Statistics*, **31**, 807–832.

Lo, Y. (2008). A likelihood ratio test of a homoscedastic normal mixture against a heteroscedastic normal mixture. *Statistics and Computing*, **18**, 233–240.

Lo, Y., Mendell, N. R. and Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, **88**, 767–778.

Lund, R. and Reeves, J. (2002). Detection of undocumented changepoints: A revision of the two-phase regression model. *Journal of Climate*, **15**, 2547–2554.

McLachlan, G., Lee, S. X. and Rathnayake, S. (2019). Finite mixture models. *Annual Review of Statistics and Its Application*, **6**, 355–378.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* John Wiley & Sons, New York.

Mendell, N. R., Thode, H. C. Jr. and Finch, S. J. (1991). The likelihood ratio test for the two-component normal mixture problem. *Biometrics*, **47**, 1143–1148.

Muggeo, V. M. R. (2008a). Modeling temperature effects on mortality: multiple segmented relationships with common break points. *Biostatistics*, **9**, 613–620.

Muggeo, V. M. R. (2008b). segmented: an R Package to Fit Regression Models with Broken-Line Relationships.

*R News*, **8/1**, 20–25. https://cran.r-project.org/doc/Rnews/

Muggeo, V. M. R., Atkins, D. C., Gallop, R. J. and Dimidjian, S. (2014). Segmented mixed models with random changepoints: a maximum likelihood approach with application to treatment for depression study. *Statistical Modelling*, **14**, 293–313.

Murphy, S. and Van der Vaart, A. (1997). Semiparametric likelihood ratio inference. *The Annals of Statistics*, **25**, 1471–1509.

Murphy, S. and Van der Vaart, A. (2000). On Profile Likelihood. *Journal of the American Statistical Association*, **95**, 449–465.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, **16**, 11–32.

Niu, Y.S., Hao, N. and Zhang, H. (2016). Multiple changepoint detection: A selective overview. *Statistical Science*, **31**, 611–623.

Oliveira-Brochado, A. and Martins, N. (2005). Assessing the number of components in mixture models: a review. *FEP Working Papers*, **194**, Universidade do Porto, Faculdade de Economia do Porto.

Owen, A. B. (1990). Empirical likelihood confidence regions. *The Annals of Statistics*, **18**, 90–120.

Owen, A. B. (1991). Empirical likelihood for linear models. *The Annals of Statistics*, **19**, 1725–1747

Pace, L. and Salvan, A. (1997). *Principles of Statistical Inference: from a Neo-Fisherian Perspective*. World Scientific Publishing, Singapore.

Page, E. S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, **42**, 523–527.

Page, E. S. (1957). On problems in which a change in a parameter occurs at an unknown point. *Biometrika*, **44**, 248–252.

Paulino, C. D. M. and Pereira, C. A. B. (1994). On identifiability of parametric statistical models. *Journal of the Italian Statistical Society*, **1**, 125–151.

Pfanzagl, J. (2017). *Mathematical Statistics: Essays on History and Methodology*. Springer-Verlag, Berlin Heidelberg.

Piegorsch, W. W. and Bailer, A. J. (1997). *Statistics for Environmental Biology and Toxicology*. Chapman & Hall, London.

Prakasa Rao, B. L. S. (1992). *Identifiability in Stochastic Models: Characterization of Probability Distributions*. Academic Press, London.

Quandt, R. E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association*, **53**, 873–880.

Quandt, R. E. (1960). Tests of the hypothesis that a linear regression system obeys two separate regimes. *Journal of the American Statistical Association*, **55**, 324–330.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Reeves, J., Chen, J., Wang, X. L., Lund, R. B. and Lu, Q. (2007). A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, **46**, 900–915.

Ritz, C. and Skovgaard, Ib M. (2005). Likelihood ratio tests in curved exponential families with nuisance parameters present only under the alternative. *Biometrika*, **92**, 507–517.

Robbins, M. W., Gallagher, C. M. and Lund, R. B. (2016). A general regression changepoint test for time series data. *Journal of the American Statistical Association*, **111**, 670–683.

Robbins, M., Gallagher, C., Lund, R. and Aue, A. (2011). Mean shift testing in correlated data. *Journal of Time Series Analysis*, **32**, 498–511.

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order Restricted Inference*. John Wiley & Sons, New York.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica*, **39**, 577–591.

Rotnitzky, A., Cox, D. R., Bottai, M. and Robins, J. (2000). Likelihood-based inference with singular information matrix. *Bernoulli*, **6**, 243–284.

Samworth, R. J. (2018). Recent progress in log-concave density estimation. *Statistical Science* **33**, 493–509.

Samworth, R. J. and Bodhisattva, S. (Eds.) (2018). Special Issue on Nonparametric Inference Under Shape Constraints. *Statistical Science*, **33**.

Scheipl, F., Greven, S. and Küchenhoff, H. (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics & Data Analysis*, **52**, 3283–3299.

Scrucca, L., Fop, M., Murphy, T. B. and Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, **8**, 289–317.

Self, S. G. and Liang, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non standard conditions. *Journal of the American Statistical Association*, **82**, 605–610.

Sen, P. K. and Silvapulle, M. J. (2002). An appraisal of some aspects of statistical inference under inequality constraints. *Journal of Statistical Planning and Inference*, **107**, 3–43.

Severini, T. (2000). *Likelihood Methods in Statistics*. Oxford University Press, Oxford.

Severini, T. (2004). A modified likelihood ratio statistic for some nonregular models. *Biometrika*, **91**, 603–612.

Shaban, S. A. (1980). Change point problems and two-phase regression: an annotated bibliography. *International Statistical Review*, **48**, 83–93.

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

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

Silvapulle, M. J. and Sen, P. K. (2005). *Constrained Statistical Inference*. John Wiley & Sons, New York.

Silvey, S. D. (1959). The Lagragian multiplier test. *The Annals of Mathematical Statistics*, **30**, 382–407.

Sinha, B., Kopylev, L. and Fox, J. (2012). Some new aspects of dose-response multi-stage models with applications. *Pakistan Journal of Statistics and Operation Research*, **8**, 441–478.

Smith, R. L. (1985). Maximum likelihood estimation in a class of non regular cases. *Biometrika*, **72**, 67–92.

Smith, R. L. (1989). A survey of non-regular problems. *Proceedings of the 47th Session of the International Statistical Institute*, Paris, August 29 – September 6, 1989, pp. 353–372.

Smith, A. M. F. and Cook, D. G. (1980). Straight lines with a change point: A Bayesian analysis of some renal transplant data. *Applied Statistics*, **29**, 180–189.

Sørensen, H. (2008). Small sample distribution of the likelihood ratio effects in the random effects model. *Journal of Statistical Planning and Inference*, **138**, 1605–1614.

Sprent, P. (1961). Some hypotheses concerning two-phase regression lines. *Biometrics*, **17**, 634–645.

Stram, D. O. and Lee, J. W. (1994). Variance components testing in the longitudinal mixed rffects model. *Biometrics*, **50**, 1171–1177.

Sun, H.-J. (1988). A FORTRAN subroutine for computing normal orthant probabilities of dimensions up to nine. *Communication in Statistics – Computation and Simulation*, **17**, 1097–1111.

Susko, E. (2013). Likelihood ratio tests with boundary constraints using data-dependent degrees of freedom. *Biometrika*, **100**, 1019–1023.

Thode, H. C. Jr., Finch, S. J. and Mendell, N. R. (1988). Simulated percentage points for the null distribution of the likelihood ratio test for a mixture of two normals. *Biometrics*, **44**, 1195–1201.

Ulm, K. W. (1991). A statistical method for assessing a threshold in epidemiological studies. *Statistics in Medicine*, **10**, 341–349.

van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical processes*. Springer, New York.

van der Vaart, A. W. (2000). *Asymtptotic Statistics*. Cambridge University Press, New York.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, **57**, 307–333.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, **20**, 595–601.

Wilks, S. S. (1938). The large sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, **1**, 60–62.

Wolak, F. A. (1987). An exact test for multiple inequality and equality constraints in the linear regression model. *Journal of the American Statistical Association*, **82**, 782–793.

Worsley, K. J. (1979). On the likelihood ratio test for a shift in location of normal populations. *Journal of the American Statistical Association*, **74**, 365–367.

Worsley, K. J. (1983). The power of likelihood ratio and cumulative sum tests for a change in a binomial probability. *Biometrika*, **70**, 455–464.

Worsley, K. J. (1986). Confidence regions and test for a change point in a sequence of exponential family random variables. *Biometrika*, **73**, 91–104.

Worsley, K. J. (1988). Exact percentage points of the likelihood-ratio test for a change-point hazard-rate model. *Biometrics*, **44**, 259–263.

Yao, Y.-C. and Davis, R.A. (1986). The asymptotic behavior of the likelihood ratio statistic for testing a shift in mean in a sequence of independent normal variates. *Sankhyā A*, **48**, 339–353.

Yu, Y. (2018). mixR: Finite Mixture Modeling for Raw and Binned Data. R package version 0.1.1. https://CRAN.R-project.org/package=mixR

Zeileis, A. (2006). Implementing a class of structural change tests: an econometric computing approach. *Computational Statistics & Data Analysis*, **50**, 2987–3008

Zeileis, A., Leisch, F., Hornik, K. and Kleiber, C. (2002). strucchange: An R package for testing for structural change in linear regression models. *Journal of Statistical Software*, **7**, 1–38.

Zhang, Y., Staicu, A.-M., and Maity, A. (2016). Testing for additivity in non-parametric regression. *Canadian Journal of Statistics*, **44**, 445–462.

Zhang, Y. (2018). lmeVarComp: Testing for a Subset of Variance Components in Linear Mixed Models. R package version 1.1. https://CRAN.R-project.org/package=lmeVarComp