A useful variant of Wilks' theorem for grouped data

Emanuele Dolera[*], Andrea Bulgarelli[†], Stefano Favaro[‡] and Alessio Aboudan[§]

Abstract

This paper provides a generalization of a classical result obtained by Wilks about the asymptotic behavior of the likelihood ratio. The new results deal with the asymptotic behavior of the joint distribution of a vector of likelihood ratios which turn out to be stochastically dependent, due to a suitable grouping of the data.

Keywords: Wilks theorem; likelihood ratio statistic; grouped data; Lévy-Prokhorov distance

1 Introduction

This paper deals with the problem of testing a composite statistical hypothesis on the basis of the analysis of several *likelihood ratio statistics* (LRS's, from now on) that are obtained after a suitable splitting of the original data set into some groups. This new statistical methodology has been devised to solve a specific problem connected with the analysis of astrophysical data, but we are confident that an abstract formulation could serve as a theoretical guide to manage similar situations arising in other branches of applied science. Thus, we mainly focus on the explanation of the sampling procedure—to be considered itself as a novelty of this work—and the allied mathematical results, deferring the description of the inspiring astrophysical setting to the end of the introduction. Furthermore, throughout the paper, we shall emphasize the theoretical aspects, because of their importance as founding elements of the methodology.

To introduce the basic elements of the problem, suppose we are observing n repeated trials of the same phenomenon, the trials being independent of each other, with the further information about the time at which any trial is performed. Hence, the theoretical description of the sample is in the form of $\big((X_1, t_1), \dots, (X_n, t_n)\big)$, where $(X_1, \dots, X_n)$ is a vector of i.i.d. random variables (r.v.'s) and $(t_1, \dots, t_n)$ belongs to $(0, +\infty)^n$. The first step consists in the use of $(t_1, \dots, t_n)$ to split the original vector $(X_1, \dots, X_n)$

[*]Address: Dipartimento di Matematica, via Ferrata 1, 27100 Pavia, Italy. E-mail: emanuele.dolera@unipv.it

[†]Address: INAF/OAS, Via Gobetti 93/3, 40129 Bologna, Italy. E-mail: bulgarelli@iasfbo.inaf.it

[‡]Address: Dipartimento di Economia e Statistica, corso Unione Sovietica 218/bis, 10134, Torino, Italy. E-mail: stefano.favaro@unito.it

[§]Address: CISAS, via Venezia 15, 35131 Padova, Italy. E-mail: alessio.aboudan@unipd.it

into P different vectors, as follows: first, fix a basic unity of time, in such a way that the whole data set corresponds to the observation of P unities of time, and then define the random vectors $\mathbf{X}^{(p)} := (X_1^{(p)}, \dots, X_{n_p}^{(p)})$, for $p = 1, \dots, P$, whose components are exactly those original X_i's for which $t_i \in (p-1, p]$. These P vectors could also be interpreted as the result of a sampling from P ordered "populations" of "individuals", which are supposed to behave with the same characteristics. Of course, such an assumption could be too restrictive in certain contexts—where, for example, the time variable could induce some form of stochastic dependence, or the probability distributions (p.d.'s) could vary from a population to another—and, consequently, their statistical treatment falls outside the present study. To complete the picture, after denoting by $\mathbb{X}$ the set of all possible realizations of any trial and endowing this set with a suitable σ-algebra $\mathscr{X}$, consider a *regular parametric model* $\{f(\cdot\ ;\ \boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\Theta}$, where Θ is an *open* subset of $\mathbb{R}^d$ and, for every $\boldsymbol{\theta} \in \Theta$, $x \mapsto f(x\ ;\ \boldsymbol{\theta})$ is a probability density function with respect to a σ-finite reference measure ν on $(\mathbb{X}, \mathscr{X})$. The unexplained concept of regular model will be made precise in Section 2. The common p.d. of the $X_j^{(p)}$'s corresponds to the density $f(\cdot\ ;\ \boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0 \in \Theta$ is the true—but unknown—value of the parameter, and the objective is to test the null hypothesis $H_0 : \boldsymbol{\theta}_0 \in \Theta_0$ against the alternative $H_1 : \boldsymbol{\theta}_0 \notin \Theta_0$, Θ_0 being a proper subset of Θ. We now formalize how the testing procedure has to be based on many LRS's, defined in terms of $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(P)}$, by means of the following principles:

A) any LRS has to be formed by *gathering the data included in G consecutive vectors* (i.e. the data observed during G unities of time), the integer G being fixed *a priori*;

B) H_0 is rejected only if *at least k* LRS's are sufficiently small, for a suitably chosen integer k.

The motivation for these two principles rests principally on the purpose to abide by a strict scientific protocol or practical, unavoidable rules. In fact, on the one hand, a prescription like A) is considered to tune the definition of the LRS's with the typical time length of those phenomena which are (supposed to be) induced by H_1. On the other hand, a prescription like B) can be justified whenever it is desirable to have *repeated manifestations* of those phenomena to accept H_1.

A real application can be found in astrophysics, as described, e.g., in [3]. In this context, every variable $X_j^{(p)}$ is associated with the detection of a single photon, the information being relative to its position in the sky map and its energy, in addition to the time of detection. To have an idea of some orders of magnitude, we confine ourselves to considering the specific field of transient emission of γ-ray astrophysical sources, where the scientific operations are focused on the detection of γ-ray flares (see [2, 4]): a representative unity of time according to which the aforesaid splitting could be performed corresponds to one hour (see [2] for technical motivations); G is usually chosen equal to 24 (one day), even if different time scales from 6 hours to some weeks are possible; the duration of a γ-ray flare is variable, and the choice of G is coherent with the fact that the research is focused on the detection of those γ-ray flares of duration of less than G unities of time; $P = N \cdot G$, where N has the size of 10^2 or 10^3 (corresponding to one or some years). The i.i.d. assumption makes sense since the region of the sky under study remains reasonably the same during the whole period of investigation. Now, to better understand and motivate the above setting, it is crucial to underline that a typical data set in this field is composed of an extraordinary huge number of observations (i.e. $n = n_1 + \cdots + n_P$ is very large), but only a very small amount of them is relevant to the investigation (in the sense that the number of photons actually ascribable to a distinguished astrophysical source is much smaller than the total number, the majority of these photons being produced by the *background*). Moreover, the relevant photons are concentrated in brief periods of typical time length of less than G unities of time (the γ-ray flares, indeed). Hence, a global analysis of the entire data set, based on a single LRS, would prove completely meaningless, since the overwhelming majority of the photons produced by the background would lead to always accept H_0. On the contrary, the application of principle A), in conjunction with a reasonable choice of G, ensures that the photons actually emitted by the source under study may be relevant with respect to a subgroup of photons detected during a period of G unities of time. A detailed description of this operational framework can be found in [1, 2], where principles A)-B) are explicitly mentioned in connection with the so-called *post-trial analysis*.

In the following subsections, we illustrate two testing procedures fulfilling principles A)-B). The former procedure (Subsection 1.1) is standard and simpler and has been used till now, being supported by a well-known mathematical framework. Nevertheless, this procedure presents a point of weakness connected with the arbitrary choice of the "origin of the time" and, to clear this hurdle, another practical methodology has been hinted at by some groups of researchers (see [1, 2]). The formal description of this latter procedure (Subsection 1.2) has revealed—at the best of our knowledge—a novelty from a statistical viewpoint. In addition, since the ensuing testing algorithm is no more supported by the current mathematical literature, we have provided the necessary new results in Section 2.

1.1 The standard approach

Letting $P = N{\cdot}G$ and the data be collected in the form of $\mathbf{x}^{(1)} := (x_1^{(1)}, \dots, x_{n_1}^{(1)}), \dots, \mathbf{x}^{(P)} := (x_1^{(P)}, \dots, x_{n_P}^{(P)})$, consider a vector of N different LRS's, say $(\Lambda_1^{(st)}, \dots, \Lambda_N^{(st)})$, where $\Lambda_i^{(st)}$ is obtained by gathering the data belonging to vectors from $(i-1)G+1$ to iG, namely

$$\Lambda_i^{(st)} := \Lambda_i^{(st)}(\mathbf{x}^{((i-1)G+1)}; \dots; \mathbf{x}^{(iG)}) := \frac{\sup_{\boldsymbol{\theta}\in\Theta_0} \prod_{p=(i-1)G+1}^{iG} \prod_{j=1}^{n_p} f(x_j^{(p)}\ ;\ \boldsymbol{\theta})}{\sup_{\boldsymbol{\theta}\in\Theta} \prod_{p=(i-1)G+1}^{iG} \prod_{j=1}^{n_p} f(x_j^{(p)}\ ;\ \boldsymbol{\theta})} \tag{1}$$

for $i = 1, \dots, N$. Notice that, in this framework, these LRS's turn out to be stochastically independent, since the N groups of G vectors just considered are disjoint. Then, reject H_0 if at least k of the $\Lambda_i^{(st)}$'s are less than some reference value α. By independence, the probability of type I error can be evaluated by means of the *binomial formula* as $\pi(k, \alpha) := \sum_{h=k}^{N} \binom{N}{h} p_\alpha^h (1-p_\alpha)^{N-h}$, where the probability p_α that a single $\Lambda_i^{(st)}$ is less than α can be approximated, with sufficiently good precision, by resorting to *Wilks' theorem*. In fact, from this well-known result, one gets the following result: if $\{f(\cdot\ ;\ \boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\Theta}$ is a regular parametric model and Θ_0 is an s-dimensional ($s \in \{1, 2, \dots, d-1\}$) sub-manifold of Θ, then, under H_0, the p.d. of $-2\log\Lambda_i^{(st)}(\mathbf{X}^{((i-1)G+1)}; \dots; \mathbf{X}^{(iG)})$ converges weakly, for every $i = 1, \dots, N$, to a standard χ^2 distribution with $d-s$ degrees of freedom, as $n_1, \dots, n_P$ go to infinity. See [12, 11] for the original formulation of Wilks' theorem, and Chapter 21 of [5], Chapter 22 of [6] or Chapter 5 of [8] for an updated mathematical treatment.

With this approach, the observed number of LRS's which are less than α could be affected by the arbitrary choice of the "origin of time", in connection with principle A). In fact, the standard subdivision of the time interval $(0, P]$ and the ensuing definition of the $\Lambda_i^{(st)}$'s yield only a partial interpretation of the data set: since the validity of H_1 is usually manifested by the observation of remarkable phenomena (the γ-ray flares in astrophysics) of typical duration of less than G unities of time, any single of such phenomena is effectively seized only in the case that both its initial and final instants belong to some time interval $((i-1)G, iG]$. On the contrary, if the initial instant belongs to some interval $((i-1)G, iG]$ and the final instant to $(iG, (i+1)G]$, it could happen that such a manifestation is not seized, both the LRS's $\Lambda_i^{(st)}$ and $\Lambda_{i+1}^{(st)}$ being possibly greater than α. In conclusion, since the final action of accepting/rejecting H_0 is determined by the comparison of a reference value $\overline{k}(\alpha, \pi_0)$—typically the smallest value of k for which $\pi(k, \alpha) \leq \pi_0$—with the actual value of LRS's which are less than α, the aforesaid weakness could distort the decision process. This difficulty is indeed attested in astrophysics, as noted in [1].

1.2 The new approach

Considering all the M groups of G consecutive vectors, the i-th of which consists of those vectors that are numbered from i to $i+G-1$, where $M = P - G + 1 = (N-1)G + 1$. Then, once the data are collected in the form of $\mathbf{x}^{(1)} := (x_1^{(1)}, \dots, x_{n_1}^{(1)}), \dots, \mathbf{x}^{(P)} := (x_1^{(P)}, \dots, x_{n_P}^{(P)})$, as before, associate a LRS with each group, obtaining the vector $\mathbf{\Lambda}^{(new)} := (\Lambda_1^{(new)}, \dots, \Lambda_M^{(new)})$ defined by

$$\Lambda_i^{(new)} := \Lambda_i^{(new)}(\mathbf{x}^{(i)}; \dots; \mathbf{x}^{(i+G-1)}) := \frac{\sup_{\boldsymbol{\theta} \in \Theta_0} \prod_{p=i}^{i+G-1} \prod_{j=1}^{n_p} f(x_j^{(p)} \mid \boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \Theta} \prod_{p=i}^{i+G-1} \prod_{j=1}^{n_p} f(x_j^{(p)} \mid \boldsymbol{\theta})} \tag{2}$$

for $i = 1, \dots, M$. At first, it is worth noticing that the random vector

$$\mathbf{\Lambda}^{(new)}(\mathbf{X}) := \Big(\Lambda_1^{(new)}(\mathbf{X}^{(1)}; \dots; \mathbf{X}^{(G)}), \dots, \Lambda_i^{(new)}(\mathbf{X}^{(i)}; \dots; \mathbf{X}^{(i+G-1)}), \dots, \Lambda_M^{(new)}(\mathbf{X}^{(M)}; \dots; \mathbf{X}^{(P)})\Big)$$

is no more formed by stochastically independent components. Therefore, our main result will deal with the *joint distribution* of $\mathbf{\Lambda}^{(new)}(\mathbf{X})$, by providing its asymptotic behavior for large values of the sample sizes $n_1, \dots, n_P$. The main difference with respect to Wilks' theorem consists of the fact that our result will not provide weak convergence towards a specific limiting distribution, but only a *merging phenomenon*, in the following sense: after fixing a probability distance to compare p.d.'s on $(\mathbb{R}^M, \mathscr{B}(\mathbb{R}^M))$, we will provide an *approximating sequence*—depending on the data only through the sample sizes $n_1, \dots, n_P$ and, above all, being completely free from the model $\{f(\cdot\ ;\ \boldsymbol{\theta})\}_{\boldsymbol{\theta} \in \Theta}$ and the choice of $\boldsymbol{\theta}_0$—such that the aforesaid distance between the p.d. of $\mathbf{\Lambda}^{(new)}(\mathbf{X})$ and the relative element of the approximating sequence goes to zero as $n_1, \dots, n_P$ go to infinity. With such a theoretical result at disposal, we can describe a more robust testing procedure which, even maintaining the basic principles A)-B), is no more affected by the dependency on the "origin of time". The new setting consists in rejecting H_0 whenever there are at least k of the $\Lambda_i^{(new)}$'s, say $\Lambda_{i_1}^{(new)}, \dots, \Lambda_{i_k}^{(new)}$, with $i_{j+1} - i_j \geq G$, which are less than α. More formally, the rejection rule corresponds to considering the event $\cup_{\substack{1 \leq i_1 < \dots < i_k \leq M \\ i_{j+1} - i_j \geq G}} \Big(\{\Lambda_{i_1}^{(new)} < \alpha\} \cap \dots \cap \{\Lambda_{i_k}^{(new)} < \alpha\}\Big)$, whose probability can be evaluated after knowing the joint distribution of $\mathbf{\Lambda}^{(new)}(\mathbf{X})$. Since our main result (Theorem 2 in Section 2) will provide an explicit approximation of this joint distribution, there are now all the elements to carry out the new procedure.

From a mere theoretical viewpoint, there is an evident improvement due to the fact that the new rejection event includes its standard counterpart, entailing that the new test turns out to be *more powerful* than the standard one. In any case, from a practical viewpoint, it is easily seen that the problem of the arbitrary choice of the time origin is now definitely overcome: indeed, it is not possible anymore to neglect a relevant manifestation (of duration of less than G unities of time) starting in the time interval $((i-1)G, iG]$ and ending in $(iG, (i+1)G]$, since there is some other time interval, in the new finer subdivision, which contains both the initial and the final instants.

2 New results

Before formulating the new mathematical results, it is worth recalling the (standard) conditions of regularity for a parametric model $\{f(\cdot\ ;\ \boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\Theta}$, that are:

a) $\forall\ x \in \mathbb{X}$, $\boldsymbol{\theta} \mapsto f(x\ ;\ \boldsymbol{\theta})$ belongs to $C^2(\Theta)$;

b) the set $\mathbb{X}_+ := \{x \in \mathbb{X} \mid f(x\ ;\ \boldsymbol{\theta}) > 0\}$ does not depend on $\boldsymbol{\theta}$ and $\nu(\mathbb{X}_+^c) = 0$;

c) for any measurable function $T : \mathbb{X} \to \mathbb{R}$ satisfying $\int_{\mathbb{X}} T(x) f(x\ ;\ \boldsymbol{\theta})\nu(\mathrm{d}x) < +\infty$ for all $\boldsymbol{\theta} \in \Theta$, derivatives of first and second order (with respect to $\boldsymbol{\theta}$) may be passed under the integral sign in $\int_{\mathbb{X}} T(x) f(x\ ;\ \boldsymbol{\theta})\nu(\mathrm{d}x)$;

d) for any $\boldsymbol{\theta}_0 \in \Theta$, there exist a measurable function $K_0 : \mathbb{X} \to [0, +\infty]$ and $\delta_0 > 0$ such that

$$\begin{aligned} \int_{\mathbb{X}} K_0(x) f(x\ ;\ \boldsymbol{\theta}_0)\nu(\mathrm{d}x) \quad &< \quad +\infty\ , \\ \sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\le\delta_0} \left|\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(x\ ;\ \boldsymbol{\theta})\right| \quad &\le \quad K_0(x) \qquad (\forall\ x \in \mathbb{X}, i, j = 1, \dots, d); \end{aligned}$$

e) the *Fisher information matrix* $\mathrm{I}(\boldsymbol{\theta}) := (\mathrm{I}_{i,j}(\boldsymbol{\theta}))_{i,j=1,\dots,d}$, given by

$$\mathrm{I}_{i,j}(\boldsymbol{\theta}) := -\int_{\mathbb{X}} \left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(x\ ;\ \boldsymbol{\theta})\right) f(x\ ;\ \boldsymbol{\theta})\nu(\mathrm{d}x)\ , \tag{3}$$

is well-defined and positive definite at every value of $\boldsymbol{\theta}$;

f) the model is identified, i.e. $\nu\left(\{x \in \mathbb{X} \mid f(x\ ;\ \boldsymbol{\theta}_1) \neq f(x\ ;\ \boldsymbol{\theta}_2)\}\right) = 0$ entails $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

In addition, we require a *maximum likelihood estimator* (MLE) actually exists as a point of Θ, meaning that such a MLE must coincide with a root of the likelihood equation. More formally, we assume that

g) $\forall\ n \ge n_0$, there exists a measurable function $\mathbf{t}_n : \mathbb{X}^n \to \Theta$ such that

$$\sup_{\boldsymbol{\theta}\in\Theta}\left[\prod_{j=1}^n f(x_j\ ;\ \boldsymbol{\theta})\right] = \prod_{j=1}^n f(x_j\ ;\ \mathbf{t}_n(x_1, \dots, x_n)) \qquad (\forall\ (x_1, \dots, x_n) \in \mathbb{X}^n)\ . \tag{4}$$

The last mathematical tool we need to formalize the concept of approximating sequence is that of probability distance. See, e.g., [7] or Chapter 2 of [10] for a comprehensive treatment of the subject. Taking cognizance that there are many distances to compare two p.d.'s on $(\mathbb{R}^l, \mathscr{B}(\mathbb{R}^l))$, each one emphasizing a peculiar aspect of the discrepancy, we now select a distinguished metric—henceforth denoted by D_l—which is particularly meaningful with respect to our problem. For the sake of definiteness, given two probability measures μ_1 and μ_2 on $(\mathbb{R}^l, \mathscr{B}(\mathbb{R}^l))$, put

$$D_l(\mu_1; \mu_2) := \inf\{\varepsilon > 0 \mid \mu_1(C) \le \mu_2(C^\varepsilon) + \varepsilon, \mu_2(C) \le \mu_1(C^\varepsilon) + \varepsilon,\ \forall\ C \in \mathscr{C}_l\}$$

where $C^\varepsilon := \{x \in \mathbb{R}^l \mid \mathrm{d}(x, C) \le \varepsilon\}$ and $\mathscr{C}_l$ stands for the class of all *convex* subsets of $\mathbb{R}^l$. It should be recalled that D_l is usually called *Lévy-Prokhorov distance*, and it is often used in the context of multidimensional extensions of the Berry-Esseen estimate, being related to the concept of weak convergence of probability measures. See, for example, [9, 10]. In any case, notice that the testing problem is concerned exclusively with probabilities of events expressible as $\{\Lambda_1^{(new)} < \alpha_1, \dots, \Lambda_M^{(new)} < \alpha_M\}$ for some numbers $\alpha_1, \dots, \alpha_M$, which involve only convex regions of $\mathbb{R}^M$.

The way is now paved for the presentation of the first new result, dealing with the *asymptotic normality* of the vector $(\hat{\boldsymbol{\theta}}_{n_1,\dots,n_G}, \dots, \hat{\boldsymbol{\theta}}_{n_M,\dots,n_P})$ of MLE's, whose components are defined by $\hat{\boldsymbol{\theta}}_{n_i,\dots,n_{i+G-1}} := \mathbf{t}_{n_i+\dots+n_{i+G-1}}(X_1^{(i)}, \dots, X_{n_i}^{(i)}; \dots; X_1^{(i+G-1)}, \dots, X_{n_{i+G-1}}^{(i+G-1)})$, for $i = 1, \dots, M$, with the same $\mathbf{t}_n$ as in (4).

Theorem 1. *If the regularity conditions* a)-g) *are valid and* $\boldsymbol{\theta}_0$ *denotes the true value of the parameter, then the joint distribution of the random vector*

$$\left(\sqrt{\sum_{k=1}^{G} n_k} \cdot (\hat{\boldsymbol{\theta}}_{n_1,\dots,n_G} - \boldsymbol{\theta}_0), \dots, \sqrt{\sum_{k=M}^{P} n_k} \cdot (\hat{\boldsymbol{\theta}}_{n_M,\dots,n_P} - \boldsymbol{\theta}_0)\right) ,$$

say $\mu_{n_1,\dots,n_P}^{(dM)}$, *meets*

$$D_{dM}\left(\mu_{n_1,\dots,n_P}^{(dM)}; \gamma^{(dM)}(\mathrm{R}_M, \mathrm{I}(\boldsymbol{\theta}_0))\right) \to 0 \tag{5}$$

as $n_1, \dots, n_P \to +\infty$, *where:*

i) $\mathrm{R}_M := \mathrm{R}_M(n_1, \dots, n_P)$ *is the* $M \times M$ *matrix whose elements* $\rho_{i,j} := \rho_{i,j}(n_1, \dots, n_P)$ *are given by*

$$\rho_{i,j}(n_1, \dots, n_P) := \begin{cases} 0 & \text{if } i, j \in \{1, \dots, M\}, |i-j| \geq G \\ \frac{\sum_{p=a(i,j)}^{b(i,j)} n_p}{\sqrt{\sum_{q=i}^{i+G-1} \sum_{l=j}^{j+G-1} n_q n_l}} & \text{if } i, j \in \{1, \dots, M\}, |i-j| < G \end{cases} \tag{6}$$

with $a(i,j) := \max\{i,j\}$ *and* $b(i,j) := \min\{i,j\} + G - 1$;

ii) $\mathrm{I}(\boldsymbol{\theta}_0)$ *is defined by* (3);

iii) $\gamma^{(dM)}(\mathrm{R}_M, \mathrm{I}(\boldsymbol{\theta}_0))$ *denotes the* dM*-dimensional Gaussian p.d. with zero means and covariance matrix*

$$\left(\begin{array}{c|c|c|c} \rho_{1,1}\mathrm{I}(\boldsymbol{\theta}_0)^{-1} & \rho_{1,2}\mathrm{I}(\boldsymbol{\theta}_0)^{-1} & \dots & \rho_{1,M}\mathrm{I}(\boldsymbol{\theta}_0)^{-1} \\ \hline \rho_{2,1}\mathrm{I}(\boldsymbol{\theta}_0)^{-1} & \rho_{2,2}\mathrm{I}(\boldsymbol{\theta}_0)^{-1} & \dots & \rho_{2,M}\mathrm{I}(\boldsymbol{\theta}_0)^{-1} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline \rho_{M,1}\mathrm{I}(\boldsymbol{\theta}_0)^{-1} & \rho_{M,2}\mathrm{I}(\boldsymbol{\theta}_0)^{-1} & \dots & \rho_{M,M}\mathrm{I}(\boldsymbol{\theta}_0)^{-1} \end{array}\right) . \tag{7}$$

Here, it is worth noticing that the covariance matrices given in (7) are actually non-singular, for any $n_1, \dots, n_P$ in $\mathbb{N}$. Then, as a consequence of Theorem 1, we can state the main result of the paper, which generalizes Wilks' theorem by dealing with the joint distribution, say $\eta_{n_1,\dots,n_P}^{(M)}$, of the random vector

$$\boldsymbol{\Xi}(\mathbf{X}) := \left(\Xi_1(\mathbf{X}^{(1)}; \dots; \mathbf{X}^{(G)}), \dots, \Xi_i(\mathbf{X}^{(i)}; \dots; \mathbf{X}^{(i+G-1)}), \dots, \Xi_M(\mathbf{X}^{(M)}; \dots; \mathbf{X}^{(P)})\right)$$

with $\Xi_i(\mathbf{X}^{(i)}; \dots; \mathbf{X}^{(i+G-1)}) := -2\log(\Lambda_i^{(new)}(\mathbf{X}^{(i)}; \dots; \mathbf{X}^{(i+G-1)}))$.

Theorem 2. *If the regularity conditions* a)-g) *are valid and* Θ_0 *is an* s*-dimensional sub-manifold of* Θ, *with* $s \in \{1, \dots, d-1\}$, *then, under* H_0,

$$D_M\left(\eta_{n_1,\dots,n_P}^{(M)}; \chi^2_{M,r}(\mathrm{R}_M)\right) \to 0 \tag{8}$$

as $n_1, \dots, n_P \to +\infty$, *where:*

i) $r := d - s$;

ii) $\chi^2_{M,r}(\mathrm{R}_M)$ *stands for the p.d. of the M-dimensional random vector*

$$\left(\sum_{h=1}^{r} Z^2_{h;1}, \sum_{h=1}^{r} Z^2_{h;2}, \dots, \sum_{h=1}^{r} Z^2_{h;M}\right) ;$$

iii) *the rM-dimensional random vector $(Z_{1;1}, \dots, Z_{r;1}, Z_{1;2}, \dots, Z_{r;2}, \dots, Z_{1;M}, \dots, Z_{r;M})$ is jointly Gaussian with zero means and covariance matrix given by*

$$\begin{cases} \mathsf{Var}(Z_{h;i}) = 1 & \text{if } h = 1, \dots, r \text{ and } i = 1, \dots, M \\ \mathsf{Cov}(Z_{h;i}, Z_{l;j}) = 0 & \text{if } h \neq l \text{ and } i, j = 1, \dots, M \\ \mathsf{Cov}(Z_{h;i}, Z_{h;j}) = 0 & \text{if } |i-j| \geq G \text{ and } h = 1, \dots, r \\ \mathsf{Cov}(Z_{h;i}, Z_{h;j}) = \rho_{i,j} & \text{if } |i-j| < G \text{ and } h = 1, \dots, r . \end{cases}$$

3 Proofs

Gathered here are the proofs of Theorems 1 and 2. They are obtained under the additional hypothesis that all the $\rho_{i,j}$ are convergent as $n_1, \dots, n_P \to +\infty$. Consequently, since D_l metrizes weak convergence, the merging phenomenon follows as a consequence of the weak convergence towards the respective limiting distributions.

3.1 Proof of Theorem 1

Start by noting that the existence of the MLE's as points of Θ, which is an open set, entails that $\hat{\boldsymbol{\theta}}_{n_i,\dots,n_{i+G-1}}$ can be expressed as a root of the likelihood equation

$$\begin{aligned} \ell'_{n_i,\dots,n_{i+G-1}}(\boldsymbol{\theta}) &:= \nabla_{\boldsymbol{\theta}} \log[L_{n_i,\dots,n_{i+G-1}}(\boldsymbol{\theta}; X_1^{(i)}, \dots X_{n_i}^{(i)}; \dots; X_1^{(i+G-1)}, \dots X_{n_{i+G-1}}^{(i+G-1)})] \\ &= \sum_{p=i}^{i+G-1} \sum_{j=1}^{n_p} \nabla_{\boldsymbol{\theta}} \log[f(X_j^{(p)} \mid \boldsymbol{\theta})] = \mathbf{0} \end{aligned}$$

for every $i = 1, \dots, M$. Under the assumptions of the theorem, it is well-known that these estimators are strongly consistent, that is $\hat{\boldsymbol{\theta}}_{n_i,\dots,n_{i+G-1}} \to \boldsymbol{\theta}_0$, $\mathsf{P}_{\boldsymbol{\theta}_0}$-a.s.. See, for example, the beginning of the proof of Theorem 18 in [6]. Now, expand $\ell'_{n_i,\dots,n_{i+G-1}}$ as

$$\ell'_{n_i,\dots,n_{i+G-1}}(\boldsymbol{\theta}) = \ell'_{n_i,\dots,n_{i+G-1}}(\boldsymbol{\theta}_0) + \int_0^1 \left\{ \sum_{p=i}^{i+G-1} \sum_{j=1}^{n_p} \mathrm{M}(X_j^{(p)}; \boldsymbol{\theta}_0 + u(\boldsymbol{\theta} - \boldsymbol{\theta}_0)) \right\} \mathrm{d}u \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$

where $\mathrm{M}(x; \mathbf{t})$ is the $d \times d$ matrix given by $\left(\frac{\partial^2}{\partial t_k \partial t_h} \log f(x \mid \mathbf{t})\right)_{k,h=1,\dots,d}$. Now, let $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{n_i,\dots,n_{i+G-1}}$, where $\hat{\boldsymbol{\theta}}_{n_i,\dots,n_{i+G-1}}$ can be any root of the likelihood equation, and divide by $\sqrt{\sum_{k=i}^{i+G-1} n_k}$ to obtain

$$\frac{1}{\sqrt{\sum_{k=i}^{i+G-1} n_k}} \ell'_{n_i,\dots,n_{i+G-1}}(\boldsymbol{\theta}_0) = \sqrt{\sum_{k=i}^{i+G-1} n_k} \cdot \mathrm{B}_{n_i,\dots,n_{i+G-1}}(\hat{\boldsymbol{\theta}}_{n_i,\dots,n_{i+G-1}} - \boldsymbol{\theta}_0)$$

with

$$\mathrm{B}_{n_i,\dots,n_{i+G-1}} := -\int_0^1 \frac{1}{\sum_{k=i}^{i+G-1} n_k} \left\{ \sum_{p=i}^{i+G-1} \sum_{j=1}^{n_p} \mathrm{M}(X_j^{(p)}; \boldsymbol{\theta}_0 + u(\hat{\boldsymbol{\theta}}_{n_i,\dots,n_{i+G-1}} - \boldsymbol{\theta}_0)) \right\} \mathrm{d}u .$$

It is well-known that $\mathrm{B}_{n_i,\dots,n_{i+G-1}} \to \mathrm{I}(\boldsymbol{\theta}_0)$, $\mathsf{P}_{\boldsymbol{\theta}_0}$-a.s., as shown, for example, in the final part of the proof of Theorem 18 in [6]. Therefore, the original problem is traced back to the determination of the limiting distribution of the Md-dimensional random vector

$$\mathbf{V}_{n_1,\dots,n_P} := \left(\frac{1}{\sqrt{\sum_{k=1}^{G} n_k}} \ell^{'}_{n_1,\dots,n_G}(\boldsymbol{\theta}_0), \dots, \frac{1}{\sqrt{\sum_{k=M}^{P} n_k}} \ell^{'}_{n_M,\dots,n_P}(\boldsymbol{\theta}_0) \right)$$

where, by definition,

$$\frac{1}{\sqrt{\sum_{k=i}^{i+G-1} n_k}} \ell^{'}_{n_i,\dots,n_{i+G-1}}(\boldsymbol{\theta}_0) = \sqrt{\sum_{k=i}^{i+G-1} n_k} \left(\frac{1}{\sum_{k=i}^{i+G-1} n_k} \sum_{p=i}^{i+G-1} \sum_{j=1}^{n_p} \boldsymbol{\Psi}(X_j^{(p)}; \boldsymbol{\theta}_0) \right)$$

for $i = 1, \dots, M$, with $\boldsymbol{\Psi}(x; \mathbf{t}) := \nabla_{\mathbf{t}} \log[f(X_j^{(p)} \mid \mathbf{t})]$. It is worth noticing, at this stage, that the random vectors $\{\boldsymbol{\Psi}(X_j^{(p)}; \boldsymbol{\theta}_0)\}_{\substack{j=1,\dots,n_p \\ p=1,\dots,P}}$ are i.i.d. under $\mathsf{P}_{\boldsymbol{\theta}_0}$, and it follows from hypothesis ii) on the model that

$$\mathsf{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\Psi}(X_j^{(p)}; \boldsymbol{\theta}_0)] = \mathbf{0} \tag{9}$$

$$\mathsf{Cov}_{\boldsymbol{\theta}_0}(\Psi^{(k)}(X_j^{(p)}; \boldsymbol{\theta}_0), \Psi^{(h)}(X_j^{(p)}; \boldsymbol{\theta}_0)) = \mathrm{I}_{k,h}(\boldsymbol{\theta}_0) \tag{10}$$

where $\Psi^{(k)}(X_j^{(p)}; \boldsymbol{\theta}_0)$ denotes the k^{th} coordinate of $\boldsymbol{\Psi}(X_j^{(p)}; \boldsymbol{\theta}_0)$. For notational convenience, define $\mathbf{S}_p := \sum_{j=1}^{n_p} \boldsymbol{\Psi}(X_j^{(p)}; \boldsymbol{\theta}_0)$ for $p = 1, \dots, P$, and note that they are independent d-dimensional random vector, under $\mathsf{P}_{\boldsymbol{\theta}_0}$. Then, the characteristic function of the random vector $\mathbf{V}_{n_1,\dots,n_P}$ is given by

$$\begin{aligned}
\Phi_{n_1,\dots,n_P}(\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_M) &= \mathsf{E}_{\boldsymbol{\theta}_0}\left[\exp\left\{ \sum_{m=1}^{M} \frac{i\boldsymbol{\xi}_m \bullet \sum_{p=m}^{m+G-1} \mathbf{S}_p}{\sqrt{\sum_{k=m}^{m+G-1} n_k}} \right\} \right] \\
&= \mathsf{E}_{\boldsymbol{\theta}_0}\left[\exp\left\{ \sum_{p=1}^{P} \mathbf{S}_p \bullet \sum_{\substack{m=1,\dots,M \\ m \le p \le m+G-1}} \frac{i\boldsymbol{\xi}_m}{\sqrt{\sum_{k=m}^{m+G-1} n_k}} \right\} \right] \\
&= \prod_{p=1}^{P} \varphi^{n_p} \left(\sum_{\substack{m=1,\dots,M \\ m \le p \le m+G-1}} \frac{\boldsymbol{\xi}_m}{\sqrt{\sum_{k=m}^{m+G-1} n_k}} \right)
\end{aligned}$$

where $\bullet$ stands for the standard scalar product in $\mathbb{R}^d$ and $\varphi(\boldsymbol{\xi}) := \mathsf{E}_{\boldsymbol{\theta}_0}[\exp\{i\boldsymbol{\xi} \bullet \boldsymbol{\Psi}(X_j^{(p)}; \boldsymbol{\theta}_0)\}]$, with $\boldsymbol{\xi} \in \mathbb{R}^d$. As in the standard proof of the multi-dimensional central limit theorem (see, for example, Proposition 5.9 and Lemma 5.10 in [**?**]), one exploits (9)-(10) to deduce

$$\begin{aligned}
&\Phi_{n_1,\dots,n_P}(\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_M) \\
&= \prod_{p=1}^{P} \left[1 - \frac{1}{2} \, {}^t\boldsymbol{\Xi}^{(p)}_{n_1,\dots,n_P}(\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_M) \mathrm{I}(\theta_0) \boldsymbol{\Xi}^{(p)}_{n_1,\dots,n_P}(\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_M) + o(\frac{1}{n_p}) \right]^{n_p}
\end{aligned}$$

with

$$\boldsymbol{\Xi}^{(p)}_{n_1,\dots,n_P}(\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_M) := \sum_{\substack{m=1,\dots,M \\ m \le p \le m+G-1}} \frac{\boldsymbol{\xi}_m}{\sqrt{\sum_{k=m}^{m+G-1} n_k}} \ .$$

Here, $o(\frac{1}{n_p})$ is complex-valued so, by taking the principal branch of the complex logarithm, one has

$$\begin{aligned}
&\mathrm{Log}[\Phi_{n_1,\dots,n_P}(\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_M)] \\
&= -\frac{1}{2} \sum_{p=1}^{P} n_p \, {}^t\boldsymbol{\Xi}^{(p)}_{n_1,\dots,n_P}(\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_M) \mathrm{I}(\theta_0) \boldsymbol{\Xi}^{(p)}_{n_1,\dots,n_P}(\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_M) + o(1) \ .
\end{aligned}$$

The former term in the RHS above is evidently a quadratic form in the $\boldsymbol{\xi}$-variables, which can be written as

$$\sum_{p=1}^{P} n_p \sum_{\substack{m=1,\dots,M \\ m\leq p\leq m+G-1}} \sum_{\substack{l=1,\dots,M \\ l\leq p\leq l+G-1}} \left(\frac{n_p}{\sqrt{\sum_{k=m}^{m+G-1} \sum_{h=l}^{l+G-1} n_k n_h}} \right) {}^t\boldsymbol{\xi}_m \mathrm{I}(\theta_0) \boldsymbol{\xi}_l$$

$$= \sum_{\substack{m,l=1,\dots,M \\ |m-l|<G}} \left(\frac{\sum_{p=a(l,m)}^{b(l,m)} n_p}{\sqrt{\sum_{k=m}^{m+G-1} \sum_{h=l}^{l+G-1} n_k n_h}} \right) {}^t\boldsymbol{\xi}_m \mathrm{I}(\theta_0) \boldsymbol{\xi}_l$$

where $a(l,m) := \max\{l,m\}$ and $b(l,m) := \min\{l,m\} + G - 1$. At this stage, taking the limit as $n_1, \dots, n_P \to +\infty$ of the above expression, one gets

$$\lim_{n_1,\dots,n_P\to+\infty} \mathrm{Log}[\Phi_{n_1,\dots,n_P}(\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_M)] = -\frac{1}{2} \sum_{\substack{m,l=1,\dots,M \\ |m-l|<G}} \rho_{l,m} \, {}^t\boldsymbol{\xi}_m \mathrm{I}(\theta_0) \boldsymbol{\xi}_l$$

for every fixed $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_M \in \mathbb{R}^d$. This fact, in view of the Lévy continuity theorem, amounts to proving that the probability distribution of $\mathbf{V}_{n_1,\dots,n_P}$, evaluated under $\mathsf{P}_{\boldsymbol{\theta}_0}$, converges weakly to the Md-dimensional normal distribution with zero means and covariance matrix given by

$$\left(\begin{array}{c|c|c|c} \rho_{1,1}\mathrm{I}(\boldsymbol{\theta}_0) & \rho_{1,2}\mathrm{I}(\boldsymbol{\theta}_0) & \dots & \rho_{1,M}\mathrm{I}(\boldsymbol{\theta}_0) \\ \hline \rho_{2,1}\mathrm{I}(\boldsymbol{\theta}_0) & \rho_{2,2}\mathrm{I}(\boldsymbol{\theta}_0) & \dots & \rho_{2,M}\mathrm{I}(\boldsymbol{\theta}_0) \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline \rho_{M,1}\mathrm{I}(\boldsymbol{\theta}_0) & \rho_{M,2}\mathrm{I}(\boldsymbol{\theta}_0) & \dots & \rho_{M,M}\mathrm{I}(\boldsymbol{\theta}_0) \end{array} \right) . \tag{11}$$

Therefore, upon observing that

$$\mathbf{V}_{n_1,\dots,n_P} =$$

$$\left(\sqrt{\sum_{k=1}^{G} n_k} \cdot \mathrm{B}_{n_1,\dots,n_G}(\hat{\boldsymbol{\theta}}_{n_1,\dots,n_G} - \boldsymbol{\theta}_0), \dots, \sqrt{\sum_{k=M}^{P} n_k} \cdot \mathrm{B}_{n_M,\dots,n_P}(\hat{\boldsymbol{\theta}}_{n_M,\dots,n_P} - \boldsymbol{\theta}_0) \right) ,$$

and that $\mathrm{B}_{n_i,\dots,n_{i+G-1}} \to \mathrm{I}(\boldsymbol{\theta}_0)$, $\mathsf{P}_{\boldsymbol{\theta}_0}$-a.s., the desired conclusion now follows, via an obvious application of the Slutsky theorem, from the above achievement on the limiting distribution of $\mathbf{V}_{n_1,\dots,n_P}$, thanks to the elementary property of normal distributions.

3.2 Proof of Theorem 2

The argumentation developed at the beginning of the proof of Theorem 22 in [6] shows that the limiting distribution of $(-2\log[\tilde{\lambda}_{n_1,\dots,n_G}], \dots, -2\log[\tilde{\lambda}_{n_M,\dots,n_P}])$, evaluated under $\mathsf{P}_{\boldsymbol{\theta}_0}$, is the same as the limiting distribution of the M-dimensional random vector $\mathbf{W}_{n_1,\dots,n_P}$ whose components are given by

$$\frac{1}{\sqrt{\sum_{k=i}^{i+G-1} n_k}} \, {}^t\ell'_{n_i,\dots,n_{i+G-1}}(\boldsymbol{\theta}^*_{n_i,\dots,n_{i+G-1}}) \mathrm{I}(\boldsymbol{\theta}_0)^{-1} \frac{1}{\sqrt{\sum_{k=i}^{i+G-1} n_k}} \ell'_{n_i,\dots,n_{i+G-1}}(\boldsymbol{\theta}^*_{n_i,\dots,n_{i+G-1}})$$

for $i = 1, \dots, M$, $\boldsymbol{\theta}^*_{n_i,\dots,n_{i+G-1}}$ denoting the MLE over Θ_0 based on the observations $\{X_j^{(p)}\}_{\substack{j=1,\dots,n_p \\ p=i,\dots,i+G-1}}$. But, exactly as in the central part of the above-mentioned proof from [6], the limiting distribution of the random vector

$$\left(\frac{1}{\sqrt{\sum_{k=1}^{G} n_k}} \ell'_{n_1,\dots,n_G}(\boldsymbol{\theta}^*_{n_1,\dots,n_G}), \dots, \frac{1}{\sqrt{\sum_{k=M}^{P} n_k}} \ell'_{n_M,\dots,n_P}(\boldsymbol{\theta}^*_{n_M,\dots,n_P}) \right)$$

turns out to be the same as the limiting distribution of the random vector with components given by

$$[\mathrm{Id}_{d\times d} - \mathrm{I}(\boldsymbol{\theta}_0)\mathrm{H}(\boldsymbol{\theta}_0)]\frac{1}{\sqrt{\sum_{k=i}^{i+G-1} n_k}}\ell'_{n_i,\dots,n_{i+G-1}}(\boldsymbol{\theta}_0)$$

for $i = 1, \dots, M$, where $\mathrm{H}(\boldsymbol{\theta}_0)$ is a $d \times d$ matrix defined as follows. Partition $\mathrm{I}(\boldsymbol{\theta}_0)$ into four matrices in such a way that

$$\mathrm{I}(\boldsymbol{\theta}_0) = \left(\begin{array}{c|c} \mathrm{G}_1(\boldsymbol{\theta}_0) & \mathrm{G}_2(\boldsymbol{\theta}_0) \\ \hline {}^t\mathrm{G}_2(\boldsymbol{\theta}_0) & \mathrm{G}_3(\boldsymbol{\theta}_0) \end{array}\right)$$

holds with $\mathrm{G}_1(\boldsymbol{\theta}_0)$ of dimension $r \times r$, $\mathrm{G}_2(\boldsymbol{\theta}_0)$ of dimension $r \times (d-r)$ and $\mathrm{G}_3(\boldsymbol{\theta}_0)$ of dimension $(d-r)\times(d-r)$, and let

$$\mathrm{H}(\boldsymbol{\theta}_0) := \left(\begin{array}{c|c} \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \mathrm{G}_3(\boldsymbol{\theta}_0)^{-1} \end{array}\right) .$$

At this stage, it is possible to exploit the form of the limiting distribution of the random vector $\mathbf{V}_{n_1,\dots,n_P}$ deduced in the proof of Theorem 1. Indeed, an application of the continuous mapping theorem entails that the limiting distribution of $(-2\log[\tilde{\lambda}_{n_1,\dots,n_G}], \dots, -2\log[\tilde{\lambda}_{n_M,\dots,n_P}])$ coincides with the probability law of the random vector with component given by

$${}^t\mathbf{Y}_i \; {}^t[\mathrm{Id}_{d\times d} - \mathrm{I}(\boldsymbol{\theta}_0)\mathrm{H}(\boldsymbol{\theta}_0)] \; \mathrm{I}(\boldsymbol{\theta}_0)^{-1} \; [\mathrm{Id}_{d\times d} - \mathrm{I}(\boldsymbol{\theta}_0)\mathrm{H}(\boldsymbol{\theta}_0)] \; \mathbf{Y}_i$$

for $i = 1, \dots, M$, where $({}^t\mathbf{Y}_1, \dots, {}^t\mathbf{Y}_M)$ is any Md-dimensional random vector having normal distribution with zero means and covariance matrix (11). Elementary linear algebra shows that

$${}^t[\mathrm{Id}_{d\times d} - \mathrm{I}(\boldsymbol{\theta}_0)\mathrm{H}(\boldsymbol{\theta}_0)]\mathrm{I}(\boldsymbol{\theta}_0)^{-1}[\mathrm{Id}_{d\times d} - \mathrm{I}(\boldsymbol{\theta}_0)\mathrm{H}(\boldsymbol{\theta}_0)] = \mathrm{I}(\boldsymbol{\theta}_0)^{-1} - \mathrm{H}(\boldsymbol{\theta}_0)$$

because $\mathrm{H}(\boldsymbol{\theta}_0)\mathrm{I}(\boldsymbol{\theta}_0)\mathrm{H}(\boldsymbol{\theta}_0) = \mathrm{H}(\boldsymbol{\theta}_0)$. To conclude, introduce the the Md-dimensional random vector $({}^t\mathbf{Z}_1, \dots, {}^t\mathbf{Z}_M)$, defined by putting $\mathbf{Z}_i := \mathrm{I}(\boldsymbol{\theta}_0)^{-1/2}\mathbf{Y}_i$, whose distribution is normal with zero means and covariance matrix equal to

$$\left(\begin{array}{c|c|c|c} \rho_{1,1}\mathrm{Id}_{d\times d} & \rho_{1,2}\mathrm{Id}_{d\times d} & \dots & \rho_{1,M}\mathrm{Id}_{d\times d} \\ \hline \rho_{2,1}\mathrm{Id}_{d\times d} & \rho_{2,2}\mathrm{Id}_{d\times d} & \dots & \rho_{2,M}\mathrm{Id}_{d\times d} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline \rho_{M,1}\mathrm{Id}_{d\times d} & \rho_{M,2}\mathrm{Id}_{d\times d} & \dots & \rho_{M,M}\mathrm{Id}_{d\times d} \end{array}\right) .$$

Upon noticing that

$$\mathrm{I}(\boldsymbol{\theta}_0)^{1/2} \; [\mathrm{I}(\boldsymbol{\theta}_0)^{-1} - \mathrm{H}(\boldsymbol{\theta}_0)] \; \mathrm{I}(\boldsymbol{\theta}_0)^{1/2} = \mathrm{P}_{d,r} := \left(\begin{array}{c|c} \mathrm{Id}_{r\times r} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right)$$

holds, the theorem is completely proved by putting

$$(\xi_{1;i}, \dots, \xi_{r;i}) := \mathrm{P}_{d,r}\mathbf{Z}_i$$

for $i = 1, \dots, M$.

References

[1] Bulgarelli, A. et al. (2012). Evaluating the maximum likelihood method for detecting short-term variability of AGILE γ-ray sources. *A&A* **540** A79

[2] Bulgarelli, A. et al. (2014). The AGILE alert system for γ-ray transients. *The Astrophysical Journal* **781** 19

[3] Cash, W. (1979). Paremeter estimation in astronomy through application of the likelihood ratio. *The Astrophysical Journal* **228** 939-947

[4] Chiang, J. (2012). Automated Science Processing for the Fermi Large Area Telescope. *Advances in Machine Learning and Data Mining for Astronomy*, CRC Press, Taylor & Francis Group, Eds.: Michael J. Way, Jeffrey D. Scargle, Kamal M. Ali, Ashok N. Srivastava, p. 41-54

[5] DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability.* Springer, New York.

[6] Ferguson, T.S. (2002). *A Course in Large Sample Theory.* Chapman and Hall/CRC, Boca Raton.

[7] Gibbs, A.L. and Su, F.E. (2002). On choosing and bounding probability metrics. *International Statistical Review* **70** 419-435

[8] Sen, P.K. and Singer, J.M. (1993). *Large Sample Methods in Statistics: An Introduction with Applications.* Chapman and Hall, New York.

[9] Senatov, V.V. (1984). An estimate of the Levy-Prokhorov distance. *Theory Prob. Appl.* **29** 109-114.

[10] Senatov, V.V. (1998). *Normal approximation: new results, methods and problems.* Modern Probability and Statistics. VSP, Utrecht.

[11] Wald, A. (1943). Test of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.* **54** 426-482

[12] Wilks, S.S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.* **9** 60-62