



Aggregated Tests Based on Supremal Divergence Estimators for Non-regular Statistical Models

Jean-Patrick Baudry^{1,3(✉)}, Michel Broniatowski¹, and Cyril Thommeret^{1,2}

¹ Sorbonne Université and Université Paris Cité, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, 75005 Paris, France

jean-patrick.baudry@sorbonne-universite.fr

² Safran Group, Paris, France

³ 4 place Jussieu, 75005 Paris, France

Abstract. A methodology is proposed to build statistical test procedures pertaining to models with incomplete information; the lack of information corresponds to a nuisance parameter in the description of the model. The supremal approach based on the dual representation of CASM divergences (or f -divergences) is fruitful; it leads to M-estimators with simple and standard limit distribution, and it is versatile with respect to the choice of the divergence. Duality approaches to divergence-based optimisation are widely considered in statistics, data analysis and machine learning; indeed, they avoid any smoothing or grouping technique which would be necessary for a more direct divergence minimisation approach for the same problem.

We are interested in a widely considered but still open problem which consists in testing the number of components in a parametric mixture. Although common, this is still a challenging problem since the corresponding model is non-regular particularly because of the true parameter lying on the boundary of the parameter space. This range of problems has been considered by many authors who tried to derive the asymptotic distribution of some statistic under boundary conditions. The present approach based on supremal divergence M-estimators makes the true parameter an interior point of the parameter space, providing a simple solution for a difficult question. To build a composite test, we aggregate simple tests.

Keywords: Non-regular models · Dual form of f -divergences · Statistical test aggregation · Number of components in mixture models

1 Dual Representation of the φ -Divergences and Tests

We consider CASM divergences (see [15] for definitions and properties):

$$D_\varphi(Q, P) = \begin{cases} \int \varphi\left(\frac{dQ}{dP}\right)dP & \text{if } Q \ll P \\ +\infty & \text{otherwise} \end{cases}$$

where Q and P are probability measures on the same probability space. Extensions to divergences between probability measures and signed measures can be found in

[22]. Dual formulations of divergences can be found in [7, 16]. Another interpretation of these formulations can be found in [9, Section 4.6]. They are widely considered in statistics, data analysis and machine learning (see e.g. [4, 20]).

As in [1], let \mathcal{F} be some class of \mathcal{B} -measurable (borelian) real valued functions and let $\mathcal{M}_{\mathcal{F}} = \{P \in \mathcal{M} : \int |f|dP < \infty, \forall f \in \mathcal{F}\}$ where \mathcal{M} is the space of probability measures. Let any $P^* \in \mathcal{M}$, which shall be the underlying true unknown probability law in a statistical context in the following sections. Assume that φ is differentiable and strictly convex. Then, for all $P \in \mathcal{M}_{\mathcal{F}}$ such that $D_{\varphi}(P, P^*)$ is finite and $\varphi'(dP/dP^*)$ belongs to \mathcal{F} , D_{φ} admits the dual representation (see Theorem 4.4 in [6]):

$$D_{\varphi}(P, P^*) = \sup_{f \in \mathcal{F}} \int f dP - \int \varphi^{\#}(f) dP^*, \tag{1}$$

where $\varphi^{\#}(x) = \sup_{t \in \mathbb{R}} tx - \varphi(t)$ is the Fenchel-Legendre convex conjugate. Moreover, the supremum is uniquely attained at $f = \varphi'(dP/dP^*)$.

This result can be used in two directions. First, a statistical model, e.g. a parametrical model $\{P_{\theta} : \theta \in \Theta\}$ with P_{θ} is absolutely continuous with respect to some dominating measure μ for any θ , naturally induces a family $\mathcal{F} = \{\varphi'(p_{\theta}/p_{\theta'}) : \theta, \theta' \in \Theta\}$. This is the main framework of this paper.

Conversely, a class of functions \mathcal{F} defines the distribution pairs P and Q that can be compared, which are these such that $\varphi'(dP/dQ) \in \mathcal{F}$. Furthermore it induces a divergence D_{φ} on these pairs. A typical example is the logistic model.

The KLm divergence is defined by the generator $\varphi : x \in \mathbb{R} \mapsto -\log x + x - 1$ and leads to the maximum likelihood estimator for both forms of estimation for the supremal estimator, once of which is defined bellow (see Remark 3.2 in [7]).

We consider in this paper the problem of testing the number of components in a mixture model. This question has been considered by various authors. [2, 10, 12, 14, 17] have considered likelihood ratio tests and showed some difficulties with those due to the fact that the likelihood ratio statistic is unbounded with respect to n . [17] prove that its distribution is driven by a $\log \log n$ term in a specific simple Gaussian mixture model. The test statistic needs to be calibrated in accordance with this result. But first, as stated by [17], the convergence to the limit distribution is extremely slow, making this result unpractical. And second, it seems very difficult to derive the corresponding term for a different model, and even more so for a general situation.

Our approach to this problem is suggested by the dual representation of the divergence. For the KLm divergence, it amounts to considering the maximum likelihood estimator itself as a test statistic instead of the usual maximum value of the likelihood function. This leads to a well-defined limiting distribution for the test statistic under the null. This holds for a class of estimators obtained by substituting KLm by any regular divergence. This approach also eliminates the curse of irregularity encountered by many authors for the problem of testing the number of components in a mixture.

Since we are interested in composite hypotheses, there is no justification in this context that the likelihood ratio test would be the best (in terms of uniform

power) as is usually considered (e.g. [8, 18]) and [13] showed what difficulties likelihood ratio tests can encounter in this context.

[8] considered tests based on an estimation of the minimum divergence between the true distribution and the null model. In we make use of the unicity of the optimiser of the dual representation of the divergence in (1) and of the supremal divergence estimator introduced by [7]. An immediate practical advantage of this choice as compared to estimating the minimum divergence is that one less optimisation is needed. Moreover [23] showed that this estimator is robust for several choices of the divergence.

Our procedure for composite hypotheses consists in the aggregation of simple tests in the spirit of [11]. [5] used a similar aggregation procedure for testing between two distributions under noisy data and obtained some control of the resulting test power.

2 Notation and Hypotheses

Let $\{f_1(\cdot; \theta_1) : \theta_1 \in \Theta_1\}$, $\Theta_1 \subset \mathbb{R}^p$, and $\{f_2(\cdot; \theta_2) : \theta_2 \in \Theta_2\}$, $\Theta_2 \subset \mathbb{R}^q$, be probability density families with respect to a σ -finite measure λ on $(\mathcal{X}, \mathcal{B})$. For some fixed open interval $]a, b[\ni 0$, let $\Theta \subset]a, b[\times \Theta_1 \times \Theta_2$, and

$$g_{\pi, \theta} = (1 - \pi)f_1(\cdot; \theta_1) + \pi f_2(\cdot; \theta_2)$$

for any $(\pi, \theta) \in \Theta$ with $\theta = (\theta_1, \theta_2)$.

Assume that $x_1, \dots, x_n \in \mathbb{R}$ have been observed and they are modelled as a realisation of the i.i.d. sample X_1, \dots, X_n which distribution $\mathbb{P}^* := g_{\pi^*, \theta^*} \cdot \lambda$ is known up to the parameters $(\pi^*, \theta^*) \in \Theta$. Our aim is to test the hypothesis $H_0 : \pi^* = 0$.

Assume that $g_{\pi, \theta} = g_{\pi^*, \theta^*} \Rightarrow \pi = \pi^*, \theta_1 = \theta_1^*$ and, if $\pi^* \neq 0$, $\theta_2 = \theta_2^*$.

Let g be a probability density with respect to λ such that $Supp(g) \subset Supp(g_{\pi, \theta})$ for any $(\pi, \theta) \in \Theta$ such that

$$\forall (\pi, \theta) \in \Theta, \int \left| \varphi' \left(\frac{g}{g_{\pi, \theta}} \right) \right| g d\lambda < \infty.$$

Let us define for any $(\pi, \theta) \in \Theta$,

$$m_{\pi, \theta} : x \in \mathcal{X} \mapsto \int \varphi' \left(\frac{g}{g_{\pi, \theta}} \right) g d\lambda - \varphi^\# \left(\frac{g}{g_{\pi, \theta}} \right) (x)$$

and assume that $(\pi, \theta) \mapsto m_{\pi, \theta}(x)$ is continuous for any $x \in \mathcal{X}$. Let us also assume that

$$\forall (\tilde{\pi}, \tilde{\theta}) \in \Theta, \exists r_0 > 0 / \forall r \leq r_0, P^* \left| \sup_{d((\tilde{\pi}, \tilde{\theta}), (\pi, \theta)) < r} m_{\pi, \theta} \right| < \infty$$

where $d(\cdot, \cdot)$ denotes the Euclidean distance and where, as usual, the operator-type notation $\mathbb{P}^* Y$ denotes the expectation—with respect to the probability measure \mathbb{P}^* —of the random variable Y .

Theorem 1. For any $(\pi^*, \theta^*) \in \Theta$

$$D_\varphi(g.\lambda, g_{\pi^*, \theta^*}.\lambda) = \sup_{(\pi, \theta) \in \Theta} P^* m_{\pi, \theta},$$

which we call the supremal form of the divergence. Moreover attainment holds uniquely at $(\pi, \theta) = (\pi^*, \theta^*)$.

Definition 1. Let \mathbb{P}_n denote the empirical measure pertaining to the sample X_1, \dots, X_n . Define

$$(\hat{\pi}, \hat{\theta}) := \arg \max_{(\pi, \theta)} \mathbb{P}_n m_{\pi, \theta}$$

the supremal estimator of (π^*, θ^*) .

The existence of $(\hat{\pi}, \hat{\theta})$ can be guaranteed by assuming that Θ is compact. When uniqueness does not hold, consider any maximizer. This class of estimators has been introduced in [7], under the name dual φ -divergence estimators.

3 Consistency of the Supremal Divergence Estimator

Let us first state the consistency of the supremal divergence estimator of the proportion and the parameters of the existing component, when the non-existing component parameters are fixed, uniformly over the latter.

Here and below, by abuse of notation, we let $\varphi'(\frac{g}{g_{\pi, \theta}})$ stand for $x \mapsto \varphi'(\frac{g(x)}{g_{\pi, \theta}(x)})$, and so on.

Remark that, for $\pi^* = 0$ and any $\theta_1^* \in \Theta_1$ and $\theta_2 \in \Theta_2$, we can unambiguously write m_{π^*, θ_1^*} for $m_{\pi^*, \theta_1^*, \theta_2}$ since the parameter θ_2 is not involved in the expression of $m_{0, \theta_1^*, \theta_2}$.

Theorem 2. Assume that $\pi^* = 0$ and let for any $\theta_2 \in \Theta_2$, $(\hat{\pi}(\theta_2), \hat{\theta}_1(\theta_2)) \in]a, b[\times \Theta_1$ such that

$$\inf_{\theta_2 \in \Theta_2} P_n m_{\hat{\pi}(\theta_2), \hat{\theta}_1(\theta_2), \theta_2} \geq P_n m_{\pi^*, \theta_1^*} - o_{P^*}(1). \tag{2}$$

Then

$$\sup_{\theta_2 \in \Theta_2} d((\hat{\pi}(\theta_2), \hat{\theta}_1(\theta_2)), (0, \theta_1^*)) \xrightarrow[n \rightarrow \infty]{P^*} 0.$$

The convergence holds a.s. in the particular case of (2) when, a.s.,

$$\forall \theta_2 \in \Theta_2, (\hat{\pi}(\theta_2), \hat{\theta}_1(\theta_2)) \in \operatorname{argmax}_{(\pi, \theta_1) \in]a, b[\times \Theta_1} P_n m_{\pi, \theta_1, \theta_2}. \tag{3}$$

4 Asymptotic Distribution of the Supremal Divergence Estimator

Under H_0 ($\pi^* = 0$), the joint asymptotic distribution of $(\hat{\pi}(\theta_2), \hat{\pi}(\theta'_2))$ is provided by the following theorem. The interior of Θ will be denoted by $\overset{\circ}{\Theta}$.

Theorem 3. *Let $\theta_2 \in \Theta_2$ and $\theta'_2 \in \Theta_2$ such that $(\pi^*, \theta_1^*, \theta_2) \in \overset{\circ}{\Theta}$ and $(\pi^*, \theta_1^*, \theta'_2) \in \overset{\circ}{\Theta}$. Write*

$$\begin{aligned}\Theta(\theta_2) &= \{(\pi, \theta_1) \in]-\infty, 1[\times \Theta_1 : (\pi, \theta_1, \theta_2) \in \Theta\} \\ \Theta(\theta'_2) &= \{(\pi, \theta_1) \in]-\infty, 1[\times \Theta_1 : (\pi, \theta_1, \theta'_2) \in \Theta\},\end{aligned}$$

and let $(\hat{\pi}, \hat{\theta}_1)$ and $(\hat{\pi}', \hat{\theta}'_1)$ be such that

$$\begin{aligned}(\hat{\pi}, \hat{\theta}_1) &\in \operatorname{argmax}_{(\pi, \theta_1) \in \Theta(\theta_2)} P_n m_{\pi, \theta_1, \theta_2} \\ (\hat{\pi}', \hat{\theta}'_1) &\in \operatorname{argmax}_{(\pi, \theta_1) \in \Theta(\theta'_2)} P_n m_{\pi, \theta_1, \theta'_2}.\end{aligned}$$

Assume that $\pi^* = 0$.

Moreover, assume that :

- $(\pi, \theta_1) \in \Theta(\theta_2) \mapsto m_{\pi, \theta_1, \theta_2}(x)$ (resp. $(\pi, \theta_1) \in \Theta(\theta'_2) \mapsto m_{\pi, \theta_1, \theta'_2}(x)$) is differentiable λ -a.e. with derivative $\psi_{\pi, \theta_1} = \left(\frac{\partial}{\partial \pi} m_{\pi, \theta_1, \theta_2} \right)_{|(\pi, \theta_1)}$ (resp. $\psi'_{\pi, \theta_1} = \left(\frac{\partial}{\partial \pi} m_{\pi, \theta_1, \theta'_2} \right)_{|(\pi, \theta_1)}$) such that $P^* \psi_{\pi^*, \theta_1^*} = 0$ (resp. $P^* \psi'_{\pi^*, \theta_1^*} = 0$).
- $(\pi, \theta_1) \in \Theta(\theta_2) \mapsto P^* \psi_{\pi, \theta_1}$ (resp. $(\pi, \theta_1) \in \Theta(\theta'_2) \mapsto P^* \psi'_{\pi, \theta_1}$) is differentiable at π^*, θ_1^* with invertible derivative matrix $H = D(P^* \psi)_{|(\pi^*, \theta_1^*)}$ (resp. $H' = D(P^* \psi')_{|(\pi^*, \theta_1^*)}$).
- $\{\psi_{\pi, \theta_1} : (\pi, \theta_1) \in \Theta(\theta_2)\}$ and $\{\psi'_{\pi, \theta_1} : (\pi, \theta_1) \in \Theta(\theta'_2)\}$ are P^* -Donsker.
- $\int (\psi_{\hat{\pi}, \hat{\theta}_1}(x) - \psi_{\pi^*, \theta_1^*}(x))^2 dP^*(x) \xrightarrow{P^*} 0$ and $\int (\psi'_{\hat{\pi}, \hat{\theta}_1}(x) - \psi'_{\pi^*, \theta_1^*}(x))^2 dP^*(x) \xrightarrow{P^*} 0$.

Assume that $H = D(P^* \psi)_{|(\pi^*, \theta_1^*)} = P^* D^2(h)_{|(\pi^*, \theta_1^*)}$ (resp. $H' = D(P^* \psi')_{|(\pi^*, \theta_1^*)} = P^* D^2(h')_{|(\pi^*, \theta_1^*)}$) with $P^* |D^2(h)_{|(\pi^*, \theta_1^*)}| < \infty$ (resp. $P^* |D^2(h')_{|(\pi^*, \theta_1^*)}| < \infty$) where $h : (\pi, \theta_1) \in \Theta(\theta_2) \mapsto m_{\pi, \theta_1, \theta_2}(x)$ (resp. $h' : (\pi, \theta_1) \in \Theta(\theta'_2) \mapsto m_{\pi, \theta_1, \theta'_2}(x)$).

Then with a_n (resp. a'_n) being the $(1, 1)$ -entry of the matrix $H_n^{-1} \cdot (P_n \psi_{\hat{\pi}, \hat{\theta}_1} \psi_{\hat{\pi}, \hat{\theta}_1}^T) \cdot H_n^{-1}$ (resp. of $H_n'^{-1} \cdot (P_n \psi'_{\hat{\pi}, \hat{\theta}_1} \psi'_{\hat{\pi}, \hat{\theta}_1}{}^T) \cdot H_n'^{-1}$)—where H_n (resp. H'_n) denotes the Hessian matrix of $(\pi, \theta_1) \mapsto P_n m_{\pi, \theta_1, \theta_2}$ (resp. of $(\pi, \theta_1) \mapsto P_n m_{\pi, \theta_1, \theta'_2}$) at the point $(\hat{\pi}, \hat{\theta}_1)$, which is supposed to be invertible with high probability—one gets

$$\begin{pmatrix} \sqrt{\frac{n}{a_n}} (\hat{\pi} - \pi^*) \\ \sqrt{\frac{n}{a'_n}} (\hat{\pi}' - \pi^*) \end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}(0, U)$$

with

$$U = \begin{pmatrix} 1 & \frac{b}{\sqrt{aa'}} \\ \frac{b}{\sqrt{aa'}} & 1 \end{pmatrix} \tag{4}$$

where b is the $(1, 1)$ -entry of the matrix $H^{-1} \cdot (\mathbb{P}\psi_{\pi^*, \theta_1^*} \psi_{\pi^*, \theta_1^*}'^T) \cdot H'^{-1}$, and a (resp. a') the $(1, 1)$ -entry of the matrix $H^{-1} \cdot (\mathbb{P}\psi_{\pi^*, \theta_1^*} \psi_{\pi^*, \theta_1^*}'^T) \cdot H^{-1}$ (resp. of $H'^{-1} \cdot (\mathbb{P}\psi_{\pi^*, \theta_1^*} \psi_{\pi^*, \theta_1^*}'^T) \cdot H'^{-1}$).

This result naturally generalises to k -tuples. The marginal result for θ actually also holds when $\pi^* > 0$ and $\theta_2 = \theta_2^*$, which is useful to control the power of the test procedure to be defined.

Let us consider as a test statistic $T_n = \sup_{\theta_2} \sqrt{\frac{n}{a_n}} \hat{\pi}$ and let us reject H_0 when T_n takes large values. It seems sensible to reduce $\hat{\pi}$ for each value of θ_2 so that, under H_0 , it is asymptotically distributed as a $\mathcal{N}(0, 1)$ and that the (reduced) values of $\hat{\pi}$ for different values of θ_2 can be compared. In practice, the asymptotic variance has to be estimated hence the substitution of $\hat{\pi}$, $\hat{\theta}_1$, and P_n for π^* , θ_1^* , and P^* in $H^{-1}P^*\psi_{\pi^*, \theta_1^*} \psi_{\pi^*, \theta_1^*}'^T H^{-1}$. This choice is justified in [19].

The Bonferoni aggregation rule is not sensible here since the tests for different values of θ_2 are obviously not independent so that such a procedure would lead to a conservative test. Hence the need in Theorem 3 for the joint asymptotic distribution to take the dependence between $\hat{\pi}$ for different values of θ_2 . This leads to the study of the asymptotic distribution of T_n which should be the distribution of $\sup W$ where W is a Gaussian process which covariance structure is given by Theorem 3. This will be proved in the forthcoming section.

5 Asymptotic Distribution of the Supremum of Supremal Divergence Estimators

H_0 is assumed to hold in this section.

It is stated that the asymptotic distribution of T_n is that of the supremum of a Gaussian process with the covariance $\frac{b}{\sqrt{aa'}}$, as in (4).

Then it is stated that the distribution of the latter can be approximated by maximising the Gaussian process with the covariance $\frac{b_n}{\sqrt{a_n a_n'}}$, where a_n , a_n' , and b_n are estimations of the corresponding quantities, on a finite grid of values for θ_2 .

Let X be the centred Gaussian process over Θ_2 with

$$\forall \theta_2, \theta_2' \in \Theta_2, r(\theta_2, \theta_2') = \text{Cov}(X_{\theta_2}, X_{\theta_2'}) = \frac{b}{\sqrt{aa'}}$$

where a and b are defined in Theorem 3.

Theorem 4. *Under general regularity conditions pertaining to the class of derivatives of m (Glivenko-Cantelli classes), we have*

$$\sqrt{\frac{n}{a_n}} (\hat{\pi}_n - \pi^*) \xrightarrow{\mathcal{L}} X.$$

This results from [21] when $\dim(\Theta_2) = 1$ and [24] when $\dim(\Theta_2) > 1$.

Theorem 5. *Under the same general regularity conditions as above, we have*

$$T_n \xrightarrow{\mathcal{L}} \sup_{\theta_2 \in \Theta_2} X(\theta_2).$$

The proof of the last result when $\dim(\Theta_2) = 1$ makes use of the fact that $\theta_2 \mapsto \hat{\pi}(\theta_2)$ is cadlag ([21]). This is a reasonable assumption, which holds in the examples which we considered. We are eager for counter-examples! When $\dim(\Theta_2) > 1$, the result holds also by [3].

Let now X^n be the centred Gaussian process over Θ_2 with

$$\forall \theta_2, \theta'_2 \in \Theta_2, \text{Cov}(X^n_{\theta_2}, X^n_{\theta'_2}) = \frac{b_n}{\sqrt{a_n a'_n}}$$

where a_n, a'_n are defined in Theorem 3 and b_n is defined analogously.

Theorem 6. *Let, for any $\delta > 0$, Θ_2^δ be a finite set such that $\forall \theta_2 \in \Theta_2, \exists \tilde{\theta}_2 \in \Theta_2^\delta / \|\theta_2 - \tilde{\theta}_2\| \leq \delta$. Then*

$$M_n^\delta = \sup_{\theta_2 \in \Theta_2^\delta} X^n_{\theta_2} \xrightarrow[\delta \rightarrow 0]{n \rightarrow \infty} M = \sup_{\theta_2 \in \Theta_2} X_{\theta_2}.$$

6 Algorithm

Our algorithm for testing that the data was sampled from a single-component mixture ($H_0 : \pi^* = 0$) against a two-component mixture ($H_1 : \pi^* > 0$) is presented in Algorithm 1.

In this algorithm, $\hat{\pi}(\theta_2)$ is defined in (2) and (3). It depends on g . This Theorems hold as long as g fulfils $\text{Supp}(g) \subset \text{Supp}(g_{\pi, \theta})$ for any $(\pi, \theta) \in \Theta$. However it has to be chosen with care. The constants in the asymptotic distribution in Theorem 3 depend on it. Moreover [23] argue that the choice of g can influence the robustness properties of the procedure.

The choice of φ is also obviously crucial (see also [23] for the induced robustness properties).

The choice of φ and g are important practical questions which are work in progress.

As already stated, the supremal estimator for the modified Kullback-Leibler divergence $\varphi : x \in \mathbb{R}^{+*} \mapsto -\log x + x - 1$ is the usual maximum likelihood estimator. In this instance the estimator does not depend on g .

Algorithm 1: Test of H_0 : one component vs H_1 : two components

Input : $\varphi, \{f_1(\cdot; \theta_1) : \theta_1 \in \Theta_1\}, \{f_2(\cdot; \theta_2) : \theta_2 \in \Theta_2\}, n, K, \Theta_2^\delta, p \in [0, 1]$

1. let $t = \sup_{\theta_2 \in \Theta_2} \sqrt{\frac{n}{a_n(\theta_2)}} \hat{\pi}(\theta_2)$
 2. for $k \in \{1, \dots, K\}$
 - (a) sample $(X_t)_{t \in \Theta_2^\delta} \sim \mathcal{N}\left(0, \left(\frac{b_n(t, t')}{\sqrt{a_n(t)a_n(t')}}\right)_{t, t' \in \Theta_2^\delta}\right)$
 - (b) let $\tilde{t}_k = \max_{t \in \Theta_2^\delta} x_t$
 3. if $t \geq \text{empirical_quantile}((\tilde{t}_k)_{k \in \{1, \dots, K\}}, 1 - p)$ reject H_0 else don't reject H_0 vs H_1 .
-

Acknowledgements. The authors gratefully acknowledge the reviewers for their valuable comments and suggestions which helped improving the article.

References

1. Al Mohamad, D.: Towards a better understanding of the dual representation of phi divergences. *Stat. Pap.* **59**(3), 1205–1253 (2016). <https://doi.org/10.1007/s00362-016-0812-5>
2. Bickel, P.J., Chernoff, H.: Asymptotic distribution of the likelihood ratio statistic in a prototypical non regular problem. In: *Statistics and Probability: A Raghu Raj Bahadur Festschrift*, pp. 83–96 (1993)
3. Bickel, P.J., Wichura, M.J.: Convergence criteria for multiparameter stochastic processes and some applications. *Ann. Math. Stat.* **42**(5), 1656–1670 (1971)
4. Birrell, J., Dupuis, P., Katsoulakis, M.A., Pantazis, Y., Rey-Bellet, L.: (f, γ)-divergences: interpolating between f-divergences and integral probability metrics. *J. Mach. Learn. Res.* **23**(1), 1816–1885 (2022)
5. Broniatowski, M., Jurečková, J., Moses, A.K., Miranda, E.: Composite tests under corrupted data. *Entropy* **21**(1), 63 (2019)
6. Broniatowski, M., Keziou, A.: Minimization of φ -divergences on sets of signed measures. *Studia Scientiarum Mathematicarum Hungarica* **43**(4), 403–442 (2006)
7. Broniatowski, M., Keziou, A.: Parametric estimation and tests through divergences and the duality technique. *J. Multivar. Anal.* **100**(1), 16–36 (2009)
8. Broniatowski, M., Miranda, E., Stummer, W.: Testing the number and the nature of the components in a mixture distribution. In: Nielsen, F., Barbaresco, F. (eds.) *GSI 2019. LNCS*, vol. 11712, pp. 309–318. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-26980-7_32
9. Broniatowski, M., Stummer, W.: Some universal insights on divergences for statistics, machine learning and artificial intelligence. In: *Geometric Structures of Information*, pp. 149–211 (2019)
10. Feng, Z., McCulloch, C.E.: Statistical inference using maximum likelihood estimation and the generalized likelihood ratio when the true parameter is on the boundary of the parameter space. *Stat. Probabil. Lett.* **13**(4), 325–332 (1992)
11. Garel, B.: Asymptotic theory of the likelihood ratio test for the identification of a mixture. *J. Stat. Plan. Infer.* **131**(2), 271–296 (2005)
12. Ghosh, J.K., Sen, P.K.: On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. Technical report, North Carolina State University. Department of Statistics (1984)

13. Hall, P., Stewart, M.: Theoretical analysis of power in a two-component normal mixture model. *J. Stat. Plan. Infer.* **134**(1), 158–179 (2005)
14. Hartigan, J.A.: Statistical theory in clustering. *J. Classif.* **2**(1), 63–76 (1985)
15. Liese, F., Vajda, I.: *Convex statistical distances*, vol. 95. Teubner (1987)
16. Liese, F., Vajda, I.: On divergences and informations in statistics and information theory. *IEEE Trans. Inf. Theory* **52**(10), 4394–4412 (2006)
17. Liu, X., Shao, Y.: Asymptotics for the likelihood ratio test in a two-component normal mixture model. *J. Stat. Plan. Infer.* **123**(1), 61–81 (2004)
18. Lo, Y., Mendell, N.R., Rubin, D.B.: Testing the number of components in a normal mixture. *Biometrika* **88**(3), 767–778 (2001)
19. McLachlan, G., Peel, D.: *Finite Mixture Models*. Wiley, New York (2000)
20. Nguyen, X., Wainwright, M.J., Jordan, M.I.: Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Trans. Inf. Theory* **56**(11), 5847–5861 (2010)
21. Pollard, D.: *Convergence of Stochastic Processes*. Springer, Heidelberg (2012). <https://doi.org/10.1007/978-1-4612-5254-2>
22. Rüschendorf, L.: On the minimum discrimination information theorem. In: *Statistical Decisions*, pp. 263–283 (1984)
23. Toma, A., Broniatowski, M.: Dual divergence estimators and tests: robustness results. *J. Multivar. Anal.* **102**(1), 20–36 (2011)
24. van der Vaart, A., Wellner, J.A.: *Weak Convergence and Empirical Processes*. Springer, Heidelberg (1996). <https://doi.org/10.1007/978-1-4757-2545-2>