TESTING FOR HOMOGENEITY IN MIXTURE MODELS

JIAYING GU, ROGER KOENKER, AND STANISLAV VOLGUSHEV

Abstract. Statistical models of unobserved heterogeneity are typically formalized as mixtures of simple parametric models and interest naturally focuses on testing for homogeneity versus general mixture alternatives. Many tests of this type can be interpreted as $C(\alpha)$ tests, as in Neyman (1959), and shown to be locally, asymptotically optimal. These $C(\alpha)$ tests will be contrasted with a new approach to likelihood ratio testing for general mixture models. The latter tests are based on estimation of general nonparametric mixing distribution with the Kiefer and Wolfowitz (1956) maximum likelihood estimator. Recent developments in convex optimization have dramatically improved upon earlier EM methods for computation of these estimators, and recent results on the large sample behavior of likelihood ratios involving such estimators yield a tractable form of asymptotic inference. Improvement in computation efficiency also facilitates the use of a bootstrap method to determine critical values that are shown to work better than the asymptotic critical values in finite samples. Consistency of the bootstrap procedure is also formally established. We compare performance of the two approaches identifying circumstances in which each is preferred.

1. Introduction

Given a simple parametric density model, $p(x|\mu)$, for iid observations, $X_1, \cdots, X_n$, there is a natural temptation to complicate the model by allowing the parameter, μ, to vary with the observation index. In the absence of other, e.g. observable covariate, information that would distinguish the observations from one another it may be justifiable to view the μ's as drawn at random. Inference for such mixture models is complicated by the enormous class of potential alternatives. Two dominant approaches to testing for homogeneity in such models exist: Neyman's $C(\alpha)$ tests and likelihood ratio tests. $C(\alpha)$ tests are particularly attractive for testing homogeneity since like their kindred score tests they do not require estimation of the model under the alternative of heterogeneity of the parameter μ. As described in Gu (2016), $C(\alpha)$ tests have a somewhat irregular, but still relatively simple asymptotic theory, and are generally easy to compute. Likelihood ratio tests, in contrast, are known to have a considerably more complicated limiting behavior, and are generally regarded as much more difficult to compute. Our primary objective here is to try to rehabilitate the reputation of the likelihood ratio test (hereafter LRT) for testing homogeneity in mixture

Version: April 25, 2017. This research was partially supported by NSF grant SES-11-53548 and Project C1 of the SFB 823 of the German Research Foundation. Part of this research was conducted while the first author was visiting the Mathematics department at Ruhr University Bochum and the third author was a visiting scholar at UIUC. They are very grateful to the UIUC Statistics and Economics departments and the Bochum Mathematics department for their hospitality. The third author also gratefully acknowledges Financial support from the DFG (grant VO1799/1-1). The authors would also like to express their appreciation to the Co-Editor and the referees for comments that led to improvements in the paper.

models by demonstrating that it is both computationally tractable and – at least under some conditions – that it has attractive power and size control properties when compared to other tests.

We will argue that recent developments in convex optimization have dramatically reduced the computational burden of the LRT approach for general, nonparametric alternatives. Following Laird (1978), prior efforts to compute the Kiefer-Wolfowitz nonparametric MLE for general nonparametric mixture models have relied upon some variant of the EM algorithm. However, Koenker and Mizera (2014) have recently shown that interior point methods for general convex optimization provide a much more efficient, and more accurate computational approach. A second impediment to the use of LRT methods for general mixture problems has been the lack of a tractable limiting distribution theory. Extending recent work of Gassiat (2002), Liu and Shao (2003) and Azaïs, Gassiat, and Mercadier (2009) we propose an easily simulated method of computing limiting critical values for the LRT statistic for testing homogeneity for Gaussian mixture models. However, we find in simulations that these limiting critical values do not serve as a good approximation in moderate samples. Instead we propose a parametric bootstrap method to determine critical values, and formally prove its consistency. Size and power performance of the bootstrap method is investigated through simulations.

There is a large and rapidly growing literature on inference for *finite* mixture models using penalized likelihood ratio methods, which can be considered an intermediate approach between $C(\alpha)$ tests and our general LRT approach based on the Kiefer-Wolfowitz nonparametric MLE. Ironically, once one restricts mixtures to discrete distributions with a finite number of support points, convexity of the log likelihood is lost, making LRT methods considerably more challenging from a computational point of view. Moreover, finite mixture models fail to satisfy certain regularity conditions that are typically required for parametric likelihood ratio tests, making their asymptotic theory challenging, see for example Cho and White (2007) and Chen, Ponomareva, and Tamer (2014). Motivated by these challenges, Chen, Chen, and Kalbfleisch (2001) have proposed penalizing the log likelihood with a log barrier penalty on the mixing weights. The penalty removes the singularity in the log likelihood that arises when mixing weights tend to zero, and leads to a relatively simple mixture of χ^2 limiting theory for the restricted LRT statistic. More recently, Chen and Li (2009), Li, Chen, and Marriott (2009) and Li and Chen (2010) have extended this approach and developed an attractive inference apparatus for restricted mixture models based on these penalized likelihood ratio methods. Kasahara and Shimotsu (2014) further extend the EM test methods to normal mixture regression models. We will incorporate these EM tests into our performance comparisons in the simulation section of the paper.

The next section provides a detailed discussion of our general approach to likelihood ratio testing based on the Kiefer-Wolfowitz nonparametric MLE (NPMLE). The following two sections briefly describe the $C(\alpha)$ and EM testing approaches. Simulation evidence on the performance of the various methods and an empirical example is reported in Section 5 and 6.

2. Likelihood Ratio Tests for Homogeneity in Mixture Models

A prerequisite for any likelihood ratio test for general mixture models must be a reliable maximum likelihood estimator for these models under the alternative of parameter heterogeneity. Lindsay (1995) offers a comprehensive overview of the vast literature on mixture models, and traces the idea of maximum likelihood estimation of a *nonparametric* mixing measure η, given random samples from the mixture density,

$$g(x) = \int p(x|\mu) d\eta(\mu), \tag{1}$$

to an *Annals* abstract of Robbins (1950). Somewhat later Kiefer and Wolfowitz (1956) provided a detailed analysis of such a NPMLE and established its consistency. Yet only with Laird (1978) did a viable computational strategy emerge for a discretized version. The EM method proposed by Laird has been employed extensively in subsequent work, notably by Heckman and Singer (1984) and Jiang and Zhang (2009), even though it has been widely criticized for its slow convergence. Recently, Koenker and Mizera (2014) have shown that the discretized version of the Kiefer-Wolfowitz estimator can be formulated as a convex optimization problem and accurately solved very efficiently by interior point methods. Recent work by Gassiat (2002) and Azaïs, Gassiat, and Mercadier (2009) has also clarified the limiting behavior of the LRT for general classes of alternatives, and taken together these developments offer a fresh opportunity to explore the viability of the LRT for inference on mixtures.

It seems ironic that many of the difficulties inherent in maximum likelihood estimation of finite parameter mixture models vanish when we consider nonparametric mixtures. The notorious multimodality of parametric likelihood surfaces is replaced by a much simpler, strictly convex optimization problem possessing a unique solution. It is of obvious concern that consideration of such a wide class of alternatives may depress the power of associated tests; we will see that while there is some loss of power when compared to more restricted parametric LRTs, the loss is typically modest, a small price to pay for power gained against a broader class of alternatives. We will also see that by comparison with $C(\alpha)$ tests that are also designed to detect general alternatives the LRT can be competitive.

2.1. Maximum Likelihood Estimation of General Mixtures.

Suppose that we have iid observations, $X_1, \cdots, X_n$ from the mixture density (1), the Kiefer-Wolfowitz NPMLE requires us to solve,

$$\min_{\eta \in \bar{\mathcal{G}}} \Big\{ -\sum_{i=1}^{n} \log g(x_i) \Big| g(x_i) = \int p(x_i|\mu) d\eta(\mu) \Big\},$$

where $\bar{\mathcal{G}}$ is the (convex) set of all mixing distributions. The problem is one of minimizing the sum of strictly convex functions subject to linear equality and inequality constraints. The dual to this (primal) convex program proves to be somewhat more tractable from a computational viewpoint, and takes the form,

$$\max_{\nu \in \mathbb{R}^n} \Big\{ \sum_{i=1}^{n} \log \nu_i \;\Big|\; \sum_{i=1}^{n} \nu_i p(x_i|\mu) \le n, \quad \text{for all } \mu \Big\}$$

See Lindsay (1983) and Koenker and Mizera (2014) for further details. This variational form of the problem may still seem rather abstract since it appears – even in the dual – that we need to check an infinite number of values of μ, for each choice of the vector, ν. However, it suffices in applications to consider a fine grid of values $\{\mu_1, \cdots, \mu_m\}$ and write the primal problem as

$$\min_{f \in \mathbb{R}^m, g \in \mathbb{R}^n} \Big\{ -\sum_{i=1}^{n} \log(g_i) \;\Big|\; Af = g, f \in \mathcal{S} \Big\}$$

where A is an n by m matrix with elements $p(x_i|\mu_j)$ and $\mathcal{S} = \{s \in \mathbb{R}^m | 1_m^\top s = 1,\ s \geq 0\}$ is the unit simplex. Thus, $\hat{f}_j$ denotes the estimated mixing density evaluated at the grid point, μ_j and $\hat{g}_i$ denotes the estimated mixture density evaluated at x_i. The dual problem in this discrete formulation becomes,

$$\max_{\nu \in \mathbb{R}^n} \Big\{ \sum_{i=1}^{n} \log \nu_i \;\Big|\; A^\top \nu \leq n 1_m, \quad \nu \geq 0 \Big\}.$$

Primal and dual solutions are immediately recoverable from the solution to either problem. Interior point methods such as those provided by PDCO of Saunders (2003) and Mosek of Andersen (2010), are capable of solving dual formulations of typical problems with $n = 200$ and $m = 300$ in less than one second. The empirical Bayes package `REBayes`, Koenker (2013), is available for download from the R repository CRAN. It is based on the `RMosek` package of Friberg (2012), and was used for all of the computations reported below. We have compared this approach with other proposals including those of Lesperance and Kalbfleisch (1992) and Groeneboom, Jongbloed, and Wellner (2008), but thus far have found nothing competitive in terms of speed and accuracy.

Solutions to the NPMLE problem of Kiefer and Wolfowitz produce estimates of the mixing measure, η, that are discrete and possess only a few mass points. A theoretical upper bound on the number of these atoms of η was established already by Lindsay (1983), but in practice the number is typically observed to be far fewer. It may seem surprising, perhaps even disturbing, that even when the true mixing distribution has a smooth density, the NPMLE of that density is discrete with only a few atoms. However, this may appear less worrying if we consider a more explicit example. Suppose that we have a location mixture of Gaussians,

$$g(x) = \int \phi(x - \mu) d\eta(\mu),$$

so we are firmly in the deconvolution business, a harsh environment notorious for its poor convergence rates. One interpretation of this is that good approximations of the mixture density g can be achieved by relatively simple discrete mixtures with only a few atoms. For many applications estimation of g is known to be sufficient: this is quite explicit for example for empirical Bayes compound decision problems where the Bayes rules are known to depend entirely on the estimated $\hat{g}$. See e.g. Efron (2011). Of course given our discrete formulation of the Kiefer-Wolfowitz problem, we can only identify the location of atoms up to the scale of the grid spacing, but we believe that the $m \approx 300$ grid points we have been using in the simulations reported below are probably adequate for most applications. For testing this assertion is reinforced by the fact that finer grids, when employed, exert a

negligible impact on the LRT statistic. Recently, Dicker and Zhao (2014) have shown that with $m = \sqrt{n}$, the Hellinger distance between $\hat{g}$ and g is bounded by $\mathcal{O}_p(\log n/\sqrt{n})$.

Given a reliable maximum likelihood estimator for the general nonparametric mixture model it is of obvious interest to know whether an effective likelihood ratio testing strategy can be developed. This question has received considerable prior attention, again Lindsay (1995) provides an authoritative overview of this literature. However, more recently work by Gassiat (2002) and Azaïs, Gassiat, and Mercadier (2009) has revealed new features of the asymptotic behavior of the likelihood ratio for mixture settings that enable one to derive asymptotic critical values for the LRT.

2.2. **Asymptotic Theory of Likelihood Ratios for General Mixtures.** Consider a parametric family of distributions that have density $p(\cdot|\mu)$ with respect to some sigma-finite measure λ and parameters from the parameter set $\Theta \subset \mathbb{R}^d$. Typically, λ is the Lebesgue measure if the data follow a continuous distribution and the counting measure if their distribution is discrete. Our aim is to test whether the i.i.d. sample $X_1, ..., X_n$ was generated from a $p(\cdot|\mu_0)$ for some $\mu_0 \in \Theta$ against the general alternative that $X_1, ..., X_n$ is generated from a mixture of the form $p_\eta(\cdot) := \int_\Theta p(\cdot|\mu)d\eta(\mu)$ for some non-degenerate distribution η on Θ (non-degenerate in the sense that η is not a one-point distribution). In order for this testing problem to make sense, we need the following mild identifiability assumption

(A0) For any probability measure η on Θ, for any $\mu_0 \in \Theta$ we have $\eta \neq \delta(\mu_0)$ (denoting by $\delta(\mu)$ the Dirac measure at the point μ) implies $\mathbb{E}[(p_\eta(X_1) - p(X_1|\mu_0))^2] > 0$.

Consider the following sets of distributions on Θ

$$\bar{\mathcal{G}} := \{\eta | \eta \text{ distribution on } \Theta, \}, \quad \mathcal{G} := \bar{\mathcal{G}} \backslash \delta(\mu_0).$$

Define the log-likelihood function corresponding to the measure η as

$$\ell_n(\eta) := \sum_{i=1}^n \log p_\eta(X_i).$$

The likelihood ratio test statistic is given by

$$L_n := \sup_{\eta \in \bar{\mathcal{G}}} \ell_n(\eta) - \sup_{\mu \in \Theta} \ell_n(\delta(\mu)).$$

To derive the asymptotic distribution of the likelihood ratio under the null, assume that the data are generated from a measure with density $p(\cdot|\mu_0)$ for some $\mu_0 \in \Theta$. Consider the decomposition

$$L_n = \sup_{\eta \in \bar{\mathcal{G}}} \ell_n(\eta) - \ell_n(\delta(\mu_0)) + \ell_n(\delta(\mu_0)) - \sup_{\mu \in \Theta} \ell_n(\delta(\mu)).$$

The second term in this decomposition can be handled by classical parametric theory. Under suitable regularity conditions we obtain [see, for instance, the proof of Theorem 16.7 in van der Vaart (1998)]

$$\sup_{\mu \in \Theta} \ell_n(\delta(\mu)) - \ell_n(\delta(\mu_0)) = \frac{1}{2} \Big\| \frac{1}{\sqrt{n}} \sum_{i=1}^n I(\mu_0)^{-1/2} \ell'(X_i|\mu_0) \Big\|^2 + o_P(1) \tag{2}$$

with $\ell'(X_i|\mu) := \nabla_\mu \log p_{\delta(\mu)}(X_i)$, and $I(\mu_0) = \mathbb{E}[\ell'(X_i|\mu_0)\ell'(X_i|\mu_0)^\top]$ being the Fisher information. Handling the first part in the decomposition is more challenging. Expansions for this term were derived in (Gassiat 2002, Liu and Shao 2003, Azaïs, Gassiat, and Mercadier 2009) under various sets of conditions. For the sake of a simple presentation we will follow Gassiat (2002). For $\eta \in \bar{\mathcal{G}}, \mu \in \Theta, \eta \neq \delta(\mu)$ let

$$s_{\eta,\mu}(x) := \Big(\frac{p_\eta(x)}{p_{\delta(\mu)}(x)} - 1\Big) \Big/ \Big\| \frac{p_\eta}{p_{\delta(\mu)}} - 1 \Big\|_{2,\delta(\mu)} \tag{3}$$

where we defined $\|f\|_{2,\eta} := (\int\int f^2(x)p(x|\mu)d\eta(\mu)d\lambda(x))^{1/2}$. For $\eta \in \mathcal{G}$ define

$$\mathbb{G}_n(\eta) := n^{-1/2} \sum_{i=1}^n s_{\eta,\mu_0}(X_i)$$

and note that by construction $\mathbb{E}[s_{\eta,\mu_0}(X_i)] = 0, \mathbb{E}[s^2_{\eta,\mu_0}(X_i)] = 1$. Now a slight modification of the proof of Theorem 3.1 in Gassiat (2002) leads to the following result for the asymptotic behavior of the LRT - for the sake of completeness a sketch of the proof is provided in the Appendix.

Theorem 2.1. *Assume $X_1, ..., X_n$ are generated from $p(\cdot|\mu_0)$, that (A0) holds and that $\mathbb{G}_n \rightsquigarrow \mathbb{G}$ in $\ell^\infty(\mathcal{G})$ for a centered Gaussian process $\mathbb{G}$. Then*

$$2\left(\sup_{\eta\in\bar{\mathcal{G}}} \ell_n(\eta) - \ell_n(\delta(\mu_0))\right) = \sup_{\eta\in\mathcal{G}} \Big(\max\Big\{\mathbb{G}_n(\eta), 0\Big\}\Big)^2 + o_P(1). \tag{4}$$

If additionally (2) holds and $\ell'(X_1|\mu_0)$ is square integrable,

$$2L_n \rightsquigarrow \sup_{\eta\in\mathcal{G}} \Big(\max\Big\{\mathbb{G}(\eta), 0\Big\}\Big)^2 - \|Y\|^2.$$

Here, $Y \sim \mathcal{N}(0, I_d)$ and $(\mathbb{G}, Y)$ is jointly centered normal with covariance taking the form $\mathbb{E}[\mathbb{G}(\eta)Y] = \mathbb{E}[s_{\eta,\mu_0}(X_1)I(\mu_0)^{-1/2}\ell'(X_1|\mu_0)], \mathrm{Cov}(\mathbb{G}(\zeta), \mathbb{G}(\eta)) = \mathbb{E}[s_{\zeta,\mu_0}(X_1)s_{\eta,\mu_0}(X_1)]$. Here, by jointly normal we mean that for any collection $\eta_1, ..., \eta_k \in \mathcal{G}$ the vector $(\mathbb{G}(\eta_1), ..., \mathbb{G}(\eta_k), Y)$ follows a centered multivariate normal distribution with the covariance described above.

Remark 2.2. We now provide a more detailed discussion of the assumption $\mathbb{G}_n \rightsquigarrow \mathbb{G}$ in $\ell^\infty(\mathcal{G})$ made in Theorem 2.1 and the corresponding limiting Gaussian process $\mathbb{G}$. In general, since $X_1, ..., X_n$ are iid, the multivariate central limit theorem implies that $\big(\mathbb{G}_n(\eta_1), ..., \mathbb{G}_n(\eta_k)\big)$ converges in distribution to a centered multivariate normal with covariance structure $\Sigma_{i,j} = \mathbb{E}[s_{\eta_i,\mu_0}(X_1)s_{\eta_j,\mu_0}(X_1)]$. In this sense, using score functions provides the most canonical way to describe the limiting process $\mathbb{G}$.

Without some additional information it seems difficult to provide a representation for $\Sigma_{i,j}$ which does not make use of score functions. However, in special cases a series expansion for $\mathrm{Cov}(\mathbb{G}(\eta_i), \mathbb{G}(\eta_j))$ can be derived. The key is to find an alternative representation for the score function $s_{\eta,\mu}$.

For illustration purposes, consider the location mixture of normal distributions with fixed variance $\sigma^2 = 1$. Assume that $\Theta = [L, U]$ for some $-\infty < L < 0 < U < +\infty$ and that the densities p take the form $p(x|\mu) = (2\pi)^{-1/2}\exp(-(x-\mu)^2/2)$ with respect to Lebesgue

measure. Without loss of generality we will assume that $\mu_0 = 0$. In this case, following the discussion in Section 3.2 of Azaïs, Gassiat, and Mercadier (2009), the likelihood ratio $\frac{p_\eta(x)}{p_{\delta(0)}(x)}$ admits the following representation

$$\frac{p_\eta(x)}{p_{\delta(0)}(x)} - 1 = \int \frac{\exp(-(x-\mu)^2/2)}{\exp(-x^2/2)} d\eta(\mu) - 1 = \int \sum_{k=1}^{\infty} \frac{\mu^k}{k!} H_k(x) d\eta(\mu)$$

where

$$H_k(x) := (-1)^k \exp(x^2/2) \frac{d^k \exp(-x^2/2)}{dx^k}$$

denote the Hermite polynomials and the series $\sum_{k=1}^{\infty} \frac{\mu^k}{k!} H_k(x)$ converges absolutely for any fixed x, μ. Now the key insight is that $\{H_k\}_{k\geq 1}$ are orthogonal polynomials with respect to $p(x|0)$ with norm given by $\mathbb{E}[H_k^2(X_1)] = k!$ for $k \geq 1$. Using dominated convergence, it is not difficult to show that integration and summation can be interchanged (recall that η has compact support), and thus

$$\frac{p_\eta(x)}{p_{\delta(0)}(x)} - 1 = \sum_{k=1}^{\infty} \frac{\int \mu^k d\eta(\mu)}{k!} H_k(x) \tag{5}$$

In other words, we have represented the likelihood ratio $\frac{p_\eta(x)}{p_{\delta(0)}(x)}$ through a series expansion in terms of $\{H_k\}_{k\geq 1}$ where the coefficients in the expansion now depend on moments of the measure η but the function $H_k(x)$ doesn't depend on η. With this representation, computing the L_2 norm with respect to $p_{\delta(0)}$ is straightforward and we obtain

$$s_{\eta,0}(x) = \frac{\sum_{k=1}^{\infty} \frac{\int \mu^k d\eta(\mu)}{k!} H_k(x)}{\big(\sum_{k=1}^{\infty} \frac{[\int \mu^k d\eta(\mu)]^2}{k!}\big)^{1/2}} =: \sum_{k=1}^{\infty} w_k(\eta) H_k(x).$$

which is again a weighted series in $\{H_k\}_{k\geq 1}$ with coefficients depending on η. This representation has several advantages. Most importantly, using general results from empirical process theory on Donsker properties of function classes given by series expansions as outlined in Chapter 2.13 of van der Vaart and Wellner (1996), one can obtain an explicit condition on the collection of weights $\{w_k(\eta) : \eta \in \mathcal{G}\}$ which ensures that $\mathbb{G}_n \rightsquigarrow \mathbb{G}$ as required by Theorem 2.1. This also gives a simple representation for the limiting process $\mathbb{G}$ through

$$\mathbb{G}(\eta) = \sum_{k=1}^{\infty} w_k(\eta) \Big(\int H_k(x)^2 p(x|0) dx \Big)^{1/2} Y_k \tag{6}$$

where $\{Y_k\}_{k\geq 1}$ are iid $\mathcal{N}(0,1)$. This representation is useful for computing asymptotic critical values and will be utilized in Section 2.3

Moreover, the orthogonality of $\{H_k\}_{k\geq 1}$ with respect to $p(x|0)$ allows for an alternative representation for the covariance structure of $\mathbb{G}$ since

$$\mathbb{E}[s_{\eta,0}(X) s_{\zeta,0}(X)] = \sum_{k=1}^{\infty} w_k(\eta) w_k(\zeta) \int H_k^2(x) p(x|0) dx = \sum_{k=1}^{\infty} k! w_k(\eta) w_k(\zeta).$$

In the Gaussian example this can be used to show that for two measure η_1, η_2

$$\mathbb{E}[\mathbb{G}(\eta_1)\mathbb{G}(\eta_2)] = \frac{\mathbb{E}[\exp(Z_1 Z_2)] - 1}{(\mathbb{E}[\exp(Z_1\tilde{Z}_1)] - 1)^{1/2}(\mathbb{E}[\exp(Z_2\tilde{Z}_2)] - 1)^{1/2}}$$

where $Z_1, \tilde{Z}_1 \sim \eta_1, Z_2, \tilde{Z}_2 \sim \eta_2$ and $Z_1, Z_2, \tilde{Z}_1, \tilde{Z}_2$ are independent.

More generally, the discussion above applies to any setting where likelihood ratios can be expanded in the form

$$\frac{p_\eta(x)}{p_{\delta(\mu_0)}(x)} - 1 = \sum_{k=1}^{\infty} v_k(\eta, \mu_0) g_k(x, \mu_0)$$

where the functions $\{g_k(\cdot, \mu_0)\}_{k\geq 1}$ are orthogonal with respect to $p(x|\mu_0)$ and the series converges in a suitable sense. For instance, for Poisson mixtures such an expansion is given in Section 3.3 of Azaïs, Gassiat, and Mercadier (2009). It takes the form

$$\frac{p_\eta(x)}{p_{\delta(\mu_0)}(x)} - 1 = \sum_{k=1}^{\infty} \frac{\int (\mu - \mu_0)^k d\eta(\mu)}{(k!\mu_0^k)^{1/2}} C_k(x|\mu_0),$$

where the functions $x \mapsto C_k(x|\mu_0)$ are orthonormal polynomials with respect to the Poisson measure with parameter μ_0, i.e. for $X_1 \sim Poisson(\mu_0)$

$$\mathbb{E}[C_k(X_1|\mu_0)] = 0, \quad \mathbb{E}[C_k(X_1|\mu_0)C_\ell(X_1|\mu_0)] = I\{k = \ell\},$$

see also Section B of the Appendix for additional technical details. Another example of mixtures with such properties are given by Binomial distributions where the expansion is in fact a finite sum (see Azaïs, Gassiat, and Mercadier (2009) Section, 3.4).

2.3. **Asymptotic Critical Values.** In order to apply the above limiting result in practice, we need to know how to obtain critical values from the asymptotic distribution. For illustrative purposes, we consider the following normal mixture example.

Example 2.3. Consider mixtures of $\mathcal{N}(\mu, 1)$ distributions and assume that $\Theta = [L, U]$ with $0 \in \Theta$. Computations in Azaïs, Gassiat, and Mercadier (2009) show that the asymptotic distribution of the LRT statistic L_n under the null of $X_i \sim \mathcal{N}(0, 1)$ i.i.d. is given by

$$D = \Big(\sup_{\eta\in\mathcal{G}} (V_\eta)_+\Big)^2 - Y_1^2$$

where $(V_\eta)_{\eta\in\mathcal{G}}$ is the Gaussian process given by

$$V_\eta := \Big(\sum_{k=1}^{\infty} \frac{Y_k \kappa_k(\eta)}{(k!)^{1/2}}\Big) \Big/ \Big(\sum_{k=1}^{\infty} \frac{\kappa_k^2(\eta)}{k!}\Big)^{1/2}$$

with $Y_1, Y_2, ...$ denoting i.i.d. $\mathcal{N}(0, 1)$ distributed random variables, $\kappa_k(\eta) := \int_\Theta \mu^k d\eta(\mu)$ and x_+ denoting the positive part of x.

There exists a simpler expression for the distribution of D. More precisely, we will demonstrate that

$$D \stackrel{\mathcal{D}}{=} \sup_{\eta \in \mathcal{G}} \Big(\Big(\Big(\sum_{k=2}^{\infty} \frac{Y_k \kappa_k(\eta)}{(k!)^{1/2}} \Big)_+ \Big)^2 \Big/ \sum_{k=2}^{\infty} \frac{\kappa_k^2(\eta)}{k!} \Big). \tag{7}$$

The detailed derivation is provided in the Appendix. Approximating the distribution function of the measure η on Θ by a discrete distribution function with masses $p_1, ..., p_N$ on a fine grid $m_1, ..., m_N$ leads to the approximation

$$D \approx \sup_{p_1,...,p_N} \Big(\Big(\Big(\sum_{j=1}^{N} p_j \sum_{k=2}^{\infty} \frac{Y_k m_j^k}{(k!)^{1/2}} \Big)_+ \Big)^2 \Big/ \sum_{i,j=1}^{N} p_i p_j \sum_{k=2}^{\infty} \frac{(m_j m_i)^k}{k!} \Big).$$

In particular, maximizing the right-hand side with respect to $p_1, ..., p_N$ under the constraints $p_i \geq 0, \sum p_i = 1$ for fixed grid $m_1, ..., m_N$ can be formulated as a quadratic optimization problem of the form

$$\min_p p^\top A p \quad \text{under} \quad p_i \geq 0,\ p^\top b = 1$$

where $p = (p_1, ..., p_N)$, $A_{ij} = \sum_{k=2}^{\infty} \frac{(m_j m_i)^k}{k!}$, $b_i = \sum_{k=2}^{\infty} \frac{Y_k m_i^k}{(k!)^{1/2}}$, if $\max_i b_i > 0$. If $\max_i b_i \leq 0$, we can set $D = 0$. This suggests a practical way of simulating critical values after replacing the infinite sum by a finite approximation and avoiding the grid point 0. Table 1 below contains simulated critical values in some particular settings. All results are based on $10,000$ simulation runs with the sums for A and b cut off at $k = 25$ and grids with 200 points equally spaced points excluding the point 0 on the interval that is specified in the first column of Table 1.

Θ	90%	95%	99%
[-1,1]	2.75	3.95	6.93
[-2,2]	3.90	5.37	8.71
[-3,3]	5.34	6.87	10.46
[-4,4]	6.38	8.32	11.91

Table 1. Simulated asymptotic critical values for the asymptotic null distribution for various choices of the set Θ.

To explore the finite sample performance of the above method we begin with an experiment to compare the critical values of the LRT of homogeneity in the Gaussian location model with the simulated asymptotic critical values in Table 1. We consider sample sizes, $n \in \{100, 500, 1000, 5000, 10000\}$ and four choices of the domain of the MLE of the mixture are considered: $\{[-j, j] : j = 1, \cdots, 4\}$. We maintain a grid spacing of 0.01 for the mixing distribution on these domains for each of these cases for the Kiefer-Wolfowitz NPMLE. Results are reported in Table 2. For the three largest sample sizes we bin the observations into 300 and 500 equally spaced bins respectively. It will be noted that the empirical critical

n	cval(.90)				cval(.95)				cval(.99)			
	[-1,1]	[-2,2]	[-3,3]	[-4,4]	[-1,1]	[-2,2]	[-3,3]	[-4,4]	[-1,1]	[-2,2]	[-3,3]	[-4,4]
100	2.09	2.69	2.80	2.80	3.07	3.70	3.97	4.06	6.43	7.58	8.31	8.55
500	2.22	2.80	2.96	2.98	3.06	3.87	4.41	4.41	5.69	7.07	7.45	7.52
1,000	2.67	3.46	3.72	3.76	3.73	4.95	5.44	5.56	7.26	8.55	9.51	9.76
5,000	2.68	3.56	3.91	3.96	3.79	4.54	4.83	5.09	6.52	8.15	8.32	8.38
10,000	2.41	3.11	3.29	3.46	3.61	4.45	4.72	4.97	6.23	7.51	7.96	8.32
∞	2.75	3.90	5.34	6.38	3.95	5.37	6.87	8.32	6.93	8.71	10.46	11.91

TABLE 2. Critical Values for LRT of Gaussian Parameter Homogeneity: The first five rows of the table report empirical critical values based on 1000 replications of the LRT based on the Kiefer-Wolfowitz estimate of the nonparametric Gaussian location mixture distribution. Results for sample sizes 5,000 and 10,000 were computed by binning the observations into 300, 500 equally spaced bins respectively. Restriction of the domain of the mixing distribution is indicated by the column labels. The last row reproduces the simulated asymptotic critical values reported in Table 1.

values are consistently smaller than those simulated from the asymptotic theory. There appears to be a tendency for the empirical critical values to increase with n, but this tendency is rather weak. This finding is perhaps not entirely surprising in view of the slow rates of convergence established elsewhere in the literature, see e.g. Bickel and Chernoff (1993) and Hall and Stewart (2005). These findings imply that our simulated asymptotic critical values are not likely to work well for size control, which motivates us to consider an alternative bootstrap based method in determining critical values in the next section.

2.4. **A Parametric Bootstrap Method for Critical Values.** The parametric bootstrap method for testing parameter homogeneity we are about to introduce is a very natural idea. In finite mixture models, similar approaches have been proposed by McLachlan (1987) and Chen and Chen (2001). However, to the best of our knowledge, this is the first time that such a bootstrap method has been formally shown to produce consistent critical values for likelihood ratio tests in mixture models.

The parametric bootstrap approach to determine critical values for the distribution of L_n is defined as follows.

(1) Compute the maximum likelihood estimator $\hat{\mu} := \text{argmax}_{\mu\in\Theta}\ell_n(\delta(\mu))$.
(2) For $b = 1, ..., B$ generate data $X_{1,n}^{(b)}, ..., X_{n,n}^{(b)} \sim p(\cdot|\hat{\mu})$ i.i.d.
(3) For $b = 1, ..., B$ denote by $L_{n,b}$ the statistic L_n computed from the sample $X_{1,n}^{(b)}, ..., X_{n,n}^{(b)}$. Compute the α-quantile $q_{n,\alpha}$ of $L_{n,1}, ..., L_{n,B}$.

The null of parameter homogeneity is rejected if $L_n > q_{n,1-\alpha}$. To prove that this bootstrap procedure leads to a valid (asymptotic) test, we need to show that $P(L_n > q_{n,1-\alpha}) \to \alpha$ if $X_1, ..., X_n$ are generated under the null. To establish this result, we need two main ingredients. First, we need to analyze the limiting properties of the LRT for data that

are generated under triangular arrays. This is done in Theorem 2.10. Second, we need to establish continuity of the limiting distribution of F_R around its $\alpha-$quantile. This is done in Theorem 2.11. Together, Theorem 2.10 and 2.11 imply consistency of the proposed bootstrap procedure.

We now require some additional notation. Fix an arbitrary sequence of points μ_n in $\Theta \subset \mathbb{R}^d$ with $\mu_n \to \mu_0 \in \Theta$ as $n \to \infty$. For $\varepsilon > 0$, define Θ^ε as the ε-enlargement of Θ with respect to Euclidean distance. Let

$$\bar{\mathcal{G}}^\varepsilon := \{\eta | \eta \text{ distribution on } \Theta^\varepsilon\}, \quad \mathcal{G}^\varepsilon := \bar{\mathcal{G}}^\varepsilon \backslash \delta(\mu_0).$$

To each measure $\eta \in \mathcal{G}$ define the measure η_n through $\eta_n(A) = \eta(A - \mu_n + \mu_0)$ for all Borel sets $A \subset \Theta$ where $A + x := \{a + x | a \in A\}$ for a set $A \subset \mathbb{R}$ and $x \in \mathbb{R}$. From now on, assume that $X_{1,n}, ..., X_{n,n}$ are i.i.d. $\sim p(\cdot|\mu_n)$ and consider the following sequence of processes indexed by $\mathcal{G}^\varepsilon$

$$\mathbb{G}_n^*(\eta) := n^{-1/2} \sum_{i=1}^n s_{\eta_n, \mu_n}(X_{i,n})$$

where the scores s_{η_n,μ_n} are defined in (3). Write $\ell_n^*(\eta) := \sum_{i=1}^n \log p_\eta(X_{i,n})$. To analyze the asymptotic behavior of $L_n^* := \sup_{\eta \in \bar{\mathcal{G}}} \ell_n^*(\eta) - \sup_{\mu \in \Theta} \ell_n^*(\delta(\mu))$, consider the decomposition

$$L_n^* = \sup_{\eta \in \bar{\mathcal{G}}} \ell_n^*(\eta) - \ell_n^*(\delta(\mu_n)) + \ell_n^*(\delta(\mu_n)) - \sup_{\mu \in \Theta} \ell_n^*(\delta(\mu)).$$

Classical results suggest that under suitable regularity conditions the second part in the above decomposition should take the form

$$\sup_{\mu \in \Theta} \ell_n^*(\delta(\mu)) - \ell_n^*(\delta(\mu_n)) = \frac{1}{2} \Big\| \frac{1}{\sqrt{n}} \sum_{i=1}^n I(\mu_n)^{-1/2} \ell'(X_{i,n}|\mu_n) \Big\|^2 + o_P(1) \tag{8}$$

provided that $\mu_n \to \mu_0$. Various conditions ensuring the above representation exist. In the Appendix, we demonstrate that μ_0 being in the interior of the parameter space together with suitable regularity conditions on the function $\mu \mapsto \ell(x|\mu)$ are sufficient to obtain this kind of expansion. Since this is not our main focus, we leave all details to the Appendix and do not go into additional details here. The main challenge is to derive an expansion for the first part of L_n^*. Such an expansion is established in Theorem 2.10 under the following set of assumptions:

(A1) Assume that

$$\Big(\mathbb{G}_n^*, \frac{1}{\sqrt{n}} \sum_{i=1}^n I(\mu_n)^{-1/2} \ell'(X_{i,n}|\mu_n) \Big) \rightsquigarrow (\mathbb{G}^*, Y)$$

in $\ell^\infty(\mathcal{G}^\varepsilon) \times \mathbb{R}^d$ where $(\mathbb{G}^*, Y)$ are jointly centered normal with $Y \sim \mathcal{N}(0, I_d)$ and covariance structure of the form,

$$\mathbb{E}[\mathbb{G}^*(\eta_1)\mathbb{G}^*(\eta_2)] = \int_{\mathbb{R}} s_{\eta_1,\mu_0}(x) s_{\eta_2,\mu_0}(x) p_{\delta(\mu_0)}(x) d\lambda(x),$$
$$\mathbb{E}[\mathbb{G}^*(\eta) Y] = \int_{\mathbb{R}} s_{\eta,\mu_0}(x) I(\mu_0)^{-1/2} \ell'(x|\mu) p_{\delta(\mu_0)}(x) d\lambda(x).$$

Additionally, assume that for $\varepsilon \downarrow 0$ we have

$$\sup_{\eta \in \mathcal{G}^\varepsilon} \inf_{\tilde{\eta} \in \mathcal{G}} |\mathbb{G}^*(\eta) - \mathbb{G}^*(\tilde{\eta})| = o_P(1). \tag{9}$$

(A2) Letting $s_{\eta,\mu,-} := \min\{0, -s_{\eta,\mu}\}$ we have that

$$\sup_{\eta \in \mathcal{G}^\varepsilon} \left| \frac{1}{n} \sum_{i=1}^n (s^2_{\eta_n,\mu_n}(X_{i,n}) - 1) \right| + \left| \frac{1}{n} \sum_{i=1}^n (s^2_{\eta_n,\mu_n,-}(X_{i,n}) - \|s_{\eta,\mu_0,-}\|^2_{2,\delta(\mu_0)}) \right| = o_P(1).$$

(A3) For every $n \in \mathbb{N}$, assume that the class of functions

$$\mathcal{F}_n := \left\{ x \mapsto s_{\eta,\mu_n}(x) \middle| \eta \in \mathcal{G} \right\}$$

admits an envelope function F_n such that $\max_{i=1,\dots,n} F_n(X_{i,n}) = o_P(n^{1/2})$.

Remark 2.4. As pointed out by a referee, conditions (A1)-(A3) do not explicitly include the assumptions that the set Θ is compact. Yet, in specific examples this compactness is known to be crucial since otherwise the nonparametric LRT may not have a sensible limiting distribution – as seen for instance in Hartigan (1985). This apparent contradiction is resolved by the fact that compactness of Θ is often needed when verifying (A1)-(A3), as is the case in our Example 2.7 and Example 2.9.

Remark 2.5. Note that the process $\mathbb{G}^*_n$ is indexed by measures η, and not by the score functions s_{η_n,μ_n} where the latter would correspond to 'classical' empirical process theory. The reason for this indexing is that the score functions s_{η_n,μ_n} depend on n. Thus indexing by score functions s_{η_n,μ_n} we would obtain an index set which depends on n, which would lead to various technical problems. On the other hand, using s_{η_n,μ_n} instead of s_{η,μ_n} in the definition of $\mathbb{G}^*_n$ is crucial since s_{η,μ_n} can be quite different for the same values of η but different μ_n. As an example of the latter, let $\mu_n = \mu_0 + 1/n$, $\tilde{\mu}_n = \mu_0 + 3/n$, $\eta = \delta(\mu_0 + \alpha)$. Then, for α small, under suitable differentiability conditions we have $s_{\eta,\delta(\mu_n)}(x) \approx sgn(\alpha - 1/n)\ell'(x|\mu_n)/\|\ell'(x|\mu_n)\|_{2,\delta(\mu_n)}$ and $s_{\eta,\delta(\tilde{\mu}_n)}(x) \approx sgn(\alpha - 3/n)\ell'(x|\tilde{\mu}_n)/\|\ell'(x|\tilde{\mu}_n)\|_{2,\delta(\tilde{\mu}_n)}$. For $\alpha \in (1/n, 3/n)$ the sign of $\alpha - 1/n$ and $\alpha - 3/n$ will differ, and this leads to different score functions. This problem does not arise if we use s_{η_n,μ_n} instead. ■

For location-shift mixtures, that is mixtures of densities of the form $p(\cdot|\mu) = p(\cdot - \mu)$, assumptions (A1)-(A3) can be considerably simplified.

Proposition 2.6. *Assume that $p(\cdot|\mu) = p(\cdot - \mu)$, the conditions of Theorem 2.1 hold with $\mathcal{G}^\varepsilon$ instead of $\mathcal{G}$, that (8) holds, and that additionally for $\gamma \downarrow 0$ we have, for $\mathbb{G}$ denoting the weak limit of $\mathbb{G}_n$ in Theorem 2.1,*

$$\sup_{\eta \in \mathcal{G}^\gamma} \inf_{\tilde{\eta} \in \mathcal{G}} |\mathbb{G}(\eta) - \mathbb{G}(\tilde{\eta})| = o_P(1). \tag{10}$$

Then conditions (A1)-(A3) hold.

The proof of Proposition 2.6 repeatedly makes use of the fact that the assumptions of Theorem 2.1 hold for $\mathcal{G}^\varepsilon$ instead of $\mathcal{G}$. In general, this can not be avoided. Intuitively, this is

due to the fact that for measures η with support in Θ the support of η_n will not necessarily be contained in Θ.

Next, we show that assumptions (A1)-(A3) are realistic and can be verified for some standard models.

Example 2.7. (Location Mixture of Gaussians) Assume that $\Theta = [a, b]$ for some $a < 0 < b$ and that the densities p take the form $f(x|\mu) = (2\pi)^{-1/2} \exp((x-\mu)^2/2)$ with respect to Lebesgue measure. Without loss of generality we will assume that $\mu_0 = 0$. In this setting, the densities have the location-scale structure described in Proposition 2.6, and thus it suffices to verify the conditions of Theorem 2.1 hold with $\mathcal{G}^\varepsilon$ instead of $\mathcal{G}$, that (8) holds, and that (10) is satisfied. Note that (8) can be established by standard arguments, the details are omitted for the sake of brevity.

The arguments from the proof of Theorem 3 in (Azaïs, Gassiat, and Mercadier 2009) yield $\mathbb{G}_n \rightsquigarrow \mathbb{G}$ in $\ell^\infty(\mathcal{G}^\varepsilon)$ where the limiting process $\mathbb{G}$ is Gaussian and has a covariance structure of the form

$$\mathbb{E}[\mathbb{G}(\eta_1)\mathbb{G}(\eta_2)] = \frac{\mathbb{E}[\exp(Z_1 Z_2)] - 1}{(\mathbb{E}[\exp(Z_1 \tilde{Z}_1)] - 1)^{1/2} (\mathbb{E}[\exp(Z_2 \tilde{Z}_2)] - 1)^{1/2}}$$

where $Z_1, \tilde{Z}_1 \sim \eta_1, Z_2, \tilde{Z}_2 \sim \eta_2$ and $Z_1, Z_2, \tilde{Z}_1, \tilde{Z}_2$ are independent. Joint asymptotic normality with Y_1 follows by standard arguments. To prove (10), consider the following construction. To each random variable Z on $[a-\varepsilon, b+\varepsilon]$ define a transformed random variable W through

$$W := ZI\{Z \in [a, b]\} + \frac{M}{M+\varepsilon} ZI\{Z \notin [a, b]\}.$$

where $M := \min(|a|, b)$. By construction, the support of W is contained in $[a, b]$. Denoting the distribution of W by $\xi_{\eta,\varepsilon}$, straightforward but tedious calculations show that

$$\sup_{\eta \in \mathcal{G}^\varepsilon} \mathbb{E}[(\mathbb{G}(\eta) - \mathbb{G}(\xi_{\eta,\varepsilon}))^2] = o(1)$$

as $\varepsilon \downarrow 0$. By the uniform continuity of the process $\mathbb{G}$ with respect to the metric $d(\eta, \xi) := (\mathbb{E}[(\mathbb{G}(\eta) - \mathbb{G}(\xi))^2])^{1/2}$ induced by its covariance [see Example 1.5.10 in (van der Vaart and Wellner 1996)], this shows that (10) also holds. ■

Remark 2.8. As pointed out by a referee, location-scale mixtures on Gaussians, i.e. mixtures of the form $p(x|\eta) = \iint p(x|\mu, \sigma) d\eta(\mu, \sigma)$ with $p(\cdot|\mu, \sigma)$ denoting the density of an $\mathcal{N}(\mu, \sigma^2)$ random variable, are also of practical interest. In such models, even identification of parameters is a very subtle issue. To illustrate this point, consider a location mixture of normals with unknown variance parameter. If the support of the location parameter is unrestricted, assumption (A0) will fail if we allow for general classes of mixtures. To see that, denote by $\eta(\tau)$ the product of an $\mathcal{N}(0, \sigma^2 - \tau^2)$ measure for location and a point mass at τ^2 for variance where $0 \leq \tau^2 \leq \sigma^2$. Then $p_{\eta(\tau)} \equiv p_{\eta(\tau')}$ for any $\tau, \tau' \in [0, \sigma]$, and setting $\tau^2 = \sigma^2$ corresponds to homogeneity. Thus (A0) does not hold. Assuming that the support for μ is restricted to a compact set, the unknown variance σ^2 and the mixing distribution can be jointly identified. We are not aware of results on identification if both, location and scale are being mixed, even if the support for both parameters is confined to compact

sets. Gaining a better understanding of identification and, provided identification holds, the behaviour of LRT in this case is a very interesting and important question. We leave this question to future research. ■

Example 2.9. (Mixture of Poisson distributions) Assume that $\Theta = [a, b]$ for some $0 < a < b$ and that the densities p take the form $p(k|\mu) = \mu^k e^{-\mu}/k!$ with respect to the counting measure on $\mathbb{N}$. Note that this model does not have the location-scale structure discussed in Proposition 2.6. Assumptions (A1)-(A3) can still be verified, and the technical details are provided in Section B of the Appendix. ■

We now state our main result.

Theorem 2.10. *Under assumptions (A0)-(A3) we have*

$$2\sup_{\eta\in\bar{\mathcal{G}}}\Big(\ell_n^*(\eta) - \ell_n^*(\delta(\mu_n))\Big) = \sup_{\eta\in\mathcal{G}}\Big(\max\Big\{\mathbb{G}_n^*(\eta), 0\Big\}\Big)^2 + o_P(1). \tag{11}$$

If additionally (8) *holds we have*

$$2\Big(\sup_{\eta\in\bar{\mathcal{G}}}\ell_n^*(\eta) - \sup_{\mu\in\Theta}\ell_n^*(\delta(\mu))\Big) \rightsquigarrow R := \sup_{\eta\in\mathcal{G}}\Big(\max(\mathbb{G}^*(\eta), 0)\Big)^2 - \|Y\|^2.$$

Intuitively, Theorem 2.10 suggests that critical values based on the parametric bootstrap should lead to an asymptotic level α test of homogeneity. However, a formal proof of this statement requires that the distribution of R, say F_R, is continuous at $F_R^{-1}(\alpha)$. The following theorem completes this last step.

Theorem 2.11. *Let the assumptions of Theorem 2.1 hold. Then the distribution of R is continuous on $(0, \infty)$ and $P(R < 0) = 0$. Provided that $B = B_n \to \infty$ we have $\limsup_{n\to\infty} P(L_n > q_{n,1-\alpha}) = \alpha$ for any α satisfying $P(R > 0) > \alpha$. Moreover, if $d = 1$ and if there exists $\eta \in \mathcal{G}$ such that $\mathbb{E}[\mathbb{G}(\eta)Y] \neq \pm 1$ we have $P(R > 0) \geq 1/4$.*

Remark 2.12. How to choose support to solve for the NPMLE is a very important practical question. For location shift models, it is easy to show that the NPMLE $\hat{\eta}$ will not have any mass points outside of the sample support. This type of result has been generalized in Lindsay (1981) to other univariate base densities that have a unique mode. In particular, suppose that for each sample point x_i, the function $\mu \mapsto p(x_i \mid \mu)$ has a unique mode at μ_i^*. Then the support of the NPMLE $\hat{\eta}$ must be contained in $[\mu_m^*, \mu_M^*]$ where μ_m^* and μ_M^* are the minimum and maximum of $(\mu_1^*, \ldots, \mu_n^*)$, respectively. This is true for many base distributions in the exponential family. For example, for mixtures of exponential distributions with mean $\exp(-\phi)$, the mode for the base density $\exp(\phi)\exp(-x_i\exp(-\phi))$ is located at $\phi_i^* = -\ln(x_i)$. Hence the support for the mixing distribution must be contained in $[\min(\ln(1/x_i)), \max(\ln(1/x_i))]$. To ensure compactness, we recommend taking the 5-th and 95-th quantile of $\mu_1^*, ..., \mu_n^*$.

Remark 2.13. For mixture models with densities of the form $p(\cdot|\mu) = p(\cdot - \mu)$ there is an alternative way of simulating quantiles of the LRT statistic. The key observation is that, assuming that we allow for an arbitrary support of the mixing distribution, the distribution of the LRT under the null does not depend on the location of the true parameter. More precisely, assume that $X_1, ..., X_n$ generated from $p(\cdot|\mu_X)$ and $Y_1, ..., Y_n$ are generated from $p(\cdot|\mu_Y)$. Then X_i has the same distribution as $Y_i - \mu_Y + \mu_X$, and for any measure η the log-likelihood $\sum_{i=1}^n \log p_\eta(X_i)$ has the same distribution as $\sum_{i=1}^n \log p_\eta(Y_i - \mu_Y + \mu_X)$, which equals the distribution of $\sum_{i=1}^n \log p_{\tilde\eta}(Y_i)$ with the measure $\tilde\eta$ defined through $\tilde\eta(A) = \eta(A - \mu_Y + \mu_X)$. This implies that the LRT statistic computed from $X_1, ..., X_n$ and the one computed $Y_1, ..., Y_n$ will have the same distribution.

Thus the following procedure provides a way to conduct an exact test for parameter homogeneity when the support of the mixing distribution is unrestricted.

(1) Repeatedly generate data $Y_1, \ldots, Y_n \sim p(\cdot|0)$ i.i.d. for B times. For each bootstrap sample, compute the LRT statistic $L_{n,b}$ for $b = 1, \ldots, B$.
(2) Compute the $1-\alpha$-quantile $q^L_{n,1-\alpha}$ of the bootstrap sample $L_{n,b}$, $b = 1, \ldots, B$.

The null of parameter homogeneity is rejected if $L_n > q^L_{n,1-\alpha}$.

Table 3 tabulates the bootstrap critical values for the null distribution of the LRT statistic for testing homogeneity of the Gaussian location parameter. B bootstrap samples of size n is generated from standard normal distribution and the critical values are found based on the empirical distribution of the corresponding LRT statistic.

	90%	95%	99%
n=100	3.14	4.60	8.12
n=200	3.15	4.48	7.21
n=500	3.44	4.69	7.84

Table 3. Bootstrap Critical Values for LRT of Homogeneity of Gaussian Location Parameter: For various sample sizes, the bootstrap critical values are found following the procedure described in Remark 2.13 with $B = 2,000$.

It is important to keep in mind that this invariance property will hold only if we consider an *unrestricted* support. In the case of Gaussian location mixtures, it is well known that the LRT statistic with mixing distributions of unbounded support diverges to infinity (see Hartigan (1985)). A more detailed analysis of this issue for some special cases of likelihood ratio tests in mixture models can be found in Azaïs, Gassiat, and Mercadier (2006) and Hall and Stewart (2005). That analysis indicates that LRT with unrestricted support can only detect local alternatives at slower rates than moment-based tests. However, the corresponding difference in rates is quite small and we compare via simulations the differences in power for using the parametric bootstrap critical values and the exact critical values for the location parameters in the Gaussian models. Results are summarized in Table 4, the power loss for reasonable sample sizes is quite modest.

To evaluate size performance of using these bootstrap critical values, we apply the LRT on a random sample $X_1, \ldots, X_n \sim \mathcal{N}(1, 1)$ for homogeneity versus general mixture on the

		90%	95%	99%
$h = 0.1$	LRT-PBS[-1,1]	0.2095	0.1180	0.0380
	LRT-PBS[-2,2]	0.2070	0.1135	0.0355
	LRT-EXT	0.1765	0.1070	0.0375
$h = 0.2$	LRT-PBS[-1,1]	0.6520	0.5120	0.2960
	LRT-PBS[-2,2]	0.6255	0.4945	0.2550
	LRT-EXT	0.5690	0.4505	0.2400
$h = 0.3$	LRT-PBS[-1,1]	0.9775	0.9615	0.8805
	LRT-PBS[-2,2]	0.9730	0.9485	0.8550
	LRT-EXT	0.9660	0.9305	0.8430

TABLE 4. Power comparison between parametric bootstrap method (denoted as LRT-PBS with stated support used for estimating the general mixture model) on restricted support and the Gaussian LRT with unrestricted support and exact critical value (denoted as LRT-EXT) as tabulated in Table 3. Simulation data is generated as $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$ with $n = 200$ and $F_\mu = \frac{2}{3}\delta_{1.5h} + \frac{1}{3}\delta_{-3h}$ for h taking values from $\{0.1, 0.2, 0.3\}$. Results are based on 2,000 repetitions and the parametric bootstrap method is based on 500 bootstrap repetitions on the stated support.

location parameter. The third row of Table 5 reports the size performance of the LRT with these tabulated bootstrap critical values. In the same table, we also report the size performance of the LRT using critical values generated from the parametric bootstrap method, the $C(\alpha)$ test and the EM test that will be discussed in the next section. ■

3. Neyman $C(\alpha)$ Tests for Mixture Models

Neyman's $C(\alpha)$ tests can be viewed as an expanded class of Rao (score) tests that accommodate general methods of estimation for nuisance parameters. In regular likelihood settings $C(\alpha)$ tests are constructed from the usual score components which consist of the first order logarithmic derivative of the likelihood. The $C(\alpha)$ tests can be shown to be asymptotically locally optimal and the associated regularity conditions for these results were originally given by Neyman (1959) and extended by Bühler and Puri (1966) employing variants of the classical Cramér conditions. In applying the $C(\alpha)$ approach to test for homogeneity in mixture models, the test statistics typically still take a simple form although their theory requires some substantial amendment due to the singularity of the score function. Gu (2016) shows that the locally asymptotic normal (LAN) apparatus of LeCam can be brought to bear to establish the large sample behavior and asymptotic optimality of the $C(\alpha)$ test for homogeneity. The LeCam approach has two salient advantages: it avoids making superfluous further differentiability assumptions on the density, and it removes any need for the symmetry assumption on the distribution of the heterogeneity that frequently appears in earlier examples of such tests. See e.g. Moran (1973) and Chesher (1984).

	$n = 100$			$n = 200$			$n = 500$		
	90%	95%	99%	90%	95%	99%	90%	95%	99%
EM	0.088	0.044	0.010	0.094	0.050	0.012	0.094	0.048	0.010
$C(\alpha)$	0.103	0.050	0.018	0.104	0.058	0.014	0.099	0.052	0.011
LRT-EXT	0.072	0.038	0.008	0.094	0.052	0.012	0.104	0.060	0.012
LRT-PBS[-1,1]	0.086	0.040	0.008	0.097	0.057	0.011	0.070	0.040	0.008
LRT-PBS[-2,2]	0.098	0.048	0.012	0.102	0.046	0.008	0.106	0.056	0.013

TABLE 5. Size Performance for Various Tests for Homogeneity of the Gaussian Location Parameter: Independent samples of different sizes are generated from $\mathcal{N}(1, 1)$. We consider test for homogeneity versus general alternative. The EM test is as proposed in Chen and Li (2009) using the R code provided on the second author's webpage `http://sas.uwaterloo.ca/~p4li/software/index.html` of the EM test for Gaussian mixture with known variance. The $C(\alpha)$ test uses critical values from $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ null distribution. LRT-EXT uses bootstrap critical values tabulated in Table 3. Results are based on 6,000 repetitions. LRT-PBS (with stated support used for estimating the general mixture model) uses parametric bootstrap critical values with 500 bootstrap repetitions on the pre-specified support for the location parameter.

The following two examples illustrate the construction of the $C(\alpha)$ test for parameter homogeneity in the Gaussian mixture model and the Poisson mixture model. Both tests lead to an over-dispersion test. In the Gaussian case, the test compares the sample variance with the variance under the null hypothesis. In the Poisson case, we reject the null of homogeneity if there exists over-dispersion in the sample variance in comparison to the sample mean.

Example 3.1. Consider testing for homogeneity in the Gaussian location mixture model with independent observations $X_i \sim \mathcal{N}(\mu_i, 1), i = 1, \cdots, n$. Assume that $\mu_i = \mu_0 + \tau\xi U_i$, for known τ, and iid $U_i \sim F$ with $\mathbb{E}U = 0$ and $\mathrm{Var}(U) = 1$. The heterogeneity in μ_i is introduced via the random variable U. We would like to test homogeneity of μ_i, $H_0 : \xi = 0$, with the location parameter μ_0 treated as a nuisance parameter. As mentioned earlier, the first-order logarithmic derivative for ξ is degenerately zero, however we can construct the test statistics using its second-order derivative, which is found to be, $\nabla_\xi^2 \log p(x|\mu_0, \xi = 0) = \tau^2((x-\mu_0)^2 - 1)$. The first-order score for the nuisance parameter μ_0 is, $\nabla_{\mu_0} \log p(x|\mu_0, \xi = 0) = (x - \mu_0)$. Note that under the null, $\mathrm{cov}(\nabla_\xi^2 \log p(X|0, \mu_0), \nabla_{\mu_0} \log p(X|0, \mu_0)) = 0$, thus the $C(\alpha)$ test statistics require no modification of the test statistics to reflect the fact that we need to estimate the nuisance parameter μ_0 and thus, we have the locally asymptotically optimal $C(\alpha)$ test as

$$Z_n = \frac{1}{\sqrt{2n}} \sum_{i=1}^{n} ((X_i - \mu_0)^2 - 1)$$

The obvious estimate for the nuisance parameter is the sample mean, and we reject the null hypothesis when $(0 \vee Z_n)^2 > c_\alpha$ where c_α is the $(1-\alpha)$ quantile of $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$. The test statistic Z_n depends on the sample variance of X. Under the general alternative model, we have $\mathrm{Var}(X) = \mathbb{E}_\mu[\mathrm{Var}(X|\mu)] + \mathrm{Var}_\mu[\mathbb{E}(X|\mu)] = 1 + \mathrm{Var}(\mu)$. Under the alternative, the magnitude of Z_n solely depends on $\sqrt{n}\,\mathrm{Var}(\mu)$.

Example 3.2. Consider now testing for homogeneity of the mean parameter in the Poisson model with independent observations $X_i \sim p(\cdot|\lambda_i), i = 1, \cdots, n$ with $p(x|\lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$. Assume that $\lambda_i = \lambda_0 \exp(\tau\xi U_i)$, for known τ, and iid $U_i \sim F$ with $\mathbb{E}U = 0$ and $\mathrm{Var}(U) = 1$. We would like to test $H_0 : \xi = 0$ with the mean parameter λ_0 treated as a nuisance parameter. The second-order score for ξ is found to be, $\nabla^2_\xi \log p(x|\lambda_0, \xi = 0) = \tau^2((x - \lambda_0)^2 - \lambda_0)$ and the first-order score for λ_0 is, $\nabla_{\lambda_0} \log p(x|\lambda_0, \xi = 0) = (x - \lambda_0)/\lambda_0$. Note that under the null, $\mathrm{cov}(\nabla^2_\xi \log p(X|\lambda_0, 0), \nabla_{\lambda_0} \log p(X|\lambda_0, 0)) = \lambda_0$. Thus, we have the locally asymptotically optimal $C(\alpha)$ test as

$$Z_n = \frac{1}{\sqrt{2n}} \sum_{i=1}^{n} \frac{((X_i - \lambda_0)^2 - \lambda_0 - (X_i - \lambda_0))}{\lambda_0}$$

The obvious estimate for the nuisance parameter λ_0 is the sample mean $\bar{X}$, which further reduces $Z_n = \frac{1}{\sqrt{2n}} \sum_{i=1}^{n} \frac{((X_i - \bar{X})^2 - \bar{X})}{\bar{X}}$ and we reject the null hypothesis when $(0 \vee Z_n)^2 > c_\alpha$. The test statistic Z_n depends on the ratio of the sample variance and sample mean of X. Under the alternative model, we have $\mathrm{Var}(X) = \mathbb{E}(\lambda) + \mathrm{Var}(\lambda)$ and $\mathbb{E}(X) = \mathbb{E}(\lambda)$. The magnitude of the test statistics Z_n under the alternative is determined by the ratio $\sqrt{n}\,\mathrm{Var}(\lambda)/\mathbb{E}(\lambda)$.

4. The EM Test of Homogeneity for Finite Mixture Models

The $C(\alpha)$ test described above is very attractive because its test statistic is easy to construct under the null model and its asymptotic theory is also relatively simple. The recently proposed EM test of Chen and Li (2009), Li, Chen, and Marriott (2009) and Li and Chen (2010) shares these nice features. The EM test employs a penalized log likelihood ratio statistic, and instead of optimizing over a general class of heterogeneous alternatives optimization is restricted to a smaller finite dimensional class. Given the mixture model (1), we consider finite mixing distributions $\eta = \sum_{h=1}^{m} \alpha_h \delta(\mu_h)$ with m distinct support points at locations $\{\mu_1, \ldots, \mu_m\}$. We are interested in testing $H_0 : m = 1$ versus $H_A : m > 1$. Rather than consider the full panoply of alternatives, attention is restricted to mixing distributions with only two points of support,

$$\Omega_2(\beta) = \{\beta\delta(\mu_1) + (1 - \beta)\delta(\mu_2) : \mu_1, \mu_2 \in I\}$$

the relative mass of the two support points, $\beta \in (0, 0.5]$, is bounded away from zero by the penalized log likelihood,

$$pl_n(\Psi) = \sum_{i=1}^{n} \log p_\Psi(X_i) + P(\beta)$$

where $\Psi \in \Omega_2(\beta)$, and $P(u) = C \log(1 - |1 - 2u|)$. The set I over which the μ's are optimized is taken to be the support of the observations in the Gaussian location mixture setting.

Optimization is carried out via the EM algorithm over the three parameters, $\{\beta, \mu_1, \mu_2\}$, and the test statistic is,

$$M_n = 2\{pl_n(\hat{\Psi}) - \sum_i \log p_{\tilde{\Psi}}(X_i)\},$$

where $\hat{\Psi}$ and $\tilde{\Psi}$ denote estimates for the model under the alternative and null, respectively. Selection of tuning parameters including initial values and stopping criteria for the EM procedure may, of course, influence performance. Penalization has the desirable effect of avoiding the singularity that would otherwise occur as $\beta \to 0$. M_n has been shown to have a $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$ limiting distribution. Testing for additional mixture components yields more complicated mixtures of χ^2's. In the next section we compare the size and power performance of our general LRT with the EM test and the $C(\alpha)$ test for different mixture models in simulations.

5. Some Simulation Evidence

To compare power of the $C(\alpha)$, the EM test and LRT to detect heterogeneity in the Gaussian location model we conducted five distinct experiments. Two were based on variants of the Chen (1995) example with the discrete mixing distribution $\eta = (1-\lambda)\delta(a + h/(1-\lambda)) + \lambda\delta(a - h/\lambda)$. In the first experiment we set $\lambda = 1/3$, as in the original Chen example, in the second experiment we set $\lambda = 1/20$ and in both experiments, a is set to be zero. The sample size is fixed at $n = 200$. We consider five tests

(i) the $C(\alpha)$ as described in Example 3.1. Under $H_0 : h = 0$, the nuisance parameter a can be estimated by the sample mean.
(ii) a parametric version of the LRT in which only the values of a and h are assumed to be unknown and the relative probabilities associated with the two mass points are known; this enables us to relatively easily find the MLE: profiling out a first, $\hat{h}$ can be estimated by separately optimizing the likelihood on the positive and negative half-line and taking the best of the two solutions; and then we can find the best pair of $(\hat{a}, \hat{h})$ that maximizes the likelihood.
(iii) the Kiefer-Wolfowitz LRT (KW-LRT) computed with equally spaced binning of 300 grid points on the support of the sample
(iv) the classical Kolmogorov-Smirnov test of normality
(v) the EM test for one component versus two components.

All of the power comparisons are based on 10,000 simulation replications. We consider 21 distinct values of h for each of the experiments equally spaced on the respective plotting regions.

In the left panel of Figure 1 we illustrate the results for the first experiment with $\lambda = 1/3$: With the location invariance property of the Gaussian mixture model, we use the bootstrap critical values in Table 3 for the nonparametric LRT. The EM test, $C(\alpha)$ and the parametric LRT are essentially indistinguishable in this experiment, and each has slightly better performance than the nonparametric LRT. All four of these tests perform substantially better than the Kolmogorov-Smirnov test. In the right panel of Figure 1 we have results of another version of the Chen example, except that now $\lambda = 1/20$, so the mixing distribution is much more skewed. Still $C(\alpha)$ does well for small values of h, but for $h \geq 0.07$ the two LRT procedures, which are now essentially indistinguishable, dominate. The performance of the

EM test lies in between the $C(\alpha)$ test and the nonparametric LRT test. Again, the KS test performance is poor compared to the other tests explicitly designed for the mixture setting.

In Figure 2 we illustrate the results of two additional experiments, both of which are based on smooth mixing distributions with densities with respect to Lebesgue measure and a sample size of $n = 200$. On the left we consider the uniform distribution on the interval $[-h, h]$. Here we can reduce the parametric LRT to optimizing over the positive half-line to compute the MLE, $\hat{h}$. This would seem to give the parametric LRT a substantial advantage over the Kiefer-Wolfowitz nonparametric MLE, however as is clear from the figure there is little difference in their performance. Again, the $C(\alpha)$ test and the EM test are somewhat better than either of the LRTs, but the difference is modest. In the right panel of Figure 2 we have a similar setup, except that now the mixing distribution is Gaussian with scale parameter h, and again the ordering is very similar to the uniform mixing case. In all of these experiments, since the asymptotic behavior of the parametric LRT is unknown, we use its empirical critical values under the null.

In the last simulation experiment on testing for homogeneity in a normal model we consider data that are generated from a two-component mixture of the form

$$(1-\alpha)N(\theta_1, 1) + \alpha N(\theta_2, 1)$$

with a very small value of α. This is the second local alternative model considered by Chen, Li, and Liu (2016). Notably, this also fits the discussion of the local alternative model on page 94 in Lindsay (1995). In the simulation, we fix $\alpha = 0.005$, $\theta_1 = \theta_0 = 0$ and $\theta_2 = b$ and conduct two sets of experiments. The first fixes $\theta_2 = -4.5$ and allows the sample size n to change and the second varies values of θ_2 for fixed sample size $n = 400$. Results are reported in Table 6. We find that in all settings, the LRT outperforms both $C(\alpha)$ and the EM test by a considerable margin, with the EM test having advantages compared to $C(\alpha)$. This suggests that for detecting small mass points away from the main bulk of the data the LRT is the method of choice. This kind of behavior is also observed in the empirical example in Section 6, where only the LRT is able to detect deviations from homogeneity.

A theoretical explanation for the findings in this experiment can be obtained by considering the likelihood expansion corresponding to a specific type of local alternative. Adopting the notation in Chen, Li, and Liu (2016) let $\alpha := \eta/\sqrt{n}$, $\theta_{1n} := \theta_0 - n^{-1/2}\tau(\frac{\eta}{1-n^{-1/2}\eta})^{1/2}$ and $\theta_{2n} := \theta_0 + \tau(\frac{1-n^{-1/2}\eta}{\eta})^{1/2} \to \theta_0 + \tau/\sqrt{\eta} \equiv \theta_2$. As shown in Chen, Li, and Liu (2016) the likelihood ratio expansion in this case takes the form

$$\frac{\eta}{\sqrt{n}} \sum_i W_i - \frac{1}{2}\frac{\eta^2}{n} \sum_i W_i^2 + o_P(1)$$

with

$$W_i = \frac{f(x_i, \theta_2) - f(x_i, \theta_0)}{f(x_i, \theta_0)} - \frac{\tau}{\sqrt{\eta}} \frac{f'(x_i, \theta_0)}{f(x_i, \theta_0)}$$

provided W_i is square integrable. Note that $W_i \approx \frac{\tau^2}{2\eta} f''(X_i, \theta_0)/f(X_i, \theta_0)$ only if θ_2 is very close to θ_0. This already suggests that the asymptotic optimality of the $C(\alpha)$ for detecting local alternatives will only continue to hold for $\tau \approx 0$. This helps to explain the clear advantages we observe for LRT and EM tests when compared to the performance of $C(\alpha)$ in these extreme cases.

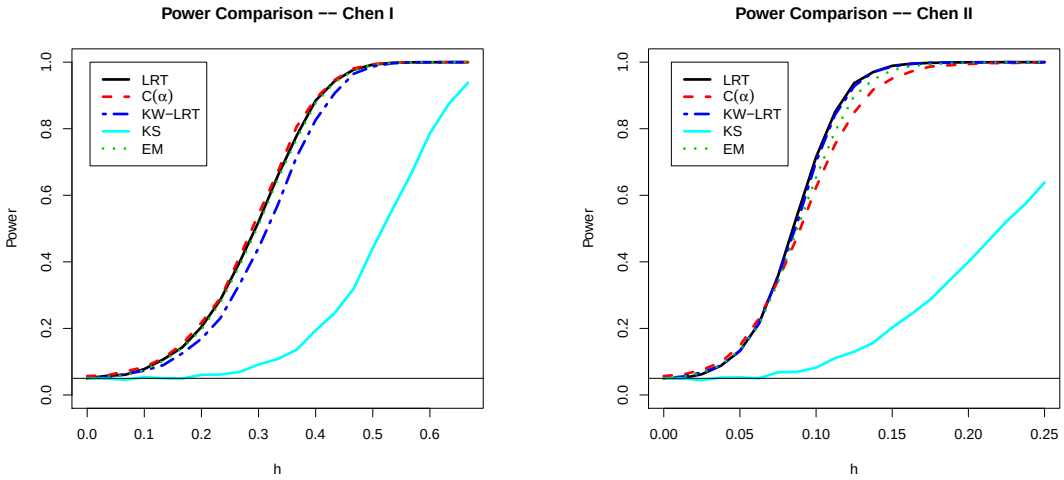

FIGURE 1. Power Comparison of Several Tests of Parameter Homogeneity: The left panel illustrates empirical power curves for four tests of parameter homogeneity for the Chen (1995) mixture with $\lambda = 1/3$, in the right panel we illustrate the power curves for the same four tests for the Chen mixture with $\lambda = 1/20$. Note that in the more extreme (right) setting, the LRTs outperform the $C(\alpha)$ test.

We also consider the power performance of the the above mentioned tests for Poisson mixture models except for the Kolmogorov-Smirnov test. Similarly to the Gaussian case, the Poisson mean parameter has the discrete mixing distribution $\eta = (1-\lambda)\delta(a\exp(h/(1-\lambda))) + \lambda\delta(a\exp((-h/\lambda))$. We consider $\lambda = 1/3$ and $\lambda = 1/20$ case and set $a = 2$ in both cases. The $C(\alpha)$ test is constructed as described in Example 3.2 with $H_0 : h = 0$ and a as the nuisance parameter. Since the Poisson distribution does not take a location shift form, we resort to the parametric bootstrap method described in Section 2.4 to determine the critical value with a bounded support on $(0, 4)$ for the mean parameter with 5,000 repetitions. To speed up the simulation, we also adopt the warp bootstrap method in Giacomini, Politis, and White (2013). Figure 3 shows the power for the $C(\alpha)$ test, the EM test and the KW-LRT for different values of h. Again, we observe similar pattern of the power curves as in the Gaussian case. For more extreme mixing distribution, the KW-LRT dominates the other two tests by quite a substantial margin.

In Figure 4 we illustrate the results for Poisson mixtures with continuous mixing distribution. In both experiments, the mean parameter is set to be $2\exp(k)$ where k has a continuous distribution. On the left, we consider k following a uniform distribution on $[0, h]$ with h taking 20 distinct equally spaced values on $[0, 0.95]$. On the right, we have k following a Gamma distribution with shape parameter h and scale parameter $1/2$ and h taking 20 distinct equally spaced values on $[0, 0.19]$. The KW-LRT performs slightly worse than $C(\alpha)$ and the EM tests for the uniform case, but dominates the other two for the Gamma case.

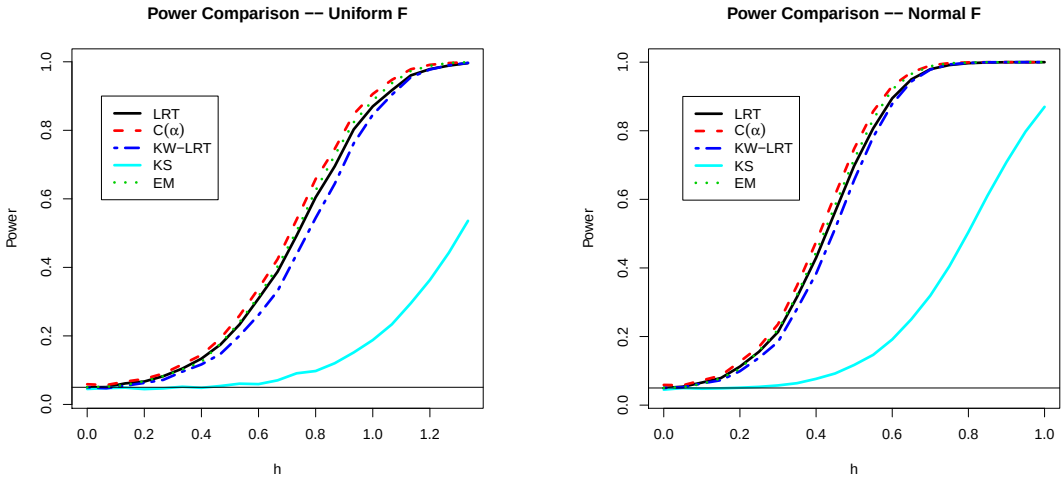

FIGURE 2. Power Comparison of Several Tests of Parameter Homogeneity: The left panel illustrates empirical power curves for four tests of parameter homogeneity for uniform mixtures of Gaussians with ϑ on $[-h, h]$, on the right panel the same four power curves are depicted for Gaussian mixtures of Gaussians with standard deviation h.

	n = 200	n = 400	n = 800	b = -6	b= -4	b = -2	b = -1
LRT	0.536	0.770	0.935	0.866	0.680	0.128	0.061
EM	0.354	0.508	0.715	0.703	0.412	0.090	0.054
$C(\alpha)$	0.296	0.412	0.578	0.635	0.329	0.093	0.060

TABLE 6. Power Comparison of Several Tests of Parameter Homogeneity for Two-component Normal Mixture Models: Results in column two to four are proportion of rejection of homogeneity using data generated from $0.995\mathcal{N}(0,1) + 0.005\mathcal{N}(-4.5,1)$ with various sample size n stated as the column names. Results in column five to eight are proportion of rejection of homogeneity using a sample of size 400 generated from $0.995\mathcal{N}(0,1) + 0.005\mathcal{N}(b,1)$ with b taking different values stated as the column names. The empirical power is based on 10,000 repetitions and LRT uses tabulated critical values of 5% nominal size.

6. EMPIRICAL EXAMPLE

We briefly revisit an application considered in Böhning, Schlattmann, and Lindsay (1992) and Chen, Li, and Liu (2016) on modeling a nutritional indicator in order to detect subclinical malnourishment. To evaluate nutritional status of children in developing countries, a standardized height score (HE/AGE) is often used. It is defined as height of the child recentered by the median and normalized by the standard deviation of heights for a reference

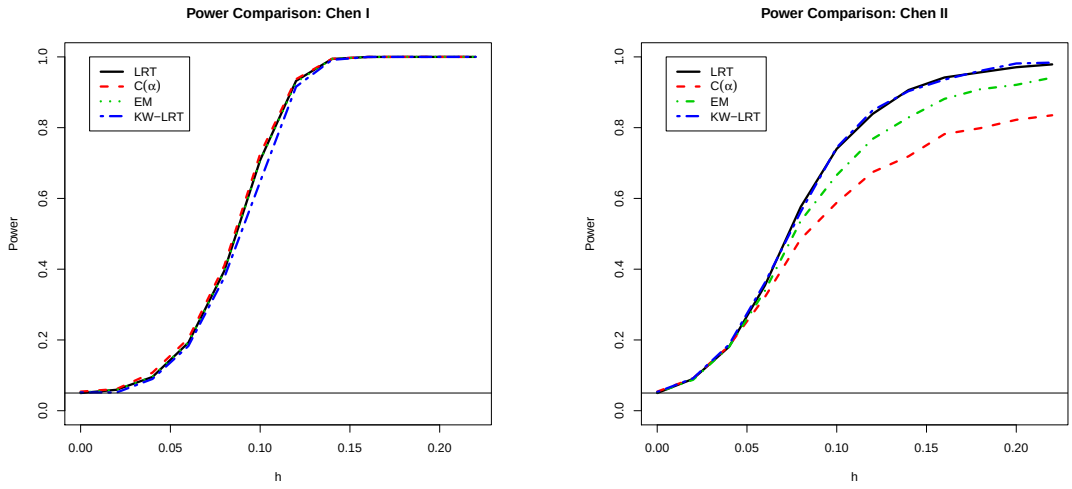

FIGURE 3. Power Comparison of Several Tests of Parameter Homogeneity for Poisson Mixture Models: The figure illustrates empirical power curves for three tests of parameter homogeneity for a discrete mixtures of Poisson. The discrete mixing distribution is specified as $F(\mu) = (1-\lambda)\delta(2\exp(h/(1-\lambda))) + \lambda\delta(2\exp(-h/\lambda))$ with $\lambda = 1/3$ in the left panel and $\lambda = 1/20$ in the right panel for h taking 21 different values. The critical values for LRT are based on the bootstrap method. The empirical power curve is based on 5,000 repetitions.

population of the same age and sex. Under the hypothesis of no malnutrition, we expect the data to follow a normal distribution with unit variance. Deviation from homogeneous normal distribution provides evidence for malnutrition of the group of children. We conduct nonparametric LRT, EM test and the $C(\alpha)$ test for homogeneity of the location parameter. Both the EM and the $C(\alpha)$ test find insufficient evidence against homogeneity, with EM test reporting a p-value close to 1 and the $C(\alpha)$ test statistic taking a value 0. In contrast, the nonparametric LRT finds strong evidence against homogeneity. Adopting the parametric bootstrap method and restricting the support to between the 5-th and 95-th percentile of the data, the nonparametric LRT statistic equals 12.77, while the parametric bootstrap critical value at 5% level equals 4.68. The nonparametric LRT using an unrestricted support and tabulated critical values leads to the same conclusion. Figure 5 shows the histogram of the data and the nonparametric MLE for the mixing distribution of the location parameter based on the estimation method described in Section 2.1. The vast majority of the mass (0.993) is allocated to the point -1.64 but we find two additional mass points at -6.19 and 6.87 with associated mass 0.005 and 0.002. Clearly, the largest data point has a mass of its own, while the mass point at -6.19 captures the very small proportion of observations at the left tail of the histogram. Although both mass points are small, they provide overwhelming evidence against homogeneity which is surprisingly not picked up by either the EM or the $C(\alpha)$ test. This sheds new light into the nature of our competing tests and illustrates that

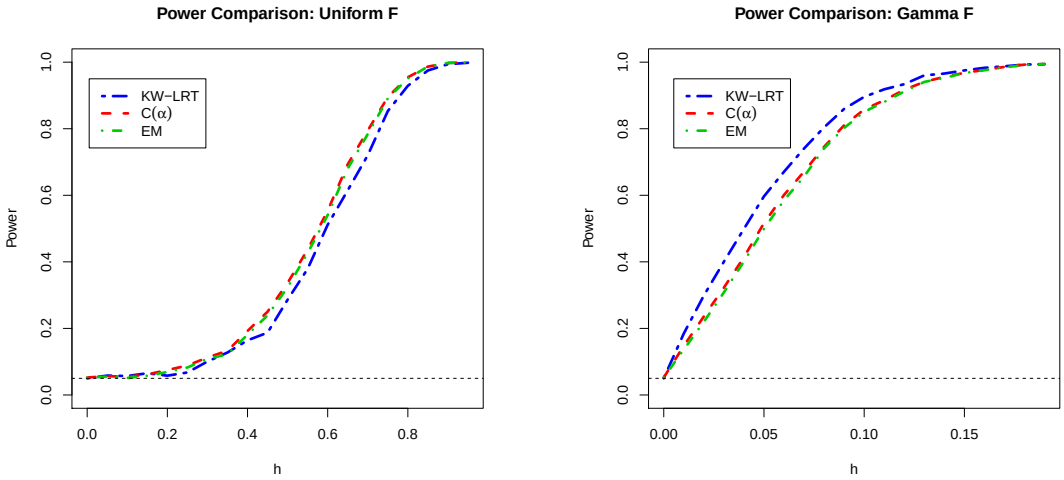

FIGURE 4. Power Comparison of Several Tests of Parameter Homogeneity for Poisson Mixture Models: The left panel illustrates empirical power curves for three tests of parameter homogeneity for uniform mixtures of Poissons with $\lambda = 2\exp(k)$ and k follows uniform distribution on $[0, h]$, on the right panel the same three power curves are depicted for Gamma mixtures of Poissons with $\lambda = 2\exp(k)$ and k follows Gamma distribution with shape parameter h and scale parameter $1/2$. Results are based on $n = 1,000$ and 5,000 simulation repetitions.

the LRT is particularly well suited to detecting deviations from the null which correspond to small mass points at extreme locations lending further support to our simulation results.

7. CONCLUSION

We have seen that the Neyman $C(\alpha)$ test provides a simple, powerful, albeit somewhat irregular, strategy for constructing tests of parameter homogeneity. In contrast, the development of likelihood ratio testing for mixture models has been somewhat inhibited by their apparent computational difficulty, as well as the complexity of their asymptotic theory. Recent developments in convex optimization have dramatically reduced the computational effort of earlier EM methods, and new theoretical developments have led to practical simulation methods for large sample critical values for the Kiefer-Wolfowitz nonparametric version of the LRT. Local asymptotic optimality of the $C(\alpha)$ test assures that it is highly competitive in many circumstances, but we have illustrated a class of examples where the LRT has a slight edge. The EM tests of Li and Chen (2010) provide an intermediate approach relying on a more restricted formulation of the likelihood. The approaches are complementary; clearly there is little point in testing for heterogeneity if there is no mechanism for estimating models under the alternative. Our LRT approach obviously provides a direct pathway to estimation of the mixture model under general alternatives. Since parametric mixture models are notoriously tricky to estimate, it is a remarkable fact that the nonparametric

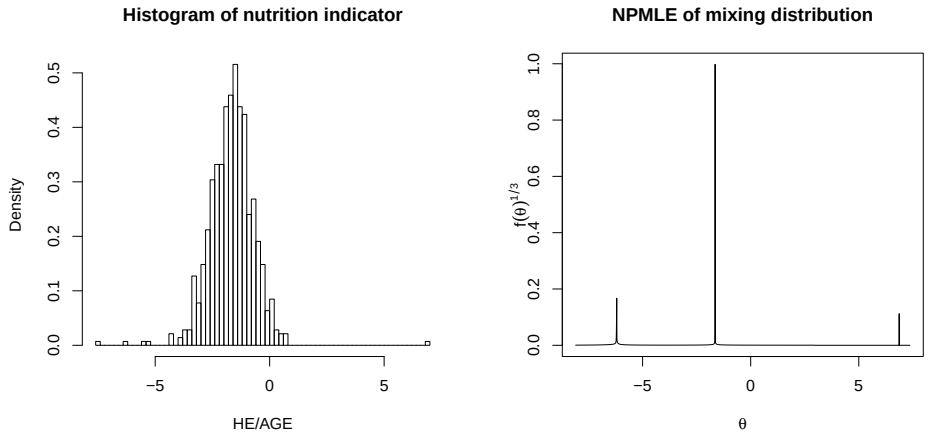

Figure 5. Thai Preschool Children Nutritional Status: The left panel plots the histogram of the HE/AGE data of size 708. The right panel depicts the Kiefer-Wolfowitz nonparametric maximum likelihood estimator of the mixing distribution for the location parameter of the normal mixture model with 1,500 grid points. The cube root of the mass associated with the support points are plotted in an effort to render the small masses more visible.

formulation of the MLE problem à la Kiefer-Wolfowitz can be solved quite efficiently – even for large sample sizes by binning – and effectively used as an alternative testing procedure. We hope that these new developments will encourage others to explore these methods.

References

Andersen, E. D. (2010): "The MOSEK Optimization Tools Manual, Version 6.0," Available from `http://www.mosek.com`.

Azaïs, J.-M., É. Gassiat, and C. Mercadier (2006): "Asymptotic distribution and local power of the log-likelihood ratio test for mixtures: bounded and unbounded cases," *Bernoulli*, 12(5), 775–799.

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

Bickel, P., and H. Chernoff (1993): "Asymptotic distribution of the likelihood ratio statistic in a prototypical nonregular problem," in *Statistics and Probability: A Raghu Raj Bahadur Festschrift*, ed. by J. Ghosh, S. Mitra, K. Parthasarathy, and B. PrakasaRao, pp. 83–96. Wiley, New Delhi.

Böhning, D., P. Schlattmann, and B. Lindsay (1992): "Computer-assisted analysis of mixtures (C.A.MAM): Statistical algorithms," *Biometrics*, 48, 283–303.

Bücher, A., H. Dette, and S. Volgushev (2011): "New estimators of the Pickands dependence function and a test for extreme-value dependence," *The Annals of Statistics*, 39(4), 1963–2006.

Bühler, W., and P. Puri (1966): "On optimal asymptotic tests of composite hypotheses with several constraints," *Probability Theory and Related Fields*, 5, 71–88.

Chen, H., and J. Chen (2001): "Large sample distribution of the likelihood ratio test for normal mixtures," *Canadian Journal of Statistics*, 29, 201–216.

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

Chen, J. (1995): "Optimal rate of convergence for finite mixture models," *The Annals of Statistics*, 23, 221–233.

CHEN, J., AND P. LI (2009): "Hypothesis Test for normal mixture models," *Annals of Statistics*, 37, 2523–2542.

CHEN, J., P. LI, AND Y. LIU (2016): "Sample-size Calculation for Tests of Homogeneity," *Canadian Journal of Statistics*, 44, 82–101.

CHEN, X., M. PONOMAREVA, AND E. TAMER (2014): "Likelihood inference in some finite mixture models," *Journal of Econometrics*, 182, 87–99.

CHESHER, A. (1984): "Testing for Neglected Heterogeneity," *Econometrica*, 52(4), 865–872.

CHO, J., AND H. WHITE (2007): "Testing for Regime Switching," *Econometrica*, 75, 1671–1720.

DICKER, L., AND S. D. ZHAO (2014): "Nonparametric Empirical Bayes and Maximum Likelihood Estimation for High-Dimensional Data Analysis," `http://arxiv.org/pdf/1407.2635`.

EFRON, B. (2011): "Tweedie's Formula and Selection Bias," *Journal of the American Statistical Association*, 106, 1602–1614.

FRIBERG, H. A. (2012): *Rmosek: The R-to-MOSEK Optimization Interface,* R package version 1.2.3.

GASSIAT, E. (2002): "Likelihood ratio inequalities with applications to various mixtures," in *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, vol. 38, pp. 897–906. Elsevier.

GIACOMINI, R., D. POLITIS, AND H. WHITE (2013): "A warp-speed method for conducting monte carlo experiments involving bootstrap estimators," *Econometric Theory*, 29(3), 567–589.

GROENEBOOM, P., G. JONGBLOED, AND J. A. WELLNER (2008): "The support reduction algorithm for computing non-parametric function estimates in mixture models," *Scandinavian Journal of Statistics*, 35, 385–399.

GU, J. (2016): "Neyman's $C(\alpha)$ Test for Unobserved Heterogeneity," *Econometric Theory*, 32(6), 1483–1522.

HALL, P., AND M. STEWART (2005): "Theoretical analysis of power in a two-component normal mixture model," *Journal of statistical planning and inference*, 134, 158–179.

HARTIGAN, J. (1985): "A failure of likelihood asymptotics for normal mixtures," in *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, ed. by L. LeCam, and R. Olshen, pp. 807–810. Wadsworth: Monterey.

HECKMAN, J., AND B. SINGER (1984): "A method for minimizing the impact of distributional assumptions in econometric models for duration data," *Econometrica*, 52, 63–132.

JIANG, W., AND C.-H. ZHANG (2009): "General maximum likelihood empirical Bayes estimation of normal means," *Annals of Statistics*, 37, 1647–1684.

KASAHARA, H., AND K. SHIMOTSU (2014): "Testing the Number of Components in Normal Mixture Regression Models," forthcoming, *Journal of American Statistical Association*.

KIEFER, J., AND J. WOLFOWITZ (1956): "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters," *The Annals of Mathematical Statistics*, 27, 887–906.

KOENKER, R. (2013): *REBayes: Empirical Bayes Estimation and Inference in R,* R package version 0.41.

KOENKER, R., AND I. MIZERA (2014): "Convex Optimization, Shape Constraints, Compound Decisions and Empirical Bayes Rules," *J. of Am. Stat. Assoc.*, 109(506), 674–685.

LAIRD, N. (1978): "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution," *Journal of the American Statistical Association*, 73, 805–811.

LEDOUX, M., AND M. TALAGRAND (1991): *Probability in Banach Spaces: isoperimetry and processes*, vol. 23. Springer Science & Business Media.

LESPERANCE, M. L., AND J. D. KALBFLEISCH (1992): "An algorithm for computing the nonparametric MLE of a mixing distribution," *Journal of the American Statistical Association*, 87, 120–126.

LI, P., AND J. CHEN (2010): "Testing the Order of a Finite Mixture," *Journal of the American Statistical Association*, 105, 1084–1092.

LI, P., J. CHEN, AND P. MARRIOTT (2009): "Non-Finite Fisher Information and Homogeneity: The EM Approach," *Biometrika*, 96, 411–426.

LINDSAY, B. (1981): "Properties of the maximum likelihood estimator of a mixing distribution," in *Statistical Distributions in Scientific Work*, ed. by G. Patil, vol. 5, pp. 95–109. Reidel.

LINDSAY, B. (1983): "The Geometry of Mixture Likelihoods: A General Theory," *Annals of Statistics*, 11, 86–94.

——— (1995): *Mixture Models: Theory, Geometry and Applications.* NSF-CBMS-IMS Conference Series in Statistics, Hayward, CA.

Liu, X., and Y. Shao (2003): "Asymptotics for likelihood ratio tests under loss of identifiability," *Annals of Statistics*, 31(3), 807–832.

McLachlan, G. (1987): "On bootstrapping likelihood ratio test statistics for the number of components in a normal mixture," *Journal of the Royal Statistical Society, Series C*, 36, 318–324.

Moran, P. (1973): "Asymptotic Properties of Homogeneity Tests," *Biometrika*, 60(1), 79–85.

Neyman, J. (1959): "Optimal Asymptotic Tests of Composite Statistical Hypotheses," in *Probability and Statistics, the Harald Cramer Volume*, ed. by U. Grenander. Wiley: New York.

Robbins, H. (1950): "A Generalization of the Method of Maximum Likelihood: Estimating a Mixing Distribution (Abstract)," *The Annals of Mathematical Statistics*, 21, 314.

Saunders, M. A. (2003): "PDCO: A Primal-Dual interior solver for convex optimization," `http://www.stanford.edu/group/SOL/software/pdco.html`.

Tsirel'son, V. (1976): "The density of the distribution of the maximum of a Gaussian process," *Theory of Probability & Its Applications*, 20(4), 847–856.

van der Vaart, A. W. (1998): *Asymptotic statistics*, vol. 3. Cambridge university press.

van der Vaart, A. W., and J. A. Wellner (1996): *Weak Convergence and Empirical Processes - Springer Series in Statistics*. Springer, New York.

Appendix A. Technical details

Sufficient conditions for the expansion in (8) Introduce the notation $\ell(x|\mu) := \log p_{\delta(\mu)}(x)$. The following conditions are sufficient to ensure that the expansion in (8) holds

(1) The true parameter μ_0 is in the interior of Θ and the following expansion holds for any μ in a neighbourhood of μ_0

$$\mathbb{E}[\ell(X_{1,n}|\mu)] = \mathbb{E}[\ell(X_{1,n}|\mu_n)] - \frac{1}{2}(\mu-\mu_n)^T I(\mu_n)(\mu-\mu_n) + r_n(\mu,\mu_n)\|\mu-\mu_n\|^2$$

where $r_n(\theta_n,\mu_n) \to 0$ if $\theta_n - \mu_n \to 0$.

(2) There exists a measurable function M such that for arbitrary μ_1, μ_2 in a neighbourhood of μ_0

$$|\ell(x|\mu_1) - \ell(x|\mu_2)| \leq M(x)\|\mu_1-\mu_2\| \quad a.s.$$

and we have $\limsup_{n\to\infty} \mathbb{E}[M^2(X_{1,n})] < \infty$ and $\mathbb{E}[M^2(X_{1,n})I\{|M(X_{1,n})| > \varepsilon\sqrt{n}\}] \to 0$ for every $\varepsilon > 0$.

(3) The function $\mu \mapsto \log p_{\delta(\mu)}(x)$ is continuously differentiable in a neighbourhood of μ_0 for all x outside of a set M_0 with $P(X_{1,n} \in M_0) = 0$ for all n. The derivative $\ell'(x|\mu)$ satisfies $\mathbb{E}\|\ell'(X_{1,n}|\tilde{\mu}_n) - \ell'(X_{1,n}|\mu_0)\|^2 \to 0$ for any $\tilde{\mu}_n \to \mu_0$.

(4) The maximum likelihood estimators $\hat{\mu}_n := \operatorname{argmax}_\mu \ell_n^*(\delta(\mu))$ satisfy $\hat{\mu}_n = \mu_n + o_P(1)$.

Assumptions (1)-(4) can be viewed as uniform generalizations of the conditions of Theorem 5.23 in van der Vaart (1998). Assumption (1)-(3) can be verified by imposing smoothness assumptions on $\mu \mapsto \ell(x|\mu)$. Details are omitted for the sake of brevity. Assumption (4) requires consistency of the maximum likelihood estimator $\hat{\mu}_n$. Note that Theorem 5.23 and other results in the aforementioned book are not directly applicable in our case since the $X_{i,n}$ are generated from a triangular array while most results in van der Vaart (1998) assume i.i.d. data or data generated under local alternatives. The proof of expansion (8) is based on very similar ideas as the proof of Theorem 5.23, and we will only outline some of the key adjustments that need to be made. First, observe that by following the arguments

of the proof of Corollary 5.53 and Theorem 5.52 in van der Vaart (1998) we obtain

$$\hat{\mu}_n - \mu_0 = O_P(n^{-1/2}). \tag{12}$$

The arguments in the corresponding proofs continue to hold for triangular arrays if we make use of the assumption $\limsup_{n\to\infty} \mathbb{E}[M^2(X_{i,n})] < \infty$. Next, consider the processes

$$\mathbb{G}_{k,n}(h) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} f_{k,n}(X_{i,n}; h) - \mathbb{E}[f_{k,n}(X_{i,n}; h)] \quad k = 1, 2$$

where $f_{1,n}(x; h) := \sqrt{n}(\ell(x|\mu_n + n^{-1/2}h) - \ell(x|\mu_n))$, $f_{2,n}(x; h) := h^T \ell'(x|\mu_n)$. We shall prove that $\sup_{\|h\|\leq 1} |\mathbb{G}_{1,n}(h) - \mathbb{G}_{2,n}(h)| = o_P(1)$. To do so, we prove that the finite-dimensional distributions (from now on fidis) of $\mathbb{G}_{1,n} - \mathbb{G}_{2,n}$ converge to zero and that the process is asymptotically tight. To prove convergence of the fidis, note that for any h with $\|h\| \leq 1$

$$\sqrt{n}(\ell(x|\mu_n + n^{-1/2}h) - \ell(x|\mu_n)) - h^T \ell'(x|\mu_n) = h^T(\ell'(x|\tilde{\mu}_n) - \ell'(x|\mu_n))$$

almost surely where $\|\tilde{\mu}_n - \mu_n\| \leq n^{-1/2}$. Now $\mu_n \to \mu_0$, thus also $\tilde{\mu}_n \to \mu_0$ and it follows that

$$\begin{aligned}\mathbb{E}\|\ell'(X_{1,n}|\tilde{\mu}_n) - \ell'(X_{1,n}|\mu_n)\|^2 \\ \leq 2\mathbb{E}\|\ell'(X_{1,n}|\tilde{\mu}_n) - \ell'(X_{1,n}|\mu_0)\|^2 + 2\mathbb{E}\|\ell'(X_{1,n}|\mu_n) - \ell'(X_{1,n}|\mu_0)\|^2 \to 0.\end{aligned}$$

This implies convergence of the fidis of $\mathbb{G}_{1,n} - \mathbb{G}_{2,n}$ to zero.

Next, we note that tightness of $\mathbb{G}_{2,n}$ is trivial to prove, so that it remains to establish tightness of $\mathbb{G}_{1,n}$. To this end, note that $\sup_{\|h\|\leq 1} |f_{1,n}(x; h)| \leq M(x)$. By Example 19.7 in van der Vaart (1998) the bracketing numbers of the class of functions $\mathcal{F}_n := \{f_{1,n}(\cdot; h) : \|h\| \leq 1\}$ satisfy

$$N_{[\,]}(\varepsilon, \mathcal{F}_n, L_2(P_{\mu_n})) \leq K\|M\|^d_{L_2(P_{\mu_n})}\varepsilon^{-d}$$

for some constants K, d which do not depend on n where the measure P_{μ_n} has density $p(\cdot|\mu_n)$ relative to λ and $\|\cdot\|_{L_2(P_{\mu_n})}$ denotes the corresponding L_2-norm. Combining this with the assumption $\mathbb{E}[M^2(X_{i,n})I\{|M(X_{i,n})| > \varepsilon\sqrt{n}\}] \to 0$ and the proof of Theorem 19.28 in van der Vaart (1998) establishes asymptotic tightness of $\mathbb{G}_{1,n}$. Combined with the arguments in the proof of Lemma 19.31 in van der Vaart (1998) this proves that for any stochastically bounded sequence $\tilde{h}_n$ we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_n(X_{i,n}; \tilde{h}_n) = o_P(1)$$

where

$$\begin{aligned}g_n(x; h) := \sqrt{n}(\ell(x|\mu_n + n^{-1/2}h) - \ell(x|\mu_n)) - h^T \ell'(x|\mu_n) \\ - \mathbb{E}[\sqrt{n}(\ell(X_{1,n}|\mu_n + n^{-1/2}h) - \ell(X_{1,n}|\mu_n)) - h^T \ell'(X_{i,n}|\mu_n)].\end{aligned}$$

This corresponds to the first equation in the proof of Theorem 5.23 in (van der Vaart 1998) if m_θ there is identified with $\ell(\cdot|\mu)$ here and $\dot{m}_\theta$ corresponds to $\ell'(\cdot|\mu)$. Together with (12) this establishes the results corresponding to the first two paragraphs in the proof of Theorem

5.23 in (van der Vaart 1998). The rest of the proof can be done completely analogously and we obtain the following two results

$$\ell_n^*(\delta(\hat{\mu}_n)) - \ell_n^*(\delta(\mu_n)) = -\frac{n}{2}(\hat{\mu}_n - \mu_n)^T I(\mu_n)(\hat{\mu}_n - \mu_n) + (\hat{\mu}_n - \mu_n)^T \sum_{i=1}^{n} \ell'(X_{i,n}|\mu_n) + o_P(1),$$

$$\sqrt{n}(\hat{\mu}_n - \mu_n) = I(\mu_n)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell'(X_{i,n}|\mu_n) + o_P(1).$$

Combining those two results yields the representation in (8).

Proof of (7) Given a measure $\eta \in \mathcal{G}, \eta \neq \delta(0)$ define $V(\eta) := \sum_{k=2}^{\infty} \frac{\kappa_k^2(\eta)}{k!}$. Also, define for $n \in \mathbb{N}$ and $\alpha \in \mathbb{R}$ the probability measure $\tilde{\eta}_n := p_n \delta_{c_n} + (1-p_n)\eta$ with $p_n := 1 - V(\eta)/n$ and $c_n := \frac{1-p_n}{p_n}(\alpha - \kappa_1(\eta))$ [the dependence of p_n, c_n on η is suppressed in the notation]. Note that for any $N > 0$ there exists $n_0 \in \mathbb{N}$ such that for $n \geq n_0$ we have $\tilde{\eta}_n \in \mathcal{G}$ for all $\alpha \in [-N, N]$. Moreover, by construction $\kappa_1(\tilde{\eta}_n) = \alpha(1-p_n)$ and

$$\kappa_k(\tilde{\eta}_n) = \kappa_k(\eta)(1-p_n) + (1-p_n)\Big(\frac{1-p_n}{p_n}\Big)^{k-1}(\alpha - \kappa_1(\eta))^k$$

for $n \in \mathbb{N}$. This implies for $n \geq n_0$ with some n_0 independent of η we have a.s.

$$\begin{aligned}\Big|\alpha Y_1 + \sum_{k=2}^{\infty} \frac{Y_k \kappa_k(\eta)}{(k!)^{1/2}} - \frac{1}{1-p_n}\sum_{k=1}^{\infty} \frac{Y_k \kappa_k(\tilde{\eta}_n)}{(k!)^{1/2}}\Big| &\leq \frac{1-p_n}{p_n}\sum_{k=2}^{\infty} \frac{|Y_k|\tilde{C}^k}{\sqrt{k!}}\Big(\frac{1-p_n}{p_n}\Big)^{k-2} \\ &\leq \frac{2\tilde{C}^2 V(\eta)}{n}\sum_{k=2}^{\infty}\frac{|Y_k|}{\sqrt{k!}}\end{aligned}$$

and

$$\Big|\alpha^2 + \sum_{k=2}^{\infty}\frac{\kappa_k^2(\eta)}{k!} - \frac{1}{(1-p_n)^2}\sum_{k=1}^{\infty}\frac{\kappa_k^2(\tilde{\eta}_n)}{k!}\Big| \leq \frac{CV(\eta)}{n}$$

for finite constants $C, \tilde{C}$ depending only on N but not on α and η [note that $\eta \in \mathcal{G}$ has support contained in $[L, U]$]. Thus for every $N < \infty, \varepsilon > 0$ there exists n_0 independent of η such that for all $n \geq n_0$ we have with probability at least $1 - \varepsilon$

$$\sup_{\eta \in \mathcal{G}} \frac{\sum_{k=1}^{\infty} \frac{Y_k \kappa_k(\eta)}{(k!)^{1/2}}}{\Big(\sum_{k=1}^{\infty}\frac{\kappa_k^2(\eta)}{k!}\Big)^{1/2}} \geq \sup_{\alpha \in [-N,N]}\sup_{\eta \in \mathcal{G}} \frac{\alpha Y_1 + \sum_{k=2}^{\infty}\frac{Y_k\kappa_k(\eta)}{(k!)^{1/2}}}{\Big(\alpha^2 + \sum_{k=2}^{\infty}\frac{\kappa_k^2(\eta)}{k!}\Big)^{1/2}} - \varepsilon.$$

Next, observe that for all $\varepsilon > 0$ there exists $N \in \mathbb{R}$ such that with probability at least $1 - \varepsilon$

$$\sup_{\alpha \in \mathbb{R}\setminus[-N,N]}\sup_{\eta \in \mathcal{G}} \frac{\alpha Y_1 + \sum_{k=2}^{\infty}\frac{Y_k\kappa_k(\eta)}{(k!)^{1/2}}}{\Big(\alpha^2 + \sum_{k=2}^{\infty}\frac{\kappa_k^2(\eta)}{k!}\Big)^{1/2}} \leq |Y_1| + \varepsilon.$$

Finally, note that

$$\sup_{\eta \in \mathcal{G}} \frac{\sum_{k=1}^{\infty} \frac{Y_k \kappa_k(\eta)}{(k!)^{1/2}}}{\Big(\sum_{k=1}^{\infty}\frac{\kappa_k^2(\eta)}{k!}\Big)^{1/2}} \geq |Y_1| \quad \text{a.s.}$$

[consider the sequence of measures $\eta_n = \delta_{sign(Y_1)/n} \in \mathcal{G}$].

Summarizing the findings above, we have shown that for any $\varepsilon > 0$ we have with probability at least $1 - 2\varepsilon$

$$\sup_{\eta \in \mathcal{G}} \frac{\sum_{k=1}^{\infty} \frac{Y_k \kappa_k(\eta)}{(k!)^{1/2}}}{\Big(\sum_{k=1}^{\infty}\frac{\kappa_k^2(\eta)}{k!}\Big)^{1/2}} \geq \sup_{\alpha \in \mathbb{R}}\sup_{\eta \in \mathcal{G}} \frac{\alpha Y_1 + \sum_{k=2}^{\infty}\frac{Y_k\kappa_k(\eta)}{(k!)^{1/2}}}{\Big(\alpha^2 + \sum_{k=2}^{\infty}\frac{\kappa_k^2(\eta)}{k!}\Big)^{1/2}} - \varepsilon.$$

By letting $\varepsilon \to 0$ the above can be turned in an almost sure inequality with no ε on the right-hand side. Finally, setting $\alpha = \kappa_1(\eta)$ we see that the converse inequality also holds almost surely. Thus we have shown that

$$\sup_{\eta\in\mathcal{G}} \frac{\sum_{k=1}^{\infty} \frac{Y_k\kappa_k(\eta)}{(k!)^{1/2}}}{\Big(\sum_{k=1}^{\infty} \frac{\kappa_k^2(\eta)}{k!}\Big)^{1/2}} = \sup_{\alpha\in\mathbb{R}}\sup_{\eta\in\mathcal{G}} \frac{\alpha Y_1 + \sum_{k=2}^{\infty} \frac{Y_k\kappa_k(\eta)}{(k!)^{1/2}}}{\Big(\alpha^2 + \sum_{k=2}^{\infty} \frac{\kappa_k^2(\eta)}{k!}\Big)^{1/2}} \quad \text{a.s.}$$

Define $\beta_k := \frac{\kappa_k(\eta)}{(k!)^{1/2}}$ and

$$g_{Y,\eta}(\alpha) := \frac{\alpha Y_1 + \sum_{k=2}^{\infty} \frac{Y_k\kappa_k(\eta)}{(k!)^{1/2}}}{\Big(\alpha^2 + \sum_{k=2}^{\infty} \frac{\kappa_k^2(\eta)}{k!}\Big)^{1/2}}.$$

Fix a realization of $Y_1, Y_2, ...$ and an $\eta \in \mathcal{G}$. Computing the derivative of $g_{Y,\eta}$ with respect to α shows that the function g has a maximum at $\alpha^* = Y_1 \frac{\sum_{k=2}^{\infty}\beta_k^2}{\sum_{k=2}^{\infty} Y_k\beta_k}$, if $\sum_{k=2}^{\infty} Y_k\beta_k > 0$ and that the supremum of $g_{Y,\eta}$ over $\alpha \in \mathbb{R}$ equals Y_1^2 if $\sum_{k=2}^{\infty} Y_k\beta_k \leq 0$. Some simple algebra shows that for $\sum_{k=2}^{\infty} Y_k\beta_k > 0$ we have

$$g_{Y,\eta}(\alpha^*) = \Big(Y_1^2 + \frac{\Big(\sum_{k=2}^{\infty} Y_k\beta_k\Big)^2}{\sum_{k=2}^{\infty}\beta_k^2}\Big)^{1/2}.$$

Thus we obtain

$$\Big(\sup_{\eta\in\mathcal{G}}\Big(\frac{\sum_{k=1}^{\infty} \frac{Y_k\kappa_k(\eta)}{(k!)^{1/2}}}{\Big(\sum_{k=1}^{\infty} \frac{\kappa_k^2(\eta)}{k!}\Big)^{1/2}}\Big)_+\Big)^2 = Y_1^2 + \sup_{\eta\in\mathcal{G}} \frac{\Big(\Big(\sum_{k=2}^{\infty} \frac{Y_k\kappa_k(\eta)}{(k!)^{1/2}}\Big)_+\Big)^2}{\sum_{k=2}^{\infty} \frac{\kappa_k^2(\eta)}{k!}}$$

and this directly implies (7) □

Proof of Theorem 2.1 The proof of the expansion in (4) is very similar to the proof of (11) in Theorem 2.8, but much simpler since the data are i.i.d. and do not form a triangular array. For this reason we will only sketch the main arguments. First, observe that the class of functions $\mathcal{F} := \{s_{\eta,\mu_0} | \eta \in \mathcal{G}\}$ is $p(\cdot|\mu_0)$-Donsker, and thus $\mathcal{F}^2$ is $p(\cdot|\mu_0)$-Glivenko-Cantelli [see Lemma 2.10.4 in van der Vaart and Wellner (1996)]. Moreover, since $\mathcal{F}$ is $p(\cdot|\mu_0)$-Donsker so is $\mathcal{F}_- := \{s_{\eta,\mu_0,-} | \eta \in \mathcal{G}^\varepsilon\}$ [apply Theorem 2.10.6 in van der Vaart and Wellner (1996)], and thus $\mathcal{F}_-^2$ is also $p(\cdot|\mu_0)$-Glivenko-Cantelli. Hence we obtain

$$\sup_{\eta\in\mathcal{G}} \Big|\frac{1}{n}\sum_{i=1}^{n}(s_{\eta,\mu_0}^2(X_i) - 1)\Big| + \Big|\frac{1}{n}\sum_{i=1}^{n}(s_{\eta,\mu_0,-}^2(X_i) - \|s_{\eta,\mu_0,-}\|_{2,\delta(\mu_0)}^2)\Big| = o_P(1).$$

Thus

$$\lim_{n\to\infty} \inf_{\eta\in\mathcal{G}\setminus\delta(\mu_0)} \frac{1}{n}\sum_{i=1}^{n} s_{\eta,\mu_0,-}^2(X_i) \geq \inf_{\eta\in\mathcal{G}} \|s_{\eta,\mu_0,-}\|_{2,\delta(\mu_0)}^2 > 0$$

where the last inequality follows by the same arguments as (5) in Gassiat (2002). Apply Inequality 1.1 from Gassiat (2002), the lower bound above, and weak convergence of $\mathbb{G}_n$ to

obtain

$$\sup_{\eta\in\mathcal{G},\ell_n(\eta)-\ell_n(\delta(\mu_0))>0}\left\|\frac{p_\eta}{p_{\delta(\mu_0)}}-1\right\|_{2,\delta(\mu_0)}=O_P(n^{-1/2}). \tag{13}$$

Next, note that

$$n^{-1}\sup_{\eta\in\mathcal{G}\setminus\delta(\mu_0)}\Big(\sum_{i=1}^n s_{\eta,\mu_0}(X_i)\Big)^2=\sup_{\eta\in\mathcal{G}}\mathbb{G}_n(\eta)^2=O_P(1). \tag{14}$$

The fact that $\mathcal{F}$ is Donsker and that $\mathbb{E}[s_{\eta,\mu_0}(X_i)]=0$ implies that there must exist an envelope function F of $\mathcal{F}$ with $\max_{i=1,..,n}F(X_i)=o_P(n^{1/2})$, this follows from Corollary 2.3.13 and Problem 2.3.4(iii) of van der Vaart and Wellner (1996). Thus there exists $\alpha_n\to\infty$ such that $\sup_{i=1,...,n}F(X_i)=o_P(\alpha_n^{-1}n^{1/2})$. For such a sequence α_n define the sets

$$M_{n1}:=\{\eta\in\mathcal{G}:\ell_n(\eta)-\ell_n(\delta(\mu_0))>0\},\quad M_{n2}:=\Big\{\eta\in\mathcal{G}:0<\Big\|\frac{p_\eta}{p_{\delta(\mu_0)}}-1\Big\|_{2,\delta(\mu_0)}\le n^{-1/2}\alpha_n^{1/2}\Big\}.$$

Note that

$$\sup_{\eta\in M_{n2}}\Big|\frac{1}{n}\sum_{i=1}^n(s^2_{\eta,\mu_0}(X_i)-1)\Big|\le\sup_{\eta\in\mathcal{G}}\Big|\frac{1}{n}\sum_{i=1}^n(s^2_{\eta,\mu_0}(X_i)-1)\Big|=o_P(1). \tag{15}$$

Now follow the arguments in the proof of Theorem 2.10 which are used to obtain (21) by replacing all instances of μ_n by μ_0, all instances of $X_{i,n}$ by X_i,, all instances of ℓ_n^* by ℓ_n and using equations (13), (14) and (15) instead of (18), (17) and (19) to arrive at the conclusion

$$\sup_{\eta\in\bar{\mathcal{G}}}\ell_n(\eta)-\ell_n(\delta(\mu_0))=\frac{1}{2}\sup_{\eta\in\mathcal{G}\setminus\delta(\mu_0)}\Big(\max\Big\{n^{-1/2}\sum_{i=1}^n s_{\eta,\mu_0}(X_i),0\Big\}\Big)^2+o_P(1). \tag{16}$$

This proves (4), and the rest of the proof follows by a standard application of the multivariate CLT. □

Proof of Theorem 2.10 The proof uses arguments from the proof of Theorem 3.1 in Gassiat (2002). Let $\gamma_n:=\|\mu_n-\mu_0\|$. Observe to each $\eta\in\mathcal{G}$ there exists $\tilde{\eta}\in\mathcal{G}^{\gamma_n}$ such that $\tilde{\eta}_n=\eta$. Thus under (A1) we have

(17)

$$n^{-1}\sup_{\eta\in\mathcal{G}\setminus\delta(\mu_n)}\Big(\sum_{i=1}^n s_{\eta,\mu_n}(X_{i,n})\Big)^2\le n^{-1}\sup_{\eta\in\mathcal{G}^{\gamma_n}}\Big(\sum_{i=1}^n s_{\eta_n,\mu_n}(X_{i,n})\Big)^2\le\sup_{\eta\in\mathcal{G}^\varepsilon}\mathbb{G}_n^*(\eta)^2=O_P(1)$$

where the first inequality holds for n sufficiently large. Moreover

$$\lim_{n\to\infty}\inf_{\eta\in\mathcal{G}\setminus\delta(\mu_n)}\frac{1}{n}\sum_{i=1}^n s^2_{\eta,\mu_n,-}(X_{i,n})\ge\lim_{n\to\infty}\inf_{\eta\in\mathcal{G}^{\gamma_n}}\frac{1}{n}\sum_{i=1}^n s^2_{\eta_n,\mu_n,-}(X_{i,n})\ge\inf_{\eta\in\mathcal{G}^\varepsilon}\|s_{\eta,\mu_0,-}\|^2_{2,\delta(\mu_0)}>0$$

where the second inequality follows by (A2) and the third inequality follows by the same arguments as (5) in Gassiat (2002). Apply Inequality 1.1 from Gassiat (2002) to obtain

$$\sup_{\eta\in\mathcal{G},\ell_n^*(\eta)-\ell_n^*(\delta(\mu_n))>0}\left\|\frac{p_\eta}{p_{\delta(\mu_n)}}-1\right\|_{2,\delta(\mu_n)}=O_P(n^{-1/2}). \tag{18}$$

By assumption (A3) there exist functions F_n such that $\sup_{\eta\in\mathcal{G}}|s_{\eta,\mu_n}(x)| \le F_n(x)$ and $\sup_{i=1,\dots,n} F_n(X_{i,n}) = o_P(n^{-1/2})$. Thus there exists $\alpha_n \to \infty$ such that $\sup_{i=1,\dots,n} F_n(X_{i,n}) = o_P(\alpha_n^{-1}n^{1/2})$. For such a sequence α_n define the sets

$$M_{n1} := \{\eta\in\mathcal{G} : \ell_n^*(\eta) - \ell_n^*(\delta(\mu_n)) > 0\}, \quad M_{n2} := \Big\{\eta\in\mathcal{G} : 0 < \Big\|\frac{p_\eta}{p_{\delta(\mu_n)}} - 1\Big\|_{2,\delta(\mu_n)} \le n^{-1/2}\alpha_n^{1/2}\Big\}.$$

From (18) we obtain that $M_{n1}\subset M_{n2}$ with probability tending to one. On the other hand a Taylor expansion of $x\mapsto\log(1+x)$ shows that

$$\begin{aligned}
&\sup_{\eta\in M_{n2}} \ell_n^*(\eta) - \ell_n^*(\delta(\mu_n)) \\
&= \sup_{\eta\in M_{n2}} \Big(\Big\|\frac{p_\eta}{p_{\delta(\mu_n)}} - 1\Big\|_{2,\delta(\mu_n)} \sum_{i=1}^n s_{\eta,\mu_n}(X_{i,n}) - \frac{1}{2}\Big\|\frac{p_\eta}{p_{\delta(\mu_n)}} - 1\Big\|^2_{2,\delta(\mu_n)} \sum_{i=1}^n s^2_{\eta,\mu_n}(X_{i,n}) \\
&\quad + \Big\|\frac{p_\eta}{p_{\delta(\mu_n)}} - 1\Big\|^2_{2,\delta(\mu_n)} \sum_{i=1}^n s^2_{\eta,\mu_n}(X_{i,n}) R\Big(\Big\|\frac{p_\eta}{p_{\delta(\mu_n)}} - 1\Big\|_{2,\delta(\mu_n)} s_{\eta,\mu_n}(X_{i,n})\Big)\Big)
\end{aligned}$$

where the remainder function R satisfies $R(u)\to 0$ for $u\to 0$. Now by the definition of α_n we have

$$\begin{aligned}
&\sup_{\eta\in M_{n2}} \sum_{i=1}^n s^2_{\eta,\mu_n}(X_{i,n}) R\Big(\Big\|\frac{p_\eta}{p_{\delta(\mu_n)}} - 1\Big\|_{2,\delta(\mu_n)} s_{\eta,\mu_n}(X_{i,n})\Big) \\
&\le \sup_{\eta\in M_{n2}} \sum_{i=1}^n s^2_{\eta,\mu_n}(X_{i,n}) R\Big(n^{-1/2}\alpha_n^{1/2} o_P(\alpha_n^{-1}n^{1/2})\Big) \\
&= o_P(1) \sup_{\eta\in M_{n2}} \sum_{i=1}^n s^2_{\eta,\mu_n}(X_{i,n}).
\end{aligned}$$

Additionally, (A2) implies that

$$\sup_{\eta\in M_{n2}} \Big|\frac{1}{n}\sum_{i=1}^n (s^2_{\eta,\mu_n}(X_{i,n}) - 1)\Big| \le \sup_{\eta\in\mathcal{G}^\varepsilon} \Big|\frac{1}{n}\sum_{i=1}^n (s^2_{\eta_n,\mu_n}(X_{i,n}) - 1)\Big| = o_P(1). \tag{19}$$

Thus we see that

$$\sup_{\eta\in M_{n2}} \ell_n^*(\eta) - \ell_n^*(\delta(\mu_n)) = \sup_{\eta\in M_{n2}} \Big(\Big\|\frac{p_\eta}{p_{\delta(\mu_n)}} - 1\Big\|_{2,\delta(\mu_n)} \sum_{i=1}^n s_{\eta,\mu_n}(X_{i,n}) - \frac{n}{2}\Big\|\frac{p_\eta}{p_{\delta(\mu_n)}} - 1\Big\|^2_{2,\delta(\mu_n)}(1+r_n)\Big)$$

where r_n does not depend on η and $r_n = o_P(1)$. Since $M_{n1}\subset M_{n2}$ with probability tending to one, and since

$$\sup_{\eta\in\bar{\mathcal{G}}} \ell_n^*(\eta) - \ell_n^*(\delta(\mu_n)) = \sup_{\eta\in M_{n1}} \ell_n^*(\eta) - \ell_n^*(\delta(\mu_n)),$$

it follows that

$$\sup_{\eta\in\bar{\mathcal{G}}} \ell_n^*(\eta) - \ell_n^*(\delta(\mu_n)) = \tag{20}$$
$$\sup_{\eta\in M_{n2}} \Big(\Big\| \frac{p_\eta}{p_{\delta(\mu_n)}} - 1 \Big\|_{2,\delta(\mu_n)} \sum_{i=1}^n s_{\eta,\mu_n}(X_{i,n}) - \frac{n}{2} \Big\| \frac{p_\eta}{p_{\delta(\mu_n)}} - 1 \Big\|^2_{2,\delta(\mu_n)} (1+r_n) \Big) + o_P(1).$$

Next observe that under (A0), for any $\eta \in \mathcal{G}\backslash\delta(\mu_n)$ we also have $\eta^t := t\eta + (1-t)\delta(\mu_n) \in \mathcal{G}$ for any $t \in (0,1)$ provided that $\mu_n \in \Theta$. Additionally, we have

$$\Big\| \frac{p_{\eta^t}}{p_{\delta(\mu_n)}} - 1 \Big\|_{2,\delta(\mu_n)} = t \Big\| \frac{p_\eta}{p_{\delta(\mu_n)}} - 1 \Big\|_{2,\delta(\mu_n)}$$

and by construction $s_{\eta^t,\mu_n} \equiv s_{\eta,\mu_n}$. Thus

$$\sup_{\eta\in M_{n2}} \Big(\Big\| \frac{p_\eta}{p_{\delta(\mu_n)}} - 1 \Big\|_{2,\delta(\mu_n)} \sum_{i=1}^n s_{\eta,\mu_n}(X_{i,n}) - \frac{n}{2} \Big\| \frac{p_\eta}{p_{\delta(\mu_n)}} - 1 \Big\|^2_{2,\delta(\mu_n)} (1+r_n) \Big)$$
$$= \sup_{\eta\in\mathcal{G}\backslash\delta(\mu_n)} \sup_{0<t\le c_n(\eta)} \Big(t \Big\| \frac{p_\eta}{p_{\delta(\mu_n)}} - 1 \Big\|_{2,\delta(\mu_n)} \sum_{i=1}^n s_{\eta,\mu_n}(X_{i,n}) - \frac{nt^2}{2} \Big\| \frac{p_\eta}{p_{\delta(\mu_n)}} - 1 \Big\|^2_{2,\delta(\mu_n)} (1+r_n) \Big)$$

where $c_n(\eta) := n^{-1/2}\alpha_n^{1/2} \Big\| \frac{p_\eta}{p_{\delta(\mu_n)}} - 1 \Big\|^{-1}_{2,\delta(\mu_n)}$. As soon as $r_n > -1$, which happens with probability tending to one, the supremum of the inner term over $t > 0$ is attained in the limit $t \to 0$ if $\sum_{i=1}^n s_{\eta,\mu_n}(X_{i,n}) \le 0$ and at

$$t_n(\eta) := \frac{n^{-1}\sum_{i=1}^n s_{\eta,\mu_n}(X_{i,n})}{(1+r_n)\Big\| \frac{p_\eta}{p_{\delta(\mu_n)}} - 1 \Big\|_{2,\delta(\mu_n)}}$$

if $\sum_{i=1}^n s_{\eta,\mu_n}(X_{i,n}) > 0$. Because of (17) it follows that $t_n(\eta) \le c_n(\eta)$ with probability tending to one, so that taken together we have

$$\sup_{\eta\in M_{n2}} \Big(\Big\| \frac{p_\eta}{p_{\delta(\mu_n)}} - 1 \Big\|_{2,\delta(\mu_n)} \sum_{i=1}^n s_{\eta,\mu_n}(X_{i,n}) - \frac{n}{2} \Big\| \frac{p_\eta}{p_{\delta(\mu_n)}} - 1 \Big\|^2_{2,\delta(\mu_n)} (1+r_n) \Big)$$
$$= \frac{1}{2(1+r_n)} \sup_{\eta\in\mathcal{G}\backslash\delta(\mu_n)} \Big(\max\Big\{ n^{-1/2} \sum_{i=1}^n s_{\eta,\mu_n}(X_{i,n}), 0 \Big\} \Big)^2 + o_P(1)$$
$$= \frac{1}{2} \sup_{\eta\in\mathcal{G}\backslash\delta(\mu_n)} \Big(\max\Big\{ n^{-1/2} \sum_{i=1}^n s_{\eta,\mu_n}(X_{i,n}), 0 \Big\} \Big)^2 + o_P(1).$$

Combining this with (20) yields

$$\sup_{\eta\in\bar{\mathcal{G}}} \ell_n^*(\eta) - \ell_n^*(\delta(\mu_n)) = \frac{1}{2} \sup_{\eta\in\mathcal{G}\backslash\delta(\mu_n)} \Big(\max\Big\{ n^{-1/2} \sum_{i=1}^n s_{\eta,\mu_n}(X_{i,n}), 0 \Big\} \Big)^2 + o_P(1). \tag{21}$$

Recall that for each $\eta \in \mathcal{G}$ there exists $\tilde{\eta} \in \mathcal{G}^{\gamma_n}$ such that $\eta = \tilde{\eta}_n$. Thus

$$
\begin{aligned}
&\Big| \sup_{\eta\in\mathcal{G}\setminus\delta(\mu_n)} \Big(\max\Big\{n^{-1/2}\sum_{i=1}^n s_{\eta,\mu_n}(X_{i,n}),0\Big\}\Big)^2 - \sup_{\eta\in\mathcal{G}} \Big(\max\Big\{n^{-1/2}\sum_{i=1}^n s_{\eta_n,\mu_n}(X_{i,n}),0\Big\}\Big)^2 \Big| \\
&\leq \sup_{\nu\in\mathcal{G}\setminus\delta(\mu_n)} \inf_{\eta\in\mathcal{G}} \Big| \Big(n^{-1/2}\sum_{i=1}^n s_{\nu,\mu_n}(X_{i,n})\Big)^2 - \Big(n^{-1/2}\sum_{i=1}^n s_{\eta_n,\mu_n}(X_{i,n})\Big)^2 \Big| \\
&\leq \sup_{\nu\in\mathcal{G}^{\gamma_n}} \inf_{\eta\in\mathcal{G}} \Big| \Big(n^{-1/2}\sum_{i=1}^n s_{\nu_n,\mu_n}(X_{i,n})\Big)^2 - \Big(n^{-1/2}\sum_{i=1}^n s_{\eta_n,\mu_n}(X_{i,n})\Big)^2 \Big| \\
&= \sup_{\nu\in\mathcal{G}^{\gamma_n}} \inf_{\eta\in\mathcal{G}} \Big| (\mathbb{G}_n^*)^2(\nu) - (\mathbb{G}_n^*)^2(\eta) \Big| \\
&\leq 2\Big(\sup_{\nu\in\mathcal{G}^{\gamma_n}} |\mathbb{G}_n^*(\nu)| \Big) \Big(\sup_{\nu\in\mathcal{G}^{\gamma_n}} \inf_{\eta\in\mathcal{G}} \Big| \mathbb{G}_n^*(\nu) - \mathbb{G}_n^*(\eta) \Big| \Big) = o_P(1)
\end{aligned} \tag{22}
$$

The $o_P(1)$ in last line above follows from assumption (A1). More precisely, note that by the Continuous Mapping Theorem applied to the map $f \mapsto \sup_{\eta\in\mathcal{G}^\varepsilon} \inf_{\tilde{\eta}\in\mathcal{G}} |f(\eta) - f(\tilde{\eta})|$ we have for any fixed $\varepsilon > 0$

$$
\sup_{\eta\in\mathcal{G}^\varepsilon} \inf_{\tilde{\eta}\in\mathcal{G}} |\mathbb{G}_n^*(\eta) - \mathbb{G}_n^*(\tilde{\eta})| \;\rightsquigarrow\; \sup_{\eta\in\mathcal{G}^\varepsilon} \inf_{\tilde{\eta}\in\mathcal{G}} |\mathbb{G}^*(\eta) - \mathbb{G}^*(\tilde{\eta})|.
$$

Thus for arbitrary $\varepsilon > 0, t > 0$ we have

$$
\limsup_{n\to\infty} P\Big(\sup_{\eta\in\mathcal{G}^{\gamma_n}} \inf_{\tilde{\eta}\in\mathcal{G}} |\mathbb{G}_n^*(\eta) - \mathbb{G}_n^*(\tilde{\eta})| \leq t \Big) \leq P\Big(\sup_{\eta\in\mathcal{G}^\varepsilon} \inf_{\tilde{\eta}\in\mathcal{G}} |\mathbb{G}^*(\eta) - \mathbb{G}^*(\tilde{\eta})| \leq t \Big),
$$

and the right-hand side can be made arbitrarily small by letting $\varepsilon \downarrow 0$. This shows that

$$
\sup_{\nu\in\mathcal{G}^{\gamma_n}} \inf_{\eta\in\mathcal{G}} \Big| \mathbb{G}_n^*(\nu) - \mathbb{G}_n^*(\eta) \Big| = o_P(1).
$$

Now equations (21), (22) yield

$$
2\sup_{\eta\in\bar{\mathcal{G}}}(\ell_n^*(\eta) - \ell_n^*(\delta(\mu_n))) = \sup_{\eta\in\mathcal{G}} \Big(\max\Big\{\mathbb{G}_n^*(\eta),0\Big\}\Big)^2 + o_P(1),
$$

and the first assertion of the theorem follows. The second assertion follows by an application of the continuous mapping theorem. □

Proof of Theorem 2.11 First we observe that $\mathbb{G}$ is the limit of $\mathbb{G}_n$ under weak convergence in $\ell^\infty(\mathcal{G} \setminus \delta(\mu_0))$ and thus tight. Next, note that $\|Y\|^2 > 0$ almost surely. On the other hand, $L_n \geq 0$ almost surely for each n. Since R is the weak limit of $2L_n$, it follows that $R \geq 0$ almost surely. Thus $\sup_\eta(\max\{\mathbb{G}(\eta),0\})^2 > 0$ almost surely, and it follows $\max(0, \sup_\eta \mathbb{G}(\eta)) = \sup_\eta \mathbb{G}(\eta)$ almost surely.

The proof of the first assertion [properties of F_R] consists of three steps. First, we show that the distribution of R is continuous on $(0,\infty)$ (**Claim 2**). Second, we provide a lower bound for $P(R > 0)$. Define

$$
F_y(t) := P\Big(\sup_\eta \mathbb{G}(\eta) \leq t \Big| Y = y \Big).
$$

We begin by proving a preliminary result.

Claim 1: *For any $y \in \mathbb{R}^d$, $F_y(\cdot)$ is continuous on $(\|y\|, \infty)$.*

Observe that by the joint normality of $(\mathbb{G}(\eta))_{\eta\in\mathcal{G}}, Y$ the conditional distribution of $(\mathbb{G}(\eta))_{\eta\in\mathcal{G}}$ given $Y = y$ is that of a tight Gaussian random element with mean $\mathbb{E}[\mathbb{G}(\eta)Y^\top]y$ and a covariance function κ that does not depend on y. Let $\tilde{\mathbb{G}}$ denote a centered Gaussian process with covariance function κ. Then the conditional distribution of $\mathbb{G}$ given $Y = y$ and the distribution of $(\tilde{\mathbb{G}}(\eta) + \mathbb{E}[\mathbb{G}(\eta)Y^\top]y)_{\eta\in\mathcal{G}}$ coincide.

Since $\tilde{\mathbb{G}}$ is a centered, tight Gaussian process, it follows by the arguments given on page 60-61 of Ledoux and Talagrand (1991) that $\sup_\eta |\tilde{\mathbb{G}}(\eta)|$ has a continuous distribution on $\mathbb{R}$ with left support point at 0, so that $P(\sup_\eta |\tilde{\mathbb{G}}(\eta)| < \varepsilon) > 0$ for all $\varepsilon > 0$. Since $P(\sup_\eta \tilde{\mathbb{G}}(\eta) < \varepsilon) \geq P(\sup_\eta |\tilde{\mathbb{G}}(\eta)| < \varepsilon)$ it follows that also $P(\sup_\eta \tilde{\mathbb{G}}(\eta) < \varepsilon) > 0$ for all $\varepsilon > 0$.

According to Tsirel'son (1976), the distribution of $\sup_\eta(\mathbb{E}[\mathbb{G}(\eta)Y^\top]y+\tilde{\mathbb{G}}(\eta))$ can only have a jump at the left endpoint of it's support and has a density to the right of that point. On the other hand, $|\mathbb{E}[\mathbb{G}(\eta)Y^\top]y| \leq \|\mathbb{E}[\mathbb{G}(\eta)Y]\|\|y\| \leq \|y\|$. Here, the second inequality follows since $\mathbb{G}(\eta), Y$ are jointly Gaussian so that there exist a_η, b_η with $(\mathbb{G}(\eta), Y) \stackrel{\mathcal{D}}{=} (a_\eta^\top Y + b_\eta Z, Y)$ for $Z \sim \mathcal{N}(0,1)$ independent of Y. As $Y \sim \mathcal{N}(0, I_d)$ we have $1 = Var(\mathbb{G}(\eta)) = \|a_\eta\|^2 + b_\eta^2 \geq \|a_\eta\|^2$ and moreover $\|\mathbb{E}[\mathbb{G}(\eta)Y]\| = \|a_\eta\|$.

Thus for $\varepsilon > 0, y \in \mathbb{R}^d$

$$P\Big(\sup_\eta\{\mathbb{E}[\mathbb{G}(\eta)Y^\top]y + \tilde{\mathbb{G}}(\eta)\} - \|y\| \leq \varepsilon\Big) = P\Big(\sup_\eta\{\mathbb{E}[\mathbb{G}(\eta)Y^\top]y - \|y\| + \tilde{\mathbb{G}}(\eta)\} \leq \varepsilon\Big)$$
$$\geq P\Big(\sup_\eta \tilde{\mathbb{G}}(\eta) \leq \varepsilon\Big) > 0.$$

Thus for all $y \in \mathbb{R}^d$ the distribution of $\sup_\eta(\mathbb{E}[\mathbb{G}(\eta)Y^\top]y + \tilde{\mathbb{G}}(\eta))$ has a density on $(\|y\|, \infty)$ and Claim 1 follows.

Claim 2: *The distribution of $(\sup_\eta \mathbb{G}(\eta))^2 - \|Y\|^2$ is continuous on $(0, \infty)$.*

Let $0 < a < b$. Then by continuity of F_y on $(\|y\|, \infty)$

$$P\Big((\sup_\eta \mathbb{G}(\eta))^2 - \|Y\|^2 \in [a,b]\Big) = \int_{\mathbb{R}} P\Big((\sup_\eta \mathbb{G}(\eta))^2 - \|Y\|^2 \in [a,b]\Big|Y = y\Big)\phi_d(y)dy$$
$$= \int_{\mathbb{R}} \Big(F_y((\|y\|^2 + b)^{1/2}) - F_y((\|y\|^2 + a)^{1/2})\Big)\phi_d(y)dy.$$

Now for $a \uparrow b > 0$ we have for every $y \in \mathbb{R}^d$ that $F_y((\|y\|^2 + b)^{1/2}) - F_y((\|y\|^2 + a)^{1/2}) \to 0$ since $(\|y\|^2 + b)^{1/2} > \|y\|^2$ is a continuity point of F_y. Thus the integral converges to zero by dominated convergence. Since $b > 0$ was arbitrary the assertion follows.

Claim 3: *For $d = 1$ $P((\sup_\eta \mathbb{G}(\eta))^2 - Y^2 > 0) \geq 1/4$.*

By assumption there exists $\eta_0 \in \mathcal{G}$ such that $|\mathbb{E}[\mathbb{G}(\eta_0)Y]| \neq 1$. Moreover,

$$P((\sup_{\eta} \mathbb{G}(\eta))^2 - Y^2 > 0) \geq P(|\mathbb{G}(\eta_0)| > |Y|) = 1/4.$$

Here, the last inequality follows since $(\mathbb{G}(\eta_0), Y)$ is a two-dimensional, centered Gaussian vector with $\mathbb{E}[(\mathbb{G}(\eta_0))^2] = \mathbb{E}[(Y)^2]$ and correlation in $(-1, 1)$.

The continuity of F_R on $(0, +\infty)$ and the bound $F_R(0) \leq 3/4$ in the case $d = 1$ follow by combining Claim 2 and Claim 3.

It remains to establish the convergence $P(L_n > q_{n,1-\alpha}) \to \alpha$ in cases where $P(R > 0) > \alpha$. Under the assumptions of the theorem, the maximum likelihood estimator $\hat{\mu}$ converges to μ_0 in probability. Arguing along subsequences, we can without loss of generality assume that the convergence takes place almost surely.

In what follows, denote by $\hat{F}_{n,B}$ the empirical distribution function of $L_{n,1}, ..., L_{n,B}$ and by F_n the true distribution function of $L_{n,1}$ conditionally on $\hat{\mu} = \mu_n$. Note that conditionally on $\hat{\mu} = \mu_n$ the quantities $L_{n,1}, ..., L_{n,B}$ constitute an i.i.d. sample from F_n. By the uniform version of the Glivenko-Cantelli Theorem [see Theorem 2.8.1 in van der Vaart and Wellner (1996)] it follows that $\sup_{t\in\mathbb{R}} |\hat{F}_{n,B}(t) - F_n(t)| \to 0$ in probability, unconditionally. Additionally, the almost sure convergence $\hat{\mu} \to \mu_0$ together with Theorem 2.10 yields weak convergence of $L_{n,1}$ to R, so that F_n converges to F_R at all continuity points of F_R almost surely. Thus we obtain that $\hat{F}_{n,B}$ converges to F_R at all continuity points of F_R in probability, and since $\hat{F}_{n,B}, F_R$ are increasing and F_R is continuous on $(0, \infty)$, $\sup_{x\in K} |\hat{F}_{n,B}(x) - F_R(x)|$ converges to zero in probability for compact $K \subset (0, \infty)$. By arguments similar to the ones given in Lemma 21.2 in van der Vaart (1998) we obtain that $\hat{q}_{n,u} = \hat{F}^{-1}_{n,B}(u) \to F^{-1}_R(u)$ in probability for all u where F^{-1}_R is continuous. Note that F^{-1}_R is increasing, and thus the set of its continuity points is dense in $[F_R(0), 1]$. Moreover, $1 - \alpha \in (F_R(0), 1)$. Thus for every $\varepsilon > 0$ there exist $1 - \alpha_1 \leq 1 - \alpha \leq 1 - \alpha_2$ such that F^{-1}_R is continuous at $1 - \alpha_1, 1 - \alpha_2$ and $|\alpha_i - \alpha| \leq \varepsilon$. By Slutzky's Lemma we obtain $L_n - \hat{F}^{-1}_{n,B}(1 - \alpha_i) \rightsquigarrow R - F^{-1}_R(1 - \alpha_i)$, and by continuity of F_R in a neighborhood of $F^{-1}_R(1 - \alpha)$ and monotonicity of $\hat{F}^{-1}_{n,B}$ it follows that

$$\begin{aligned} 1 - \alpha_1 &= P(R - F^{-1}_R(1 - \alpha_1) \leq 0) \leq \liminf_{n\to\infty} P(L_n \leq \hat{q}_{n,1-\alpha}) \leq \limsup_{n\to\infty} P(L_n \leq \hat{q}_{n,1-\alpha}) \\ &\leq P(R - F^{-1}_R(1 - \alpha_2) \leq 0) = 1 - \alpha_2. \end{aligned}$$

Since α_i above can be chosen to be arbitrarily close to α the claim follows. □

Proof of Proposition 2.6 Note that the special structure of $p(\cdot|\mu)$ implies that $X_{1,n} \overset{\mathcal{D}}{=} X_1 - \mu_0 + \mu_n$ [recall that $X_{1,n} \sim p(\cdot|\mu_n), X_1 \sim p(\cdot|\mu_0)$]. On the other hand

$$p_{\eta_n}(x) = \int p(x - \mu) d\eta_n(\mu) = \int p(x - \mu + \mu_0 - \mu_n) d\eta(\mu) = p_\eta(x + \mu_0 - \mu_n).$$

Thus also $s_{\eta_n,\mu_n}(x) = s_{\eta,\mu}(x + \mu_0 - \mu_n)$ and in particular $s_{\eta_n,\delta(\mu_n)}(X_{1,n}) \overset{\mathcal{D}}{=} s_{\eta,\delta(\mu_0)}(X_1)$. This in turn implies that for any measure $\eta \in \mathcal{G}^\varepsilon$ we have by definition $\mathbb{G}^*_n(\eta) \overset{\mathcal{D}}{=} \mathbb{G}(\eta)$. Assuming

that μ_0 is an interior point of Θ, similar computations show that $\ell'(X_{i,n}|\mu_n) \stackrel{\mathcal{D}}{=} \ell'(X_i|\mu_0)$ and $\|\ell'(\cdot|\mu_n)\|_{2,\delta(\mu_n)} = \|\ell'(\cdot|\mu_0)\|_{2,\delta(\mu_0)}$. Thus, the first part of (A1) follows.

To verify assumption (A2), observe that $\mathbb{G}_n$ can be identified with the empirical process based on the observations $X_1, ..., X_n$ and indexed by the class of functions $\mathcal{F} := \{s_{\eta,\mu_0}|\eta \in \mathcal{G}^\varepsilon\}$. Weak convergence of $\mathbb{G}_n$ implies that the class $\mathcal{F}$ is $p(\cdot|\mu_0)$-Donsker, and thus $\mathcal{F}^2$ is $p(\cdot|\mu_0)$-Glivenko-Cantelli [see Lemma 2.10.4 in (van der Vaart and Wellner 1996)]. Moreover, since $\mathcal{F}$ is $p(\cdot|\mu_0)$-Donsker so is $\mathcal{F}_- := \{s_{\eta,\mu_0,-}|\eta \in \mathcal{G}^\varepsilon\}$ [apply Theorem 2.10.6 in (van der Vaart and Wellner 1996)], and thus $\mathcal{F}_-^2$ is also $p(\cdot|\mu_0)$-Glivenko-Cantelli. This shows that (A2) holds.

For assumption (A3), note that for every $\eta \in \mathcal{G}$ there exists $\tilde{\eta} \in \mathcal{G}^\varepsilon$ with $\tilde{\eta}_n = \eta$ provided that $\|\mu_n - \mu_0\| \leq \varepsilon$. Thus $s_{\eta_n,\mu_n}(x) = s_{\eta,\mu_0}(x + \mu_0 - \mu_n)$ implies that for any $x \in \mathbb{R}$

$$\sup_{f\in\mathcal{F}_n} |f(x)| \leq \sup_{\eta\in\mathcal{G}^\varepsilon} |s_{\eta_n,\mu_n}(x)| = \sup_{\eta\in\mathcal{G}^\varepsilon} |s_{\eta,\mu_0}(x + \mu_0 - \mu_n)|.$$

Thus if F is an envelope for $\mathcal{F}^\varepsilon := \{s_{\eta,\mu_0}|\eta \in \mathcal{G}^\varepsilon\}$ then $F_n(\cdot) := F(\cdot + \mu_0 - \mu_n)$ is an envelope for $\mathcal{F}_n$. On the other hand, the fact that $\mathcal{F}^\varepsilon$ is Donsker and that $\mathbb{E}[s_{\eta,\mu_0}(X_i)] = 0$ implies that there must exist an envelope function F of $\mathcal{F}^\varepsilon$ with $\max_{i=1,..,n} F(X_i) = o_P(n^{1/2})$, this follows from Corollary 2.3.13 and Problem 2.3.4(iii) of van der Vaart and Wellner (1996). Moreover, $F_n(X_{i,n}) \stackrel{\mathcal{D}}{=} F(X_i)$ and thus (A3) follows. □

Appendix B. Verification of Assumptions (A1) - (A3) for Poisson Mixtures

Assume that $\Theta = [a, b]$ for some $0 < a < b$ and that the densities p take the form $p(x|\mu) = \mu^x e^{-\mu}/x!$ with respect to the counting measure on $\mathbb{N}$. As stated in Section 3.3 of Azaïs, Gassiat, and Mercadier (2009), the likelihood ratios have the following representation

$$\frac{p_{\eta_n}(x)}{p_{\delta(\mu_n)}(x)} - 1 = \sum_{k=1}^{\infty} \frac{k\mathbb{E}[(Z-\mu_n)^k]}{(k!\mu_n^k)^{1/2}} \frac{C_k(x|\mu_n)}{k} =: \sum_{k=1}^{\infty} a_k(\eta_n, \mu_n)\frac{C_k(x|\mu_n)}{k} \tag{23}$$

where $Z \sim \eta_n$. Here, the functions $x \mapsto C_k(x|\mu_n)$ are polynomials of order k which are given by

$$C_k(x|\mu_n) := \frac{\mu_n^{k/2}}{(k!)^{1/2}} \Big[\frac{d^k}{dz^k}\Big(\frac{z}{\mu_n}\Big)^x \exp(-z + \mu_n)\Big]_{z=\mu_n}.$$

The functions $(x \mapsto C_k(x|\mu_n))_{k\in\mathbb{N}}$ are centered and orthonormal with respect to $P_{\delta(\mu_n)}$, i.e. for $k, \ell \in \mathbb{N}$

$$\mathbb{E}[C_k(X_{1,n}|\mu_n)] = 0, \quad \mathbb{E}[C_k(X_{1,n}|\mu_n)C_\ell(X_{1,n}|\mu_n)] = I\{k = \ell\}. \tag{24}$$

In particular, we have that

$$1 = \mathbb{E}[C_k^2(X_{1,n}|\mu_n)] = \sum_{u\geq 0} C_k^2(u|\mu_n)e^{-\mu_n}\mu_n^u/u! \geq C_k^2(x|\mu_n)e^{-\mu_n}\mu_n^x/x! \quad \forall\ x \in \mathbb{N}_0$$

so that the series in (23) converges pointwise. The score functions s_{η_n,μ_n} can be represented as

$$s_{\eta_n,\mu_n}(x) = \sum_{k=1}^{\infty} \frac{a_k(\eta_n, \mu_n)C_k(x|\mu_n)}{kw(\eta_n, \mu_n)}, \quad w(\eta_n, \mu_n) := \Big(\sum_{\ell=1}^{\infty} \ell^{-2}a_\ell^2(\eta_n, \mu_n)\Big)^{1/2}. \tag{25}$$

For $L \geq 2$, define the approximating function

$$s^{(L)}_{\eta_n,\mu_n}(x) = \sum_{k=1}^{L} \frac{a_k(\eta_n,\mu_n)C_k(x|\mu_n)}{kw^{(L)}(\eta_n,\mu_n)}, \quad w^{(L)}(\eta_n,\mu_n) := \Big(\sum_{\ell=1}^{L} \ell^{-2} a_\ell^2(\eta_n,\mu_n)\Big)^{1/2}.$$

Obviously, the function $x \mapsto s^{(L)}_{\eta_n,\mu_n}(x)$ is a polynomial of degree L. Later, we will prove the following identities holding for $L \geq 2$, some finite n_0 and a constant C independent of n, η_n, μ_n, μ_0

$$\sup_{\eta\in\mathcal{G}^\varepsilon} \sup_{n\geq n_0} \Big|\frac{w^{(L)}(\eta_n,\mu_n)}{w(\eta_n,\mu_n)} - 1\Big| \leq CL^{-1}, \quad \sup_{\eta\in\mathcal{G}^\varepsilon} \Big|\frac{w^{(L)}(\eta,\mu_0)}{w(\eta,\mu_0)} - 1\Big| \leq CL^{-1}, \tag{26}$$

$$\sum_{k\geq 2} a_k^2(\eta_n,\mu_n) \leq Ca_2^2(\eta_n,\mu_n). \tag{27}$$

Additionally, for any fixed k one obtains by straightforward calculations

$$\sup_{\eta\in\mathcal{G}^\varepsilon} |a_k(\eta_n,\mu_n) - a_k(\eta,\mu_0)| \to 0, \quad n\to\infty, \tag{28}$$

and for any fixed $L \geq 2$ [this will be proved later]

$$\sup_{\eta\in\mathcal{G}^\varepsilon} \Big|\frac{w^{(L)}(\eta_n,\mu_n)}{w^{(L)}(\eta,\mu_0)} - 1\Big| \to 0, \quad n\to\infty. \tag{29}$$

Assumption (A3) can be verified by a straightforward extension of the arguments in the proof of Theorem 4 of Azaïs, Gassiat, and Mercadier (2009). Details are omitted for the sake of brevity. In the proofs that follow, we will repeatedly use (A3).

Verification of Assumption (A1). To establish assertion (A1), it suffices to prove asymptotic tightness of the process $\mathbb{G}_n^*$ in $\ell^\infty(\mathcal{G}_\varepsilon)$ and that weak convergence

$$\Big(\mathbb{G}_n^*(\eta_1),...,\mathbb{G}_n^*(\eta_k), \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \|\ell'(\cdot|\mu_0)\|^{-1}_{2,\delta(\mu_0)}\ell'(X_{i,n}|\mu_n)\Big) \rightsquigarrow (\mathbb{G}(\eta_1),...,\mathbb{G}(\eta_k),Y_1)$$

holds for any fixed collection of measures $\eta_1,....\eta_k$. The weak convergence above follows by straightforward arguments, and we will only provide the details for establishing tightness. To prove asymptotic tightness of $\mathbb{G}_n^*$, we will prove that $\mathbb{G}_n^* \rightsquigarrow \mathbb{G}$. For $L \geq 2$ define

$$\mathbb{G}^{(L)}(\eta) := \sum_{k=1}^{L} \frac{a_k(\eta,\mu_0)Z_k}{kw^{(L)}(\eta,\mu_0)}, \quad \mathbb{G}(\eta) := \sum_{k=1}^{\infty} \frac{a_k(\eta,\mu_0)Z_k}{kw(\eta,\mu_0)}$$

where $Z_1, Z_2, ...$ i.i.d. $\sim \mathcal{N}(0,1)$. In what follows, define for an arbitrary function $f : \mathbb{R} \to \mathbb{R}$ with $\mathbb{E}|f(X_{1,n})| < \infty$

$$\mathbb{F}_n f := \frac{1}{n^{1/2}}\sum_{i=1}^{n}(f(X_{i,n}) - \mathbb{E}[f(X_{i,n})]).$$

Note that by construction $\mathbb{G}_n^*(\eta) = \mathbb{F}_n s_{\eta_n,\mu_n}$. By an application of Lemma B.1 from Bücher, Dette, and Volgushev (2011), weak convergence of $\mathbb{G}_n^*$ to $\mathbb{G}$ follows from the following three claims:

(i) For every $L \geq 2$ we have $(\mathbb{F}_n s^{(L)}_{\eta_n,\mu_n})_{\eta\in\mathcal{G}^\varepsilon} \rightsquigarrow (\mathbb{G}^{(L)})_{\eta\in\mathcal{G}^\varepsilon}$ as $n\to\infty$.

(ii) $\mathbb{G}^{(L)} \rightsquigarrow \mathbb{G}$ as $L \to \infty$.

(iii) For every $\delta > 0$ we have [with P^* denoting outer probability]

$$\lim_{L\to\infty} \limsup_{n\to\infty} P^*\Big(\sup_{\eta\in\mathcal{G}^\varepsilon} |\mathbb{F}_n s^{(L)}_{\eta_n,\mu_n} - \mathbb{F}_n s_{\eta_n,\mu_n}| > \delta\Big) = 0.$$

For a proof of (iii) note that

$$\begin{aligned}\mathbb{F}_n s_{\eta_n,\mu_n} - \mathbb{F}_n s^{(L)}_{\eta_n,\mu_n} =& \Big(1 - \frac{w(\eta_n,\mu_n)}{w^{(L)}(\eta_n,\mu_n)}\Big) \sum_{k=1}^{\infty} \frac{a_k(\eta_n,\mu_n)}{k w(\eta_n,\mu_n)} \mathbb{F}_n C_k(\cdot|\mu_n) \\ &+ \frac{w(\eta_n,\mu_n)}{w^{(L)}(\eta_n,\mu_n)} \sum_{k=L+1}^{\infty} \frac{a_k(\eta_n,\mu_n)}{k w(\eta_n,\mu_n)} \mathbb{F}_n C_k(\cdot|\mu_n) \\ =:& A^{(L)}_n(\eta_n,\mu_n) + B^{(L)}_n(\eta_n,\mu_n).\end{aligned}$$

The first term in the above decomposition can be bounded as follows

$$\begin{aligned}\sup_{\eta\in\mathcal{G}^\varepsilon} |A^{(L)}_n(\eta_n,\mu_n)| =& \sup_{\eta\in\mathcal{G}^\varepsilon} \Big|\Big(1 - \frac{w(\eta_n,\mu_n)}{w^{(L)}(\eta_n,\mu_n)}\Big) \sum_{k=1}^{\infty} \frac{a_k(\eta_n,\mu_n)}{k w(\eta_n,\mu_n)} \mathbb{F}_n C_k(\cdot|\mu_n)\Big| \\ \le& CL^{-1}\Big(\sum_{k=1}^{\infty} \frac{(\mathbb{F}_n C_k(\cdot|\mu_n))^2}{k^2}\Big)^{1/2} \sup_{\eta\in\mathcal{G}^\varepsilon} \Big(\sum_{k=1}^{\infty} \frac{a_k^2(\eta_n,\mu_n)}{w^2(\eta_n,\mu_n)}\Big)^{1/2} \\ \le& CL^{-1}\Big(\sum_{k=1}^{\infty} \frac{(\mathbb{F}_n C_k(\cdot|\mu_n))^2}{k^2}\Big)^{1/2} \sup_{\eta\in\mathcal{G}^\varepsilon} \Big(\frac{a_1^2(\eta_n,\mu_n) + Ca_2^2(\eta_n,\mu_n)}{a_1^2(\eta_n,\mu_n) + a_2^2(\eta_n,\mu_n)/4}\Big)^{1/2} \\ \le& \tilde{C}L^{-1}\Big(\sum_{k=1}^{\infty} \frac{(\mathbb{F}_n C_k(\cdot|\mu_n))^2}{k^2}\Big)^{1/2},\end{aligned}$$

where the first inequality follows from (26) and the second inequality from (27). Since $\mathbb{E}[(\mathbb{F}_n C_k(\cdot|\mu_n))^2] = 1$ for all $k \in \mathbb{N}$ by the orthonormality of the $(C_k(\cdot|\mu_n))_{k\in\mathbb{N}}$, we obtain

$$\lim_{L\to\infty} \limsup_{n\to\infty} \mathbb{E}\Big| \sup_{\eta\in\mathcal{G}^\varepsilon} A^{(L)}_n(\eta_n,\mu_n)\Big|^2 = 0.$$

By similar arguments as above we also obtain the bound

$$\begin{aligned}\sup_{\eta\in\mathcal{G}^\varepsilon} |B^{(L)}_n(\eta_n,\mu_n)| &\le C_1\Big(\sum_{k=L+1}^{\infty} \frac{(\mathbb{F}_n C_k(\cdot|\mu_n))^2}{k^2}\Big)^{1/2} \sup_{\eta\in\mathcal{G}^\varepsilon} \Big(\frac{w(\eta_n,\mu_n)}{w^{(L)}(\eta_n,\mu_n)}\Big) \\ &\le C_2\Big(\sum_{k=L+1}^{\infty} \frac{(\mathbb{F}_n C_k(\cdot|\mu_n))^2}{k^2}\Big)^{1/2}\end{aligned}$$

where the last inequality holds for n sufficiently large. Thus

$$\lim_{L\to\infty} \limsup_{n\to\infty} \mathbb{E}\Big| \sup_{\eta\in\mathcal{G}^\varepsilon} B^{(L)}_n(\eta_n,\mu_n)\Big|^2 \le \lim_{L\to\infty} C_2 \sum_{k=L+1}^{\infty} \frac{1}{k^2} = 0.$$

and assertion (iii) follows. Assertion (ii) can be proved by similar arguments with Z_k replacing $\mathbb{F}_n C_k(\cdot|\mu_n)$ and the arguments are omitted for brevity. For the proof of assertion

(i), note that for any fixed L it is easy to verify that

$$(\mathbb{F}_n C_1(\cdot|\mu_n), ..., \mathbb{F}_n C_L(\cdot|\mu_n)) \ \rightsquigarrow \ (Z_1, ..., Z_L).$$

To see this, recall that the $C_k(\cdot|\mu_n)$ are polynomials and that for $\mu_n \to \mu_0$ the coefficients of $C_k(\cdot|\mu_n)$ converge to those of $C_k(\cdot|\mu_0)$. Weak convergence of $(\mathbb{F}_n s^{(L)}_{\eta_n,\mu_n})_{\eta\in\mathcal{G}^\varepsilon}$ follows by the extended continuous mapping theorem [see Theorem 1.11.1 in van der Vaart and Wellner (1996)] applied to the maps [to verify the conditions of the continuous mapping theorem, make use (28)-(29)]

$$g_n : (x_1, ..., x_L) \mapsto \Big(\sum_{k=1}^{L} \frac{a_k(\eta_n, \mu_n) x_k}{k w^{(L)}(\eta_n, \mu_n)} \Big)_{\eta\in\mathcal{G}^\varepsilon}, \quad g : (x_1, ..., x_L) \mapsto \Big(\sum_{k=1}^{L} \frac{a_k(\eta, \mu_0) x_k}{k w^{(L)}(\eta, \mu_0)} \Big)_{\eta\in\mathcal{G}^\varepsilon}.$$

Thus (i)-(iii) are established and we see that weak convergence of $\mathbb{G}_n$ holds and the limiting Gaussian process $\mathbb{G}$ has the following covariance structure (this follows after some calculations)

$$\mathbb{E}[\mathbb{G}(\eta_1)\mathbb{G}(\eta_2)] = \frac{\mathbb{E}[\exp((Z_1-\mu)(Z_2-\mu)/\mu)] - 1}{(\mathbb{E}[\exp((Z_1-\mu)(\tilde{Z}_1-\mu)/\mu)] - 1)^{1/2} (\mathbb{E}[\exp((Z_2-\mu)(\tilde{Z}_2-\mu)/\mu)] - 1)^{1/2}}$$

where $Z_1, \tilde{Z}_1 \sim \eta_1, Z_2, \tilde{Z}_2 \sim \eta_2$ and $Z_1, Z_2, \tilde{Z}_1, \tilde{Z}_2$ are independent. Equation (9) can be proved by arguments similar to those in Example 2.7. Thus we have established (A1).

Verification of condition (A2). Consider the following decomposition

$$\begin{aligned}
&\mathbb{E} \sup_{\eta\in\mathcal{G}^\varepsilon} \Big| \frac{1}{n} \sum_{i=1}^{n} s^2_{\eta_n,\mu_n}(X_{i,n}) - (s^{(L)}_{\eta_n,\mu_n})^2(X_{i,n}) \Big| \\
=&\mathbb{E} \sup_{\eta\in\mathcal{G}^\varepsilon} \Big| \frac{1}{n} \sum_{i=1}^{n} [s_{\eta_n,\mu_n}(X_{i,n}) - s^{(L)}_{\eta_n,\mu_n}(X_{i,n})][s_{\eta_n,\mu_n}(X_{i,n}) + s^{(L)}_{\eta_n,\mu_n}(X_{i,n})] \Big| \\
\le&\mathbb{E}\Big[\Big(\sup_{\eta\in\mathcal{G}^\varepsilon} \frac{1}{n} \sum_{i=1}^{n} [s_{\eta_n,\mu_n}(X_{i,n}) - s^{(L)}_{\eta_n,\mu_n}(X_{i,n})]^2 \Big)^{1/2} \\
&\quad \times \Big(\sup_{\eta\in\mathcal{G}^\varepsilon} \frac{1}{n} \sum_{i=1}^{n} [s_{\eta_n,\mu_n}(X_{i,n}) + s^{(L)}_{\eta_n,\mu_n}(X_{i,n})]^2 \Big)^{1/2} \Big] \\
&\le \mathbb{E}\Big[\sup_{\eta\in\mathcal{G}^\varepsilon} \frac{1}{n} \sum_{i=1}^{n} [s_{\eta_n,\mu_n}(X_{i,n}) - s^{(L)}_{\eta_n,\mu_n}(X_{i,n})]^2 \Big] \mathbb{E}\Big[\sup_{\eta\in\mathcal{G}^\varepsilon} \frac{1}{n} \sum_{i=1}^{n} [s_{\eta_n,\mu_n}(X_{i,n}) + s^{(L)}_{\eta_n,\mu_n}(X_{i,n})]^2 \Big].
\end{aligned} \tag{30}$$

Moreover, for n sufficiently large and some constants $C_2, \tilde{C}$ we obtain by arguments similar to the ones in the proof of

$$\begin{aligned}
&\sup_{\eta\in\mathcal{G}^\varepsilon} |s_{\eta_n,\mu_n}(X_{i,n}) - s^{(L)}_{\eta_n,\mu_n}(X_{i,n})| \\
\leq &\sup_{\eta\in\mathcal{G}^\varepsilon} \Big|1 - \frac{w(\eta_n,\mu_n)}{w^{(L)}(\eta_n,\mu_n)}\Big| \Big| \sum_{k=1}^{\infty} \frac{a_k(\eta_n,\mu_n)}{kw(\eta_n,\mu_n)} C_k(X_{i,n}|\mu_n)\Big| \\
&+ \sup_{\eta\in\mathcal{G}^\varepsilon} \Big|\frac{w(\eta_n,\mu_n)}{w^{(L)}(\eta_n,\mu_n)}\Big| \Big| \sum_{k=L+1}^{\infty} \frac{a_k(\eta_n,\mu_n)}{kw(\eta_n,\mu_n)} C_k(X_{i,n}|\mu_n)\Big| \\
\leq &\sup_{\eta\in\mathcal{G}^\varepsilon} \Big|1 - \frac{w(\eta_n,\mu_n)}{w^{(L)}(\eta_n,\mu_n)}\Big| \Big| \sum_{k=1}^{\infty} \frac{a_k^2(\eta_n,\mu_n)}{w^2(\eta_n,\mu_n)}\Big|^{1/2} \Big| \sum_{k=1}^{\infty} \frac{C_k^2(X_{i,n}|\mu_n)}{k^2}\Big|^{1/2} \\
&+ \sup_{\eta\in\mathcal{G}^\varepsilon} \Big|\frac{w(\eta_n,\mu_n)}{w^{(L)}(\eta_n,\mu_n)}\Big| \Big| \sum_{k=L+1}^{\infty} \frac{a_k^2(\eta_n,\mu_n)}{w^2(\eta_n,\mu_n)}\Big|^{1/2} \Big| \sum_{k=L+1}^{\infty} \frac{C_k^2(X_{i,n}|\mu_n)}{k^2}\Big|^{1/2} \\
\leq &\tilde{C}L^{-1}\Big| \sum_{k=1}^{\infty} \frac{C_k^2(X_{i,n}|\mu_n)}{k^2}\Big|^{1/2} + C_2\Big| \sum_{k=L+1}^{\infty} \frac{C_k^2(X_{i,n}|\mu_n)}{k^2}\Big|^{1/2}
\end{aligned}$$

where the last inequality follows from (26) and (27). The last identity shows that for some constant C_3 and n sufficiently large

$$\mathbb{E}\sup_{\eta\in\mathcal{G}^\varepsilon} |s_{\eta_n,\mu_n}(X_{i,n}) - s^{(L)}_{\eta_n,\mu_n}(X_{i,n})|^2 \leq C_3\Big(L^{-2} + \sum_{k=L+1}^{\infty} \frac{1}{k^2}\Big). \tag{31}$$

Combining (A3) with (30) and (31) shows that

$$\limsup_{n\to\infty} \mathbb{E}\sup_{\eta\in\mathcal{G}^\varepsilon} \Big|\frac{1}{n}\sum_{i=1}^{n} s^2_{\eta_n,\mu_n}(X_{i,n}) - (s^{(L)}_{\eta_n,\mu_n})^2(X_{i,n})\Big| \leq C_4\Big(L^{-2} + \sum_{k=L+1}^{\infty} \frac{1}{k^2}\Big). \tag{32}$$

Next, observe that by construction we have $\mathbb{E}[(s^{(L)}_{\eta_n,\mu_n})^2(X_{i,n})] = 1$ for all $n \in \mathbb{N}, L \geq 2, \eta \in \mathcal{G}^\varepsilon$. Moreover simple arguments show that for every fixed $k, l \in \mathbb{N}$

$$\frac{1}{n}\sum_{i=1}^{n} C_k(X_{i,n}|\mu_n)C_l(X_{i,n}|\mu_n) \xrightarrow{P} I\{k = l\}.$$

By the extended continuous mapping theorem [see Theorem 1.11.1 in van der Vaart and Wellner (1996)] applied to the maps

$$\begin{aligned}
g_n : (x_{kl})_{k,l=1,\ldots,L} &\mapsto \Big(\sum_{k,l=1}^{L} \frac{a_k(\eta_n,\mu_n)a_l(\eta_n,\mu_n)x_{kl}}{kl(w^{(L)}(\eta_n,\mu_n))^2}\Big)_{\eta\in\mathcal{G}^\varepsilon} \\
g : (x_{kl})_{k,l=1,\ldots,L} &\mapsto \Big(\sum_{k,l=1}^{L} \frac{a_k(\eta,\mu_0)a_l(\eta,\mu_0)x_{kl}}{kl(w^{(L)}(\eta,\mu_0))^2}\Big)_{\eta\in\mathcal{G}^\varepsilon}
\end{aligned}$$

it follows that for every $L \geq 2$

$$\sup_{\eta \in \mathcal{G}^\varepsilon} \Big| \frac{1}{n} \sum_{i=1}^n ((s^{(L)}_{\eta_n,\mu_n})^2(X_i) - 1) \Big| = o_P(1).$$

Combining this with (32) proves the first part of assertion (A2). To establish the second part of (A2), note that for $x, y \in \mathbb{R}$ we have $|x_- - y_-| \leq |x - y|$. Thus

$$\begin{aligned}
&\sup_{\eta \in \mathcal{G}^\varepsilon} \Big| \frac{1}{n} \sum_{i=1}^n s^2_{\eta_n,\mu_n,-}(X_{i,n}) - (s^{(L)}_{\eta_n,\mu_n,-})^2(X_{i,n}) \Big| \\
&\leq \sup_{\eta \in \mathcal{G}^\varepsilon} \Big| \frac{1}{n} \sum_{i=1}^n (s_{\eta_n,\mu_n,-}(X_{i,n}) - s^{(L)}_{\eta_n,\mu_n,-}(X_{i,n}))^2 \Big|^{1/2} \Big| \frac{1}{n} \sum_{i=1}^n (s_{\eta_n,\mu_n,-}(X_{i,n}) + s^{(L)}_{\eta_n,\mu_n,-}(X_{i,n}))^2 \Big|^{1/2} \\
&\leq \sup_{\eta \in \mathcal{G}^\varepsilon} \Big\{ \Big| \frac{1}{n} \sum_{i=1}^n (s_{\eta_n,\mu_n}(X_{i,n}) - s^{(L)}_{\eta_n,\mu_n}(X_{i,n}))^2 \Big|^{1/2} \\
&\qquad \times \Big| \frac{4}{n} \sum_{i=1}^n 4(s_{\eta_n,\mu_n}(X_{i,n}))^2 + (s_{\eta_n,\mu_n}(X_{i,n}) - s^{(L)}_{\eta_n,\mu_n}(X_{i,n}))^2 \Big|^{1/2} \Big\}.
\end{aligned}$$

This combined with (31) and (A3) yields

$$\limsup_{n \to \infty} \mathbb{E} \sup_{\eta \in \mathcal{G}^\varepsilon} \Big| \frac{1}{n} \sum_{i=1}^n s^2_{\eta_n,\mu_n,-}(X_{i,n}) - (s^{(L)}_{\eta_n,\mu_n,-})^2(X_{i,n}) \Big| \leq C_4 \Big(L^{-2} + \sum_{k=L+1}^\infty \frac{1}{k^2} \Big). \tag{33}$$

Thus it suffices to show that for each fixed L

$$\sup_{\eta \in \mathcal{G}^\varepsilon} \Big| \frac{1}{n} \sum_{i=1}^n (s^{(L)}_{\eta_n,\mu_n,-})^2(X_{i,n}) - \| s^{(L)}_{\eta_n,\mu_n,-} \|^2_{2,\delta(\mu_n)} \Big| = o_P(1) \tag{34}$$

and that

$$\lim_{L \to \infty} \limsup_{n \to \infty} \sup_{\eta \in \mathcal{G}^\varepsilon} \Big| \| s^{(L)}_{\eta_n,\mu_n,-} \|^2_{2,\delta(\mu_n)} - \| s_{\eta,\mu_0,-} \|^2_{2,\delta(\mu_0)} \Big| = 0. \tag{35}$$

To prove (34), define $y^{(L)}(x) := (1, ..., x^L)$ and observe that there exists a constant C [note that $s^{(L)}_{\eta_n,\mu_n}(x)$ is a polynomial in x of degree L] such that

$$\begin{aligned}
&\sup_{\eta \in \mathcal{G}^\varepsilon} \Big| \frac{1}{n} \sum_{i=1}^n (s^{(L)}_{\eta_n,\mu_n,-})^2(X_{i,n}) - \| s^{(L)}_{\eta_n,\mu_n,-} \|^2_{2,\delta(\mu_n)} \Big| \\
&\qquad \leq \sup_{b \in \mathbb{R}^{L+1}, \|b\| \leq C} \Big| \frac{1}{n} \sum_{i=1}^n (b^T Y^{(L)}(X_{i,n}))^2 I\{ b^T Y^{(L)}(X_{i,n}) \leq 0 \} \\
&\qquad\qquad - \mathbb{E}[(b^T Y^{(L)}(X_{i,n}))^2 I\{ b^T Y^{(L)}(X_{i,n}) \leq 0 \}] \Big|.
\end{aligned}$$

Weak convergence to zero of the right-hand side can be proved after observing that the class of functions $\{ y \mapsto (b^T y)^2 I\{ b^T y \leq 0 \} : \|b\| \leq C \}$ is VC and has an envelope G function which satisfies $\sup_{n \geq n_0} \mathbb{E} G^2(Y^{(L)}(X_{i,n})) < \infty$ for some $n_0 < \infty$. Thus convergence of the right-hand side above to zero follows from Theorem 2.8.1 in van der Vaart and Wellner (1996).

Next, let us prove (35). We begin by proving

$$\limsup_{n\to\infty}\sup_{\eta\in\mathcal{G}^\varepsilon}\left|\|s^{(L)}_{\eta_n,\mu_n,-}\|^2_{2,\delta(\mu_n)}-\|s^{(L)}_{\eta,\mu_0,-}\|^2_{2,\delta(\mu_n)}\right|+\left|\|s^{(L)}_{\eta,\mu_0,-}\|^2_{2,\delta(\mu_n)}-\|s^{(L)}_{\eta,\mu_0,-}\|^2_{2,\delta(\mu_0)}\right|=0 \tag{36}$$

for every fixed $L\geq 2$. Convergence to zero of $\sup_{\eta\in\mathcal{G}^\varepsilon}\left|\|s^{(L)}_{\eta_n,\mu_n,-}\|^2_{2,\delta(\mu_n)}-\|s^{(L)}_{\eta,\mu_0,-}\|^2_{2,\delta(\mu_n)}\right|$ follows from the fact that, for $V_n\sim Pois(\mu_n)$, we have for some sequence $\alpha_n=o(1)$

$$\begin{aligned}
&\sup_{\eta\in\mathcal{G}^\varepsilon}\left|\|s^{(L)}_{\eta_n,\mu_n,-}\|^2_{2,\delta(\mu_n)}-\|s^{(L)}_{\eta,\mu_0,-}\|^2_{2,\delta(\mu_n)}\right|\\
&\leq \sup_{\|a-b\|\leq\alpha_n,\|a\|\leq C,\|b\|\leq C}\mathbb{E}\Big|(b^TY^{(L)}(V_n))^2I\{b^TY^{(L)}(V_n)\leq 0\}\\
&\qquad\qquad -(a^TY^{(L)}(V_n))^2I\{a^TY^{(L)}(V_n)\leq 0\}\Big|\\
&\leq 2C\alpha_n\mathbb{E}[\|Y^{(L)}(V_n)\|^4]=o(1) \qquad (37)
\end{aligned}$$

where the last inequality follows from $|x_-^2-y_-^2|\leq(|x|+|y|)(|x|-|y|)$. Similarly, letting $V_0\sim Pois(\mu_0)$, the second term can be bounded by

$$\begin{aligned}
&\sup_{\eta\in\mathcal{G}^\varepsilon}\left|\|s^{(L)}_{\eta,\mu_0,-}\|^2_{2,\delta(\mu_n)}-\|s^{(L)}_{\eta,\mu_0,-}\|^2_{2,\delta(\mu_0)}\right|\\
&\leq \sup_{b\in\mathbb{R}^{L+1},\|b\|\leq C}\Big|\mathbb{E}[(b^TY^{(L)}(V_n))^2I\{b^TY^{(L)}(V_n)\leq 0\}]-\mathbb{E}[(b^TY^{(L)}(V_0))^2I\{b^TY^{(L)}(V_0)\leq 0\}]\Big|.
\end{aligned}$$

Covering $B:=\{b\in\mathbb{R}^{L+1}:\|b\|\leq C\}$ with a finite number of balls of radius ε one can reduce the above problem to showing that

$$\mathbb{E}[(b^TY^{(L)}(V_n))^2I\{b^TY^{(L)}(V_n)\leq 0\}]\to\mathbb{E}[(b^TY^{(L)}(V_0))^2I\{b^TY^{(L)}(V_0)\leq 0\}]$$

for any fixed $b\in B$. Observe that V_n converges weakly to V. The continuous mapping theorem implies that $(b^TY^{(L)}(V_n))^2I\{b^TY^{(L)}(V_n)\leq 0\}\rightsquigarrow(b^TY^{(L)}(V_0))^2I\{b^TY^{(L)}(V_0)\leq 0\}$, and by uniform integrability of the sequence $(b^TY^{(L)}(V_n))^2I\{b^TY^{(L)}(V_n)\leq 0\}$ this implies convergence of the first moment. Together with (37) this establishes (36). Finally, the convergence

$$\lim_{L\to\infty}\sup_{\eta\in\mathcal{G}^\varepsilon}\left|\|s^{(L)}_{\eta,\mu_0,-}\|^2_{2,\delta(\mu_0)}-\|s_{\eta,\mu_0,-}\|^2_{2,\delta(\mu_0)}\right|=0$$

can be proved by similar arguments as (33) with $n^{-1}\sum_i$ replaced by the expectation, the details are omitted for the sake of brevity. This completes the proof of Assumption (A2).

Verification of (26)-(29) We begin by noting that for $Z\sim\eta_n$ with η_n having support contained in $[m,M]$ it follows that $|Z-\mu_n|^k\leq M^{k-2}(Z-\mu_n)^2$ for $k\geq 3$. Thus, as soon as $\mu_n\in[m,M]$, which is the case for n sufficiently large, we have

$$\sum_{k\geq 2}a_k^2(\eta_n,\mu_n)=\sum_{k\geq 2}\frac{k^2(\mathbb{E}[(Z-\mu_n)^k])^2}{k!\mu_n^k}\leq(\mathbb{E}[(Z-\mu_n)^2])^2\sum_{k\geq 2}\frac{k^2M^{2k-4}}{k!m^k}\leq C(\mathbb{E}[(Z-\mu_n)^2])^2.$$

This shows (27). Next, observe that

$$\Big(\frac{w(\eta_n,\mu_n)}{w^{(L)}(\eta_n,\mu_n)}\Big)^2=\frac{\sum_{\ell=1}^\infty\ell^{-2}a_\ell^2(\eta_n,\mu_n)}{\sum_{\ell=1}^L\ell^{-2}a_\ell^2(\eta_n,\mu_n)}=1+\frac{\sum_{\ell=L+1}^\infty\ell^{-2}a_\ell^2(\eta_n,\mu_n)}{\sum_{\ell=1}^L\ell^{-2}a_\ell^2(\eta_n,\mu_n)}.$$

Now for $Z \sim \eta_n$ with η_n having support contained in $[m, M]$ we have as soon as $\mu_n \in [m, M]$

$$0 \leq \frac{\sum_{\ell=L+1}^{\infty} \ell^{-2} a_\ell^2(\eta_n, \mu_n)}{\sum_{\ell=1}^{L} \ell^{-2} a_\ell^2(\eta_n, \mu_n)} \leq \frac{\sum_{k=L+1}^{\infty} \frac{(\mathbb{E}[(Z-\mu_n)^k])^2}{k! \mu_n^k}}{\frac{(\mathbb{E}[(Z-\mu_n)^2])^2}{2\mu_n^2}} \leq 2M^2 \sum_{k \geq L+1} \frac{M^{2k-4}}{k! m^k} \leq CL^{-1}.$$

The first part of (26) follows, and the second part of (26) can be established by exactly the same arguments. Finally, for $\tilde{Z} \sim \eta$

$$\Big(\frac{w^{(L)}(\eta_n, \mu_n)}{w^{(L)}(\eta, \mu_0)} \Big)^2 = \frac{\sum_{k=1}^{L} \frac{(\mathbb{E}[(Z-\mu_n)^k])^2}{k! \mu_n^k}}{\sum_{k=1}^{L} \frac{(\mathbb{E}[(\tilde{Z}-\mu_0)^k])^2}{k! \mu_0^k}}$$

and by construction $\mathbb{E}[(Z - \mu_n)^k] = \mathbb{E}[(\tilde{Z} - \mu_0)^k]$ for all $k \in \mathbb{N}$. Now (29) follows since $\max_{k=1,..,L} |(\mu_n/\mu_0)^k - 1| \to 0$ as $n \to \infty$. This completes all proofs for the Poisson case. □