Available online at www.sciencedirect.com
SCIENCE DIRECT®

Journal of Multivariate Analysis 97 (2006) 1409 – 1436

Journal of Multivariate Analysis

www.elsevier.com/locate/jmva

An estimation method for the Neyman chi-square divergence with application to test of hypotheses

M. Broniatowski[a,*,1], S. Leorato[b]

[a] LSTA, Université Paris 6, 175 Rue du Chevaleret, 75013 Paris, France
[b] Dip. di Statistica, Probabilitá e Statistiche Applicate, University of Rome "La Sapienza", P.le A. Moro, 5 00185 Roma, Italy

Received 13 August 2004
Available online 29 March 2006

Abstract

We propose a new definition of the Neyman chi-square divergence between distributions. Based on convexity properties and duality, this version of the χ^2 is well suited both for the classical applications of the χ^2 for the analysis of contingency tables and for the statistical tests in parametric models, for which it is advocated to be robust against *outliers*.

We present two applications in testing. In the first one, we deal with goodness-of-fit tests for finite and infinite numbers of linear constraints; in the second one, we apply χ^2-methodology to parametric testing against contamination.

AMS 1991 subject classification: 62F03; 62F10; 62F30

Keywords: Chi-square divergence; Hypothesis testing; Linear constraints; Empirical likelihood; Marginal distributions; Contamination models; Fenchel–Legendre transform; Outliers

1. Introduction

The scope of this paper is to introduce new insights on a χ^2-type criterion and to promote it as a tool for testing. In contrast with the classical approach (see e.g. [12]), the definition of the χ^2 criterion to be adopted here is the so-called Neyman χ^2 in the terminology of [9]. It will be argued that this χ^2 is adequate for various types of models, discrete or continuous, parametric or not. Based on convexity arguments a new form for the χ^2 distance between two distributions on

* Corresponding author.
E-mail addresses: mbr@ccr.jussieu.fr (M. Broniatowski), samantha.leorato@uniroma1.it (S. Leorato).

1 This work was partly supported by fund Ateneo di Padova, 2003–2004, Italy.

0047-259X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2006.02.001

$\mathbb{R}^d$ is derived. It allows for a direct plug-in technique which defines good test statistics avoiding the bias inherent either to grouping or to smoothing. Also, the optimization of the Neyman χ^2 distance between a given distribution and any set of distributions Ω reduces to the optimization of a linear functional on Ω; this property is not shared by other criterions such as the Kullback–Leibler, Hellinger, likelihood, the standard χ^2 or others, and plays in favor of this χ^2 for numerical simplicity. Some statistical difficult problems can be solved using this χ^2. We will consider a classical example, namely a test for contamination, which can be stated as follows. Given an observed data set, can we claim for homogeneity, or do we have to consider that some proportion of the data is generated by some exogeneous contamination mechanism? This is a two components mixture problem with two nonhomogeneous components, and the focus is on the test for the weight of the contaminating distribution at the border 0 of the domain [0, 1]. More classical problems will also be handled, such as goodness of fit, comparing the empirical distribution with the set of distributions satisfying a number of linear constraints; in this respect we will provide information for the maximal number of such constraints (with respect to the sample size) that will allow for such test based on this χ^2-statistics. Also, in these examples the advantage of using this version of the χ^2 is clear: the test statistics always exists, in contrast with empirical likelihood approaches for the same problems, and its calculation is simple; further its limit distribution is classical.

For categorized distributions $p := (p_1, \ldots, p_k)$ and $q := (q_1, \ldots, q_k)$ sharing the same support the Neyman χ^2-distance writes $\chi^2(q, p) := \sum_{i=1}^k \frac{(q_i - p_i)^2}{p_i}$, the averaged square relative error committed when substituting p by q. Note that the classical χ^2 is just $\chi^2(p, q)$ in our definition. From now on we will simply call χ^2 what is usually intended as the Neyman χ^2. In order to handle general models an adequate formula is given by

$$\chi^2(Q, P) = \begin{cases} \int \left(\frac{dQ - dP}{dP}\right)^2 dP, & Q \text{ is a.c. w.r.t. } P, \\ \infty & \text{otherwise}, \end{cases} \tag{1}$$

where P is a probability measure, and Q is an element of $\mathcal{M}$, the vector space of all finite signed measures on $\mathbb{R}^d$. This definition may seem strange, but consider the above-mentioned contamination problem: P is the unknown distribution of the data, and Q runs in the whole family of contaminated (or uncontaminated) distributions, hence with density $(1-\lambda) f_\theta(x) + \lambda r_\eta(x)$; here f is the density of the "good" data, r is the contaminating density and λ is the rate of contamination. Both θ and η are unknown. In order to get a properly defined test for $\lambda = 0$ it is necessary to make this null hypothesis an interior point in the set of all contaminated densities; this forces the model to include negative values for λ, which turns to justify definition (1) with Q in $\mathcal{M}$. Note also, as will be advocated further on (see Section 5), that this example is in favor of a Neyman χ^2 approach instead of the likelihood ratio one.

Some robustness argument plays in favor of the Neyman χ^2. A notion of robustness against model contamination is found in Lindsay [14] and in Jimenez and Shao [10] which provide an instrument to compare estimators associated to different divergences. Although their argument deals with finite support models and is related to *estimation* procedures, it can be invoqued for testing. A test statistic should bear some robustness properties with respect to *outliers* (excedances of data in or outside the range of the variable) or *inliers* (missing data or subsampling in the range of the variable). For finitely supported models a family of minimum divergence test statistics writes $\Phi(\Omega, P_n) = \inf_{Q \in \Omega} \sum_{i \in S} Q(i) \Phi\left(\frac{P_n(i)}{Q(i)}\right)$, where S is the common support of all distributions in Ω, and Φ is a strictly convex function from $(0, +\infty)$ onto $\mathbb{R}$ with $\Phi(1) = 0$. In the case of the Neyman χ^2, $\Phi(x) = (x-1)^2/2x$. The robustness properties of a statistical estimate are captured by the *Residual Adjustment Function* A_Φ (*RAF*). In the case of the Neyman χ^2, $A(x) = \frac{x}{1+x}$. The robustness of the test statistics against outliers or inliers is a consequence of the behavior of the RAF at $+\infty$ or at -1. The smallest its variations at $+\infty$ and -1, the most robust the test statistics. *Minimum Hellinger distance* test provides a good compromise since both variations are small (see the discussion in [2,10]). When the model might be subject to outlier contaminations only, as will be advocated in the present paper for contamination models, then minimum Neyman χ^2-divergence test behaves better than all classical minimum divergence estimates (Hellinger, Kullback–Leibler, Likelihood, Pearson χ^2, etc.), as easily seen through the calculation of RAF functions following [10].

This paper is organized as follows. Section 2 presents what is needed for the definition of the dual form of the Neyman χ^2. We adapt general results obtained in [8] for differentiable divergences, and state explicit results for sake of completeness. Section 3 presents the estimates together with general convergence results. Section 4 presents specific examples in the range of tests of fit, and also includes some results on Neyman χ^2 goodness-of-fit tests based on sieves. In Section 5, we handle the aforementioned test for contamination. Most proofs are postponed to Section 6.

2. A new definition of the χ^2 statistics

2.1. The dual form of the χ^2

Denote $\mathcal{M}_1$ the set of all probability measures on $\left(\mathbb{R}^d, \mathcal{B}\right)$. The χ^2 distance defined on $\mathcal{M}$ for fixed P in $\mathcal{M}_1$ through (1) is a convex function. As such it is the upper envelope of its support hyperplanes, a statement which we now develop, adapting [8]. Let $\mathcal{F}$ be some class of $\mathcal{B}$-measurable real valued functions f (bounded or unbounded) defined on $\mathbb{R}^d$. Denote $\mathcal{B}_b$ the set of all bounded $\mathcal{B}$-measurable real-valued functions defined on $\mathbb{R}^d$, and $\langle \mathcal{F} \cup \mathcal{B}_b \rangle$ the linear span of $\mathcal{F} \cup \mathcal{B}_b$. Consider the real vector space

$$\mathcal{M}_\mathcal{F} := \left\{ Q \in \mathcal{M} \ \text{ such that } \int |f|\, d|Q| < \infty, \ \text{for all } f \text{ in } \mathcal{F} \right\},$$

in which $|Q|$ denotes the total variation of the signed finite measure Q. When $\mathcal{F} = \mathcal{B}_b$, then $\mathcal{M}_\mathcal{F} = \mathcal{M}$. On $\mathcal{M}_\mathcal{F}$ the $\tau_\mathcal{F}$ topology is the coarsest which makes all mappings $Q \in \mathcal{M}_\mathcal{F} \mapsto \int f\, dQ$ continuous when f belongs to $\mathcal{F} \cup \mathcal{B}_b$.

The hyperplanes in $\mathcal{M}_\mathcal{F}$ are described through

Proposition 1. *Equipped with the $\tau_\mathcal{F}$-topology, $\mathcal{M}_\mathcal{F}$ is a Hausdorff locally convex topological vector spaces. Its topological dual space is the set of all mappings $Q \mapsto \int f\, dQ$ when f belongs to $\langle \mathcal{F} \cup \mathcal{B}_b \rangle$.*

The function $Q \in [\mathcal{M}_\mathcal{F}; \tau_\mathcal{F}] \mapsto \chi^2(Q, P)$ is lower semi-continuous (l.s.c.); see [8], Proposition 2.2. This allows to use the Legendre Fenchel theory in order to represent the mapping $Q \longmapsto \chi^2(Q, P)$ as the upper envelope of its support hyperplanes.

On $\langle \mathcal{F} \cup \mathcal{B}_b \rangle$ define the Fenchel–Legendre transform of $\chi^2(\cdot, P)$

$$T(f, P) := \sup_{Q \in \mathcal{M}_\mathcal{F}} \int f\, dQ - \chi^2(Q, P) = \int f\, dP + \frac{1}{4} \int f^2\, dP \tag{2}$$

for all $f \in \langle \mathcal{F} \cup \mathcal{B}_b \rangle$, see e.g. [1, Chapter 4].

Lemma 1. *It holds*

$$\chi^2(Q, P) = \sup_{f \in \langle \mathcal{F} \cup \mathcal{B}_b \rangle} \int f \, dQ - T(f, P). \tag{3}$$

The function $f^* = 2\left(\frac{dQ}{dP} - 1\right)$ is the supremum in (3) as a consequence of Theorem 3.4 in [8]. When f^* belongs to $\mathcal{F}$ then

$$\chi^2(Q, P) = \sup_{f \in \mathcal{F}} \int m_f(x) \, dP(x), \tag{4}$$

where

$$m_f(x) := \int f \, dQ - \left(f(x) + \frac{1}{4} f^2(x)\right).$$

Call (4) the *dual representation* of the χ^2.

2.2. χ^2 projections

Let Ω be a set of measures in $\mathcal{M}$. A statistic for the test H0 : $P \in \Omega$ against H1 : $P \notin \Omega$ is an estimate of $\inf_{Q \in \Omega} \chi^2(Q, P)$. First say that a measure Q^* in Ω is a χ^2-projection of P on Ω if $\chi^2(Q^*, P) < \infty$ and for all measure Q in Ω, $\chi^2(Q^*, P) \leqslant \chi^2(Q, P)$.

Using the dual form of the χ^2 it holds

$$\chi^2(\Omega, P) = \inf_{Q \in \Omega} \chi^2(Q, P) = \inf_{Q \in \Omega} \sup_{f \in \mathcal{F}} \int m_f(x) \, dP(x). \tag{5}$$

Specifying Theorem 3.3 in [8] to the modified χ^2 divergence the existence and characterization of the χ^2 projection of P on a general set Ω in $\mathcal{M}_{\mathcal{F}}$ are captured in the following statement.

Proposition 2. *If there exists some Q^* in Ω such that $\chi^2(Q^*, P) < \infty$ and for all Q in Ω*

$$q^* \in L_1(Q) \quad \textit{and} \quad \int q^* \, dQ^* \leqslant \int q^* dQ,$$

where $q^* = \frac{dQ^*}{dP}$, then Q^* is the χ^2-projection of P on Ω.

Conversely, if Ω is convex and P has projection Q^* on Ω, then, for all Q in Ω, q^* belongs to $L_1(P)$ and *$\int q^* \, dQ^* \leqslant \int q^* \, dQ$.*

Theorem 2.6 in [8] adapted to the Neyman χ^2 divergence assets that any p.m. P which satisfies $\int |f| \, dP < \infty$ for all f in $\mathcal{F}$ has a projection on any closed set in $\mathcal{M}_{\mathcal{F}}$ equipped with the $\tau_{\mathcal{F}}$ topology.

The link with the dual representation stated above will now be established. Indeed, the set of measures Ω often bears enough information to specify $\mathcal{F}$. We will consider two important cases. In both cases, $\mathcal{F}$ will be defined as the smallest class of functions containing $f^* := 2\left(\frac{dQ^*}{dP} - 1\right)$ when both Q^* and P are unknown.

(a) Consider the case when both P and Q belong to a same parametric model indexed by some class Θ, so that $dP = p_\alpha \, d\mu$ and $dQ = p_\theta \, d\mu$ for some dominating measure μ; then f^* belongs to the class of functions $\mathcal{F} := \left\{2\left(\frac{p_\theta}{p_\alpha} - 1\right), \alpha, \theta \in \Theta\right\}$ a class of functions which is well defined. Also, for $dQ = p_\theta \, d\mu$, and when $\chi^2(Q, P)$ is finite, then Q belongs to $\mathcal{M}_\mathcal{F}$ (see Lemma 3.2 in [8]). The contamination problem is precisely of this kind: P is the unknown distribution of the data (contaminated or not) and Ω is the set of uncontaminated distributions. Hence, f^* belongs to $\mathcal{F} := \left\{2\left(\frac{f_\theta}{(1-\lambda)f_\alpha + \lambda r_\eta} - 1\right), \alpha, \theta \in \Theta,\ \eta \in H,\ \lambda \in \Lambda\right\}$.

(b) Consider the problems of existence and characterization of χ^2-projections of some p.m. P on linear set Ω of measures in $\mathcal{M}$ defined by an arbitrary family of linear constraints. Let $\mathcal{G}$ denote a collection (finite or infinite, countable or not) of real-valued functions defined on $\left(\mathbb{R}^d, \mathcal{B}\right)$. The class $\mathcal{G}$ is assumed to contain the function $1_{\mathbb{R}^d}$. The set Ω is defined by

$$\Omega := \left\{Q \in \mathcal{M} \text{ such that } \int_{\mathbb{R}^d} dQ = 1 \text{ and } \int_{\mathbb{R}^d} g\, dQ = 0, \text{ for all } g \text{ in } \mathcal{G} \backslash \left\{1_{\mathbb{R}^d}\right\}\right\}. \tag{6}$$

The following result states the explicit form of Q^*, the χ^2-projection of P on Ω, when it exists. Adapting [8, Theorem 3.4] for the modified χ^2 divergence we have

Theorem 1. *If there exists a measure Q^* in Ω such that $q^* := \frac{dQ^*}{dP}$ belongs to $\langle\mathcal{G}\rangle$ then Q^* is the χ^2-projection of P on Ω. Reciprocally, let Q^* be the χ^2-projection of P on Ω; then q^* belongs to $\overline{\langle\mathcal{G}\rangle}$, the closure of $\langle\mathcal{G}\rangle$ in $L_1(\mathbb{R}^d, |Q^*|)$.*

If $\mathcal{G}$ is a finite class of functions, then the vector space $\langle\mathcal{G}\rangle$ is closed in $L_1(\mathbb{R}^d, |Q^*|)$. Therefore, it holds

Corollary 1. *Let $\mathcal{G} := \left\{1_{\mathbb{R}^d}, g_1, \ldots, g_l\right\}$ be a collection of measurable functions on $\mathbb{R}^d$. Let Q^* in Ω such that $\chi^2(Q^*, P) < \infty$. Then Q^* is the χ^2 projection of P on Ω iff there exists some vector $c := (c_0, \ldots, c_l) \in \mathbb{R}^{l+1}$ such that*

$$\frac{dQ^*}{dP}(x) = c_0 + \sum_{i=1}^{l} c_i g_i(x) \quad (|Q^*|\text{-}a.e.). \tag{7}$$

The definition of the class $\mathcal{F}$ is deduced from the above corollary, setting

$$\mathcal{F} := \langle\mathcal{G}\rangle = \left\{c_0 + \sum_{i=1}^{l} c_i g_i(x), \text{ with } c_0, \ldots, c_l \text{ in } \mathbb{R}\right\} \tag{8}$$

which turns out to be a parametric class of functions.

Corollary 1 provides the form of the projection Q^* of P on $\mathcal{M}_\mathcal{F}$ but does not give a description of Q^+, the projection of P on $\mathcal{M}_1$, if it exists. The following example shows that this is an argument in favor of the definition of the χ^2 distance on subsets of $\mathcal{M}$.

Example 1. Let P be the uniform distribution on [0, 1] and $\mathcal{F} := \left\{1_{[0,1]}, I_d\right\}$, where I_d is the identity function. Consider

$$\Omega := \left\{Q \in \mathcal{M}_\mathcal{F} \text{ such that } \int dQ = 1 \text{ and } \int (x - 1/4)\, dQ(x) = 0\right\}.$$

The signed finite measure Q^* defined by $dQ^*(x) = \left(\frac{5}{2} - 3x\right) dP(x)$ belongs to Ω; it has same support as P and by Corollary 1 it is the χ^2-projection of P on Ω. The projection Q^+ of P on $\Omega \cap \mathcal{M}_1$ exists and is unique, since Ω is closed in the relative topology on $\mathcal{M}_1$. Therefore, we also have $dQ^+(x) = \left(\frac{5}{2} - 3x\right) dP(x)$ on its support which proves that Q^+ cannot have the same support as P, since $dQ^+/dP(x)$ takes negative values for x between $\frac{5}{6}$ and 1. Therefore, (7) does not provide a definite description of the projection in this case. This example proves that a properly defined χ^2 statistics defined in order to assert whether P lays in Ω should be defined by (1) with Ω in $\mathcal{M}$ and not in $\mathcal{M}_1$.

2.3. A minimax result

We close this section by noting that in many cases the inf and sup operations in (5) commute, namely

$$\chi^2(\Omega, P) = \inf_{Q \in \Omega} \sup_{f \in \mathcal{F}} \int m_f(x)\, dP(x) = \sup_{f \in \mathcal{F}} \inf_{Q \in \Omega} \int m_f(x)\, dP(x). \tag{9}$$

For example (9) holds when Ω is defined through (6) and $\mathcal{F}$ through (8). We refer to Theorem 20 in [7] for a general version of this result. Also, the contamination model developed in Section 5 provides an example of the usefulness of (9) as well as a way to prove it in parametric models.

3. The definition of the estimator

3.1. The estimator χ_n^2

Formula (1) is not suitable for statistical purposes as such. Indeed, suppose that we are interested in testing whether P is in some class Ω of distributions with absolutely continuous component. Let $X = (X_1, \ldots, X_n)$ be an i.i.d. sample with unknown distribution P. Let $P_n := \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ be the empirical measure pertaining to X. Then, for all $Q \in \Omega$, the χ^2 distance defined in (1) between Q and P_n is infinite. Therefore, no plug-in technique can lead to a definite statistic in this usual case; smoothing can be used in the spirit of [2,16], but our method uses the properties of χ^2 projections in full force.

Formula (4) is suitable for a plug-in estimate of $\chi^2(Q, P)$ through

$$\chi_n^2(Q, P) := \sup_{f \in \mathcal{F}} \int m_f(x)\, dP_n(x).$$

This estimate is defined also when Q is continuous. When both Q and P share the same finite support then it clearly coincides with the classical estimate of the Neyman χ^2 for suitably chosen classes $\mathcal{F}$.

For example with $\Omega := \left\{Q \in \mathcal{M} \text{ such that } \int 1_{A_i}(x)\, dQ(x) = \int 1_{A_i}(x)\, dQ_0(x)\right\}$ for some p.m. Q_0 and some partition $(A_i)_i$ of the common support of all measures Q and P, it holds, with $\mathcal{F} := \left\{x \to 1_{A_i}(x), i\right\}$ $\inf_{Q \in \Omega} \sup_{f \in \mathcal{F}} \int m_f(x)\, dP_n(x) = \chi_n^2(Q, P) := \sum_{i=1}^k \frac{(Q_0(A_i) - P_n(A_i))^2}{P_n(A_i)}$, where $P_n(A_i)$ is the frequency of the cell i in the sample X.

Optimizing on Ω yields a natural estimate of $\chi^2(\Omega, P)$ through

$$\chi_n^2(\Omega, P) := \inf_{Q \in \Omega} \sup_{f \in \mathcal{F}} \int m_f(x)\, dP_n(x). \tag{10}$$

These estimates may seem cumbersome. However, in the case when we are able to reduce the class $\mathcal{F}$ to a reasonable degree of complexity, they perform quite well and can be used for testing $P \in \Omega$ against $P \notin \Omega$. As seen in the preceding section this is indeed the case when Ω is defined through a finite number of linear constraints, as will be exemplified in the next section.

When (9) holds we may define an other test statistic through

$$\overline{\chi}_n^2(\Omega, P) = \sup_{f \in \mathcal{F}} \inf_{Q \in \Omega} \int f\, dQ - T(f, P_n). \tag{11}$$

In many cases $\chi_n^2(\Omega, P)$ and $\overline{\chi}_n^2(\Omega, P)$ share the same asymptotic behavior. Indeed they even coincide when Ω is defined through linear constraints. In parametric cases, their asymptotic distribution under any alternative is the same, as proved in Section 5 for the contamination problem. The limiting properties of $\overline{\chi}_n^2$ under any alternative can be obtained following closely the proof of Theorem 3.6 in [6]. Denote $\mathcal{H} := \left\{ \left(f + \frac{1}{4} f^2\right),\ f \in \mathcal{F} \right\}$. Note that compactness in sup norm of the class $\mathcal{H}$ in the proof of this latest theorem is only required in order to ensure that the sup in (11) is actually reached at some point in $\mathcal{H}$. Since this fact may occur without this assumption, typically when Ω is defined through linear constraints, we state the limit distribution of the test statistic under H1 as follows.

Theorem 2. *When $\mathcal{H}$ is a Glivenko–Cantelli class of functions, then $\overline{\chi}_n^2(\Omega, P)$ converges to $\chi^2(\Omega, P)$ almost surely as n tends to infinity, whenever* H0 *or* H1 *holds. Assume now that* H1 *holds and P has a unique projection Q^* in Ω, and that $\mathcal{H}$ is a functional Donsker class. Assume that for each n,* arg $\sup_{f \in \mathcal{F}} \inf_{Q \in \Omega} \int f\, dQ - T(f, P_n)$ *exists. Denote $f^* := 2\left(\frac{dQ^*}{dP} - 1\right)$ and $g^* = f^* + \frac{1}{4} f^{*2}$. Then*

$$\sqrt{n}\left(\overline{\chi}_n^2(\Omega, P) - \chi^2(\Omega, P)\right)$$

is asymptotically a centered gaussian random variable with variance $E_P\left(\left(f^* + \frac{1}{4} f^{*2}\right)^2 (X)\right) - E_P\left(\left(f^* + \frac{1}{4} f^{*2}\right)(X)\right)^2$, where X has law P.

Since $\overline{\chi}_n^2(Q, P) = \chi_n^2(Q, P)$ the above theorem also provides the asymptotic distribution of $\sqrt{n}\left(\chi_n^2(Q, P) - \chi^2(Q, P)\right)$.

The asymptotic distribution of χ_n^2 or $\overline{\chi}_n^2$ under H0, i.e. when P belongs to Ω, cannot be obtained in a general frame and must be derived in accordance with the context; this will be the focus of the next sections.

Let Λ_n be the set of all measures in $\mathcal{M}$ with support $(X_1, \ldots, X_n)$. When Ω is as in (6) with $\mathcal{G}$ a finite class of functions, then $\Omega \cap \Lambda_n$ is nonvoid. An estimate of $\chi^2(\Omega, P)$ can be defined through $\chi_n^2(\Omega, P) = \inf_{Q \in \Omega \cap \Lambda_n} \chi^2(Q, P_n) = \chi^2(\Omega, P_n)$. This device is the "Generalized Empirical Likelihood" (GEL) paradigm (see [18] and references therein). Section 4 develops duality approaches in the GEL setting. The estimation of Q^* results as the unconstrained solution of a linear system of equations, which argues in favor of the χ^2 approach for these problems. An approach by sieves is adopted when Ω is defined through an infinite number of linear constraints. Similarly as in Theorem 2, we provide some information on the distribution of the test statistics under H1 (when P does not belong to Ω), which is not addressed in the current literature on Empirical Likelihood methods, and which proves the consistency of the proposed test procedure.

We apply the above results to the case of a test of fit, where $\Omega = \{P_0\}$ is a fixed probability measure.

When $\Omega \cap \Lambda_n$ is void some smoothing technique has been proposed, following [2], substituting P_n by some regularized version; see [17]. The duality approach avoids smoothing. An example is presented in Section 5, devoted to the test for contamination.

4. Test of a set of linear constraints

Let $\mathcal{G}$ be a countable family of real-valued functions g_i defined on $\mathbb{R}^d$, $\{a_i\}_{i=1}^{\infty}$ a real sequence and

$$\Omega := \left\{ Q \in \mathcal{M} \text{ such that } \int g_i \, dQ(x) = a_i, \ i \geqslant 1 \right\}. \tag{12}$$

We assume that Ω is not void. In accordance with (6) the function $g_0 := 1$ belongs to $\mathcal{G}$ with $a_0 = 1$. Let $X_1, \ldots, X_n$ be an i.i.d. sample with common distribution P.

We intend to propose a test for H0 : $P \in \Omega$ vs. H1 : $P \notin \Omega$. In the entire section, the test statistics is $\chi^2(\Omega \cap \Lambda_n, P_n)$, following the GEL approach.

We first consider the case when $\mathcal{G}$ is a finite collection of functions, and next extend the results to the infinite case. For notational convenience, we write Pf for $\int f \, dP$ whenever defined.

4.1. Finite number of linear constraints

Consider the set Ω defined in (12) with $card\{\mathcal{G}\} = k$. Introduce the estimate of $\chi^2(\Omega, P)$, namely

$$\chi_n^2 := \inf_{Q \in \Omega \cap \Lambda_n} \chi^2(Q, P_n). \tag{13}$$

Embedding the projection device in $\mathcal{M} \cap \Lambda_n$ instead of $\mathcal{M}_1 \cap \Lambda_n$ yields to a simple solution for the optimum in (13), since no inequality constraints will be used. Also, the topological context is simpler than as mentioned in the previous section since the projection of P_n belongs to $\mathbb{R}^n$. Our approach differs from the GEL through the use of the dual representation (10), which provides consistency of the test procedure.

The set $\Omega \cap \Lambda_n$ is a convex closed subset in $\mathbb{R}^n$. The projection of P_n on $\Omega \cap \Lambda_n$ exists and is unique. This is in strong contrast with the usual empirical likelihood approach. Consider the case when Ω is the set of all probability measures with mean a. The empirical likelihood projection of P_n on Ω exists iff a is in the interior of the convex envelope of the X_i's. Otherwise the projection should put negative masses on some X_i, a contradiction with the likelihood approach; see [19, p. 106]. Such cases play in favor of the χ^2 approach on signed measures in order to perform tests of hypotheses. The next subsections provide limit properties of $\chi_n^2(\Omega, P)$. Let $\mathcal{F} := \langle \mathcal{G} \rangle$.

4.1.1. Notation and basic properties

Let Q_0 be any fixed measure in Ω. By (5)

$$\begin{aligned}\chi^2(\Omega, P) &= \sup_{f \in \mathcal{F}} (Q_0 - P) f - \frac{1}{4} P f^2 \\ &= \sup_{a_0, a_1, \ldots, a_k} \sum_{i=1}^{k} a_i (Q_0 - P) g_i - \frac{1}{4} P \left(\sum_{i=1}^{k} a_i g_i + a_0 \right)^2\end{aligned} \tag{14}$$

since, for Q in Ω and for all g in $\mathcal{G}$, $Qg = Q_0 g$, and

$$\chi_n^2 = \sup_{a_0,a_1,\ldots,a_k} \sum_{i=1}^{k} a_i \left(Q_0 - P_n\right) g_i - \frac{1}{4} P_n \left(\sum_{i=1}^{k} a_i g_i + a_0\right)^2 . \tag{15}$$

The infinite dimensional optimization problem in (13) thus reduces to a $(k+1)$-dimensional one, much easier to handle. The statistic χ_n^2 is the distance between P_n and Q_n^*, its projection on Ω, and Q_n^* has same support as P_n.

Optimizing in (14) and (15), $\chi^2(\Omega, P)$ and χ_n^2 get written through a quadratic form.

Define the vectors $\underline{v}_n$ and $\underline{v}$ by

$$\begin{aligned}
\underline{v}_n{}' &= \underline{v}(\mathcal{G}, P_n)' = \left\{\left(Q_0 - P_n\right) g_1, \ldots, \left(Q_0 - P_n\right) g_k\right\},\\
\underline{v}' &= \underline{v}(\mathcal{G}, P)' = \left\{\left(Q_0 - P\right) g_1, \ldots, \left(Q_0 - P\right) g_k\right\},\\
\underline{\gamma}_n &= \underline{\gamma}_n(\mathcal{G}) = \sqrt{n}\left\{\left(P_n - P\right) g_1, \ldots, \left(P_n - P\right) g_k\right\} = \sqrt{n}\left(\underline{v} - \underline{v}_n\right).
\end{aligned} \tag{16}$$

Let S be the covariance matrix of $\underline{\gamma}_n$. Write S_n for the empirical version of S, obtained substituting P by P_n in all entries of S.

Proposition 3. *Let Ω be as in* (12) *and let card $\{\mathcal{G}\}$ be finite. We then have*

$$\chi_n^2 = \underline{v}_n' S_n^{-1} \underline{v}_n,$$

$$\chi^2(\Omega, P) = \underline{v}' S^{-1} \underline{v}.$$

The statistics χ_n^2 is a consistent estimate of $\chi^2(\Omega, P)$. Indeed, we have

Proposition 4. *Assume that $\chi^2(\Omega, P)$ is finite. Let $\mathcal{G}$ be a finite class of functions as in* (12) *and assume that* $\max_{g\in\mathcal{G}} Pg^2$ *is finite. Then χ_n^2 tends to $\chi^2(\Omega, P)$ a.s.*

4.1.2. Asymptotic distribution of the test statistic

Write

$$n\chi_n^2 = \sqrt{n}\underline{v}_n' S^{-1} \sqrt{n}\underline{v}_n + \sqrt{n}\underline{v}_n' \left(S_n^{-1} - S^{-1}\right) \sqrt{n}\underline{v}_n.$$

We then have

Theorem 3. *Let Ω be defined by* (12) *and $\mathcal{G}$ be a finite class of linearly independent functions such that* $\max_{g\in\mathcal{G}} Pg^2$ *is finite. Set $k = card\{\mathcal{G}\}$. Then, under* H0,

$$n\chi_n^2 \xrightarrow{d} chi\,(k)\,,$$

where chi (k) denotes a chi-square distribution with k degrees of freedom.

Proof. For P in Ω, $\sqrt{n}\underline{v}_n = \underline{\gamma}_n$. Therefore $n\chi_n^2 = \underline{\gamma}_n{}' S^{-1} \underline{\gamma}_n + \underline{\gamma}_n' \left(S_n^{-1} - S^{-1}\right) \underline{\gamma}_n$.

By continuity of the mapping $h\left(\underline{y}\right) = \underline{y}' S^{-1} \underline{y}$, $\underline{\gamma}_n' S^{-1} \underline{\gamma}_n$ has a limiting $chi\,(k)$ distribution.

It remains to prove that the second term is negligible. Indeed, again from

$$\left(\underline{\gamma}_n\right)'\left(S_n^{-1}-S^{-1}\right)\left(\underline{\gamma}_n\right)\leqslant const\cdot k\left\|\left|S^{1/2}S_n^{-1}S^{1/2}-I\right|\right\|$$

it is enough to show that $\left\|\left|S^{1/2}S_n^{-1}S^{1/2}-I\right|\right\|$ is $o_P(1)$; see (38) in the proof of Proposition 4. □

Turn now to the distribution of the statistics under H1. Suppose that all functions g in $\mathcal{G}$ are bounded. Since the optimizing vector $a^{(n)} := (a_0^{(n)}, \ldots, a_k^{(n)})$ in (15) converges to the corresponding maximizing vector in (14) we may assume that for n large we can identify a compact set in $\mathbb{R}^{k+1}$ that contains all the $a^{(n)}$'s. The projection of P_n on Ω exists and when all functions g in $\mathcal{G}$ satisfy $Pg^2 < \infty$, the conditions in Theorem 2 are fulfilled, providing the asymptotic power of the test, as well as a way to infer the minimal sample size required in order to reach a given power under some alternative. The procedure is asymptotically consistent as observed by the difference of the normalizing factors in the asymptotic distributions of χ_n^2 under H0 and H1.

4.2. *Infinite number of linear constraints, an approach by sieves*

In various cases Ω is defined through a countable collection of linear constraints. An example is presented in Section 4.3. Suppose that Ω is defined as in (12), with $\mathcal{G}$ a countable class of functions

$$\mathcal{G}=\left\{g_\alpha:\mathbb{R}^d\to\mathbb{R},\ \ \alpha\in A\right\},$$

where *card* $(A)=\infty$. Thus $\Omega=\{Q\in\mathcal{M}$ such that $Qg=Q_0g,\ g\in\mathcal{G}\}$, for some Q_0 in $\mathcal{M}$.

Assume that the projection Q^* of P on Ω exists. Then, by Theorem 1 $f^*\in\overline{\langle\mathcal{G}\rangle}$.

We approximate $\mathcal{G}$ through a suitable increasing sequence of classes of functions $\mathcal{G}_n$ with finite cardinality $k=k(n)$ increasing with n. Each $\mathcal{G}_n$ induces a subset Ω_n in Ω as follows.

Let

$$\mathcal{G}_n\subseteq\mathcal{G}_{n+1}\subset\mathcal{G}\quad\text{for all }\ n\geqslant 1 \tag{17}$$

and

$$\mathcal{G}=\bigcup_{n\geqslant 1}\mathcal{G}_n$$

and let

$$\Omega_n=\{Q:Qg=Q_0g,\ \ g\in\mathcal{G}_n\}.$$

We thus have $\Omega_n\supseteq\Omega_{n+1},\ n\geqslant 1$ and $\Omega=\bigcap_{n\geqslant 1}\Omega_n$.

The idea of determining the projection of a measure P on a set Ω through an approximating sequence of sets—or sieve—has been introduced in this setting in [22].

Theorem 4 (*see Teboulle and Vajda [22]*). *With the above notation, define Q_n^* the projection of P on Ω_n. Then*

$$\lim_{n\to\infty}\left\|f^*-f_n^*\right\|_{L_1(P)}=\lim_{n\to\infty}\left\|\frac{dQ^*}{dP}-\frac{dQ_n^*}{dP}\right\|_{L_1(P)}=0. \tag{18}$$

When $\sup_{g\in\mathcal{G}} \sup_x g(x) < \infty$ then (18) implies

$$\lim_{n\to\infty} \chi^2(\Omega_n, P) = \chi^2(\Omega, P). \tag{19}$$

The above result suggests that we can build a sequence of estimators of $\chi^2(\Omega, P)$ letting $k = k(n) = card\, \mathcal{G}_n$ grow to infinity together with n. Define

$$\chi^2_{n,k} = \sup_{f\in\langle\mathcal{G}_n\rangle} (Q_0 - P_n) f - \frac{1}{4} P_n f^2.$$

In the following subsection, we consider conditions on $k(n)$ entailing the asymptotic normality of the suitably normalized sequence of estimates $\chi_{n,k}$ when P belongs to Ω, i.e. under H0.

4.2.1. Convergence in distribution under H0

As a consequence of Theorem 3, $n\chi^2_{n,k}$ tends to infinity with probability 1 as $n \to \infty$.

We consider the statistic

$$\frac{n\chi^2_{n,k} - k}{\sqrt{2k}} \tag{20}$$

which will be seen to have a nondegenerate distribution as $k(n)$ tends to infinity together with n.

We assume from now on that $\mathcal{G}$ is a Donsker class of functions; see e.g. [24]. We need some improved version on the rate of gaussian approximations for empirical processes indexed by $\mathcal{G}$. This leads to the following notion.

For some $a > 0$, let δ_n be a decreasing sequence with $\delta_n = o\left(n^{-a}\right)$. We assume that $\mathcal{G}$ is *Komlós–Major–Tusnády* (*KMT*) with respect to P, with rate δ_n ($\mathcal{G} \in KMT\,(\delta_n; P)$), which is to say that there exists a version $B^0_n(.)$ of P-Brownian bridges such that for any $t > 0$ it holds

$$\Pr\left\{\sup_{g\in\mathcal{G}} \left|\sqrt{n}(P_n - P)g - B^0_n(g)\right| \geqslant \delta_n (t + b\log n)\right\} \leqslant ce^{-\theta t}, \tag{21}$$

where the positive constants b, c and θ depend on $\mathcal{G}$ only. We refer to [4,15,5,11] for examples of classical KMT classes, together with calculations of rates.

From (21) it follows that

$$\sup_{g\in\mathcal{G}} \left|\gamma_n(g) - B^0_n(g)\right| = O_P(\delta_n \log n) \tag{22}$$

where, with the same notation as in the finite case (see (16)), i.e. $\gamma_n(g) = \sqrt{n}(P_n - P)g$ is the empirical process indexed by $g \in \mathcal{G}$.

For any n, set $\underline{\gamma}_{n,k} = \underline{\gamma}_n(\mathcal{G}_n)$ (resp., $\underline{B}^0_{n,k}$) the k-dimensional vector resulting from the projection of the empirical process γ_n (resp., of the P-Brownian bridge B^0_n) defined on $\mathcal{G}$ to the subset $\mathcal{G}_n$. Then, if $\mathcal{G}_n = \left\{g^{(n)}_1, \ldots, g^{(n)}_k\right\}$, $\underline{\gamma}_{n,k} = \left\{\gamma_n(g^{(n)}_1), \ldots, \gamma_n(g^{(n)}_k)\right\}$ and $\underline{B}^0_{n,k} = \left\{B^0_n\left(g^{(n)}_1\right), \ldots, B^0_n\left(g^{(n)}_k\right)\right\}$. Denote S_k the covariance matrix of the vector $\underline{\gamma}_{n,k}$ and $S_{n,k}$ its empirical covariance matrix. Let $\lambda_{1,k}$ be the smallest eigenvalue of S_k. Call an envelope for $\mathcal{G}$ a function G such that $|g| \leqslant G$ for all g in $\mathcal{G}$.

Theorem 5. *Let $\mathcal{G}$ have an envelope G with $PG < \infty$ and be KMT $(\delta_n; P)$ for some sequence $\delta_n \downarrow 0$. Define further a sequence $\{\mathcal{G}_n\}_{n\geqslant 1}$ of classes of linearly independent functions satisfying* (17).

Moreover, let $k = k(n) = card\, \mathcal{G}_n$ *satisfy*

$$\lim_{n\to\infty} k = \infty,$$

$$\lim_{n\to\infty} \lambda_{1,k}^{-1/2} k^{1/2} \delta_n \log n = 0, \tag{23}$$

$$\lim_{n\to\infty} \lambda_{1,k}^{-1} k^{3/2} n^{-1/2} = 0. \tag{24}$$

Then under H0

$$\frac{n\chi^2_{n,k} - k}{\sqrt{2k}} \xrightarrow{d} N(0, 1).$$

In both conditions (23) and (24) the value of $\lambda_{1,k}$ appears, which cannot be estimated without any further hypothesis on the structure of the class $\mathcal{G}$. However, for concrete problems, once defined $\mathcal{G}$ it is possible to give bounds for $\lambda_{1,k}$, depending on k. This will be shown in the next subsection, for a particular class of goodness-of-fit tests.

4.3. Application: testing marginal distributions

Let P be an unknown distribution on $\mathbb{R}^d$ with density bounded by below. We consider goodness-of-fit tests for the marginal distributions $P_1, \ldots, P_d$ of P on the basis of an i.i.d. sample $(X_1, \ldots, X_n)$.

Let, thus, $Q^0_1, \ldots, Q^0_d$ denote d distributions on $\mathbb{R}$. The null hypothesis writes H0: $P_j = Q^0_j$ for $j = 1, \ldots, d$. That is to say that we simultaneously test goodness of fit of the marginal laws $P_1, \ldots, P_d$ to the laws $Q^0_1, \ldots, Q^0_d$. Through the transform $P'(y_1, \ldots, y_d) = P\left(\left(Q^0_1\right)^{-1}(y_1), \ldots, \left(Q^0_d\right)^{-1}(y_d)\right)$ we can restrict the analysis to the case when all p.m's have support $[0, 1]^d$ and marginal laws uniform in [0, 1] under H0. So without loss of generality, we write Q_0 for the uniform distribution on [0, 1].

Bickel et al. [3] focused on the estimation of linear functionals of the probability measure subject to the knowledge of the marginal laws in the case of r.v.'s with a.c. distribution, letting the number of cells grow to infinity.

Under H0 all c.d.f's of the margins of P are uniform on [0, 1]. Introduce the corresponding class of functions, namely the characteristic functions of sets of the form $[0, u)$ on each of the d axes, namely

$$\mathcal{G} := \left\{ 1_{u,j} : [0,1]^d \longrightarrow \{0, 1\},\ j = 1, \ldots, d,\ u \in [0, 1] \right\},$$

where $1_{u,j}(x_1, \ldots, x_d) = \begin{cases} 1, & x_j \leqslant u \\ 0, & x_j > u \end{cases}.$

Let Ω be the set of all p.m's on $[0, 1]^d$ with uniform marginals, i.e.

$$\Omega = \left\{ Q \in M_1\left([0,1]^d\right) \text{ such that } Qg = \int_{[0,1]^d} g(x)\,dx \quad \text{for all } g \text{ in } \mathcal{G} \right\}. \tag{25}$$

The test writes H0 : P belongs to Ω vs. H1 : P does not belong to Ω.

This set Ω has the form (12), where $\mathcal{G}$ is the class of characteristic functions of intervals, a KMT class with rate $\delta_n = n^{-1/2}$ (see [15]).

We now build the family $\mathcal{G}_n$ satisfying (17). Let $m = m(n)$ tend to $+\infty$ with n. Let $0 = u_0 < u_1 < \cdots < u_m < u_{m+1} = 1$ and $\left\{\mathcal{U}^{(n)}\right\}$ be the $m \cdot d$ points in $[0, 1]^d$ with coordinates in $\{u_1, \ldots, u_m\}$.

Let $\mathcal{G}_n$ denote the class of characteristic functions of the d-dimensional rectangles $[\underline{0}, \underline{u}]$ for $\underline{u} \in \mathcal{U}^{(n)}$. Hence $card\{\mathcal{G}_n\} = k = m \cdot d$, and

$$\mathcal{G}_n=\left\{1_{u_i,j}\colon [0,1]^d \to \{0,1\},\ j=1,\ldots,d,\ u_i\in(0,1),\ u_i<u_{i+1},\ i=0,\ldots,m\right\}, \tag{26}$$

which satisfies (17).

In order to establish a lower bound for the smallest eigenvalue of S_k, say $\lambda_{1,k}$, the volumes of the cells in the grid defined by the $u_i^{(n)}$ should not shrink too rapidly to 0. Suppose that the intervals $(u_i, u_{i+1}]$ are such that

$$0 < \lim_{n\to\infty} \inf \min_{i=0,\ldots,m} k\,(u_{i+1} - u_i) \leqslant \lim_{n\to\infty} \sup \max_{i=0,\ldots,m} k\,(u_{i+1} - u_i) < \infty. \tag{27}$$

Remark 1. Condition for the sequence $\mathcal{G}_n$ to converge to $\mathcal{G}$ coincides with (F2) and (F3) in [3].

The order of magnitude of $\lambda_{1,k}$ is obtained through a one-to-one transformation of the class $\mathcal{G}_n$ which leaves invariant the smallest eigenvalue of S_k. We simply turn from the class of characteristic functions of sets $[0, u_i)$ to characteristic functions of sets $[u_i, u_{i+1})$ on each of the axes. We order these dm functions in a natural order, namely the m first ones for the grid of the first axis, and so on.

Let P belong to Ω. Let us then write the matrix S_k. We have $P1_{u_i,j} = Q_0 1_{u_i,j} = u_i$ for $i = 1, \ldots, m$ and $j = 1, \ldots, d$. Set $P1_{u_i,j}1_{u_l,h} = P\left(X_j \leqslant u_i,\ X_h \leqslant u_l\right)$, for every $h, j = 1, \ldots, d$ and $l, i = 1, \ldots, m$. When $j = h$ then $P1_{u_i,j}1_{u_l,j} = P\left(X_j \leqslant u_i \wedge u_l\right) = u_i \wedge u_l$.

The generic term of S_k writes

$$\begin{aligned} s_k(u,v) &= s_k\left((j-1)m+i,(h-1)m+l\right) = P1_{u_i,j}1_{u_l,h} - P1_{u_i,j}P1_{u_l,h} \\ &= \begin{cases} u_i - u_i^2 & \text{if } j=h,\ i=l, \\ P(X_j \leqslant u_i \wedge u_l) - u_i u_l & \text{if } j=h,\ i\neq l, \\ P(X_j \leqslant u_i, X_h \leqslant u_l) - u_i u_l & \text{if } j\neq h. \end{cases} \end{aligned}$$

We make use of the class of functions

$$\begin{aligned} \mathcal{G}_n^\delta &= \left\{1_{u_i,j} - 1_{u_{i-1},j},\ i=1,\ldots,m,\ \ 1_{u_0,j}=0,\ j=1\ldots,d\right\} \\ &= \left\{1_{A_i^j},\ i=1\ldots,m,\ j=1,\ldots,d\right\}. \end{aligned}$$

Let S_k^δ be the covariance matrix of the vector $\underline{\gamma}_n^\delta = \underline{\gamma}_n(\mathcal{G}_n^\delta)$ and define the vectors $\underline{v}^\delta$ and $\underline{v}_n^\delta$ similarly as in (16).

S_k^δ has $((j-1)m+i, (h-1)m+l)$th component equal to $P_{A_i^j A_l^h} - P_{A_i^j} P_{A_l^h}$, which is

$$\begin{cases} p_i - p_i^2 & \text{if } j=h,\ i=l, \\ -p_i p_l & \text{if } j=h,\ i\neq l, \\ P(u_{i-1} \leqslant X_j < u_i, u_{l-1} \leqslant X_h < u_l) - p_i p_l & \text{if } j\neq h, \end{cases}$$

where we have written $p_i = P(u_{i-1} \leqslant X_j < u_i) = u_i - u_{i-1}$, for all $j = 1, \ldots, d$.

Also, χ^2 (and χ_n^2) can be written using $\mathcal{G}_n^\delta$ instead of $\mathcal{G}_n$:

$$\chi^2(\Omega, P) = \underline{v}' S_k^{-1} \underline{v} = \left(\underline{v}^\delta\right)' \left(S_k^\delta\right)^{-1} \left(\underline{v}^\delta\right).$$

Let M be the diagonal d-block matrix with all diagonal blocks equal to the unit inferior triangular $(m \times m)$ matrix. Then $\underline{v} = M\underline{v}^\delta$ and $S_k = MS_k^\delta M'$.

Thus $\underline{v}^{\delta\prime}(S_k^\delta)^{-1}\underline{v}^\delta = \underline{v}'(M')^{-1}M'S_k^{-1}M(M)^{-1}\underline{v} = \chi^2$. Similar arguments yield $\chi_n^2 = \underline{v}_n^{\delta\prime}(S_{n,k}^\delta)^{-1}\underline{v}_n^\delta$.

Let $\lambda_{1,\delta}$ be the minimum eigenvalue of S_k^δ. Since M has all eigenvalues equal to one, the following chain holds:

$$\lambda_{1,k} \leqslant \min_x \frac{x'S_k x}{\|x\|^2} \leqslant \min_y \frac{y'S_k^\delta y}{\|y\|^2} \max_x \frac{\|Mx\|^2}{\|x\|^2} = \lambda_{1,\delta} \leqslant \min_y \frac{y'S_k y}{\|y\|^2} \max_x \frac{\|M^{-1}x\|^2}{\|x\|^2} = \lambda_{1,k}$$

which is to say that the minimum eigenvalues of S_k^δ and S_k are equal. As a consequence we obtain

Lemma 2. *Suppose that $P \in \Omega$ has density on $[0, 1]^d$ bounded from below by $\alpha > 0$. Then the smallest eigenvalue of S_k^δ, and consequently $\lambda_{1,k}$, is bounded below by $\alpha p_{m+1} \min_{1\leqslant i\leqslant m} p_i$.*

We will now consider the covariance matrix of $\underline{\gamma}_n^\delta$ under H0, when the underlying distribution is Q_0^d, the uniform distribution on $[0, 1]^d$. Denote this matrix S_k^0. We then have

Lemma 3. *If $P = Q_0^d$, then*

(i) *$S_k^0 = D^{1/2}(I - V)D^{1/2}$, where D and V are both diagonal block matrices with diagonal blocks equal to diag $\{p_i\}_{i=1,\dots,m}$ and to $U = \left\{\sqrt{p_i p_l}\right\}_{i=1,\dots,m,\, l=1,\dots,m}$, respectively.*

(ii) *The $(m \times m)$ matrix U has eigenvalues equal to*

$$\lambda_U = \begin{cases} (1 - p_{m+1}) = \sum_{i=1}^m p_i & \text{with cardinality } 1, \\ 0 & \text{with cardinality } m-1. \end{cases}$$

Moreover $(I - U)^{-1} = (I + \frac{1}{p_{m+1}}U)$.

(iii) *For any eigenvalue λ of S_k^0 it holds*

$$p_{m+1} \min_{1\leqslant i\leqslant m} p_i \leqslant \lambda \leqslant \max_{1\leqslant i\leqslant m} p_i.$$

From Theorem 5 and using (27) in order to evaluate $p_{m+1} \min_{1\leqslant i\leqslant m} p_i$, together with the fact that the class $\mathcal{G}$ is KMT with rate $\delta_n = n^{-1/2}$ we obtain

Theorem 6. *Let (27) hold. Assume that H0 holds with Ω defined by (25). Assume that P has a density bounded by below by some positive number. Let, further, $k = d \cdot m(n)$ be a sequence such that $\lim_{n\to\infty} k = \infty$ and $\lim_{n\to\infty} k^{7/2}n^{-1/2} = 0$.*

Then $\frac{n\chi_{n,k}^2 - k}{\sqrt{2k}} = \frac{n\underline{\gamma}_{n,k}' S_{n,k}^{-1} \underline{\gamma}_{n,k} - k}{\sqrt{2k}}$ has limiting normal standard distribution.

Remark 2. A lower bound for the density of P also appears in [3] (see their condition (P3)).

In the last part of this section, we show that conditions in Theorem 6 can be weakened for small values of d. We consider the case when $d = 2$; for larger values of d, see Remark 3.

For clearness define $p_{ij} := P(A_i^1 \times A_j^2)$ and $N_{ij} := nP_n(A_i^1 \times A_j^2)$, where the events A_i^h, $h = 1, 2, i = 1, \dots, m$ are as above. The marginal distributions will be denoted $p_{i,\cdot} = p_{\cdot,i} = p_i$ and the empirical marginal distributions by $N_{i,\cdot}/n$ and $N_{\cdot,i}/n$.

Turning back to the proof of Theorem 5 we see that condition (24) is used in order to ensure that $m^{-1/2}\underline{\gamma}'_{n,k}(S_{n,k}^{-1} - S_k^{-1})\underline{\gamma}_{n,k}$ goes to 0 in probability as n tends to infinity, while condition (23) implies the convergence of $\frac{\underline{\gamma}'_{n,k} S_k^{-1} \underline{\gamma}_{n,k} - 2m}{\sqrt{4m}}$ to the standard normal distribution.

Sharp inequalities pertaining to the binomial probabilities yield the following improvement upon Theorem 6.

Theorem 7. *Let* (27) *hold. Assume that* H0 *holds and P is as in Lemma* 2. *Let $m := m(n)$ with* $\lim_{n\to\infty} m = \infty$ *and* $\lim_{n\to\infty} m^{3/2}n^{-1/2}\log n = 0$. *Then*

$$\frac{n\chi^2_{n,2m} - 2m}{\sqrt{4m}} \longrightarrow N(0, 1).$$

Remark 3. The preceding arguments carry over to the case $d > 2$ and yield to the condition $\lim_{n\to\infty} m^{d+1/2}n^{-1/2}\log n = 0$. However for $d \geqslant 6$ this ultimate condition is stronger than (24).

5. A test for contamination

In this section, we address the contamination test problem sketched in Section 1. Such problems have been considered in the recent literature; see [13] and references therein. A contamination model can be written as

$$\mathcal{C}^+ := \left\{ p(x) = (1-\lambda) f_\theta(x) + \lambda r_\eta(x), 0 \leqslant \lambda \leqslant 1, \theta \in \Theta, \eta \in H \right\}, \tag{28}$$

where f_θ is the density of the uncontaminated data, r_η is the contaminating density and λ is the rate of contamination, usually close to 0. All densities are considered with respect to some common dominating measure μ. The parameter θ belongs to a subset Θ of $\mathbb{R}^d$, and the set H is finite dimensional. In contrast with [13] we do not assume that the parameter θ of the distribution of the "true data" is known; further the kind of alternative which we consider seems somehow more natural, since it writes simply $\lambda \neq 0$. The above model is clearly nonidentifiable when $\lambda = 0$ since any value of η characterizes the same distribution. The identifiability condition which is assumed here is

$$\frac{f_\alpha}{(1-\lambda) f_\theta + \lambda r_\eta} = 1 \text{ iff } \lambda = 0 \text{ and } \alpha = \theta. \tag{29}$$

Lack of identifiability may lead to difficulties in estimation and in test, typically for mixture models. Contamination does not enter in this class of problems, since the two components belong to quite different families of distributions. The density r_η models outliers. We assume that

$$\sup_{\eta \in H} \int \frac{f_\theta f_{\theta'}}{r_\eta}\, d\mu \text{ is finite for all } \theta \text{ and } \theta' \text{ in } \Theta \tag{30}$$

which amounts to say that r_η is heavy-tailed with respect to f_θ for all η and θ. Also usual integrability conditions allowing the change of order of differentiation with respect to the parameters up to the third order and integration will be assumed. For a real values function g let $m_{g,\alpha}(x) := \int g f_\alpha \, d\mu - \left(g(x) + \frac{1}{4}g^2(x)\right)$ whenever defined.

The hypotheses to be tested are H0: $\lambda = 0$ vs. H1: $\lambda \neq 0$. Let

$$\Omega := \{f_\alpha, \alpha \in \Theta\},$$

so that H0 holds iff p belongs to Ω.

The test statistic is an estimate of the χ^2 distance from p to Ω. Likelihood ratio methods are not valid in this case due to failure of the standard regularity conditions under which the asymptotic theory is based; see [23], Section 5.4. We also advocate in favor of the χ^2 since it is robust against *outliers*. Indeed, the choice of the class r_η to model the "heavy tail" behavior of the data is usually problematic and we cannot accept a test procedure which would be too sensitive with respect to this choice.

Following our definition of the χ^2 divergence we extend the model (28) defining

$$\mathcal{C} := \left\{p(x) = (1-\lambda) f_\theta(x) + \lambda r_\eta(x), \lambda \in \Lambda, \theta \in \Theta, \eta \in H\right\},$$

where $\Lambda = [a, 1]$ and a is a negative number. This makes $\mathcal{C}$ a subset of $\mathcal{M}$ and $\lambda = 0$ is an interior point in Λ.

Let

$$\mathcal{F}_\alpha := \left\{g = 2\left(\frac{f_\alpha}{(1-\lambda) f_\theta + \lambda r_h} - 1\right), \theta \in \Theta,\ \lambda \in \Lambda, h \in H\right\}. \tag{31}$$

By (30) $\int |g| \, f_\alpha \, d\mu$ is finite for all g in $\mathcal{F}_\alpha$. Following (4) the divergence $\chi^2(\Omega, p)$ writes

$$\chi^2(\Omega, p) = \inf_{\alpha \in \Theta} \chi^2_n(f_\alpha, p) = \inf_{\alpha \in \Theta} \sup_{g \in \mathcal{F}_\alpha} \int m_{g,\alpha}(x) \, dP(x). \tag{32}$$

The estimate is defined analogously through

$$\chi^2_n(\Omega, p) = \inf_{\alpha \in \Theta} \sup_{g \in \mathcal{F}_\alpha} \int m_{g,\alpha}(x) \, dP_n(x).$$

We consider now the asymptotic distribution of $\chi^2_n(\Omega, p)$ under H0.

When $p = f_{\theta_0}$ belongs to Ω then

$$\chi^2(\Omega, p) = \chi^2(f_{\theta_0}, f_{\theta_0}) = 0,$$

is estimated by

$$\chi^2_n(\Omega, f_{\theta_0}) = \inf_{\alpha} \sup_{g \in \mathcal{F}_\alpha} \int m_{g,\alpha} \, dP_n =: \inf_{\boldsymbol{\alpha}} \sup_{\mathbf{t}} \int m(\mathbf{t}, \boldsymbol{\alpha}) \, dP_n,$$

where $\boldsymbol{\alpha} = (\alpha, 0, \eta)$with α in Θ and η runs in H; $\mathbf{t} := (\theta, \lambda, h)$ defines a function g in $\mathcal{F}_\alpha$; see (31). Obviously, all possible values of η define the same function $m(\mathbf{t}, \boldsymbol{\alpha})$. Define $\mathbf{t}_n(\boldsymbol{\alpha}) := \arg\sup_{\mathbf{t}} \int m(\mathbf{t}, \boldsymbol{\alpha}) \, dP_n$ so that $\chi^2_n(\Omega, f_{\theta_0}) = \inf_{\boldsymbol{\alpha}} \int m(\mathbf{t}_n(\boldsymbol{\alpha}), \boldsymbol{\alpha}) \, dP_n$. Denote $\boldsymbol{\alpha}_n$ the optimizer of this expression so that

$$\chi^2_n(\Omega, f_{\theta_0}) = \int m(\mathbf{t}_n(\boldsymbol{\alpha}_n), \boldsymbol{\alpha}_n) \, dP_n.$$

Accordingly, $\mathbf{t}(\boldsymbol{\alpha})$ and $\boldsymbol{\alpha}$ are defined as above with P (with density f_{θ_0}) substituting P_n. The vectors of parameters $\mathbf{t}_n(\boldsymbol{\alpha}_n)$ and $\boldsymbol{\alpha}_n$ satisfy the set of equations

$$\begin{aligned}
&\frac{\partial}{\partial \mathbf{t}} \int m(\mathbf{t}, \boldsymbol{\alpha})_{\mathbf{t}_n(\boldsymbol{\alpha}_n),\boldsymbol{\alpha}_n}\, dP_n = 0,\\
&\frac{\partial}{\partial \boldsymbol{\alpha}} \int m(\mathbf{t}(\boldsymbol{\alpha}), \boldsymbol{\alpha})_{\boldsymbol{\alpha}_n,\boldsymbol{\alpha}_n}\, dP_n = 0,
\end{aligned} \tag{33}$$

where the derivatives are considered with respect to α in the second line and with respect to θ, λ and h in the first. The second line in (33) writes $\frac{\partial}{\partial \boldsymbol{\alpha}} \int m(\mathbf{t}, \boldsymbol{\alpha})_{\boldsymbol{\alpha}_n,\boldsymbol{\alpha}_n}\, dP_n + \frac{\partial}{\partial \boldsymbol{\alpha}} \mathbf{t}(\boldsymbol{\alpha})_{\boldsymbol{\alpha}_n} \frac{\partial}{\partial \mathbf{t}} \int m(\mathbf{t}, \boldsymbol{\alpha})_{\mathbf{t}_n(\boldsymbol{\alpha}_n),\boldsymbol{\alpha}_n}\, dP_n = 0$, which, together with the first line in (33) yields the somewhat simpler system of equations

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\alpha}} \int m(\mathbf{t}, \boldsymbol{\alpha})_{\mathbf{t}_n(\boldsymbol{\alpha}_n),\boldsymbol{\alpha}_n}\, dP_n = 0,\\
&\frac{\partial}{\partial \mathbf{t}} \int m(\mathbf{t}, \boldsymbol{\alpha})_{\mathbf{t}_n(\boldsymbol{\alpha}_n),\boldsymbol{\alpha}_n}\, dP_n = 0.
\end{aligned}$$

We now state the asymptotic distribution of the test statistics under H0. The following classical hypotheses are needed for usual derivation under the integral sign. The proof of the theorem in the spirit of Wilk's Theorem. Denote (H) the set of hypotheses

(1) The supports of f_θ and r_η do not depend on the parameters.
(2) The sets Θ and H are compact.
(3) The function $m(\mathbf{t}, \boldsymbol{\alpha})$ is derivable at order 3.
(4) There exists a neighborhood of $(\mathbf{t}_0, \boldsymbol{\alpha}_0) := ((\theta_0, 0, h), (\theta_0, 0, \eta))$ such that all these derivatives are bounded by $f_{\theta_0}\, d\mu$-integrable functions uniformly upon h.

Note that η plays no role in the above assumption (4).

Theorem 8. *Under the above notation and assumption* (H), *under* H0, *the asymptotic distribution of* $2n\chi_n^2$ *is Chi*(1), *a chi-square distribution with* 1 *degree of freedom.*

We now deduce the limit distribution of the test statistic under H1.

It can be obtained under two different sets of hypotheses pertaining to the class of functions $m_{g,\alpha}$. Assume (D)

p has a unique projection f_{α^*} on Ω

(33) has a solution for each large n.

Denote $g^* := 2\left(\frac{f_{\alpha^*}}{p} - 1\right)$, where $\alpha^* := \arg\min_\alpha \sup_{g\in\mathcal{F}_\alpha} \int m_{g,\alpha} p\, d\mu$. The classical approach, using Taylor expansions and classical CLT in the same vein as in the proof of Theorem 8, yields

Theorem 9. *When* (H) *and* (D) *hold then under* (H1) *the limit distribution of* $\sqrt{n}\left(\chi_n^2(\Omega, p) - \chi^2(\Omega, p)\right)$ *is normal with zero mean and variance* $\int \left(g^* + \frac{1}{4}g^{*2}\right)^2 p\, d\mu - \left(\int \left(g^* + \frac{1}{4}g^{*2}\right) p\, d\mu\right)^2$.

The second approach uses the minimax property (9). First observe that, introducing $\mathcal{F} := \bigcup_{\alpha\in\Theta} \mathcal{F}_\alpha$, $\sup_{g\in\mathcal{F}_\alpha} \int m_{g,\alpha}(x)\, dP(x) = \sup_{g\in\mathcal{F}} \int m_{g,\alpha}(x)\, dP(x)$ since the sup is indeed reached on $\mathcal{F}_\alpha$. This allows to commute the sup and inf operations in (32), replacing $\mathcal{F}_\alpha$ by $\mathcal{F}$, in accordance with our claim in Section 1, (9).

Proposition 5. *Assume that both* Θ *and* H *are compact. Then*

$$\chi^2(\Omega, p) = \inf_{\alpha\in\Theta} \sup_{g\in\mathcal{F}} \int m_{g,\alpha}(x)\, dP(x) = \sup_{g\in\mathcal{F}} \inf_{\alpha\in\Theta} \int m_{g,\alpha}(x)\, dP(x). \tag{34}$$

Proof. A generic function g in $\mathcal{F}$ writes $g = 2\left(\frac{f_{\theta_2}}{(1-\lambda) f_{\theta_1}+\lambda r_h} - 1\right)$ for some θ_1, θ_2 in Θ, η in H and λ in Λ.

On one hand it holds $\inf_{\alpha\in\Theta} \sup_{g\in\mathcal{F}} \int m_{g,\alpha}(x)\, dP(x) \geqslant \sup_{g\in\mathcal{F}} \inf_{\alpha\in\Theta} \int m_{g,\alpha}(x)\, dP(x)$. For the reverse inequality, it holds, for all α in Θ,

$$\begin{aligned}\int m_{g,\alpha} p\, d\mu &= \left\{\int 2\frac{f_{\theta_2}}{(1-\lambda) f_{\theta_1} + \lambda r_\eta} \frac{f_\alpha}{p} p\, d\mu \right. \\ &\quad \left. - \int \left(\frac{f_{\theta_2}}{(1-\lambda) f_{\theta_1} + \lambda r_\eta}\right)^2 p\, d\mu + 1\right\} \\ &= -\int \left(\frac{f_{\theta_2}}{(1-\lambda) f_{\theta_1} + \lambda r_\eta} - \frac{f_\alpha}{p}\right)^2 p\, d\mu + \int \left(\frac{f_\alpha}{p} - 1\right)^2 p\, d\mu \\ &\leqslant \int \left(\frac{f_\alpha}{p} - 1\right)^2 p\, d\mu = \chi^2(f_\alpha, p)\end{aligned}$$

and $\sup_g \int m_{g,\alpha} p\, d\mu = \chi^2(f_\alpha, p)$, and $g^* := \arg\sup_g \int m_{g,\alpha} p\, d\mu$ is defined through $g^* := 2\left(\frac{f_\alpha}{p} - 1\right)$. Therefore $\inf_\alpha \sup_g \int m_{g,\alpha} p\, d\mu = \inf_\alpha \chi^2(f_\alpha, p) = \chi^2(\Omega, p)$. Furthermore by continuity and compactness arguments, there exists α^+ such that $\sup_g \inf_\alpha \int m_{g,\alpha} p\, d\mu = \chi^2(f_{\alpha^+}, p)$. Hence $\chi^2(f_{\alpha^+}, p) = \sup_g \inf_\alpha \int m_{g,\alpha} p\, d\mu \geqslant \inf_\alpha \sup_g \int m_{g,\alpha} p\, d\mu = \chi^2(\Omega, p)$, which concludes the proof. Furthermore $\alpha^+ = \alpha^*$ by identifiability. $\square$

Following (11) introduce an auxiliary estimate

$$\overline{\chi_n^2(\Omega, p)} := \sup_{g\in\mathcal{F}} \inf_{\alpha\in\Theta} \int m_{g,\alpha}(x)\, dP_n(x).$$

Both statistics $\overline{\chi_n^2(\Omega, p)}$ and $\chi_n^2(\Omega, p)$ are close to each other; indeed it holds.

Lemma 4. *Assume that* $\mathcal{H} := \left\{\left(g + \frac{1}{4}g^2\right), g \in \mathcal{F}\right\}$ *is a Glivenko–Cantelli class of functions. Then both* $\overline{\chi_n^2(\Omega, p)}$ *and* $\chi_n^2(\Omega, p)$ *converge to* $\chi^2(\Omega, p)$.

The asymptotic distribution of $\sqrt{n}\left(\overline{\chi_n^2(\Omega, p)} - \chi^2(\Omega, p)\right)$ follows from Theorem 2. This allows to state the limit distribution of $\chi_n^2(\Omega, p)$.

Theorem 10. *Assume* (D). *Assume further that both* Θ *and* H *are compact and* $\mathcal{H} := \left\{\left(g+\frac{1}{4}g^2\right), g \in \mathcal{F}\right\}$ *is a Donsker class of functions. Then* $\sqrt{n}\left(\chi_n^2(\Omega, p) - \chi^2(\Omega, p)\right)$ *converges to a centered normal distribution with variance* $\int \left(g^* + \frac{1}{4}g^{*2}\right)^2 p\, d\mu - \left(\int \left(g^* + \frac{1}{4}g^{*2}\right) p\, d\mu\right)^2$.

6. Proofs

6.1. Proof of Proposition 3

(i) Differentiating the function in (14) with respect to a_s, $s = 0, 1, \ldots, k$ yields

$$a_0 = -\sum_{i=1}^{k} a_i P_n g_i \tag{35}$$

for $s = 0$, while for $s > 0$

$$(Q_0 - P_n)\, g_s = \frac{1}{2}\left(a_0 P_n g_s + \sum_{i=1}^{k} a_i P_n g_i g_s\right). \tag{36}$$

Substituting (35) in the last display, and setting $\underline{a}' = \{a_1, a_2, \ldots, a_k\}$, we get

$$(Q_0 - P_n)\, g_s = \frac{1}{2}\sum_{i=1}^{k} a_i \left(P_n g_i g_s - P_n g_s P_n g_i\right),$$

i.e.

$$2\underline{v}_n = S_n \underline{a} \tag{37}$$

Set $f_n^* = \arg\max_{\mathcal{F}}(Q_0 - P_n) f - \frac{1}{4} P_n f^2$. For every h in $\langle \mathcal{G} \rangle$, $(Q_0 - P_n)\, h - \frac{1}{2} P_n h f_n^* = 0$. Set $h := f_n^*$ to obtain $(Q_0 - P_n)\, f_n^* = \frac{1}{2} P_n (f_n^*)^2$.

It then follows, using (35) and (37),

$$\chi_n^2 = \left[(Q_0 - P_n)\, f_n^* - \frac{1}{4} P_n (f_n^*)^2\right] = \frac{1}{4} P_n (f_n^*)^2$$

$$= \frac{1}{4} P_n \left(\sum_{i=1}^{k} a_i g_i - \sum_{i=1}^{k} a_i P_n g_i\right)^2 = \frac{1}{4}\underline{a}' S_n \underline{a} = \underline{v}_n' S_n^{-1} \underline{v}_n.$$

(ii) The proof is similar to the above one

6.2. Proof of Proposition 4

From Proposition 3,

$$\left|\chi_n^2 - \chi^2\left(\Omega, P\right)\right| = \left|\underline{v}_n' S_n^{-1} \underline{v}_n - \underline{v} S^{-1} \underline{v}\right|$$

$$= \left|\underline{v}_n' \left(S_n^{-1} - S^{-1}\right) \underline{v}_n\right| + \left|\underline{v}_n' S^{-1} \underline{v}_n - \underline{v}' S^{-1} \underline{v}\right|.$$

For $\underline{x}$ in $\mathbb{R}^k$ denote $\left\|\underline{x}\right\|$ the euclidean norm. Over the space of matrices $k \times k$ introduce the algebraic norm $|||A||| = \sup_{\|\underline{x}\| \leqslant 1} \frac{\|A\underline{x}\|}{\|\underline{x}\|} = \sup_{\|\underline{x}\|=1} \left\|A\underline{x}\right\|$. All entries of A satisfy $|a\,(i, j)| \leqslant |||A|||$. Moreover, if $|\lambda_1| \leqslant |\lambda_2| \leqslant \cdots \leqslant |\lambda_k|$ are the eigenvalues of A, $|||A||| = |\lambda_k|$. Observe further that, if for all (i, j), $|a\,(i, j)| \leqslant \varepsilon$, then, for any $\underline{x} \in \mathbb{R}^k$, such that $\left\|\underline{x}\right\| = 1$, $\left\|A\underline{x}\right\|^2 = \sum_{i=1}^{k}\left(\sum_j a\,(i, j)\, x_j\right)^2$ $\leqslant \sum_i \sum_j a\,(i, j)^2 \left\|\underline{x}\right\|^2 \leqslant k^2 \varepsilon^2$, i.e. $|||A||| \leqslant k\varepsilon$.

For the first term in the RHS of the above display

$$A := \underline{v}_n{}' \left(S_n^{-1} - S^{-1}\right) \underline{v}_n = \underline{v}_n{}' S^{-1/2} \left(S^{1/2} S_n^{-1} S^{1/2} - I\right) S^{-1/2} \underline{v}_n$$
$$\leqslant \left\| \underline{v}_n{}' S^{-1/2} \right\| \left\| S^{1/2} S_n^{-1} S^{1/2} - I \right\| \leqslant const.k \left\| S^{1/2} S_n^{-1} S^{1/2} - I \right\|.$$

Hence, if $B := \left\| S^{1/2} S_n^{-1} S^{1/2} - I \right\|$ tends to 0 a.s., so does A.

First note that

$$S_n^{-1} = (S + S_n - S)^{-1} = S^{-1/2} \left(I + S^{-1/2} (S_n - S) S^{-1/2}\right)^{-1} S^{-1/2}$$
$$= S^{-1/2} \left[I + \sum_{h=1}^{\infty} \left(S^{-1/2} (S - S_n) S^{-1/2} \right)^h \right] S^{-1/2}.$$

Hence

$$S^{1/2} S_n^{-1} S^{1/2} - I = \sum_{h=1}^{\infty} \left(S^{-1/2} (S - S_n) S^{-1/2} \right)^h,$$

which entails

$$\left\| S^{1/2} S_n^{-1} S^{1/2} - I \right\| = \left\| \sum_{h=1}^{\infty} \left(S^{-1/2} (S - S_n) S^{-1/2} \right)^h \right\| \leqslant \sum_{h=1}^{\infty} \| S - S_n \|^h \left\| S^{-1/2} \right\|^{2h}$$
$$= O_P \left(\lambda_1^{-1} k \sup_{i,j} |s_n (i, j) - s (i, j)| \right),$$

where λ_1 is the smallest eigenvalue of S.

Since

$$C := \sup_{i,j} |s_n (i, j) - s (i, j)| \leqslant \sup_{i,j} \left| (P_n - P) g_i g_j \right| + \sup_i |(P_n - P) g_i| (P_n + P) \max_{g \in \mathcal{G}} |g| \tag{38}$$

the LLN implies that C tends to 0 a.s. which in turn implies that B tends to 0.

For the second term, $\left| \underline{v}_n{}' S^{-1} \underline{v}_n - \underline{v}' S^{-1} \underline{v} \right| = \left| \left(\underline{v}_n + \underline{v} \right)' S^{-1} \left(n^{-1/2} \underline{\gamma}_n \right) \right|$ tends to 0 by LLN.

6.3. Proof of Theorem 5

By Proposition 3

$$\frac{n\chi_{n,k}^2 - k}{\sqrt{2k}} = \frac{\left(\underline{B}_{n,k}^0\right)' S_k^{-1} \underline{B}_{n,k}^0 - k}{\sqrt{2k}} + 2 (2k)^{-1/2} \left(\underline{B}_{n,k}^0\right)' S_k^{-1} \left(\underline{\gamma}_{n,k} - \underline{B}_{n,k}^0\right)$$
$$+ (2k)^{-1/2} \left(\underline{\gamma}_{n,k} - \underline{B}_{n,k}^0\right)' S_k^{-1} \left(\underline{\gamma}_{n,k} - \underline{B}_{n,k}^0\right)$$
$$+ (2k)^{-1/2} \underline{\gamma}_{n,k}' \left(S_{n,k}^{-1} - S_k^{-1}\right) \underline{\gamma}_{n,k}$$
$$= A + B + C + D.$$

The first term above can be written as

$$A = \frac{\left(\underline{B}^0_{n,k}\right)' S_k^{-1} \underline{B}^0_{n,k} - k}{\sqrt{2k}} = \frac{\sum_{i=1}^k \left(Z_i^2 - EZ_i^2\right)}{\sqrt{kVarZ_i^2}}$$

which converges to the standard normal distribution by the CLT applied to the i.i.d. standard normal r.v's Z_i.

As to the term C it is straightforward that $C = o(B)$. From the proof of Theorem 3, D goes to zero if $\lambda_{1,k}^{-1} k^{1/2} \left(\sup_{i,j} \left| s_{n,k}(i,j) - s_k(i,j) \right|\right) \left\| \underline{\gamma}'_{n,k} S_k^{-1/2} \right\|^2 = o_P(1)$. Since, using (38) and (17), $\sup_{i,j} \left| s_{n,k}(i,j) - s_k(i,j) \right| \leqslant \sup_{f,g \in \mathcal{F}} |(P_n - P)\, fg| + \sup_{f \in \mathcal{F}} |(P_n - P)\, f|\, |(P_n + P)\, F| = O_P\left(n^{-1/2}\right)$, and considering that $\left\| \underline{\gamma}'_{n,k} S_k^{-1/2} \right\|^2 = O_P(k)$, (24) yields $D = o(1)$.

For B, $\left| \left(\underline{B}^0_{n,k}\right)' S_k^{-1} \left(\underline{\gamma}_{n,k} - \underline{B}^0_{n,k}\right) \right| \leqslant \left\| \left(\underline{B}^0_{n,k}\right)' S_k^{-1/2} \right\| \left\| S_k^{-1/2} \left(\underline{\gamma}_{n,k} - \underline{B}^0_{n,k}\right) \right\| = \sqrt{\sum_{i=1}^k Z_i^2}$ $\left\| S_k^{-1/2} \left(\underline{\gamma}_{n,k} - \underline{B}^0_{n,k}\right) \right\|$ where, as used in A, Z_i^2 are i.i.d. with a chi-square distribution with 1 degree of freedom. Hence $\sqrt{\sum_{i=1}^k Z_i^2} = O_P\left(k^{1/2}\right)$. Further

$$\begin{aligned}
\left\| S_k^{-1/2} \left(\underline{\gamma}_{n,k} - \underline{B}^0_{n,k}\right) \right\| &\leqslant \left\| S_k^{-1/2} \right\| \cdot \left\| \underline{\gamma}_{n,k} - \underline{B}^0_{n,k} \right\| \\
&\leqslant \lambda_{1,k}^{-1/2} k^{1/2} \sqrt{\frac{1}{k} \sum_{i=1}^k \left(\underline{\gamma}_{n,k}(f_i) - \underline{B}^0_{n,k}(f_i)\right)^2} \\
&\leqslant \lambda_{1,k}^{-1/2} k^{1/2} \sup_{f \in \mathcal{F}} \left| \gamma_n(f) - B_n^0(f) \right|
\end{aligned}$$

from which $B = O_P\left(\lambda_{1,k}^{-1/2} k^{1/2} \delta_n \log n\right) = o_P(1)$ if (23) holds. We have used the fact that P belongs to Ω in the last evaluation of B.

Remark 4. Under H1, using the relation $\underline{v}_n = \underline{v} - n^{-1/2} \underline{\gamma}_{n,k}$, we can write

$$\begin{aligned}
\frac{n\chi^2_{n,k} - k}{\sqrt{2k}} &= (2k)^{-1/2} \left(\underline{\gamma}'_{n,k} S_{n,k}^{-1} \underline{\gamma}_{n,k} - k\right) + (2k)^{-1/2} \left(n \underline{v}' S_{n,k}^{-1} \underline{v} - 2\sqrt{n} \underline{\gamma}'_{n,k} S_{n,k}^{-1} \underline{v}\right) \\
&= (2k)^{-1/2} \left(n \underline{v}' S_{n,k}^{-1} \underline{v} - 2\sqrt{n} \underline{\gamma}'_{n,k} S_{n,k}^{-1} \underline{v}\right) + O_P(1),
\end{aligned}$$

where the $O_P(1)$ term captures $(2k)^{-1/2} \left(\underline{\gamma}'_{n,k} S_{n,k}^{-1} \underline{\gamma}_{n,k} - k\right)$ that coincides with the test statistic $\frac{n\chi^2_{n,k} - k}{\sqrt{2k}}$ under H0. We can bound the first term from below by

$$\begin{aligned}
&(2k)^{-1/2} n \underline{v}' S_k^{-1} \underline{v} \left(1 - O_P\left(\left\|\left| S_k^{1/2} S_{n,k}^{-1} S_k^{1/2} - I \right|\right\|\right)\right) \\
&-(2n/k)^{1/2} \left\| \underline{\gamma}'_{n,k} S_k^{-1/2} \right\| \left\| \underline{v}' S_k^{-1/2} \right\| \left(1 + O_P\left(\left\|\left| S_k^{1/2} S_{n,k}^{-1} S_k^{1/2} - I \right|\right\|\right)\right) \\
&= O_P\left(n k^{-1/2}\right) - O_P(n^{1/2}).
\end{aligned}$$

Hence, if (23) and (24) are satisfied then the test statistic is asymptotically consistent also for the case of an infinite number of linear constraints.

6.4. Proof of Lemma 2

Write $s_k^\delta(u,v)$ for the (u,v)th element of S_k^δ. We have, for $P \in \Omega$, i.e. if $Pg = Q_0^d f$, for every g in $\mathcal{G}_n^\delta$

$$s_k^\delta((j-1)m+i,(h-1)m+l) = s_k^\delta(u,v) = Pg_u g_v - Pg_u Pg_v$$
$$= P\left(g_u - Q_0^d g_u\right)\left(g_v - Q_0^d g_v\right) = P\left(\overline{g_u g_v}\right),$$

where $\overline{g_u} = g_u - Q_0^d g_u$. For each vector $\underline{a} \in \mathbb{R}^{d\cdot m}$ it then holds

$$\begin{aligned}\underline{a}' S_k^\delta \underline{a} &= \sum_{u=1}^{dm}\sum_{v=1}^{dm} a_u a_v P\left(\overline{g_u g_v}\right) = P\left(\sum_{u=1}^{dm} a_u \overline{g_u}\right)^2\\ &= \int_{[0,1]^d}\left(\sum_{u=1}^{dm} a_u\overline{g_u}\right)^2 dP \geqslant \alpha\int_{[0,1]^d}\left(\sum_{u=1}^{dm} a_u\overline{g_u}\right)^2 dQ_0^d\\ &= \alpha\,\underline{a}'\left\{Q_0^d\left(g_u - Q_0^d g_u\right)\left(g_v - Q_0^d g_v\right)\right\}_{u,v}\underline{a} = \alpha\,\underline{a}' S_k^0 \underline{a}.\end{aligned}$$

On the other hand, the preceding inequality implies

$$\lambda_{1,k} = \inf_{\underline{a}} \frac{\underline{a}' S_k^\delta \underline{a}}{\left\|\underline{a}\right\|^2} \geqslant \alpha \inf_{\underline{a}} \frac{\underline{a}' S_k^0 \underline{a}}{\left\|\underline{a}\right\|^2} = \alpha \min_{1\leqslant i\leqslant m} p_i. \tag{39}$$

6.5. Proof of Lemma 3

(i) This can be checked easily through some calculation.

(ii) First, notice that

$$U^2 = (1-p_{m+1})U. \tag{40}$$

Formula (40) implies that at least one eigenvalue equals $(1-p_{m+1})$. On the other hand, summing up all diagonal entries in U we get $trace(U) = \sum_{i=1}^m p_i = 1 - p_{m+1}$. This allows to conclude that there can be only one eigenvalue equal to $1-p_{m+1}$ while the others must be zero.

For the second statement, use $(I-U)^{-1} = I + \sum_{h=1}^\infty U^h$ together with (40) to obtain $(I-U)^{-1} = I + U\sum_{h=1}^\infty (1-p_{m+1})^h = U + \frac{1}{p_{m+1}}U$.

(iii) For any eigenvalue λ of S_k^0 we have

$$\lambda \leqslant \lambda_{k,k} = \left\|S_k^0\right\| \leqslant \left\|D^{1/2}\right\|^2 \left\|(I-V)\right\| = \max_{1\leqslant i\leqslant m} p_i\left(1 - \inf_{\|\underline{x}\|=1}\underline{x}' V \underline{x}\right) = \max_{1\leqslant i\leqslant m} p_i$$

since the eigenvalues of V coincide with the eigenvalues of U with order multiplied by d.

For the opposite inequality consider

$$\begin{aligned}\lambda^{-1} &\leqslant \lambda_{1,k}^{-1} = \left\|S_k^{0^{-1}}\right\| \leqslant \left\|D^{-1}\right\|\left\|\left(I + \frac{1}{p_{m+1}}V\right)\right\|\\ &= \left(\max_{1\leqslant i\leqslant m} p_i^{-1}\right)\left(1 + \frac{1}{p_{m+1}}(1-p_{m+1})\right) = \left(\min_{1\leqslant i\leqslant m} p_i\right)^{-1} p_{m+1}^{-1}.\end{aligned}$$

6.6. Proof of Theorem 7

Let

$$\mathcal{Q} = \left\{ Q \in M_1([0,1]^2) : \sum_{i=1}^{m+1} q_{i,j} = p_{\cdot,j} = u_{j+1} - u_j, : j = 1, \ldots, m+1; \right.$$
$$\left. \sum_{j=1}^{m+1} q_{i,j} = p_{i,\cdot} = u_{i+1} - u_i, : i = 1, \ldots, m+1 \right\}.$$

Lemma 5. *Under* H0 *it holds*

$$n\chi^2_{n,k} = \min_{Q\in\mathcal{Q}} \sum_{i=1}^{m+1}\sum_{j=1}^{m+1} \frac{(nq_{i,j} - N_{i,j})^2}{N_{i,j}} \mathbb{I}_{N_{i,j}>0}, \tag{41}$$

$$\underline{\gamma}'_{n,k} S_k^{-1} \underline{\gamma}_{n,k} = \min_{Q\in\mathcal{Q}} \sum_{i=1}^{m+1}\sum_{j=1}^{m+1} \frac{(nq_{i,j} - N_{i,j})^2}{np_{i,j}}. \tag{42}$$

Proof. We prove (41), since the proof of (42) is similar. Following [3] the RHS in (41) is

$$\sum_{i=1}^{m+1}\sum_{j=1}^{m+1} N_{i,j}(a_i + b_j)^2,$$

where the vectors a and $b \in \mathbb{R}^{m+1}$ are solutions of the equations

$$a_i \frac{N_{i,\cdot}}{n} = p_i - \frac{N_{i,\cdot}}{n} - \sum_{j=1}^{m+1} b_j \frac{N_{i,j}}{n}, \quad i = 1, \ldots, m+1,$$

$$b_j \frac{N_{\cdot,j}}{n} = p_j - \frac{N_{\cdot,j}}{n} - \sum_{i=1}^{m+1} a_i \frac{N_{i,j}}{n}, \quad j = 1, \ldots, m+1.$$

Let $\underline{a} = (\widetilde{a}_1, \ldots, \widetilde{a}_m, \tilde{b}_1, \ldots, \tilde{b}_m)$ be the coefficients in Eq. (37). Making use of Eqs. (35) and (36) we obtain, using the class $\mathcal{G}$ in place of $\mathcal{G}_n$ in the definition of $\chi^2_{n,k}$,

$$\begin{aligned}
\widetilde{a}_i &= 2\left(a_i - a_{m+1}\right), \quad i = 1, \ldots, m, \\
\widetilde{b}_j &= 2\left(b_j - b_{m+1}\right), \quad j = 1, \ldots, m, \\
\widetilde{a}_0 &= 2\left(a_{m+1} + b_{m+1}\right).
\end{aligned}$$

From the proof of Proposition 3 we get, setting $\delta_{i,j} = 1$ for $i = j$ and 0 otherwise,

$$\begin{aligned}
\chi^2_{n,k} &= \frac{1}{4}\underline{a}' S_{n,k} \underline{a} \\
&= \frac{1}{4n} \sum_{i=1}^{m}\sum_{j=1}^{m} \Big[\widetilde{a}_i \widetilde{a}_j \left(N_{i,\cdot}\delta_{i,j} - N_{i,\cdot}N_{j,\cdot}/n\right) + \tilde{b}_i \tilde{b}_j \left(N_{\cdot,i}\delta_{i,j} - N_{\cdot,i}N_{\cdot,j}/n\right) \\
&\quad + 2\tilde{a}_i \tilde{b}_j \left(N_{i,j} - N_{i\cdot}N_{\cdot,j}/n\right) \Big],
\end{aligned}$$

which, after some algebra yields

$$n\chi^2_{n,k} = \sum_{i=1}^{m+1}\sum_{j=1}^{m+1} N_{i,j}(a_i + b_j)^2. \qquad \square$$

Let us turn to the proof of Theorem 7. Set $k := 2m$. It is enough to prove $\frac{n\chi^2_{n,k} - \underline{\gamma}'_{n,k} S_k^{-1} \underline{\gamma}_{n,k}}{\sqrt{4m}} = o_P(1)$.

Denote $\hat{P}$ and $\overline{P}$ the minimizers of (41) and (42) in $\mathcal{Q}$. Let $\hat{p}_{i,j}$ and $\overline{p}_{i,j}$ denote the respective probabilities of cells.

We write

$$\begin{aligned} n\chi^2_{n,k} - \underline{\gamma}'_{n,k} S_k^{-1} \underline{\gamma}_{n,k} &\leqslant \sum_{i=1}^{m+1}\sum_{j=1}^{m+1} \left(N_{i,j} - n\overline{p}_{i,j}\right)^2 \left(\frac{1}{N_{i,j}} - \frac{1}{np_{i,j}}\right) \\ &\leqslant \max_{i,j}\left(\frac{np_{i,j}}{N_{i,j}} - 1\right) \underline{\gamma}'_{n,k} S_k^{-1} \underline{\gamma}_{n,k} \end{aligned}$$

and

$$\begin{aligned} n\chi^2_{n,k} - \underline{\gamma}'_{n,k} S_k^{-1} \underline{\gamma}_{n,k} &\geqslant \min_{Q\in\mathcal{Q}} \sum_{i=1}^{m+1}\sum_{j=1}^{m+1} \frac{\left(N_{i,j} - nq_{i,j}\right)^2}{np_{i,j}} \left(\frac{np_{i,j}}{N_{i,j}} - 1\right) \\ &\geqslant -\max_{i,j}\left|\frac{np_{i,j}}{N_{i,j}} - 1\right| \underline{\gamma}'_{n,k} S_k^{-1} \underline{\gamma}_{n,k}. \end{aligned}$$

Whenever

$$\sqrt{m}\max_{i,j}\left|\frac{np_{i,j}}{N_{i,j}} - 1\right| \xrightarrow{P} 0 \tag{43}$$

holds, then the above inequalities yield $\frac{n\chi^2_{n,k} - \underline{\gamma}'_{n,k} S_k^{-1} \underline{\gamma}_{n,k}}{\sqrt{4m}} = o_P\left(\frac{\underline{\gamma}'_{n,k} S_k^{-1} \underline{\gamma}_{n,k}}{m}\right) = o_P\left(\frac{\underline{\gamma}'_{n,k} S_k^{-1} \underline{\gamma}_{n,k} - 2m}{\sqrt{4m}}\frac{2}{\sqrt{m}} + 1\right) = o_P(1)$, which proves the claim.

We now prove (43). We proceed as in Lemma 2 in [3], using inequalities (10.3.2) in [21]. Let $B_n \sim Bin(n, p)$. Then, for $t > 1$,

$$\Pr\left(\frac{np}{B_n} \geqslant t\right) \leqslant \exp\{-np\, h(1/t)\} \quad \text{and} \quad \Pr\left(\frac{B_n}{np} \geqslant t\right) \leqslant \exp\{-np\, h(t)\}, \tag{44}$$

where $h(t) = t\log t - t + 1$ is a positive function.

Since $N_{i,j} \sim Bin(n, p_{i,j})$,

$$\begin{aligned} \Pr\left\{\max_{i,j}\left(\frac{np_{i,j}}{N_{i,j}} - 1\right) \geqslant \frac{t}{\sqrt{m}}\right\} &\leqslant \sum_{i=1}^{m+1}\sum_{j=1}^{m+1} \Pr\left\{\frac{np_{i,j}}{N_{i,j}} \geqslant \frac{t}{\sqrt{m}} + 1\right\} \\ &\leqslant \sum_{i=1}^{m+1}\sum_{j=1}^{m+1} \exp\left\{-np_{i,j} h\left(1/\left(1 + t/\sqrt{m}\right)\right)\right\} \end{aligned}$$

$$\text{(by (27) and by } p_{i,j} > \alpha p_{i,\cdot}p_{\cdot,j}) \leqslant (m+1)^2 \exp\left\{-\frac{c\alpha n}{\log n}(\log n)\frac{h\left(1/(1 + t/\sqrt{m})\right)}{m^2}\right\}.$$

For $x = 1+\varepsilon$, $h(x) = O(\varepsilon^2)$. Therefore, using (23) with $k = 2m$, for every $M > 0$ there exists n large enough that

$$\alpha c \frac{n}{\log n} m^{-2} h\left(1 + \frac{-t/\sqrt{m}}{1+t/\sqrt{m}}\right) \geqslant M$$

and consequently $\Pr\left\{\max_{i,j}\left(\frac{np_{i,j}}{N_{i,j}} - 1\right) \geqslant \frac{t}{\sqrt{m}}\right\}$ goes to 0.

To get convergence to zero of $\Pr\left\{\max_{i,j}\left(1 - \frac{np_{i,j}}{N_{i,j}}\right) \geqslant \frac{t}{\sqrt{m}}\right\} = \Pr\left\{\max_{i,j} \frac{N_{i,j}}{np_{i,j}} \geqslant \frac{1}{1-t/\sqrt{m}}\right\}$, the second inequality in (44) is used in a similar way.

6.7. Proof of Lemma 4

It holds $\left|\overline{\chi_n^2(\Omega, p)} - \chi^2(\Omega, p)\right| \leqslant \sup_g |\inf_\alpha \int m_{g,\alpha}(x)\, dP_n(x) - \inf_\alpha \int m_{g,\alpha}(x)\, dP(x)| = \sup_g \left|\int \left(g + \frac{1}{4}g^2\right) d(P_n - P)\right|$ which converges to 0 under the condition of the lemma.

By the GC property, $\inf_\alpha \sup_{g\in\mathcal{F}_\alpha} \int m_{g,\alpha}(x)\, dP_n(x) \leqslant \sup_{g\in\mathcal{F}_\alpha} \int m_{g,\alpha}(x)\, dP_n(x) = \sup_{g\in\mathcal{F}_\alpha} \int m_{g,\alpha}(x)\, dP(x) + R_n(\alpha)$ with $\lim_{n\to\infty} R_n(\alpha) = 0$, implying $\limsup_{n\to\infty} \chi_n^2(\Omega, p) \leqslant \chi^2(\Omega, p)$. Further $\sup_{g\in\mathcal{F}_\alpha} \int m_{g,\alpha}(x)\, dP_n(x) \geqslant \int m_{g,\alpha}(x)\, dP_n(x) + R_n(\alpha, g)$ for all α and g in $\mathcal{F}_\alpha$, and $\lim_{n\to\infty} \sup_{g\in\mathcal{F}_\alpha} R_n(\alpha, g) = 0$. Hence $\liminf_{n\to\infty} \chi_n^2(\Omega, p) \geqslant \chi^2(\Omega, p) + \liminf_{n\to\infty} \inf_\alpha \sup_{g\in\mathcal{F}_\alpha} R_n(\alpha, g)$. Since $R_n(\alpha, g) = \int m_{g,\alpha}(x) d\,(P_n - P) = -\int \left(g + \frac{1}{4}g^2\right) d\,(P_n - P)$ and $\left\{g + \frac{1}{4}g^2, g \in \mathcal{F}\right\}$ is GC, $\liminf_{n\to\infty} \chi_n^2(\Omega, p) \geqslant \chi^2(\Omega, p)$, which concludes the proof.

6.8. Proof of Theorem 8

First note that the two first components of $\mathbf{t}(\theta_0)$ are θ_0 and 0, and its third component is arbitrary in H (since $\int m(t(\theta_0), \theta_0) f_{\theta_0}\, d\mu = 0$ implies $\int g^2 f_{\theta_0}\, d\mu = 0$, implying the claim).

Also for all η in H, the gradient vector $\left(\frac{\partial}{\partial \mathbf{t}} m(\mathbf{t}, \alpha), \frac{\partial}{\partial \alpha} m(\mathbf{t}, \alpha)\right)$ at $(\mathbf{t}(\theta_0), \theta_0)$ writes $\left(\frac{\partial}{\partial \theta} m(t(\theta_0), \theta_0), \frac{\partial}{\partial \lambda} m(t(\theta_0), \theta_0), 0, \frac{\partial}{\partial \alpha} m(t(\theta_0), \theta_0)\right)$, which we write as a $2d + 1$ nonnull entries vector. Also, the hessian matrix of $m(t, \alpha)$ at $(\mathbf{t}(\theta_0), \theta_0)$ has all entries in the row and column η equal to 0. We consider it a $2d + 1$ lines and $2d + 1$ columns matrix $\mathbf{H}$ with nonnull entries. Denote $C := \int \mathbf{H} f_{\theta_0}\, d\mu$ and C_n the matrix $\int \mathbf{H}\, dP_n$. We need now to prove that the $d + 1$ first entries of $\mathbf{t}_n(\boldsymbol{\alpha}_n)$ tend to $(\theta_0, 0)$ as n increases. This result together with compactness of the set H and hypotheses (H) will entail that all remainder terms in the subsequent Taylor expansions go to 0. Let $\mathbf{t}_n(\boldsymbol{\alpha}_n) := (\theta_n, \lambda_n, h_n)$.

Lemma 6. *Under* H0, $\lim_{n\to\infty} \lambda_n = 0$ *and* $\lim_{n\to\infty} \theta_n = \theta_0$ *a.s.*

Proof. Denote p by f_{θ_0}. First note that for any $\boldsymbol{\alpha}$ there exists a unique $\mathbf{t}_n(\boldsymbol{\alpha}) := \arg\sup_{\mathbf{t}} \int m(\mathbf{t}, \boldsymbol{\alpha})\, dP_n$. Indeed $g \to m_{g,\alpha} := \int g f_\alpha\, d\mu - \left(g - \frac{1}{4}g^2\right)$ is a strictly concave mapping and so is $g \to \int m_{g,\alpha}\, dP_n$. Hence there exists a (possibly nonunique) optimizer of $\int m_{g,\alpha}\, dP_n$ on $\mathcal{F}_\alpha$, say g^+. Identifiability then implies existence of $\mathbf{t}_n(\boldsymbol{\alpha})$, the vector of parameters which defines g^+. By Lemma 4, $\chi_n^2(f_{\theta_0}, f_{\theta_0}) \to 0$ a.s. Since $\boldsymbol{\Theta}$ and H are compact sets there exists $\{n_k\} \subset \{n\}$ and $\mathbf{t} := (\theta, \lambda, h)$ such that $\lim_{k\to\infty} \mathbf{t}_{n_k}(\boldsymbol{\alpha}_{n_k}) = \mathbf{t}$. Since $\sup_{\mathbf{t}} \left|\int m(\mathbf{t}, \boldsymbol{\alpha})\, dP_n - \int m(\mathbf{t}, \boldsymbol{\alpha}) f_{\theta_0}\, d\mu\right| \to 0$ and since $\int m(\mathbf{t}, \boldsymbol{\alpha}) f_{\theta_0}\, d\mu$ has a unique maximizer $\mathbf{t}(\boldsymbol{\alpha})$ it follows that the first two components of $\mathbf{t}$ and $\mathbf{t}(\boldsymbol{\alpha})$ coincide. It therefore follows that

$$\begin{aligned}\chi^2_{n_k} &= \int m(\mathbf{t}_{n_k}(\boldsymbol{\alpha}_{n_k}), \boldsymbol{\alpha}_{n_k})\, dP_{n_k} \\ &\to \int m(\mathbf{t}, \boldsymbol{\alpha}) f_{\theta_0}\, d\mu = \chi^2(f_\alpha, f_{\theta_0}) = 0 = \chi^2(f_{\theta_0}, f_{\theta_0}),\end{aligned}$$

which implies $\alpha = \theta_0$. Therefore $\int m(\mathbf{t}, f_{\theta_0}) f_{\theta_0}\, d\mu = 0$, implying $g = 0$, which in turn (due to (29)) entails $\lambda = 0$. Therefore, all limiting points of the sequence λ_n equal 0, which amounts to say that λ_n tends to 0. Furthermore, the first component of $\mathbf{t}(\boldsymbol{\alpha})$ equals $\alpha = \theta_0$. $\square$

As a first consequence of the above lemma and (H), $C_n = C + o_p(1)$. By a Taylor expansion we get

$$-\int \left(\frac{\partial}{\partial \mathbf{t}} m(\mathbf{t}, \alpha), \frac{\partial}{\partial \alpha} m(\mathbf{t}, \alpha)\right)_{(\mathbf{t}(\theta_0), \theta_0)} dP_n = C a_n + o_p(1), \tag{45}$$

where $a_n := (\theta_n - \theta_0, \lambda_n, \alpha_n - \theta_0)'$. By the above display and the multidimensional CLT, $\sqrt{n}\, a_n = O_p(1)$. We develop a simple expression for $\chi^2_n(\Omega, f_{\theta_0})$, through usual second-order Taylor expansion, in the form

$$\begin{aligned}\chi^2_n(\Omega, f_{\theta_0}) = &\int m(\mathbf{t}(\boldsymbol{\theta}_0), \boldsymbol{\theta}_0)\, dP_n + \int \left(\left(\frac{\partial}{\partial \mathbf{t}} m(\mathbf{t}, \alpha), \frac{\partial}{\partial \alpha} m(\mathbf{t}, \alpha)\right)_{(\mathbf{t}(\theta_0), \theta_0)}\right)' a_n \\ &+ \frac{1}{2} a_n' C a_n + o_p(1/n) = -\frac{1}{2} (a_n)' C (a_n) + o_p(1/n)\end{aligned}$$

from which $2n\chi^2_n(\Omega, f_{\theta_0}) = -\frac{1}{2}\left(\sqrt{n} a_n\right)' C \left(\sqrt{n} a_n\right) + o_p(1)$. The matrix C is block-diagonal, since it can easily be seen that $\int \frac{\partial^2}{\partial \mathbf{t} \partial \boldsymbol{\alpha}} m(\mathbf{t}, \alpha)_{(\mathbf{t}(\theta_0), \theta_0)} f_{\theta_0}\, d\mu = 0$. Also $\int \frac{\partial^2}{\partial \boldsymbol{\alpha}^2} m(\mathbf{t}, \alpha)_{(\mathbf{t}(\theta_0), \theta_0)} f_{\theta_0}\, d\mu = -2E\left[\left(\frac{\partial}{\partial \alpha} \log f_\alpha\right)'_{(\mathbf{t}(\theta_0), \theta_0)} \left(\frac{\partial}{\partial \alpha} \log f_\alpha\right)_{(\mathbf{t}(\theta_0), \theta_0)}\right]$ and $\int \frac{\partial^2}{\partial \mathbf{t}^2} m(\mathbf{t}, \alpha)_{(\mathbf{t}(\theta_0), \theta_0)} f_{\theta_0}\, d\mu = -2E\left[\left(\frac{\partial}{\partial \mathbf{t}} \log g_{\mathbf{t}}\right)'_{(\mathbf{t}(\theta_0), \theta_0)} \left(\frac{\partial}{\partial \mathbf{t}} \log g_{\mathbf{t}}\right)_{(\mathbf{t}(\theta_0), \theta_0)}\right]$, with $g_{\mathbf{t}} := (1 - \lambda) f_\theta + \lambda r_\eta$. Let $U_n := \sqrt{n} \int \frac{\partial}{\partial \alpha} m(\mathbf{t}, \alpha)_{(\mathbf{t}(\theta_0), \theta_0)}\, dP_n$, $V_n := \sqrt{n} \int \frac{\partial}{\partial \mathbf{t}} m(\mathbf{t}, \alpha)_{(\mathbf{t}(\theta_0), \theta_0)}\, dP_n$, $A := \int \frac{\partial^2}{\partial \alpha^2} m(\mathbf{t}, \alpha)_{(\mathbf{t}(\theta_0), \theta_0)} f_{\theta_0}\, d\mu$ and $B := \int \frac{\partial^2}{\partial \mathbf{t}^2} m(\mathbf{t}, \alpha)_{(\mathbf{t}(\theta_0), \theta_0)} f_{\theta_0}\, d\mu$ where the entries pertaining to the derivatives with respect to η have been canceled. It then holds, using (45)

$$2n\chi^2_n(\Omega, f_{\theta_0}) = -U_n' A U_n + V_n' B V_n + o_p(1).$$

Also, A is minus twice the Fisher information I_{θ_0} pertaining to α in the uncontaminated model at point θ_0, and B is minus twice the Fisher information matrix pertaining to $\mathbf{t}$ in the contaminated model at point $\mathbf{t}(\theta_0)$ (irrelevantly upon η). Some classical evaluation will now provide the asymptotic distribution of $2n\chi^2_n(\Omega, f_{\theta_0})$, in the spirit of the proof of Wilk's Theorem.

We express V_n in terms of U_n as follows. For any $\boldsymbol{\alpha}$ it holds

$$\frac{\partial}{\partial \boldsymbol{\alpha}} m(\mathbf{t}(\boldsymbol{\alpha}), \boldsymbol{\alpha}) =: G(\mathbf{t}(\boldsymbol{\alpha}))' \frac{\partial}{\partial \mathbf{t}} m(\mathbf{t}, \boldsymbol{\alpha})_{\mathbf{t}(\boldsymbol{\alpha}), \boldsymbol{\alpha}} + \frac{\partial}{\partial \boldsymbol{\alpha}} m(\mathbf{t}, \boldsymbol{\alpha})_{\mathbf{t}(\boldsymbol{\alpha}), \boldsymbol{\alpha}} = 0, \tag{46}$$

which yields $U_n = -G(\theta_0)'V_n$. The multivariate CLT and Slutsky's theorem yield $A^{-1} = G(\theta_0)B^{-1}G(\theta_0)'$. Hence

$$2n\chi_n^2(\Omega, f_{\theta_0}) = V_n' \left[B - G(\theta_0)'AG(\theta_0)\right] V_n + o_p(1).$$

Let $S := \left[B - G(\theta_0)'AG(\theta_0)\right]$ and $V := Var\ V_n = -2B^{-1}$. Then $SVG(\theta_0)' = -2\left[B - G(\theta_0)' AG(\theta_0)\right] B^{-1}G(\theta_0)' = 0$.

Hence $(SV)\,(SV) = SV\left(-2\left[B - G(\theta_0)'AG(\theta_0)\right]B^{-1}\right) = SV\left[I_{d+1} - G(\theta_0)'AG(\theta_0)B^{-1}\right] = SV$, showing that SV is idempotent. Therefore, the asymptotic distribution of $2n\chi_n^2(\Omega, f_{\theta_0})$ is χ^2 whose d.f. equals $rank\,SV$ (see e.g. [20, Corollary 2.2, p. 58]). Set $H := \left(B^{-1}G(\theta_0)'\right)(AG(\theta_0)) =: LM$. Then $MLM = L$, which is to say that L is a pseudo-inverse of M. By Lemma 1, p. 12 in [20], $rank\ SV = rank\ (I_{d+1} - H) = q - r$, where q is the number of columns of L and r its rank. Hence $q = d$, and $rank\ L = rank\ G(\theta_0)' = d - 1$, where the restriction on the rank stems from (46). This concludes the proof.

References

[1] D. Aze, Elements d' analyse convexe et variationnelle, Ellipses, 1997.
[2] R. Beran, Minimum hellinger distance estimates for parametric models, Ann. Statist. 5 (1977) 445–463.
[3] P.J. Bickel, Y. Ritov, J.A. Wellner, Efficient estimation of linear functionals of a probability measure p with known marginal distributions, Ann. Statist. 19 (1991) 1316–1346.
[4] I.S. Borisov, On the accuracy of the approximation of empirical random fields, Theoret. Probab. Appl. 26 (1981) 632–633.
[5] J. Bretagnolle, P. Massart, Hungarian construction from the nonasymptotic viewpoint, Ann. Probab. 17 (1989) 239–256.
[6] M. Broniatowski, Estimation of the Kullback–Leibler divergence, Math. Methods Statist. 12 (4) (2003) 391–409.
[7] M. Broniatowski, A minimax result for divergence projections on locally convex sets of measures, Rendiconti Circ. Mat. Univ. Palermo (2006), to appear.
[8] M. Broniatowski, A. Kéziou, Optimization of phi-divergences on sets of signed measures, Studia Sci. Math. Hungar., 2006, to appear.
[9] N. Cressie, T.R.C. Read, Multinomial goodness-of-fit tests, J. Roy. Statist. Soc. Ser. B 46 (3) (1984) 440–464 MR 86k:62072.
[10] R. Jiménez, Y. Shao, On robustness and efficiency of minimum divergence estimators, Test 10 (2) (2001) 241–248 MR 1 881 138.
[11] V.I. Koltchinskii, Komlos—Major–Tusnady approximation for the general empirical process and haar expansions of classes of functions, J. Theoret. Probab. 7 (1994) 73–118.
[12] H.O. Lancaster, The Chi-squared Distribution, Wiley, New York, 1969 MR 40 #6667.
[13] M. Lembani, O. Pons, Likelihood tests in contamination models, Bernouilli 5 (4) (1999) 705–719.
[14] B.G. Lindsay, Efficiency versus robustness: the case for minimum Hellinger distance and related methods, Ann. Statist. 22 (2) (1994) 1081–1114 MR 95k:62083.
[15] P. Massart, Strong approximation for multivariate empirical and related processes, via KMT reconstructions, Ann. Probab. 17 (1989) 266–291.
[16] D. Morales, L. Pardo, I. Vajda, Asymptotic divergence of estimates of discrete distributions, J. Statist. Plann. Inference 38 (1995) 347–369.
[17] D. Morales, L. Pardo, I. Vajda, Some new statistics for testing hypotheses in parametric models, J. Multivariate Anal. 62 (1997) 137–168.
[18] W. Newey, R. Smith, Higher order properties of GMM and generalized empirical likelihood estimators, Cemmap Working Paper CWP04/03, 2003.
[19] A.B. Owen, Empirical likelihood confidence regions, Ann. Statist. 18 (1990) 90–120.
[20] S.R. Searle, Linear Models, Wiley, New York, 1971.
[21] G.R. Shorack, J.A. Wellner, Empirical Processes with Applications to Statistics, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, Wiley, New York, 1986 MR 88e:60002.

[22] M. Teboulle, I. Vajda, Convergence of best phi-entropy estimates, IEEE Trans. Inform. Theory 39 (1993) 297–301.
[23] D.M. Titterington, A.F.M. Smith, U.E. Makov, Statistical Analysis of Finite Mixture Distributions, Wiley, New York, 1985.
[24] A. Van der Waart, J. Wellner, Weak Convergence of Empirical Processes, Springer, Berlin, 1996.