Electronic Journal of Statistics
ISSN: 1935-7524

Large-Sample Properties of Unsupervised Estimation of the Linear Discriminant Using Projection Pursuit

Una Radojičić [1], **Klaus Nordhausen** [1,2] **and Joni Virta** [3]

[1] *Institute of Statistics & Mathematical Methods in Economics, Vienna University of Technology, Austria. e-mail:* [*] una.radojicic@tuwien.ac.at

[2] *Department of Mathematics and Statistics, University of Jyväskylä, Finland. e-mail:* klaus.k.nordhausen@jyu.fi

[3] *Department of Mathematics and Statistics, University of Turku, Finland. e-mail:* joni.virta@utu.fi

Abstract: We study the estimation of the linear discriminant with projection pursuit, a method that is unsupervised in the sense that it does not use the class labels in the estimation. Our viewpoint is asymptotic and, as our main contribution, we derive central limit theorems for estimators based on three different projection indices, skewness, kurtosis, and their convex combination. The results show that in each case the limiting covariance matrix is proportional to that of linear discriminant analysis (LDA), a supervised estimator of the discriminant. An extensive comparative study between the asymptotic variances reveals that projection pursuit gets arbitrarily close in efficiency to LDA when the distance between the groups is large enough and their proportions are reasonably balanced. Additionally, we show that consistent unsupervised estimation of the linear discriminant can be achieved also in high-dimensional regimes where the dimension grows at a suitable rate to the sample size, for example, $p_n = o(n^{1/3})$ is sufficient under skewness-based projection pursuit. We conclude with a real data example and a simulation study investigating the validity of the obtained asymptotic formulas for finite samples.

Keywords and phrases: Clustering, Kurtosis, Skewness, Linear discriminant analysis, Projection pursuit.

Contents

[3] The work of Joni Virta was supported by the Academy of Finland (Grant 335077)

1. Introduction

Classification and clustering are two central themes in modern data analysis and can be seen, respectively, as the supervised and unsupervised versions of the same problem: In classification, the group memberships, or labels, of the training data points are known and the objective is to use the training data to form a classification rule for future observations, some of the standard methods including, e.g., linear discriminant analysis, support vector machines, and random forests, see [19]. Whereas in clustering, no labels for the data points are known but we postulate that a reasonable grouping exists and aim to find it, with, e.g., k-means clustering or spectral clustering, see [19, 63].

In this paper, we work under a clustering context and the assumption that the data admit a natural grouping but that their labels are indeed unknown to us. In their seminal work, [50] studied in this setting the use of *projection pursuit* (PP), a general family of methods searching for a projection direction that maximizes the value of the so-called projection index, see, e.g., [25, 10, 8, 16] and the references therein. Namely, denoting the within-class covariance matrix by $\mathbf{\Sigma}$ and the two group means by $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$, [50] established that using kurtosis as the projection index in projection pursuit allows the unsupervised estimation of the projection direction $\boldsymbol{\theta} := \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$ that is used in linear discriminant analysis to construct the optimal Bayes classifier, in the full absence of any label information. In other words, projection pursuit essentially allows conducting LDA in a unsupervised fashion to recover the subspace that optimally separates the two groups. Afterward, various clustering methods can then be applied to the projected data to conduct efficient clustering.

While very interesting, the result of [50] raises a natural question regarding the efficiency of the procedure. Namely, how much does one lose by not knowing the labels and relying on projection pursuit compared to using LDA to recover the same direction $\boldsymbol{\theta}$ when the group memberships are known? This is the main question we study in the current paper, working, for simplicity, under the assumption of two-group normal mixtures. Our approach is asymptotic in nature and we perform the comparison through the limiting covariance matrices of the estimators in question. In particular, we show that the limiting covariance matrices of projection pursuit and LDA are proportional, allowing us to conduct the comparisons simply through the corresponding constants of proportionality. Interestingly, the ratios of these constants depend on the model parameters only

through the mixing proportion and the squared Mahalanobis distance (MD) between the group means, $\tau := (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. In particular, the ratios do not depend directly on the dimension p of the data. As our second contribution, we show that projection pursuit can be used to consistently estimate the optimal direction even in the high-dimensional regime where $p \equiv p_n \to \infty$ as long as the dimension grows at a suitable rate to the sample size n.

The present work can thus be seen as a continuation to the classical literature on confirmatory (or inferential) projection pursuit. However, unlike the present study that focuses on asymptotic efficiencies, the main question in confirmatory PP is typically to assess whether the found projections reflect actual structures in the population and are not simply artifacts caused by noise. For example, to assess the significance of the obtained results, [20] uses a bootstrap-like procedure and compares the observed value of the projection index to the one obtained by applying the procedure to the corresponding Gaussian data. Similar problems are discussed also in, e.g., [57, 45, 35]. To illustrate the importance of assessing the significance of the obtained results, [13] give cautionary examples in which they show how exploratory projection pursuit can always find structures when applied to sparse, high-dimensional data.

Besides the confirmatory projection pursuit, asymptotic results for general projection indices have been derived earlier also in the context of *independent component analysis* (ICA), see, e.g., [49, 15, 44, 62]. In ICA, one assumes that the observed random p-vector $\mathbf{x}$ is an *independent component (IC) model*, i.e., there exists a full rank $p \times p$ matrix $\boldsymbol{\Gamma}$ such that $\boldsymbol{\Gamma}\{\mathbf{x} - \mathrm{E}(\mathbf{x})\}$ has independent components. The IC model is a rather wide family of distributions and, in particular, contains our model of choice, the multivariate normal mixture. This, apparently novel, result is given as Lemma B.1 in the Appendix A and reveals that the multivariate normal mixture decomposes into $p-1$ independent standard normals and a non-Gaussian component which corresponds, up to the scale and sign, to the optimal linear discriminant projection of the data. This connection between the two models implies that the results of the current paper are intimately related to [62] who considered (in the context of ICA) the same projection indices as we do here. However, we remark that our contributions surpass those of [62] in two critical regards: 1) [62] derived only the asymptotic variances of the ICA parameters (ignoring their covariances), whereas we give the full limiting distribution of the estimated projection direction. Besides completing the asymptotic story, knowledge of the full distribution is crucial concerning the comparison of PP and LDA as it reveals that the limiting covariance matrix of PP is exactly proportional to the limiting covariance of LDA, see Theorems 1–4 later on. 2) From a technical viewpoint, the derivation of the convergence rates of the estimators was in [62] left implicit and our proofs provide a rigorous treatment of this. In particular, to guarantee well-defined Taylor expansions of the objective functions, we need to establish the almost sure convergence of the projection pursuit estimates and, as far as we are aware, such results have not been given previously either in ICA or PP-literature. Finally, we note that the link between the two models furthermore implies that any projection index capable of recovering independent components in an IC model can, in this setting,

be used to recover the optimal linear discriminant (assuming the index yields distinct values for the mixture and Gaussian noise). For example, [53] show that any increasing function of a subadditive squared dispersion measure can be used for such a purpose.

While kurtosis is the most popular choice for the projection index in projection pursuit, also several alternatives are commonly used. In particular, skewness is a somewhat standard choice, see, for example, [39], and was shown in [37] to have the same property of being able to find the optimal projection direction without the label information as possessed by kurtosis. As such, we study also skewness-based projection pursuit in the current work. We also note that, prior to their use in projection pursuit, both skewness and kurtosis have a rich history as test statistics when testing for (multivariate) normality. For example, [42] first introduced the maximal kurtosis and skewness obtained by projections as test statistics when testing for multivariate normality and, using Monte Carlo methods, compared the power of the obtained tests to other common tests for normality against various alternatives, including several two-dimensional Gaussian mixtures. Distributional properties of the statistics introduced by [42] were later studied in [41] and [6], under the null hypotheses of multivariate normality and elliptical symmetry, respectively. See also the conjecture by [39] that the limiting distribution of the maximal skewness attainable by a linear combination of normal variables is skew-normal.

Despite their ubiquitousness, as shown by [50, 37], for both kurtosis and skewness there exist particular values of the mixing proportion under which the two indices are unable to recover the optimal projection direction (for example, skewness fails to produce a consistent estimate of $\boldsymbol{\theta}$ when the two groups have equal proportions). These drawbacks can be mitigated by combining both cumulants into a single projection index, in a form of a weighted linear combination. This combined projection index was first proposed in [28] and some of its distributional properties were discussed in [35]. Our results show that with a proper choice of weighting, a rather efficient unsupervised competitor for the LDA-based supervised estimator can be obtained with the combined index. Indeed, in the extreme case where the distance between the group means is large enough and the group sizes are reasonably balanced, projection pursuit is able to achieve efficiency arbitrarily close to LDA. As remarked in the previous paragraph, the asymptotic properties of the hybrid index have been studied also earlier, in the context of independent component analysis, in [62].

We note that despite the theoretical guarantees of projection pursuit, the most common unsupervised method for revealing clusters is still arguably PCA, see, e.g., [27]. However, it is also well known that PCA does not, in general, yield a consistent estimator of the linear discriminant direction. A standard example demonstrating this is the extreme case where the within-group covariance matrix $\boldsymbol{\Sigma}$ is heavily concentrated on a direction orthogonal to the difference of the group means $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$. In such a case, the projections of the two group means onto the first principal component direction overlap, making clustering based on the direction impossible. Hence, due to its unreliability in estimating the linear discriminant, PCA cannot truly be seen as a unsupervised estimator of

the separating direction and, as such, we do not include it in the comparisons in the current paper. However, we have still included, for completeness, equivalent asymptotic results for PCA as we state for the other methods, and these are given in Appendix A.

In recent years there has been a large amount of work on parameter estimation in Gaussian mixture models [67, 34, 32, 71, 23, 29, 22], particularly in high-dimensional settings and by the EM-algorithm, see for example [70, 64, 68, 69, 11] and references therein. It is worth mentioning how, in general, methods for parameter estimation in Gaussian mixture models can also be used for unsupervised estimation of the linear discriminant in this setting, with the potential estimator being the plug-in estimator in which the group means and the common covariance are estimated by the method of choice. Some statistical guarantees and properties of EM-based estimators of parameters in Gaussian mixtures can be found in, e.g., [5, 47], and are beyond the scope of this paper. However, under the real data example of Section 6, we compare the results obtained by the projection pursuit approach to those obtained using the EM-algorithm when estimating the optimal linear discriminant.

The rest of the manuscript is organized as follows. In Section 2 we derive the asymptotic behavior of three estimators of the linear discriminant direction: LDA and kurtosis- and skewness-based projection pursuit. A short comparison of the results is also presented. In Section 3, we give the corresponding results for projection pursuit based on a weighted combination of skewness and kurtosis and conduct a more extensive set of asymptotic comparisons between all considered methods. In Section 4 we show that both kurtosis- and skewness-based PP, as well as the PP based on the convex combination of those produce consistent estimators of the linear discriminant in the high-dimensional setting, where both the sample size n and the dimension p diverge to infinity in a suitable ratio. Simulation studies exploring both the finite-sample performance of the methods and the applicability of our asymptotic results to practice are given in Section 5, while the performance and the applicability of presented methods to a real data example, as well as the comparison to the PCA, are given in Section 6. Finally, we conclude with some discussion in Section 7. All proofs of the technical results are postponed to Appendix B.

2. Estimation of the linear discriminant

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Throughout the following, we assume that the $(p+1)$-dimensional pair $(\mathbf{x}, y)$ obeys the following model:

$$y \sim \text{Ber}(\alpha_1) \quad \text{and} \quad \mathbf{x} \mid y \sim \mathcal{N}_p\{y\boldsymbol{\mu}_1 + (1-y)\boldsymbol{\mu}_2, \boldsymbol{\Sigma}\}, \tag{1}$$

for $0 < \alpha_1 < 1$, $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^p$, $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, and a full rank $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$. The marginal distribution of $\mathbf{x}$ is then the multivariate normal mixture,

$$\mathbf{x} \sim \alpha_1 \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + \alpha_2 \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}),$$

where $\alpha_2 := 1 - \alpha_1$. Under model (1), the classification of $\mathbf{x}$ is usually based on its projection onto the linear discriminant direction $\boldsymbol{\theta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. This projection direction is optimal in the sense that the optimal Bayes classifier (having the minimal miss-classification rate out of all classifiers) depends on the data only through the projection $\boldsymbol{\theta}'\mathbf{x}$, see, e.g., [43].

Our objective throughout the paper is the estimation of the standardized projection direction $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ (the scale of the projection direction is irrelevant, meaning that the unit length constraint is without loss of generality). As described in Section 1, we will consider two types of estimators, unsupervised ones which use only the random vector $\mathbf{x}$ (a sample from its distribution) in the estimation, and an supervised one which bases the estimation on the full pair $(\mathbf{x}, y)$. The supervised method is allowed more information in the estimation and is, naturally, expected to provide a more efficient estimator, a fact that is verified by our comparisons later on.

2.1. Supervised estimation of the linear discriminant

If we have a sample $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ from the distribution of the full pair $(\mathbf{x}, y)$ available, the standard estimator of $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$ is the plug-in estimator (which is also its MLE, up to the scaling of the pooled covariance matrix) used in standard LDA. That is, using the notation,

$$\bar{\mathbf{x}}_{n1} := \frac{1}{\sum_{i=1}^n y_i} \sum_{i=1}^n y_i \mathbf{x}_i, \quad \bar{\mathbf{x}}_{n2} := \frac{1}{\sum_{i=1}^n (1-y_i)} \sum_{i=1}^n (1-y_i)\mathbf{x}_i,$$

$$\mathbf{S}_n := \frac{1}{n-2} \left\{ \sum_{i=1}^n y_i (\mathbf{x}_i - \bar{\mathbf{x}}_1)(\mathbf{x}_i - \bar{\mathbf{x}}_1)' + \sum_{i=1}^n (1-y_i)(\mathbf{x}_i - \bar{\mathbf{x}}_2)(\mathbf{x}_i - \bar{\mathbf{x}}_2)' \right\},$$

we consider the estimator,

$$\mathbf{w}_n := \mathbf{S}_n^{-1}(\bar{\mathbf{x}}_{n2} - \bar{\mathbf{x}}_{n1}).$$

Asymptotic results for LDA are very standard in the literature, see for example [3]. However, these results are usually given in the case of fixed group sizes, whereas in our model the group sizes are determined by the indicator variables $y_1, \ldots, y_n$ and are, as such, random. Hence, as far as we know, the following theorem is, if not particularly groundbreaking in its conclusions, a novel one.

Theorem 1. *Under model (1), we have, as* $n \to \infty$,

$$\sqrt{n}(\boldsymbol{w}_n/\|\boldsymbol{w}_n\| - \boldsymbol{\theta}/\|\boldsymbol{\theta}\|) \rightsquigarrow \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Psi}_U),$$

where

$$\boldsymbol{\Psi}_U := \left(\frac{1+\beta\tau}{\|\boldsymbol{\theta}\|^2\beta}\right)\left(\boldsymbol{I}_p - \frac{\boldsymbol{\theta}\boldsymbol{\theta}'}{\|\boldsymbol{\theta}\|^2}\right)\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{I}_p - \frac{\boldsymbol{\theta}\boldsymbol{\theta}'}{\|\boldsymbol{\theta}\|^2}\right),$$

$\beta := \alpha_1\alpha_2$ *and* $\tau := (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$.

The form of the limiting covariance matrix in Theorem 1 is rather simple and inspection of the proof of the result reveals that the involved projection matrices onto the orthogonal complement of the direction $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ are simply consequences of the standardization of the estimator to unit length. Note also that the scalar factor in front can be written as $1/(\|\boldsymbol{\theta}\|^2\beta)+(\boldsymbol{\theta}/\|\boldsymbol{\theta}\|)'\boldsymbol{\Sigma}(\boldsymbol{\theta}/\|\boldsymbol{\theta}\|)$, the two summands of which have the following rough interpretations: If the groups are imbalanced, β is small, making the first summand large and inflating the asymptotic variance. Similarly, if the data exhibit a large amount of variation in the direction of the optimal discriminant direction, i.e., $(\boldsymbol{\theta}/\|\boldsymbol{\theta}\|)'\boldsymbol{\Sigma}(\boldsymbol{\theta}/\|\boldsymbol{\theta}\|)$ is large, the second term increases the magnitude of the asymptotic variance.

2.2. Unsupervised estimation of the linear discriminant

Kurtosis-based projection pursuit

Let $\delta_1 := 1/2 - 1/\sqrt{12}$, $\delta_2 := 1/2 + 1/\sqrt{12}$ and $\tilde{\mathbf{x}} := \mathbf{x} - \mathrm{E}(\mathbf{x})$. The kurtosis $\kappa : \mathbb{S}^{p-1} \to \mathbb{R}$ of the projection of $\mathbf{x}$ on a given direction $\mathbf{u} \in \mathbb{S}^{p-1}$ is then defined as,

$$\kappa(\mathbf{u}) = \frac{\mathrm{E}\{(\mathbf{u}'\tilde{\mathbf{x}})^4\}}{[\mathrm{E}\{(\mathbf{u}'\tilde{\mathbf{x}})^2\}]^2}.$$

The fact that projection pursuit based on kurtosis is Fisher consistent for the linear discriminant under normal mixtures was first shown in [50, Corollary 2]. However, the successful use of their result in practice requires knowing something about the mixing proportion α_1. Namely, if $\alpha_1 \in (\delta_1, \delta_2)$ then the linear discriminant $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ is found as the minimizer of κ, whereas if $\alpha_1 \in (0, \delta_1) \cup (\delta_2, 1)$ then $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ is found as the maximizer of κ. Naturally, as a workaround, one could in practice always search for both the minimizer and the maximizer of κ but, even in this case, it might be non-trivial to recognize the linear discriminant amongst the two. Thus, to obtain a truly unsupervised estimator, we propose instead using the squared *excess kurtosis* $\{\kappa(\mathbf{u}) - 3\}^2$ as an objective function. Indeed, the next lemma reveals that the squared excess kurtosis yields a Fisher consistent estimate of the linear discriminant, apart from the degenerate cases $\alpha_1 \in \{\delta_1, \delta_2\}$ where excess kurtosis vanishes, without the need to choose between minimization and maximization.

Lemma 1. *Given model (1),*

1) *if $\alpha_1 \notin \{\delta_1, \delta_2\}$, then the function $\boldsymbol{u} \mapsto \{\kappa(\boldsymbol{u}) - 3\}^2$ is uniquely maximized by $\pm\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$,*
2) *if $\alpha_1 \in \{\delta_1, \delta_2\}$, then $\{\kappa(\boldsymbol{u}) - 3\}^2 = 0$ for all $\boldsymbol{u} \in \mathbb{S}^{p-1}$.*

Moving next to study the asymptotic properties of κ, let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a random sample from the marginal distribution of $\mathbf{x}$ in the model (1). The sample counterpart of κ is

$$\kappa_n : \mathbb{S}^{p-1} \to \mathbb{R}, \quad \kappa_n(\mathbf{u}) = \frac{(1/n)\sum_{i=1}^n (\mathbf{u}'\tilde{\mathbf{x}}_i)^4}{\{(1/n)\sum_{i=1}^n (\mathbf{u}'\tilde{\mathbf{x}}_i)^2\}^2},$$

where $\tilde{\mathbf{x}}_i := \mathbf{x}_i - \bar{\mathbf{x}}$. If $n \geq p$ the denominator of the random function κ_n is a.s. positive, making κ_n well-defined and an estimator for $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ is then obtained as any maximizer of $\mathbf{u} \mapsto \{\kappa_n(\mathbf{u}) - 3\}^2$ (note that if $n \geq p$, a maximizer exists almost surely due to the compacity of $\mathbb{S}^{p-1}$). The following theorem shows that any sequence of such maximizers has a limiting normal distribution. Note that the need to include the "corrective" signs s_n in Theorem 2 stems from the sign-invariance of the objective function (which also causes the existence of two maximizers in Lemma 1).

Theorem 2. *Given model (1), assume that $\alpha_1 \notin \{\delta_1, \delta_2\}$ and let $\boldsymbol{u}_n$ be any sequence of maximizers of $\boldsymbol{u} \mapsto \{\kappa_n(\boldsymbol{u}) - 3\}^2$. Then, there exists a sequence of signs $s_n \in \{-1, 1\}$ such that, as $n \to \infty$,*

1) $s_n \boldsymbol{u}_n \to \boldsymbol{\theta}/\|\boldsymbol{\theta}\|$, almost surely.
2) $\sqrt{n}(s_n \boldsymbol{u}_n - \boldsymbol{\theta}/\|\boldsymbol{\theta}\|) \rightsquigarrow \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Psi}_\kappa)$, where

$$\boldsymbol{\Psi}_\kappa := C_\kappa \boldsymbol{\Psi}_U,$$

and

$$C_\kappa := \frac{6 + 24\beta\tau + 9\beta(1-2\beta)\tau^2 + \beta(1-3\beta)\tau^3}{\beta\tau^3(1-6\beta)^2},$$

with $\boldsymbol{\Psi}_U$, β and τ as in Theorem 1.

The limiting covariance matrices in Theorems 1 and 2 are proportional, the only difference being the factor C_κ. This makes their comparisons in Subsection 2.3 particularly straightforward. However, even without the formal comparisons, it is evident that the kurtosis-based estimator has a clear flaw in that it fails to be consistent for the mixing proportions $\alpha_1 \in \{\delta_1, \delta_2\}$ (for these values of α_1, we have $1 - 6\beta = 0$ in the denominator of C_κ in Theorem 2). And even though these are only two points in the continuum $(0, 1)$, the continuity of C_κ in α_1 outside of these points implies that the estimator is highly inefficient for values of α_1 near δ_1 or δ_2. Hence, we will next discuss an alternative estimator that is consistent when $\alpha_1 \in \{\delta_1, \delta_2\}$ (at the price of lacking consistency in another point).

Skewness-based projection pursuit

To complement the kurtosis-based projection pursuit, we next consider skewness -based projection pursuit. Note that, despite its dependency on lower moments, this form of PP is less studied in the literature (see the references in Section 1).

The skewness of the projection of $\mathbf{x}$ on a given direction $\mathbf{u} \in \mathbb{S}^{p-1}$ is measured by the objective function $\gamma : \mathbb{S}^{p-1} \to \mathbb{R}$ defined as,

$$\gamma(\mathbf{u}) = \frac{\mathrm{E}\{(\mathbf{u}'\tilde{\mathbf{x}})^3\}}{[\mathrm{E}\{(\mathbf{u}'\tilde{\mathbf{x}})^2\}]^{3/2}},$$

where again $\tilde{\mathbf{x}} = \mathbf{x} - \mathrm{E}(\mathbf{x})$. Similarly to kurtosis, also with skewness it is more convenient to work with its squared value. The next lemma presents the Fisher consistency of the corresponding estimator and reveals that the mixing proportion $1/2$ plays the role of the proportions δ_1, δ_2 for skewness. The reason for this is intuitively clear as, under the choice $\alpha_1 = 1/2$, the normal mixture is perfectly symmetrical, explaining the vanishing of the skewness. The result appeared originally as Proposition 1 in [37] but we give, for completeness, a proof in Appendix B.

Lemma 2. *Given model (1),*

1) *if $\alpha_1 \neq 1/2$, then the function $\boldsymbol{u} \mapsto \gamma(\boldsymbol{u})^2$ is uniquely maximized by $\pm\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$,*
2) *if $\alpha_1 = 1/2$, then $\gamma(\boldsymbol{u})^2 = 0$ for all $\boldsymbol{u} \in \mathbb{S}^{p-1}$.*

Finally, we derive in Theorem 3 below the strong consistency and limiting distribution of the corresponding sample estimator, obtained through the maximization of the square of the sample skewness, defined as,

$$\gamma_n : \mathbb{S}^{p-1} \to \mathbb{R}, \quad \gamma_n(\mathbf{u}) = \frac{(1/n)\sum_{i=1}^n (\mathbf{u}'\tilde{\mathbf{x}}_i)^3}{\{(1/n)\sum_{i=1}^n (\mathbf{u}'\tilde{\mathbf{x}}_i)^2\}^{3/2}}.$$

Theorem 3. *Given model (1), assume that $\alpha_1 \neq 1/2$ and let $\boldsymbol{u}_n$ be any sequence of maximizers of $\boldsymbol{u} \mapsto \gamma_n(\boldsymbol{u})^2$. Then, there exists a sequence of signs $s_n \in \{-1, 1\}$ such that, as $n \to \infty$,*

1) *$s_n \boldsymbol{u}_n \to \boldsymbol{\theta}/\|\boldsymbol{\theta}\|$, almost surely.*
2) *$\sqrt{n}(s_n \boldsymbol{u}_n - \boldsymbol{\theta}/\|\boldsymbol{\theta}\|) \rightsquigarrow \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Psi}_\gamma)$, where*

$$\boldsymbol{\Psi}_\gamma := C_\gamma \boldsymbol{\Psi}_U,$$

and

$$C_\gamma := \frac{2 + 6\beta\tau + \beta\tau^2}{\beta\tau^2(1 - 4\beta)}$$

with $\boldsymbol{\Psi}_U$, β, τ as in Theorem 1.

Interestingly, also the limiting covariance of the skewness-based estimator is proportional to that of LDA, meaning that the main object of interest in the result is the factor C_γ. These factors will be compared in the next section to make statements about the relative asymptotic efficiencies (ARE) of the estimators under various scenarios where, for two unbiased estimators with proportional covariance matrices, the asymptotic relative efficiency of one over the other is calculated as the inverse of the ratio of the corresponding proportionality constants.

2.3. Asymptotic comparison of the three estimators

Theorems 1, 2 and 3 show that the limiting distributions of the supervised and unsupervised estimators of $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ all have proportional covariance matri-

ces. Thus, their efficiencies may be compared simply through the corresponding constants of proportionality which depend on the problem parameters only through the mixing proportion ($\beta = \alpha_1\alpha_2$) and the degree of separation between the two groups, as measured by the (squared) Mahalanobis distance $\tau = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. The relative asymptotic efficiencies (pair-wise ratios of the constants) of the unsupervised estimators vs. the supervised estimator (LDA) are simply C_κ^{-1} and C_γ^{-1} where the values of the constants are given in Theorems 2 and 3. Especially the former expression is somewhat complicated for arbitrary β and τ but both simplify greatly if we consider the case where the Mahalanobis distance τ is large. That is, letting $\tau \to \infty$, the relative asymptotic efficiencies are simply

$$\mathrm{Eff}_\kappa = \frac{(1-6\beta)^2}{1-3\beta} \quad \text{and} \quad \mathrm{Eff}_\gamma = 1 - 4\beta.$$

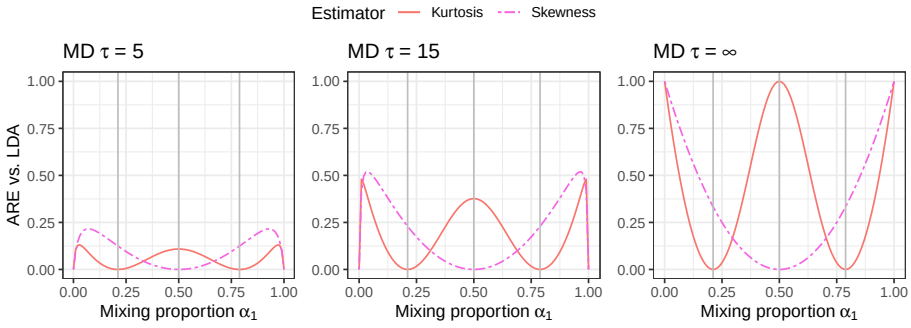

FIG 1. *The relative asymptotic efficiency of the unsupervised vs. supervised estimation of* $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ *under different Mahalanobis distances* τ *between the two group means. The three gray vertical reference lines mark those values of* α_1 *where kurtosis* $(1/2 \pm 1/\sqrt{12})$ *or skewness* $(1/2)$ *of all projections is a constant.*

Figure 1 plots the relative efficiencies as a function of the first mixing proportion α_1 in the cases $\tau = 5, 15$ and $\tau \to \infty$. The plots verify that, for any practical value of the Mahalanobis distance, LDA is always asymptotically highly superior to both unsupervised methods. However, in the extreme case where the two groups are well-separated to an arbitrarily large degree, we see, in particular, that the kurtosis estimator is asymptotically equally efficient to LDA in the balanced case $\alpha_1 = 1/2$, and also in the limits $\alpha_1 \to 0$ and $\alpha_1 \to 1$ (although, in these cases, the actual asymptotic covariance matrices themselves grow without bounds).

Figure 1 also shows that, depending on the Mahalanobis distance, around the point $\alpha_1 \approx 0.30$ there is a mixing proportion for which κ and γ are asymptotically equally efficient. Figure 2 plots these proportions (and their mirror images on the upper half of the region $(0, 1)$) as a function of the Mahalanobis distance. The plot reveals that the region of mixing proportions for which κ is

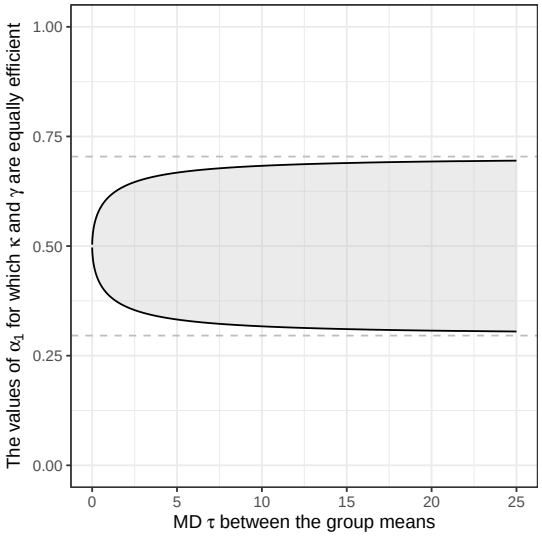

FIG 2. *The two solid curves trace the values of the mixing proportion α_1 (y-axis) for which κ and γ are asymptotically equally efficient as a function of the Mahalanobis distance between the two group means (x-axis). The shaded region represents the values of α_1 for which κ is the superior choice over γ. The two horizontal dashed lines indicate the limits $1/2 \pm 1/\sqrt{24}$ of the two curves.*

asymptotically superior choice to γ (the gray inner region) gets wider as the groups get more well-separated, finally approaching the region $1/2 \pm 1/\sqrt{24}$ in the limit $\tau \to \infty$ (the two horizontal dashed lines).

3. Convex combination of skewness and kurtosis

3.1. Theoretical properties

Based on Figure 1, the objective functions κ and γ produce, especially for well-separated groups, fairly efficient estimators of the linear discriminant in the absence of any grouping information. However, for this, it is crucial to have at least an approximate idea of the mixing proportion α_1 in order to choose the more efficient of the two objective functions and to avoid the points where a particular estimator becomes completely inefficient (for example, if the groups are close to being balanced, one wants to use kurtosis as skewness contains no information when $\gamma = 1/2$, see Lemma 2 and Figure 1). As the mixing proportion is rarely known in practice, this makes the procedure difficult to implement.

A natural way to overcome this weakness is to, instead of choosing between γ and κ, use them both simultaneously, through a objective function $\eta : \mathbb{S}^{p-1} \to \mathbb{R}$ that is a convex combination of the two squared cumulants,

$$\eta(\mathbf{u}) = \eta(\mathbf{u}; w_1) := w_1 \gamma(\mathbf{u})^2 + w_2 \{\kappa(\mathbf{u}) - 3\}^2,$$

where $w_1, w_2 \geq 0$, $w_1 + w_2 = 1$. Naturally, the cases $w_1 = 0$ and $w_2 = 0$ simply correspond to the two individual objective functions and, hence, we will in the following assume that $w_1, w_2 > 0$. The next result, following straightforwardly from Lemmas 1 and 2, shows that η indeed allows the completely unsupervised recovery of the linear discriminant regardless of the mixing proportion $\alpha_1 \in (0, 1)$.

Lemma 3. *Given model (1), η is uniquely maximized by $\pm\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$.*

The sample version of the hybrid objective function is $\eta_n : \mathbb{S}^{p-1} \to \mathbb{R}$, defined as $\eta_n(\mathbf{u}) = \eta_n(\mathbf{u}; w_1) := w_1\gamma_n(\mathbf{u})^2 + w_2\{\kappa_n(\mathbf{u}) - 3\}^2$, and we next give the limiting behavior of its maximizer. Unsurprisingly, the resulting limiting covariance matrix is up to a multiplicative constant equal to the previous ones.

Theorem 4. *Given model (1), let $\boldsymbol{u}_n$ be any sequence of maximizers of η_n. Then, there exists a sequence of signs $s_n \in \{-1, 1\}$ such that, as $n \to \infty$,*

1) $s_n\boldsymbol{u}_n \to \boldsymbol{\theta}/\|\boldsymbol{\theta}\|$, almost surely.
2) $\sqrt{n}(s_n\boldsymbol{u}_n - \boldsymbol{\theta}/\|\boldsymbol{\theta}\|) \rightsquigarrow \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Psi}_\eta)$, where

$$\boldsymbol{\Psi}_\eta := C_\eta \boldsymbol{\Psi}_U,$$

and

$$C_\eta = \frac{(1+\beta\tau)(1-4\beta)\left(9w_1^2(1+\beta\tau)(2+6\beta\tau+\beta\tau^2)\right)}{\beta\tau^2\{3w_1(1+\beta\tau)(1-4\beta)+4w_2\tau(1-6\beta)^2\}^2} + \frac{(1+\beta\tau)(1-4\beta)\left(24w_1w_2\tau^2\beta(1-6\beta)(6+\tau)\right)+16w_2^2\tau(1-6\beta)^2\Delta}{\beta\tau^2\{3w_1(1+\beta\tau)(1-4\beta)+4w_2\tau(1-6\beta)^2\}^2},$$

where $\Delta := 6 + 24\beta\tau + 9\beta(1-2\beta)\tau^2 + \beta(1-3\beta)\tau^3$ and $\boldsymbol{\Psi}_U$, β, τ are as in Theorem 1.

The constant of proportionality C_η in Theorem 4 is again rather complicated but simplifies in the limit $\tau \to \infty$ to the more manageable, if not intuitive, form,

$$\frac{9w_1^2\beta^2(1-4\beta) + 24w_1w_2\beta(1-4\beta)(1-6\beta) + 16w_2^2(1-6\beta)^2(1-3\beta)}{\{3w_1\beta(1-4\beta) + 4w_2(1-6\beta)^2\}^2}.$$

3.2. Asymptotic comparisons

We next investigate how the efficiency of the hybrid estimator compares to its competitors. Figure 3 shows the relative asymptotic efficiency of the hybrid estimator vs. LDA as a function of the mixing proportion α_1 for the same values of τ as in Figure 1 and for $w_1 \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. Note that the extreme cases $w_1 = 0$ and $w_1 = 1$ are equivalent to using the individual objective functions $\mathbf{u} \mapsto \{\kappa(\mathbf{u}) - 3\}^2$ and $\mathbf{u} \mapsto \gamma(\mathbf{u})^2$, respectively. The curves show somewhat erratic behavior around the points δ_1, δ_2 where kurtosis vanishes but otherwise seem to convey a clear message: inside the interval (δ_1, δ_2) the hybrid estimator is, in general, a superior choice over the individual estimators, whereas

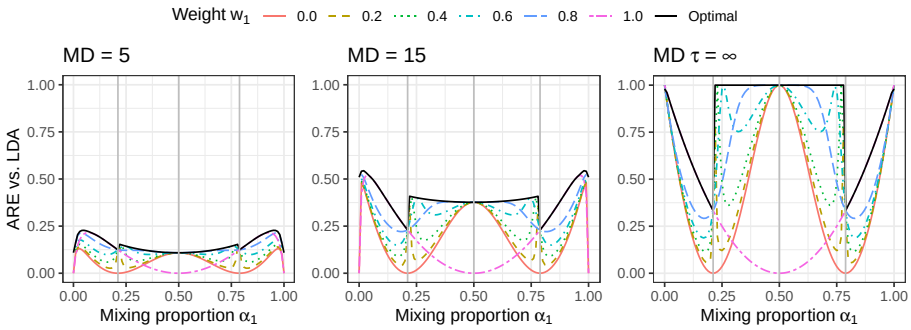

FIG 3. *The relative asymptotic efficiency of the hybrid estimator vs. LDA under different Mahalanobis distances τ between the two group means. The color and type of the lines denote the weighting parameter w_1. The solid black line titled "optimal" traces the efficiencies obtained using the optimal choice of weighting for a given pair (α_1, τ).*

outside of the interval the choice $w_1 = 1$ (corresponding to using skewness only) is preferable over the hybrid estimator.

To obtain a "universal" value of w_1 that yields (in some sense) on average the most efficient estimator over all α_1, we compute with numerical integration the "average efficiency" $A(w_1, \tau)$ of the weight w_1 for a given value of τ as the area between the x-axis and the corresponding efficiency curve. For example, $A(0, 15)$ is the area under the red solid curve in the middle panel of Figure 3. Figure 4 then plots the weights w_1 yielding the maximal value of $A(w_1, \tau)$ as a function of the Mahalanobis distance τ and reveals that, regardless of the separation of the groups, one should optimally consider weights only in the range around $\tau \in (0.725, 0.825)$. This conclusion is rather predictable as kurtosis is based on a higher moment than skewness, meaning that the latter should be given a larger weight in order to obtain a "balanced" combination. As a further interesting observation, when $\tau \to 0$, i.e., when the mixture model approaches the multivariate normal model, the limit of the optimal weight seems to approach the value 0.8, which is the exact weighting used in the Jarque-Bera test statistic for testing normality, $(n/6)\gamma_n^2 + (n/24)(\kappa_n - 3)^2$ [26]. Moreover, the same weighting was also recommended by [28] as an approximation to an entropy-based index.

Whereas Figure 4 aims to obtain a single universally useful value of w_1, in the optimal situation one would always use the particular weighting yielding the highest relative asymptotic efficiency for a given combination of mixing proportion and Mahalanobis distance. The optimal weights are plotted as a function of (α_1, τ) in the heatmap of Figure 5, the most striking features of which are the discontinuities at the horizontal lines $\alpha_1 = \delta_1$, $\alpha_1 = 1/2$ and $\alpha_1 = \delta_2$. These are caused by the fact that the coefficient C_η in Theorem 4 becomes a constant function of w_1 at each of these values of α_1 (where either the skewness or excess kurtosis vanishes). Consequently, there is no unique maximizer w_1 at these points and to emphasize their nature we have chosen to color them in Figure 5 with the corresponding extreme color (e.g., black for $\alpha_1 = 1/2$ where

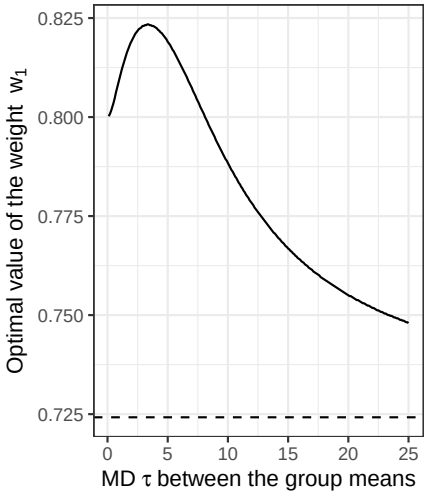

FIG 4. *The weight w_1 yielding the maximal area under the corresponding efficiency curve as a function of the Mahalanobis distance τ. The horizontal dashed line indicates the limit of the curve as $\tau \to \infty$ (approximately 0.7242).*

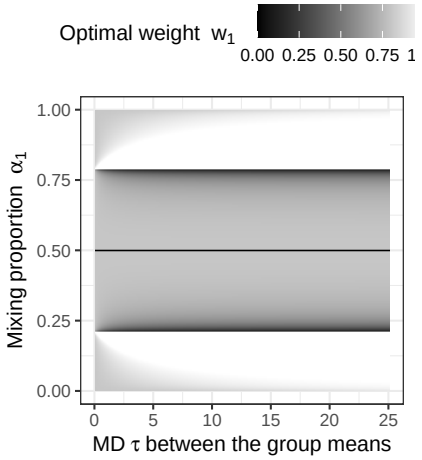

FIG 5. *The heatmap shows as a function of (α_1, τ) the weighting w_1 yielding the highest relative asymptotic efficiency compared to LDA. The three discontinuities correspond to the lines $\alpha_1 = \delta_1$, $\alpha_1 = 1/2$ and $\alpha_1 = \delta_2$.*

skewness carries no information). However, more puzzling are the differing limits when approaching the points δ_1, δ_2 from below and above. Essentially, when one approaches either of these points from inside the interval, the highest efficiency is obtained by focusing all weight on kurtosis which seems very counter-intuitive as in the limit kurtosis carries no information at all. This behaviour is visualized still in more detail in Figure 6 which plots the relative asymptotic efficiency C_η^{-1} as a function of w_1 for $\tau = 5$ and $\alpha_1 - \delta_1 =: \varepsilon \in \{0.001, 0.002, 0.005, 0.010\}$. The weight achieving the maximal efficiency indeed approaches zero as $\varepsilon \to 0$. Algebraically, it is easy to see what is happening: For $\tau \to \infty$ and $\beta \approx 1/6$, the approximation of C_η, obtained by ignoring the terms of order $(1-6\beta)^2$ and higher is $C_\eta \approx \frac{1}{1-4\beta} + \frac{8}{3}\frac{1-w_1}{w_1}\frac{1-6\beta}{\beta(1-4\beta)}$. For $\alpha_1 \to \delta_1$ from the inside of the interval we have $1 - 6\beta < 0$, which yields that C_η is minimized for $w_1 \to 0$. Similarly, for $\alpha_1 \to \delta_1$ from the outside of the interval we have $1 - 6\beta > 0$, which yields that $C_\eta \geq 0$ and shows that it is minimized for $w_1 = 1$. Also, as $\beta \to 1/6$, no matter from which side, C_η converges to constant in w_1, implying that there is no discontinuity in the efficiency value itself. No such behavior is observed for $\alpha_1 \to 0.5$ and the reason for this is that $1 - 4\beta \geq 0$, implying that no sign change occurs when passing the critical value $\alpha_1 = 1/2$.

The efficiencies achieved by the optimal weighting are shown by the solid black line in Figure 5 and indicate that the hybrid estimator is able to reach satisfying levels of efficiency, particularly when the mixing proportion lies in the

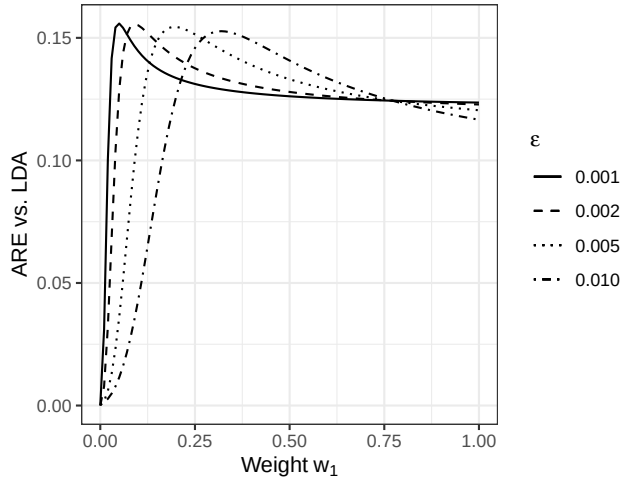

FIG 6. *Relative asymptotic efficiencies of the hybrid estimator as a function of* w_1 *when* $\tau = 5$ *and* $\alpha_1 = \delta_1 + \varepsilon$*, where* $\varepsilon \in \{0.001, 0.002, 0.005, 0.010\}$*. The value of the weight* w_1 *achieving the maximal efficiency can be seen to approach zero when* $\varepsilon \to 0$*.*

interval (δ_1, δ_2). Indeed, in the limit $\tau \to \infty$, within the interval there always exists a weighting that reaches efficiency equal to LDA, as evidenced by the right-most panel of Figure 5. On the other hand, outside of the interval (δ_1, δ_2), LDA is still, even in the limit $\tau \to \infty$, a superior choice. We conjecture that the reason for this critical difference in behavior inside and outside of the interval is that when the value of α_1 is extreme, one of the groups is small, making the pin-pointing of the optimal direction difficult in general, but even more so for the unsupervised methods which have no class information available. However, it is not clear why the particular points δ_1, δ_2 serve as the cut-off values for this behavior.

Finally, note that the discontinuities make the use of the optimal choice of weighting somewhat difficult in practice, as, if one's prior information/guess on the value of the mixing proportion α_1 is even slightly off, relying on the seemingly optimal choice can in the worst case lead to relative efficiency close to zero. Moreover, recall that the previous experiments were asymptotical in nature and do not necessarily reflect the behavior of the method under sample sizes encountered in practical situations. Hence, we suggest using a "safe" universal value of w_1, most preferably falling in the interval $(0.725, 0.825)$ identified in conjunction with Figure 4. For example, Figure 3 shows that the value $w_1 = 0.80$ delivers, for finite τ, performance not far behind the optimal choice for any α_1. However, if one is reasonably certain about the value of α_1 (which, optimally, is far away from δ_1, δ_2) and has n sufficiently large, resorting to the optimal choice is, of course, also possible.

4. High-dimensional projection pursuit

In this section, we study projection pursuit in a high-dimensional regime where both the sample size n and the dimension $p \equiv p_n$ diverge to infinity in a suitable rate. That is, we will work with the n-indexed sequence of high-dimensional centered normal mixtures,

$$\mathbf{x}_n \sim \alpha_1 \mathcal{N}_{p_n}(-\alpha_2 \mathbf{h}_n, \mathbf{\Sigma}_n) + \alpha_2 \mathcal{N}_{p_n}(\alpha_1 \mathbf{h}_n, \mathbf{\Sigma}_n), \tag{2}$$

where the parameters $p_n \in \mathbb{N}$, $\mathbf{h}_n \in \mathbb{R}^{p_n}$ and $\mathbf{\Sigma}_n \in \mathbb{R}^{p_n \times p_n}$ are all functions of the sample size n, and where $\alpha_1, \alpha_2 \in (0,1)$ are taken to be fixed. The notation $\mathbb{S}^{p-1}$ refers to the unit sphere in $\mathbb{R}^p$ and $\|\mathbf{\Sigma}_n\|_2$ denotes the spectral norm of the matrix $\mathbf{\Sigma}_n$. The specific forms for the two locations guarantee that $\mathbf{x}_n$ has zero mean and is without loss of generality.

For simplicity, we work with the assumption that the location of the data is known (and equals zero), allowing us to consider non-centered quantities in our objective functions. For a fixed n, the population and sample skewness-based objective functions are thus $\gamma_{n0}^2 : \mathbb{S}^{p_n-1} \to \mathbb{R}$ and $\gamma_n^2 : \mathbb{S}^{p_n-1} \to \mathbb{R}$, with,

$$\gamma_{n0}(\mathbf{u}) = \frac{\mathrm{E}\{(\mathbf{u}'\mathbf{x}_n)^3\}}{[\mathrm{E}\{(\mathbf{u}'\mathbf{x}_n)^2\}]^{3/2}} \quad \text{and} \quad \gamma_n(\mathbf{u}) = \frac{\frac{1}{n}\sum_{i=1}^n (\mathbf{u}'\mathbf{x}_{ni})^3}{\{\frac{1}{n}\sum_{i=1}^n (\mathbf{u}'\mathbf{x}_{ni})^2\}^{3/2}},$$

while the population and sample kurtosis-based objective functions are $(\kappa_{n0} - 3)^2 : \mathbb{S}^{p_n-1} \to \mathbb{R}$ and $(\kappa_n - 3)^2 : \mathbb{S}^{p_n-1} \to \mathbb{R}$, with,

$$\kappa_{n0}(\mathbf{u}) = \frac{\mathrm{E}\{(\mathbf{u}'\mathbf{x}_n)^4\}}{[\mathrm{E}\{(\mathbf{u}'\mathbf{x}_n)^2\}]^2} \quad \text{and} \quad \kappa_n(\mathbf{u}) = \frac{\frac{1}{n}\sum_{i=1}^n (\mathbf{u}'\mathbf{x}_{ni})^4}{\{\frac{1}{n}\sum_{i=1}^n (\mathbf{u}'\mathbf{x}_{ni})^2\}^2}.$$

Furthermore, denote for a fixed n, the population and sample hybrid objective functions $\eta_{n0} : \mathbb{S}^{p_n-1} \to \mathbb{R}$ and $\eta_n : \mathbb{S}^{p_n-1} \to \mathbb{R}$, with,

$$\begin{aligned} \eta_{n0}(\mathbf{u}; w_1) &= w_1 \gamma_{n0}^2(\mathbf{u}) + (1-w_1)\{\kappa_{n0}(\mathbf{u}) - 3\}^2 \quad \text{and} \\ \eta_n(\mathbf{u}; w_1) &= w_1 \gamma_n^2(\mathbf{u}) + (1-w_1)\{\kappa_n(\mathbf{u}) - 3\}^2, \end{aligned}$$

respectively. Note that, indeed, also the population objective functions are now indexed by n since our model evolves with the growing sample size. By Theorems 3, 4 and Lemma 3, the unique maximizers of $\mathbf{u} \mapsto \gamma_{n0}^2(\mathbf{u})$, $\mathbf{u} \mapsto \{\kappa_{n0}(\mathbf{u}) - 3\}^2$, and $\mathbf{u} \mapsto \eta_{n0}(\mathbf{u}; w_1)$, $w_1 \in (0,1)$, for $\alpha_1 \neq \alpha_2$, $\alpha_1 \neq 1/2 \pm 1/\sqrt{12}$ and $\alpha_1 \in (0,1)$, respectively, are now $\pm\boldsymbol{\theta}_n$ where $\boldsymbol{\theta}_n := \mathbf{\Sigma}_n^{-1}\mathbf{h}_n / \|\mathbf{\Sigma}_n^{-1}\mathbf{h}_n\|$.

As our main results in this section, we show that projection pursuits using all three considered projection indices produce consistent estimates of $\boldsymbol{\theta}_n$ in the high-dimensional regime (2) where the dimension $p_n \to \infty$, as long as its growth rate is sufficiently slow compared to n and the model parameters are bounded in size from both above and below. Note that, since the target parameter $\boldsymbol{\theta}_n$ is indexed by n, by its "consistent estimate" we mean that the angle between our estimator $\mathbf{u}_n$ and $\boldsymbol{\theta}_n$ gets arbitrarily small in the sense of convergence of probability, $|\mathbf{u}_n'\boldsymbol{\theta}_n| - 1 \to_p 0$ as $n \to \infty$.

Our technique of proof is essentially a high-dimensional version of the standard M-estimator argument where additional care has been taken to accommodate the n-indexed model (2).

Theorem 5. *Let $\boldsymbol{x}_{n1}, \ldots, \boldsymbol{x}_{nn}$ be a random sample (a triangular array) from the model (2) where $\alpha_1 \neq \alpha_2$ and assume that there exists $C_1, C_2 > 0$ such that, for all n,*

$$1/C_1 \leq \|\boldsymbol{h}_n\| \leq C_1, \quad \|\boldsymbol{\Sigma}_n\|_2 \leq C_2, \quad \|\boldsymbol{\Sigma}_n^{-1}\|_2 \leq C_2.$$

Assume further that,

$$p_n \to \infty \qquad \text{and} \qquad p_n = o(n^{1/3}).$$

Then any sequence $\boldsymbol{u}_n$ of maximizers of $\boldsymbol{u} \mapsto \gamma_n^2(\boldsymbol{u})$ satisfies,

$$|\boldsymbol{u}_n' \boldsymbol{\theta}_n| - 1 \to_p 0,$$

as $n \to \infty$.

The equivalent statement to Theorem 5 holds for kurtosis-based projection pursuit as well and, since the proof of this is exactly analogous to that of Theorem 5, we have refrained from including it in Appendix B.

Theorem 6. *Let $\boldsymbol{x}_{n1}, \ldots, \boldsymbol{x}_{nn}$ be a random sample (a triangular array) from the model (2) where $\alpha_1 \neq \alpha_2$ and assume that there exists $C_1, C_2 > 0$ such that, for all n,*

$$1/C_1 \leq \|\boldsymbol{h}_n\| \leq C_1, \quad \|\boldsymbol{\Sigma}_n\|_2 \leq C_2, \quad \|\boldsymbol{\Sigma}_n^{-1}\|_2 \leq C_2.$$

Assume further that,

$$p_n \to \infty \qquad \text{and} \qquad p_n = o(n^{1/4}).$$

Then any sequence $\boldsymbol{u}_n$ of maximizers of $\boldsymbol{u} \mapsto \{\kappa_n(\boldsymbol{u}) - 3\}^2$ satisfies,

$$|\boldsymbol{u}_n' \boldsymbol{\theta}_n| - 1 \to_p 0,$$

as $n \to \infty$.

Theorems 5 and 6 imply the equivalent statement for the hybrid estimator as well, whose proof we again omit.

Theorem 7. *Let $\boldsymbol{x}_{n1}, \ldots, \boldsymbol{x}_{nn}$ be a random sample (a triangular array) from the model (2) where $\alpha_1 \neq \alpha_2$ and assume that there exists $C_1, C_2 > 0$ such that, for all n,*

$$1/C_1 \leq \|\boldsymbol{h}_n\| \leq C_1, \quad \|\boldsymbol{\Sigma}_n\|_2 \leq C_2, \quad \|\boldsymbol{\Sigma}_n^{-1}\|_2 \leq C_2.$$

Assume further that,

$$p_n \to \infty \qquad \text{and} \qquad p_n = o(n^{1/4}).$$

Then any sequence $\boldsymbol{u}_n$ *of maximizers of* $\boldsymbol{u} \mapsto \eta_n(\boldsymbol{u}; w_1)$, $0 < w_1 < 1$, *satisfies,*

$$|\boldsymbol{u}_n' \boldsymbol{\theta}_n| - 1 \to_p 0,$$

as $n \to \infty$.

We note that while some of the assumptions of Theorems 5 - 7 may seem counter-intuitive given the nature of the problem (e.g., one might expect that the problem would get easier when the distance $\|\mathbf{h}_n\|$ between the groups increases), they are in fact needed for controlling the third and fourth moments of the distribution. Similarly, the growth rates $p_n = o(n^{1/3})$ and $p_n = o(n^{1/4})$ are indeed a consequence of using third and fourth moment based objective functions, respectively.

Finally, a natural continuation to the above consistency results would be to derive limiting distributions for the corresponding quantities $|\mathbf{u}_n' \boldsymbol{\theta}_n| - 1$, in order to allow efficiency comparisons also in the high-dimensional case. However, this task is beyond our current scope and thus left for future work.

5. Simulations

The three projection pursuit estimators considered here have been discussed in the context of ICA in detail in [62] where also fixed point algorithms for their computations are described. For our purpose here we can use their deflation-based algorithms when only one direction is to be extracted. Projection pursuit is considered notoriously prone to local optima and therefore it is known that good initial values for such algorithms are crucial. [62] suggest to use initial values based on a simple ICA method called FOBI [12]. This is also suitable in our context as the normal mixture is a sub-model of the IC model, see Lemma B.1 in Appendix B. The algorithms of [62] are implemented in the package ICtest [48], which we will use in the following together with R 3.6.1 [58]. Further details about the software used are contained in the appendix.

Let $\mathbf{u}_n$ be any of the estimators of $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ discussed in Section 2. The accuracy of the estimator can in simulations be measured through the inner product $\mathbf{u}_n' \boldsymbol{\theta}/\|\boldsymbol{\theta}\|$, which, by the Cauchy-Schwarz inequality, achieves the absolute value one if and only if the two vectors are parallel. In the continuation, we call the presented inner product the "Maximal similarity index" (MSI). The following lemma presents the limiting distribution of this performance measure.

Lemma 4. *Let the unit length vector* $\boldsymbol{u}_n$ *satisfy* $\sqrt{n}(s_n \boldsymbol{u}_n - \boldsymbol{\theta}/\|\boldsymbol{\theta}\|) \rightsquigarrow \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Psi})$ *for some sequence of signs* $s_n \in \{-1, 1\}$ *and limiting covariance matrix* $\boldsymbol{\Psi}$. *Then, as* $n \to \infty$,

$$2n(1 - s_n \boldsymbol{u}_n' \boldsymbol{\theta}/\|\boldsymbol{\theta}\|) \rightsquigarrow \boldsymbol{z}' \boldsymbol{\Psi} \boldsymbol{z}, \tag{3}$$

where the random vector $\boldsymbol{z}$ *obeys the p-variate standard normal distribution. Moreover, the expected value of the right-hand side of* (3) *is* $\mathrm{tr}(\boldsymbol{\Psi})$.

Note that the sign correction in Lemma 4 can be incorporated in practice by choosing the sign s_n such that the quantity $s_n \mathbf{u}_n' \boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ is positive. In the simulations, we will evaluate the performances of methods through the left-hand side of (3). By Lemma 4 the average of this criterion over several replicates should be close to the trace of the limiting covariance matrix of the corresponding estimator, for sample size n large enough. Hence, the simulations also serve to "verify" our asymptotic results.

In the following simulations four projection pursuit (PP) directions have been calculated: kurtosis based (obtained by maximization of $(\kappa_n - 3)^2$), skewness based (obtained by maximization of γ_n^2), "safe" hybrid estimator (obtained by maximization of η_n for $w_1 = 0.8$) and "optimal" hybrid estimator (obtained by maximization of η_n for $w_1 = w_1(\alpha_1, \tau)$ which maximizes the relative asymptotic efficiency of the hybrid estimator w.r.t. LDA). Sign s_n is chosen such that $s_n \mathbf{u}_n' \boldsymbol{\theta}/||\boldsymbol{\theta}|| \geq 0$. As discussed, the performances of the four presented PP directions $\mathbf{u}_n$ are evaluated using the maximal similarity index $\mathbf{u}_n' \boldsymbol{\theta}/||\boldsymbol{\theta}||$ from above.

For the first simulation setting, the means of maximal similarity indices $s_n \mathbf{u}_n' \boldsymbol{\theta}/||\boldsymbol{\theta}||$ in the simulations are obtained using 1000 random samples in each setting. In each (τ, α_1, n)-setting, where the Mahalanobis distance between the group means $\tau = 1, 2, \ldots, 20$, mixing proportion $\alpha_1 = 0.05, 0.1, \ldots, 0.45, 0.5$ and sample size $n = 500, 1000, 2000, 4000, 8000, 16000, 32000$, $m = 1000$ random samples are generated from a 10-dimensional normal mixture $\alpha_1 \mathcal{N}_{10}(\mathbf{0}, \boldsymbol{\Sigma}) + (1 - \alpha_1)\mathcal{N}_{10}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a covariance matrix with autoregression AR(1) structure with $\rho = 0.6$, and $\boldsymbol{\mu}$ is in each setting chosen randomly such that $\boldsymbol{\mu}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} = \tau$.

The heatmaps of the MSI-values in Figure 7 show that for moderate sample sizes ($n \geq 2000$), both hybrid estimators estimate the optimal LDA direction very well. It is also visible that kurtosis and skewness based PP directions perform very badly when α_1 is near their corresponding discontinuity points. The hybrid estimators suffer from the same problem when α_1 is near $1/2 - 1/\sqrt{12}$. The hybrid estimator with optimal w_1 is performing worse when α_1 is approaching $1/2 - 1/\sqrt{12}$ from the inside the interval $(1/2 - 1/\sqrt{12}, 1/2 + 1/\sqrt{12})$. It is important to recall, that the criterion for choosing the optimal weight w_1 is an asymptotic one and thus might not perform well in small sample settings. Furthermore, to calculate the optimal weight w_1 for the hybrid estimator, one needs to know both the Mahalanobis distance τ between the group means and the mixing proportion α_1, which is a rather unrealistic requirement in practice. Luckily, the hybrid estimator with the "safe" weight $w_1 = 0.8$ shows a very good performance in this simulation study and is therefore recommended in cases where knowledge of α_1 and τ is lacking. Another observation based on this simulation is that for small sample sizes skewness based PP seems to be preferable. This might be due to the fact that moments of order three are easier to estimate than moments of order four.

The corresponding heatmap of the standard deviation of MSI can be found in the Appendix, Figure C.1, and shows that for sample size and distance between the group means moderately large, deviation of the MSI to the corresponding

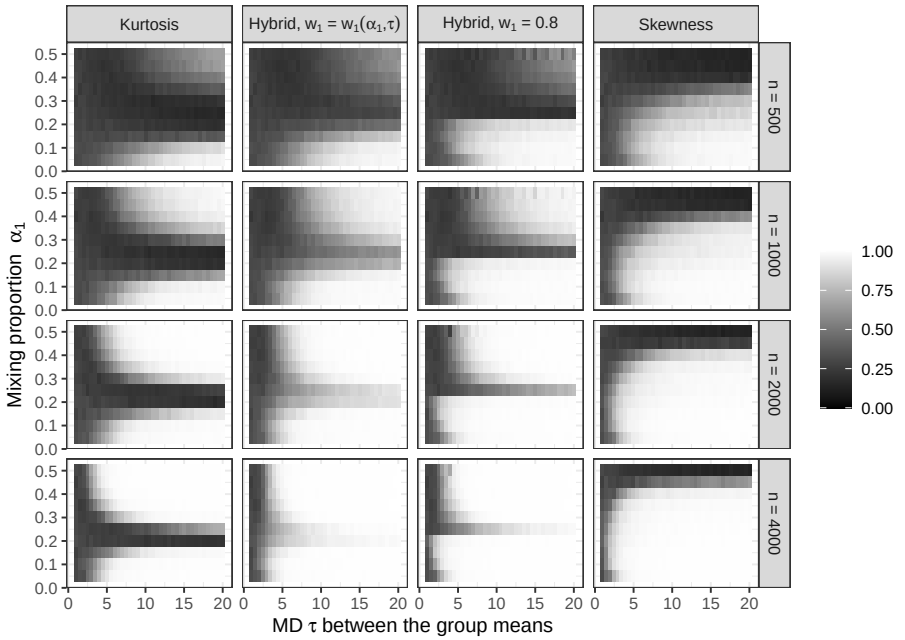

FIG 7. *Average values of the MSI* $s_n \boldsymbol{u}_n'\boldsymbol{\theta}/||\boldsymbol{\theta}||$ *as a function of Mahalanobis distance* τ *between the group means and mixing proportion* α_1*, where* $\boldsymbol{u}_n$ *is one of the four estimators discussed above.*

mean, which is very close to the optimal value of 1, is negligible, for most of the values of the mixing proportion α_1. Heatmaps of mean and standard deviation of the MSI in Figures C.3 and C.2 show that for large sample sizes ($n = 8000, 16000, 32000$) MSI is virtually 1.

In the next simulation, the theoretic results are to be confirmed by exploiting the results of Lemma 4. For that purpose we select three values for $\tau \in \{1, 5, 10\}$, to represent hardly, moderately, and clearly separated clusters, respectively. Then we simulate, for sample sizes $n = 500, 1000, 2000, 4000, 8000, 16000, 32000$, from a three-variate Gaussian mixture model as specified above with $\boldsymbol{\Sigma} = \mathbf{I}_3 + \mathbf{1}_3$, where $\mathbf{1}_3$ is matrix of ones, $\boldsymbol{\mu}_1 = \mathbf{0}$ and $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_2(\tau)$ is for each τ chosen such that Mahalanobis distance between the means is equal to τ, i.e., $\boldsymbol{\mu}_2(1) = (0.68, -0.55, 0.6)'$, $\boldsymbol{\mu}_2(5) = (0.81, -2.24, -0.36)'$, $\boldsymbol{\mu}_2(10) = (3.06, 1.6, -1.11)'$. For $\alpha_1 = 0.1, 0.2, \ldots, 0.5$ we compute then for sample sizes $n = 500, 1000, 2000, 4000, 8000, 16000, 32000$ the means of the $2n(1 - s_n \boldsymbol{u}_n^\top)\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ based on 2000 repetitions. The question is then whether those averages for the presented methods stabilize for the cases they are expected to work. In order not to clutter the figure we computed only for $\alpha_1 = 0.1$ trace $\mathrm{tr}(\boldsymbol{\Psi})$, for the corresponding matrix $\boldsymbol{\Psi}$. Figure 8 shows the results of this simulation and confirms the corresponding theoretic findings from above. The less separated the clusters are the more difficult the estimation and even for $n = 32000$ observations there is no sta-

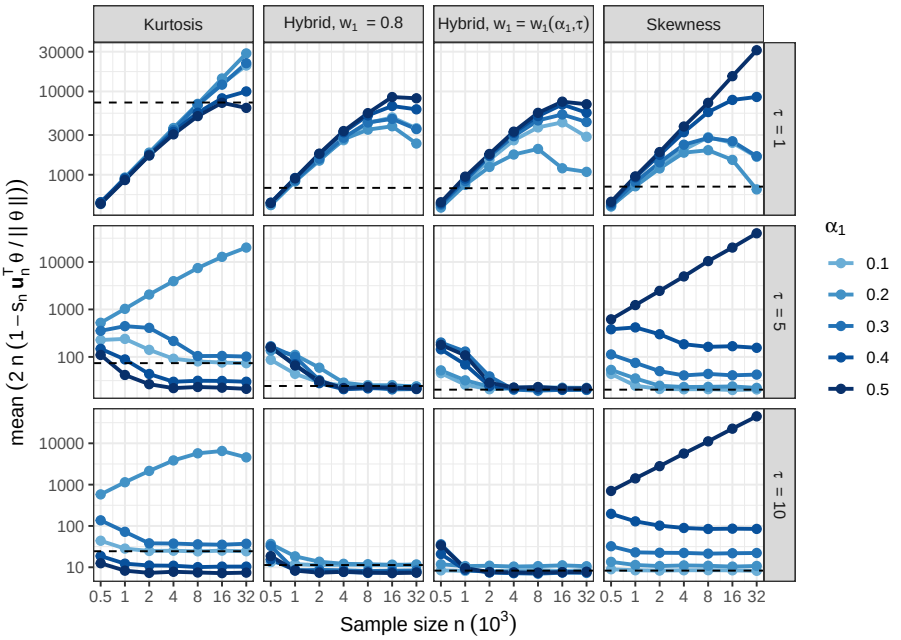

FIG 8. *Average values of the MSI* $s_n \boldsymbol{u}_n' \boldsymbol{\theta}/||\boldsymbol{\theta}||$ *as a function of* n*, for mixing proportion* $\alpha_1 \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ *and Mahalanobis distance between the group means* $\tau \in \{1, 5, 10\}$*, where* $\boldsymbol{u}_n$ *is one of the four estimators discussed above.*

bilization visible. But the more separated the two clusters are, the faster the stabilization. Also, the closer α_1 is to the critical value $1/2 - 1/\sqrt{12}$, the worse is kurtosis based PP. Skewness based PP similarly is better the more skewed the distribution and clearly does not work in the symmetric case. Both hybrid estimators show excellent performance in this setting. It is also clearly visible that the empirical lines for the mixing proportion $\alpha_1 = 0.1$ correspond to the theoretically computed dashed lines given that the groups are separated enough and we assume for $\tau = 1$ the line would be reached for much larger sample sizes. Though we show the theoretic lines only for one mixing proportion the behavior is similar for all others naturally with the exception of skewness not working in the symmetric case.

Principal component analysis (PCA) can also be seen as a projection method where the variance is maximized. PCA is arguably the most popular dimension reduction method and is often used before clustering. While skewness and kurtosis can be related to mixtures the variance does not have the same connection with the discriminant direction as the other cumulants. A theoretic consideration of when PCA can be used to estimate the discriminant is in Appendix A. Here we show an example where PCA fails.

Figure 9 visualizes a sample of size $n = 100$ from our Gaussian mixture model

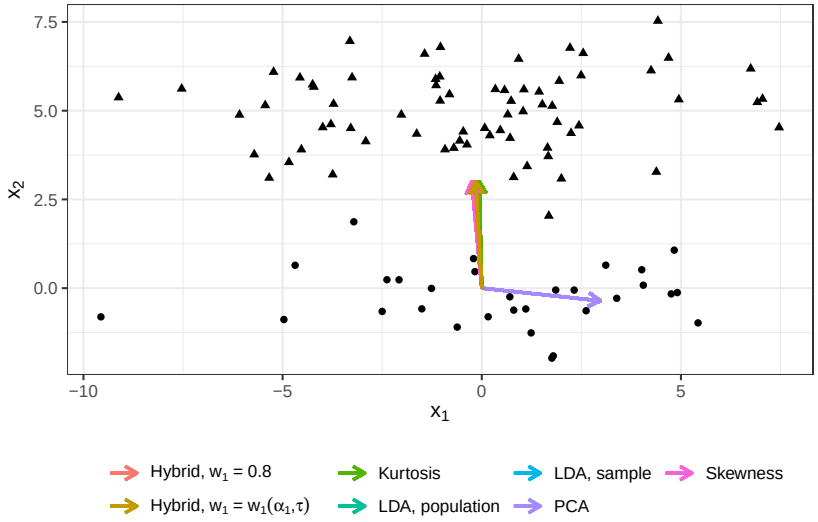

FIG 9. *PP direction $\boldsymbol{u}_n$ based on LDA (here denoted as "LDA, sample"), PCA, and one of the four estimators discussed above.*

with

$$\boldsymbol{\mu}_1 = (0,0)', \quad \boldsymbol{\mu}_2 = (0,5)', \quad \boldsymbol{\Sigma} = \begin{pmatrix} 10 & 0.3 \\ 0.3 & 1 \end{pmatrix}$$

and mixing proportion $\alpha_1 = 0.3$. The figure contains then the direction $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ of the population LDA as well as its estimate together with our four PP methods considered in this section and the direction of the first principal component. As it is clearly visible here, there is no big difference between the methods except for PCA which points in a direction that contains no information for the separation of the two groups.

6. Real data example

To evaluate the performance of the hybrid estimator in a real data set we consider the finance data set available in the R package Rmixmod [36], which consists of 889 records of companies where based on four numeric summary statistics, it should be decided if the company is financially healthy or not, where the information is provided in the data set.

The scatter plot matrix is given in the Appendix as Figure C.4 and shows no clear clusters. As a reference we compute for the data set LDA and then compare this supervised estimate via the estimate $s_n \mathbf{u}_n' \boldsymbol{\theta}_n / \|\boldsymbol{\theta}_n\|$ of the MSI to our hybrid estimator for different weights, to PCA and to an estimate obtained by fitting a two-component Gaussian mixture model via the expectation - maximization (EM) algorithm. Namely, since the nature of the presented problem is essentially

clustering, it is only natural to consider also the classification by EM-algorithm [17]. We consider the EM-algorithm implemented in R package mclust [56], where it is being initialized using the initial partitions from model-based hierarchical agglomerative clustering. By assuming the model (1), the EM-algorithm estimates the parameters of the Gaussian mixture; covariance matrix $\boldsymbol{\Sigma}_{EM}$ and group means $\boldsymbol{\mu}_{1,EM}$ and $\boldsymbol{\mu}_{2,EM}$. Then, the estimated mean and the covariance matrix can be used in order to estimate the linear discriminant direction

$$\mathbf{u}_{EM} = \boldsymbol{\Sigma}_{EM}^{-1}(\boldsymbol{\mu}_{2,EM} - \boldsymbol{\mu}_{1,EM})/\|\boldsymbol{\Sigma}_{EM}(\boldsymbol{\mu}_{2,EM} - \boldsymbol{\mu}_{1,EM})\|.$$

We further refer to the estimator $\mathbf{u}_{EM}$ of the LDA direction as the mclust-based estimate. Figure 10 shows obtained MSI values for the discussed estimators. The figure clearly shows that as long as enough weight is given to kurtosis, the hybrid estimator based PP clearly outperforms both PCA and mclust based estimators, while its performance is poor if skewness gets too much weight. This is not surprising as the amount of healthy (457) and bankrupt (432) companies is almost equal. The weight of 0.8 gives again a good performance. Nevertheless, even though for most values of w_1, and especially for suggested $w_1 \in [0.7, 0.8]$, hybrid estimators clearly outperform both PCA and mclust based estimators, the achieved MSI values of around 0.5 are not ideal. Such performance can be explained by the low sample size and that the cluster centers are not that far apart, as is shown in the boxplots of Figure C.5 in the Appendix C, which also indicates that the results obtained by mclust based estimation are not satisfactory.

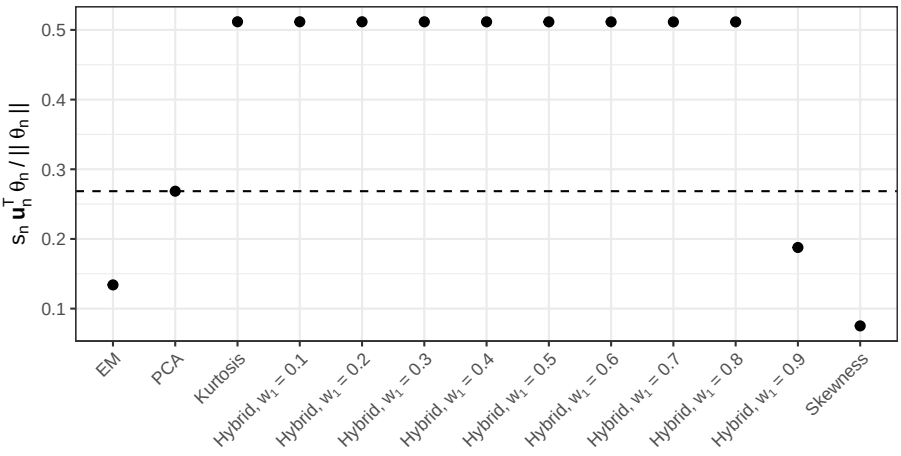

FIG 10. *MSI $s_n \boldsymbol{u}_n \boldsymbol{\theta}_n / \|\boldsymbol{\theta}_n\|$ for the finance data, where $\boldsymbol{u}_n$ is the direction based on PCA and PP estimators for $w_1 = 0, 0.1, 0.2, \ldots, 0.9, 1$. $\boldsymbol{\theta}_n$ is the direction based on LDA.*

7. Discussion

In this paper, we conducted an asymptotic comparison of two popular estimators of the linear discriminant direction, LDA, and projection pursuit based on skewness and kurtosis. For the latter, we proposed using the weighted combination of kurtosis and skewness as the projection index (giving the individual cumulants as special cases). Both the theoretical results and simulations indicate that, with a suitable choice of weighting, such projection pursuit achieves reasonably good performance compared to LDA (e.g., around 15% relative asymptotic efficiency if the Mahalanobis distance between the groups is 5, see Figure 3), considering it operates in the complete absence of any label information. Moreover, in the extreme case of balanced and infinitely well-separated groups, projection pursuit is able to reach asymptotic efficiency equal to LDA with an optimal choice of weighting.

The use of our optimal weighting results is difficult in practice by the discontinuities around the mixing proportions δ_1, δ_2 observed in Section 3, see Figure 5. As such, unless one is sure that the mixing proportion is not in these regions, our recommendation is to use a universal choice of weighting, anything between 0.7 and 0.8 (as the weight for skewness) likely being a good choice.

At first, we thought that the discontinuities, and the surprising recommendation to favor kurtosis just outside the interval (δ_1, δ_2), might be caused by the uneven robustness properties of skewness and kurtosis in the objective function. Namely, being based on fourth moments, kurtosis is more affected by outliers than skewness (despite the standardization with second moments). Hence, we also considered using the "balanced" objective function,

$$\eta^*(\mathbf{u}) := w_1\gamma(\mathbf{u})^{8/3} + w_2\{\kappa(\mathbf{u}) - 3\}^2,$$

in an attempt to put skewness and kurtosis on an equal footing. However, the asymptotic properties of η^* (not shown here) turn out to be essentially the same as for η, including the discontinuities which are also observed for it. Note also that the discontinuous behavior was observed also in [62], where the normal mixture model was studied using independent component analysis.

Similarly one could extend our considerations here to many other PP indices as well, which often are modifications of skewness or kurtosis (see e.g. [24]) or otherwise motivated to be useful in clustering or structure detection, see, for example, [13, 16] and references therein for alternative indices. These indices are however often computationally expensive and therefore much less popular than skewness and kurtosis.

Finally, besides projection pursuit, there exist also other unsupervised estimators of the linear discriminant. For example, it is known that the linear discriminant can be reconstructed using invariant coordinate selection (ICS) [59] where two scatter matrices are jointly diagonalized. Especially when using the regular covariance matrix and the scatter matrix of fourth moments in this context as, for example, suggested in [1, 51], would allow a theoretic comparison (actually, this combination corresponds to the FOBI-method mentioned in Sec-

tion 5). Comparisons of LDA to other supervised and unsupervised classification methods are given for example in [9, 17].

Another prospective line of work is the extension of our asymptotic results to location mixtures of Gaussians with proportional covariances. Such a model was, for example, considered in [52] while studying the multivariate-outlier detection problem, as well as in [40]. [54] argues how this model is particularly difficult to analyze from an outlier-detection point of view since it corresponds to a situation where the outliers form a cluster with the same shape as the bulk of the data. Another natural extension is then in the direction of mixtures of elliptical distributions, as [50] indeed showed that projection pursuit yields a Fisher consistent estimator of the linear discriminant also in the case of general elliptical families. Additionally, [38] studied estimation of linear discriminant using skewness-based projection pursuit in mixtures of two symmetric distributions with unequal means and proportional covariances. Similarly, another possible extension is to the case of multiple groups instead of just two or to groups with unequal covariance matrices.

Appendix A: Equivalent results for PCA

While PCA manages to capture the linear discriminant direction only under very specific conditions, and cannot thus be reasonably seen as a "unsupervised" estimator of it, we still give for it in the following, for completeness, equivalent results to the ones in Sections 2 and 3. The first result, detailing conditions required for the Fisher consistency of PCA, is *qualitatively* well-known in the literature (see, e.g., Section 9.1 in [27]), but, as far as we know, the exact eigenvalue bound is novel.

Lemma A.1. *Given model* (1)*, the following two are equivalent:*

i) The vector $\boldsymbol{h} := \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ *is an eigenvector of* $\boldsymbol{\Sigma}$ *and, denoting the corresponding eigenvalue with* ϕ*, the second-to-largest eigenvalue* $\phi_2\{\mathrm{Cov}(\boldsymbol{x})\}$ *of* $\mathrm{Cov}(\boldsymbol{x})$ *satisfies*

$$\phi_2\{\mathrm{Cov}(\boldsymbol{x})\} < \phi(1 + \beta\tau),$$

where β *and* τ *are as in Theorem 2.*

ii) The unique leading unit length eigenvectors of $\mathrm{Cov}(\boldsymbol{x})$ *are* $\pm\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$.

Lemma A.1 states that for the first PC to recover the discriminant direction, it is necessary that the difference $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ between the group means is an eigenvector of $\boldsymbol{\Sigma}$. However, it is not necessary for it to be the leading eigenvector but instead, roughly, the more well-separated the groups are (large Mahalanobis distance τ) and the more balanced the groups are (large β), the smaller the corresponding eigenvalue can be relative to the rest of the spectrum. Note also that in the spherical case, $\boldsymbol{\Sigma} \propto \mathbf{I}_p$, the first part of condition i) in Lemma A.1 is trivially satisfied.

Asymptotic results for PCA are also well-known, see, e.g., [2, 14], and the following theorem details the strong consistency and the limiting normality of the first PC in our particular scenario. For completeness, we provide a proof.

Theorem A.1. *Given model (1), assume that the condition i) (or, equivalently, ii)) holds and let $\boldsymbol{u}_n$ be any sequence of leading unit-length eigenvectors of the sample covariance matrix $\boldsymbol{C}_n$ of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Then, there exists a sequence of signs $s_n \in \{-1, 1\}$ such that, as $n \to \infty$,*

1) $s_n \boldsymbol{u}_n \to \boldsymbol{\theta}/\|\boldsymbol{\theta}\|$, almost surely.
2) $\sqrt{n}(s_n \boldsymbol{u}_n - \boldsymbol{\theta}/\|\boldsymbol{\theta}\|) \rightsquigarrow \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Psi}_{\mathrm{PCA}})$, where

$$\boldsymbol{\Psi}_{\mathrm{PCA}} := \left(\frac{1+\beta\tau}{\|\boldsymbol{\theta}\|^2}\right) \boldsymbol{M}^{\dagger} \left[\tau(\boldsymbol{\Sigma} - \tau^{-1}\boldsymbol{h}\boldsymbol{h}') + (1+\beta\tau)\{\kappa(\boldsymbol{\theta}) - 1\}\boldsymbol{h}\boldsymbol{h}'\right] \boldsymbol{M}^{\dagger},$$

where $\boldsymbol{M} := \boldsymbol{\Sigma} + \beta\boldsymbol{h}\boldsymbol{h}' - \lambda_1 \boldsymbol{I}_p$, λ_1 is the eigenvalue of $\mathrm{Cov}(\boldsymbol{x})$ corresponding to the eigenvector $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$, $\boldsymbol{M}^{\dagger}$ denotes the Moore-Penrose pseudoinverse of $\boldsymbol{M}$ and $\kappa(\boldsymbol{\theta})$ is the kurtosis of $\boldsymbol{x}$ in the direction $\boldsymbol{\theta}$.

It is evident from part 2) of Theorem A.1 that the limiting covariance matrix of the PCA-based estimator is not proportional to the four others derived in Theorems 1, 2, 3 and 4. However, proportionality is reached in the special case where the group covariance matrix is spherical, $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_p$, for some $\sigma^2 > 0$. In this case, $\mathbf{h}/\|\mathbf{h}\| = \boldsymbol{\theta}/\|\boldsymbol{\theta}\|$, the Moore-Penrose pseudoinverses in Theorem A.1 equal $(\boldsymbol{\Sigma} + \beta\mathbf{h}\mathbf{h}' - \phi\mathbf{I}_p)^{\dagger} = -1/(\beta\|\mathbf{h}\|^2)(\mathbf{I}_p - \mathbf{h}\mathbf{h}'/\|\mathbf{h}\|^2)$ and the limiting covariance matrix $\boldsymbol{\Psi}_{\mathrm{PCA}}$ can be expressed as

$$\boldsymbol{\Psi}_{\mathrm{PCA}} = \frac{1}{\tau\beta}\left(\frac{1+\beta\tau}{\|\boldsymbol{\theta}\|^2\beta}\right)\left(\mathbf{I}_p - \frac{\boldsymbol{\theta}\boldsymbol{\theta}'}{\|\boldsymbol{\theta}\|^2}\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{I}_p - \frac{\boldsymbol{\theta}\boldsymbol{\theta}'}{\|\boldsymbol{\theta}\|^2}\right).$$

Comparison to Theorem 1 now reveals that the relative asymptotic efficiency of PCA vs. LDA equals $\tau\beta$, showing, in particular, that in the balanced case with $\alpha_1 = \alpha_2$ PCA surpasses LDA in asymptotic efficiency as soon as the Mahalanobis distance between the groups is greater than 4. Moreover, in the limit $\tau \to \infty$, PCA is infinitely more efficient than LDA regardless of the mixing proportion. This counterintuitive result is, of course, not something one should rely on in practice, as the conditions required to achieve the situation are being very restrictive.

Appendix B: Proofs of technical results

Lemma B.1. *Let $\boldsymbol{x} \sim \alpha_1\mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + \alpha_2\mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, where $\alpha_1, \alpha_2 > 0$, $\alpha_1 + \alpha_2 = 1$, $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^p$, $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ is full rank. Then $\boldsymbol{x}$ is an independent component model, i.e., there exists an invertible matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times p}$ such that $\boldsymbol{\Gamma}\{\boldsymbol{x} - \mathrm{E}(\boldsymbol{x})\}$ has independent components.*

Proof of Lemma B.1. We have

$$\mathbf{x} - \mathrm{E}(\mathbf{x}) \sim \alpha_1\mathcal{N}_p(-\alpha_2\mathbf{h}, \boldsymbol{\Sigma}) + \alpha_2\mathcal{N}_p(\alpha_1\mathbf{h}, \boldsymbol{\Sigma}),$$

where $\mathbf{h} := \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$. Let then $\boldsymbol{\Gamma} := \mathbf{U}'\boldsymbol{\Sigma}^{-1/2}$ where $\mathbf{U}$ is an orthogonal matrix satisfying $\mathbf{U}'\boldsymbol{\Sigma}^{-1/2}\mathbf{h} \propto \mathbf{e}_1$ and $\mathbf{e}_1$ is the first canonical basis vector of $\mathbb{R}^p$ (such an $\mathbf{U}$ always exists as $\boldsymbol{\Sigma}$ is full rank and $\mathbf{h} \neq \mathbf{0}$). Now,

$$\boldsymbol{\Gamma}\{\mathbf{x} - \mathrm{E}(\mathbf{x})\} \sim \alpha_1 \mathcal{N}_p(-\alpha_2 b\mathbf{e}_1, \mathbf{I}_p) + \alpha_2 \mathcal{N}_p(\alpha_1 b\mathbf{e}_1, \mathbf{I}_p),$$

for some $b \neq 0$. The result now follows by writing out the density function of $\boldsymbol{\Gamma}\{\mathbf{x} - \mathrm{E}(\mathbf{x})\}$ and observing that it factors into a product of the density of a univariate Gaussian mixture and the densities of $p-1$ univariate Gaussians with zero means. □

Proof of Theorem 1. The estimator $\mathbf{w}$ is translation invariant, meaning that we may, without loss of generality, assume that $\mathrm{E}(\mathbf{x}) = \mathbf{0}$. Under this, the model (1) takes the form

$$y \sim \mathrm{Ber}(\alpha_1) \quad \text{and} \quad \mathbf{x} \mid y \sim \mathcal{N}_p\{-y\alpha_2\mathbf{h} + (1-y)\alpha_1\mathbf{h}, \boldsymbol{\Sigma}\}, \tag{B.1}$$

where $\mathbf{h} := \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$. Concurrently, $\mathbf{x} \sim \alpha_1\mathcal{N}_p(-\alpha_2\mathbf{h}, \boldsymbol{\Sigma}) + \alpha_2\mathcal{N}_p(\alpha_1\mathbf{h}, \boldsymbol{\Sigma})$ and $\mathrm{Cov}(\mathbf{x}) = \boldsymbol{\Sigma} + \beta\mathbf{h}\mathbf{h}'$.

We begin by deriving asymptotic linearizations for $\bar{\mathbf{x}}_{n2} - \bar{\mathbf{x}}_{n1}$. Let in the following $\beta := \alpha_1\alpha_2$. By LLN, $\bar{y}_n \to_p \alpha_1$ and $(1/n)\sum_i y_i\mathbf{x}_i \to_p -\beta\mathbf{h}$. Hence, the relation $\bar{y}_n\bar{\mathbf{x}}_{n1} = (1/n)\sum_i y_i\mathbf{x}_i$ shows that $\bar{\mathbf{x}}_{n1} \to_p -\alpha_2\mathbf{h}$. We further have the expansion,

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\mathbf{x}_i + \beta\mathbf{h}\right) = \sqrt{n}\,(\bar{y}_n - \alpha_1)\,\bar{\mathbf{x}}_{n1} + \alpha_1\sqrt{n}(\bar{\mathbf{x}}_{n1} + \alpha_2\mathbf{h}),$$

which, by CLT, shows that $\bar{\mathbf{x}}_{n1}$ is asymptotically normal,

$$\alpha_1\sqrt{n}(\bar{\mathbf{x}}_{n1} + \alpha_2\mathbf{h}) = \sqrt{n}\left(\frac{1}{n}\sum_i y_i\mathbf{x}_i + \beta\mathbf{h}\right) + \sqrt{n}(\bar{y}_n - \alpha_1)\alpha_2\mathbf{h} + o_p(1).$$

One can similarly show that,

$$\alpha_2\sqrt{n}(\bar{\mathbf{x}}_{n2} - \alpha_1\mathbf{h}) = \sqrt{n}\left\{\frac{1}{n}\sum_i (1-y_i)\mathbf{x}_i - \beta\mathbf{h}\right\} + \sqrt{n}(\bar{y}_n - \alpha_1)\alpha_1\mathbf{h} + o_p(1).$$

Defining $\mathbf{a}_{n1} := \sqrt{n}\{(1/n)\sum_i y_i\mathbf{x}_i + \beta\mathbf{h}\}$, $\mathbf{a}_{n2} := \sqrt{n}\{(1/n)\sum_i (1-y_i)\mathbf{x}_i - \beta\mathbf{h}\}$ and $a_{n3} := \sqrt{n}(\bar{y}_n - \alpha_1)$, the previous two can be written as $\alpha_1\sqrt{n}(\bar{\mathbf{x}}_{n1} + \alpha_2\mathbf{h}) = \mathbf{a}_{n1} + a_{n3}\alpha_2\mathbf{h} + o_p(1)$ and $\alpha_2\sqrt{n}(\bar{\mathbf{x}}_{n2} - \alpha_1\mathbf{h}) = \mathbf{a}_{n2} + a_{n3}\alpha_1\mathbf{h} + o_p(1)$. The two in combination yield the desired linearization,

$$\beta\sqrt{n}(\bar{\mathbf{x}}_{n2} - \bar{\mathbf{x}}_{n1} - \mathbf{h}) = \alpha_1\mathbf{a}_{n2} - \alpha_2\mathbf{a}_{n1} + (\alpha_1 - \alpha_2)a_{n3}\mathbf{h} + o_p(1).$$

We then derive a similar expansion for the pooled covariance matrix $\mathbf{S}_n$. It is straightforwardly seen that $\sum_{i=1}^n y_i(\mathbf{x}_i - \bar{\mathbf{x}}_{n1})(\mathbf{x}_i - \bar{\mathbf{x}}_{n1})' = \sum_i y_i\mathbf{x}_i\mathbf{x}_i' - \sum_i y_i\bar{\mathbf{x}}_1\bar{\mathbf{x}}_1'$. This together with the equivalent formula for the second group yields

$$\mathbf{S}_n = \frac{1}{n-2}\left\{\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i' - \sum_{i=1}^n y_i\bar{\mathbf{x}}_{n1}\bar{\mathbf{x}}_{n1}' - \sum_{i=1}^n (1-y_i)\bar{\mathbf{x}}_{n2}\bar{\mathbf{x}}_{n2}'\right\}.$$

Since $\sqrt{n}(\mathbf{S}_n - \boldsymbol{\Sigma}) = \sqrt{n}[\{(n-2)/n\}\mathbf{S}_n - \boldsymbol{\Sigma}] + o_p(1)$, we have the expansion,

$$\begin{aligned}\sqrt{n}(\mathbf{S}_n - \boldsymbol{\Sigma}) = \sqrt{n}&\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i' - (\boldsymbol{\Sigma} + \beta\mathbf{h}\mathbf{h}')\right\}\\ &- \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\bar{\mathbf{x}}_{n1}\bar{\mathbf{x}}_{n1}' - \alpha_2\beta\mathbf{h}\mathbf{h}'\right)\\ &- \sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}(1-y_i)\bar{\mathbf{x}}_{n2}\bar{\mathbf{x}}_{n2}' - \alpha_1\beta\mathbf{h}\mathbf{h}'\right\} + o_p(1).\end{aligned}$$

The second term above expands as,

$$\begin{aligned}\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\bar{\mathbf{x}}_{n1}\bar{\mathbf{x}}_{n1}' - \alpha_2\beta\mathbf{h}\mathbf{h}'\right) =& \alpha_2^2 a_{n3}\mathbf{h}\mathbf{h}' - \beta\sqrt{n}(\bar{\mathbf{x}}_{n1} + \alpha_2\mathbf{h})\mathbf{h}'\\ &- \beta\mathbf{h}\sqrt{n}(\bar{\mathbf{x}}_{n1} + \alpha_2\mathbf{h})' + o_p(1),\end{aligned}$$

and the third as,

$$\begin{aligned}\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}(1-y_i)\bar{\mathbf{x}}_{n2}\bar{\mathbf{x}}_{n2}' - \alpha_1\beta\mathbf{h}\mathbf{h}'\right\} =& - a_{n3}\alpha_1^2\mathbf{h}\mathbf{h}' + \beta\sqrt{n}(\bar{\mathbf{x}}_{n2} - \alpha_1\mathbf{h})\mathbf{h}'\\ &+ \beta\mathbf{h}\sqrt{n}(\bar{\mathbf{x}}_{n2} - \alpha_1\mathbf{h})' + o_p(1).\end{aligned}$$

Denoting then $\mathbf{A}_{n4} := \sqrt{n}\{(1/n)\sum_i \mathbf{x}_i\mathbf{x}_i' - (\boldsymbol{\Sigma} + \beta\mathbf{h}\mathbf{h}')\}$, the linearizations derived earlier for $\bar{\mathbf{x}}_{n1}$ and $\bar{\mathbf{x}}_{n2}$ allow us to write,

$$\begin{aligned}\sqrt{n}(\mathbf{S}_n - \boldsymbol{\Sigma}) = \mathbf{A}_{n4} &+ a_{n3}(\alpha_2 - \alpha_1)\mathbf{h}\mathbf{h}' + \alpha_2\mathbf{a}_{n1}\mathbf{h}' + \alpha_2\mathbf{h}\mathbf{a}_{n1}' - \alpha_1\mathbf{a}_{n2}\mathbf{h}'\\ &- \alpha_1\mathbf{h}\mathbf{a}_{n2}' + o_p(1).\end{aligned}$$

The above in particular shows that $\mathbf{S}$ is asymptotically normal. Hence, the relation

$$\mathbf{0} = \sqrt{n}(\mathbf{S}_n\mathbf{S}_n^{-1} - \mathbf{I}_p) = \sqrt{n}(\mathbf{S}_n - \boldsymbol{\Sigma})\mathbf{S}_n^{-1} + \boldsymbol{\Sigma}\sqrt{n}(\mathbf{S}_n^{-1} - \boldsymbol{\Sigma}^{-1}),$$

gives $\sqrt{n}(\mathbf{S}_n^{-1} - \boldsymbol{\Sigma}^{-1}) = -\boldsymbol{\Sigma}^{-1}\sqrt{n}(\mathbf{S}_n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1} + o_p(1)$.

We are now equipped to derive the limiting distribution of the optimal direction $\mathbf{w}_n = \mathbf{S}_n^{-1}(\bar{\mathbf{x}}_{n2} - \bar{\mathbf{x}}_{n1})$. Recalling that $\boldsymbol{\theta} = \boldsymbol{\Sigma}^{-1}\mathbf{h}$, we have, by the calculus of $o_p(1)$ and $O_p(1)$ sequences,

$$\begin{aligned}\sqrt{n}(\mathbf{w}_n - \boldsymbol{\theta}) &= \sqrt{n}(\mathbf{S}_n^{-1} - \boldsymbol{\Sigma}^{-1})\mathbf{h} + \boldsymbol{\Sigma}^{-1}\sqrt{n}(\bar{\mathbf{x}}_{n2} - \bar{\mathbf{x}}_{n1} - \mathbf{h}) + o_p(1)\\ &= -\boldsymbol{\Sigma}^{-1}\sqrt{n}(\mathbf{S}_n - \boldsymbol{\Sigma})\boldsymbol{\theta} + \boldsymbol{\Sigma}^{-1}\sqrt{n}(\bar{\mathbf{x}}_{n2} - \bar{\mathbf{x}}_{n1} - \mathbf{h}) + o_p(1).\end{aligned}$$

Hence,

$$\begin{aligned}\beta\boldsymbol{\Sigma}\sqrt{n}(\mathbf{w}_n - \boldsymbol{\theta}) &= -\beta\mathbf{A}_{n4}\boldsymbol{\theta} - a_{n3}\beta(\alpha_2 - \alpha_1)\mathbf{h}\mathbf{h}'\boldsymbol{\theta} - \alpha_2\beta\mathbf{a}_{n1}\mathbf{h}'\boldsymbol{\theta}\\ &\quad - \alpha_2\beta\mathbf{h}\mathbf{a}_{n1}'\boldsymbol{\theta} + \alpha_1\beta\mathbf{a}_{n2}\mathbf{h}'\boldsymbol{\theta} + \alpha_1\beta\mathbf{h}\mathbf{a}_{n2}'\boldsymbol{\theta}\\ &\quad + \alpha_1\mathbf{a}_{n2} - \alpha_2\mathbf{a}_{n1} + a_{n3}(\alpha_1 - \alpha_2)\mathbf{h} + o_p(1)\\ &= -\beta\mathbf{A}_{n4}\boldsymbol{\theta} - a_{n3}(\alpha_2 - \alpha_1)(\beta\mathbf{h}'\boldsymbol{\theta} + 1)\mathbf{h} - \alpha_2(\beta\mathbf{h}'\boldsymbol{\theta} + 1)\mathbf{a}_{n1}\\ &\quad + \alpha_1(\beta\mathbf{h}'\boldsymbol{\theta} + 1)\mathbf{a}_{n2} - \alpha_2\beta\mathbf{h}\mathbf{a}_{n1}'\boldsymbol{\theta} + \alpha_1\beta\mathbf{h}\mathbf{a}_{n2}'\boldsymbol{\theta} + o_p(1).\end{aligned}$$

By the definitions of $\mathbf{a}_{n1}, \mathbf{a}_{n2}, a_{n3}$ and $\mathbf{A}_{n4}$ and CLT, the limiting covariance matrix of $\beta\boldsymbol{\Sigma}\sqrt{n}(\mathbf{w}-\boldsymbol{\theta})$ is

$$\begin{aligned}\mathrm{Cov}\{ &-\beta(\boldsymbol{\theta}'\mathbf{x})\mathbf{x} - (\alpha_2-\alpha_1)\Delta y\mathbf{h} - \alpha_2\Delta y\mathbf{x} + \alpha_1\Delta(1-y)\mathbf{x} - \alpha_2\beta y(\boldsymbol{\theta}'\mathbf{x})\mathbf{h} \\ &+ \alpha_1\beta(1-y)(\boldsymbol{\theta}'\mathbf{x})\mathbf{h}\},\end{aligned}$$

where $\Delta := \beta\lambda+1$ and $\lambda := \boldsymbol{\theta}'\mathbf{h}$. The covariance matrix is a sum of a total of 36 terms, which we next compute one-by-one. We use the notation $\gamma := \alpha_1^3 + \alpha_2^3$.

(1, 1): $\beta^2(1+3\beta\lambda+\beta(\gamma-\beta)\lambda^2)\mathbf{h}\mathbf{h}' + \beta^2\lambda(1+\beta\lambda)\boldsymbol{\Sigma}$.
(1, 2): $\beta^2(\alpha_2-\alpha_1)^2(\beta\lambda+1)\lambda\mathbf{h}\mathbf{h}'$.
(1, 3): $-\beta^2\alpha_2\lambda(\beta\lambda+1)\boldsymbol{\Sigma} - \beta^2\alpha_2(\beta\lambda+1)\{1+\lambda\alpha_2(\alpha_2-\alpha_1)\}\mathbf{h}\mathbf{h}'$.
(1, 4): $-\beta^2\alpha_1\lambda(\beta\lambda+1)\boldsymbol{\Sigma} - \beta^2\alpha_1(\beta\lambda+1)\{1+\lambda\alpha_1(\alpha_1-\alpha_2)\}\mathbf{h}\mathbf{h}'$.
(1, 5): $\beta^3\alpha_2\lambda\{\lambda\alpha_2(\alpha_1-\alpha_2)-2\}\mathbf{h}\mathbf{h}'$.
(1, 6): $\beta^3\alpha_1\lambda\{\lambda\alpha_1(\alpha_2-\alpha_1)-2\}\mathbf{h}\mathbf{h}'$.
(2, 1): $\beta^2(\alpha_2-\alpha_1)^2(\beta\lambda+1)\lambda\mathbf{h}\mathbf{h}'$.
(2, 2): $(\alpha_2-\alpha_1)^2(\beta\lambda+1)^2\beta\mathbf{h}\mathbf{h}'$.
(2, 3): $-(\alpha_2-\alpha_1)(\beta\lambda+1)^2\alpha_2^2\beta\mathbf{h}\mathbf{h}'$.
(2, 4): $(\alpha_2-\alpha_1)\alpha_1^2(\beta\lambda+1)^2\beta\mathbf{h}\mathbf{h}'$.
(2, 5): $-(\alpha_2-\alpha_1)(\beta\lambda+1)\alpha_2^2\beta^2\lambda\mathbf{h}\mathbf{h}'$.
(2, 6): $(\alpha_2-\alpha_1)(\beta\lambda+1)\alpha_1^2\beta^2\lambda\mathbf{h}\mathbf{h}'$.
(3, 1): $-\beta^2\alpha_2\lambda(\beta\lambda+1)\boldsymbol{\Sigma} - \beta^2\alpha_2(\beta\lambda+1)\{1+\lambda\alpha_2(\alpha_2-\alpha_1)\}\mathbf{h}\mathbf{h}'$.
(3, 2): $-(\alpha_2-\alpha_1)(\beta\lambda+1)^2\alpha_2^2\beta\mathbf{h}\mathbf{h}'$.
(3, 3): $\alpha_2\beta(\beta\lambda+1)^2\boldsymbol{\Sigma} + \alpha_2^4\beta(\beta\lambda+1)^2\mathbf{h}\mathbf{h}'$.
(3, 4): $-\beta^3(\beta\lambda+1)^2\mathbf{h}\mathbf{h}'$.
(3, 5): $\alpha_2(\beta\lambda+1)\beta^2(1+\lambda\alpha_2^3)\mathbf{h}\mathbf{h}'$.
(3, 6): $-\beta^4\lambda(\beta\lambda+1)\mathbf{h}\mathbf{h}'$.
(4, 1): $-\beta^2\alpha_1\lambda(\beta\lambda+1)\boldsymbol{\Sigma} - \beta^2\alpha_1(\beta\lambda+1)\{1+\lambda\alpha_1(\alpha_1-\alpha_2)\}\mathbf{h}\mathbf{h}'$.
(4, 2): $(\alpha_2-\alpha_1)\alpha_1^2(\beta\lambda+1)^2\beta\mathbf{h}\mathbf{h}'$.
(4, 3): $-\beta^3(\beta\lambda+1)^2\mathbf{h}\mathbf{h}'$.
(4, 4): $\alpha_1\beta(\beta\lambda+1)^2\boldsymbol{\Sigma} + \alpha_1^4\beta(\beta\lambda+1)^2\mathbf{h}\mathbf{h}'$.
(4, 5): $-\beta^4\lambda(\beta\lambda+1)\mathbf{h}\mathbf{h}'$.
(4, 6): $\alpha_1(\beta\lambda+1)\beta^2(1+\lambda\alpha_1^3)\mathbf{h}\mathbf{h}'$.
(5, 1): $\beta^3\alpha_2\lambda\{\lambda\alpha_2(\alpha_1-\alpha_2)-2\}\mathbf{h}\mathbf{h}'$.
(5, 2): $-(\alpha_2-\alpha_1)(\beta\lambda+1)\alpha_2^2\beta^2\lambda\mathbf{h}\mathbf{h}'$.
(5, 3): $\alpha_2(\beta\lambda+1)\beta^2(1+\lambda\alpha_2^3)\mathbf{h}\mathbf{h}'$.
(5, 4): $-\beta^4\lambda(\beta\lambda+1)\mathbf{h}\mathbf{h}'$.
(5, 5): $\alpha_2\beta^3\lambda(1+\lambda\alpha_2^3)\mathbf{h}\mathbf{h}'$.
(5, 6): $-\beta^5\lambda^2\mathbf{h}\mathbf{h}'$.
(6, 1): $\beta^3\alpha_1\lambda\{\lambda\alpha_1(\alpha_2-\alpha_1)-2\}\mathbf{h}\mathbf{h}'$.
(6, 2): $(\alpha_2-\alpha_1)(\beta\lambda+1)\alpha_1^2\beta^2\lambda\mathbf{h}\mathbf{h}'$.
(6, 3): $-\beta^4\lambda(\beta\lambda+1)\mathbf{h}\mathbf{h}'$.
(6, 4): $\alpha_1(\beta\lambda+1)\beta^2(1+\lambda\alpha_1^3)\mathbf{h}\mathbf{h}'$.
(6, 5): $-\beta^5\lambda^2\mathbf{h}\mathbf{h}'$.
(6, 6): $\alpha_1\beta^3\lambda(1+\lambda\alpha_1^3)\mathbf{h}\mathbf{h}'$.

Summing the previous terms, we obtain $\beta(1+\beta\lambda)\boldsymbol{\Sigma} + \beta^2\mathbf{h}\mathbf{h}'$. Hence, the the limiting covariance of $\sqrt{n}(\mathbf{w}_n-\boldsymbol{\theta})$ is $(\boldsymbol{\theta}'\mathbf{h}+1/\beta)\boldsymbol{\Sigma}^{-1} + \boldsymbol{\theta}\boldsymbol{\theta}'$.

Finally, the Jacobian of the map $\boldsymbol{\theta} \mapsto \boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ is $(\|\boldsymbol{\theta}\|^2\mathbf{I}_p - \boldsymbol{\theta}\boldsymbol{\theta}')/\|\boldsymbol{\theta}\|^3$ and the delta method then implies that the scaled direction $\mathbf{w}/\|\mathbf{w}\|$ has the limiting covariance matrix,

$$\begin{aligned}\boldsymbol{\Psi}_U &:= (\|\boldsymbol{\theta}\|^2\mathbf{I}_p - \boldsymbol{\theta}\boldsymbol{\theta}')/\|\boldsymbol{\theta}\|^3\{(\boldsymbol{\theta}'\mathbf{h} + 1/\beta)\boldsymbol{\Sigma}^{-1} + \boldsymbol{\theta}\boldsymbol{\theta}'\}(\|\boldsymbol{\theta}\|^2\mathbf{I}_p - \boldsymbol{\theta}\boldsymbol{\theta}')/\|\boldsymbol{\theta}\|^3 \\ &= \left(\frac{\boldsymbol{\theta}'\boldsymbol{\Sigma}\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|^2} + \frac{1}{\beta\|\boldsymbol{\theta}\|^2}\right)\left(\mathbf{I}_p - \frac{\boldsymbol{\theta}\boldsymbol{\theta}'}{\|\boldsymbol{\theta}\|^2}\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{I}_p - \frac{\boldsymbol{\theta}\boldsymbol{\theta}'}{\|\boldsymbol{\theta}\|^2}\right).\end{aligned}$$

□

Before proving results regarding the unsupervised estimators, we establish two auxiliary lemmas.

Lemma B.2. *Let* $\boldsymbol{A} = \sum_{j=2}^{p}\lambda_j\boldsymbol{w}_j\boldsymbol{w}_j' + \boldsymbol{w}_1\boldsymbol{w}_1'\boldsymbol{C} \in \mathbb{R}^{p\times p}$, *where* $\lambda_2, \ldots, \lambda_p \in \mathbb{R}$, $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p$ *constitute an orthonormal set of vectors and* $\boldsymbol{C} \in \mathbb{R}^{p\times p}$ *is a symmetric positive definite matrix. Then* $\boldsymbol{A}$ *is invertible and*

$$\boldsymbol{A}^{-1} = \boldsymbol{B}^{\dagger} + (\boldsymbol{w}_1'\boldsymbol{C}\boldsymbol{w}_1)^{-1}\boldsymbol{w}_1\boldsymbol{w}_1'(\boldsymbol{I}_p - \boldsymbol{C}\boldsymbol{B}^{\dagger}),$$

where $\boldsymbol{B}^{\dagger} = \sum_{j=2}^{p}\lambda_j^{-1}\boldsymbol{w}_j\boldsymbol{w}_j$ *is the Moore-Penrose pseudoinverse of the matrix* $\boldsymbol{B} := \sum_{j=2}^{p}\lambda_j\boldsymbol{w}_j\boldsymbol{w}_j$.

Proof of Lemma B.2. Observe first that $\mathbf{B}\mathbf{B}^{\dagger} = \mathbf{B}^{\dagger}\mathbf{B} = \mathbf{I}_p - \mathbf{w}_1\mathbf{w}_1'$. Then, we compute the product of the two matrices to be,

$$\begin{aligned}&(\mathbf{B} + \mathbf{w}_1\mathbf{w}_1'\mathbf{C})\left\{\mathbf{B}^{\dagger} + (\mathbf{w}_1'\mathbf{C}\mathbf{w}_1)^{-1}\mathbf{w}_1\mathbf{w}_1'(\mathbf{I}_p - \mathbf{C}\mathbf{B}^{\dagger})\right\} \\ =&\mathbf{I}_p - \mathbf{w}_1\mathbf{w}_1' + \mathbf{w}_1\mathbf{w}_1'\mathbf{C}\mathbf{B}^{\dagger} + \mathbf{w}_1\mathbf{w}_1'(\mathbf{I}_p - \mathbf{C}\mathbf{B}^{\dagger}) \\ =&\mathbf{I}_p.\end{aligned}$$

The opposite product can be verified to equal identity in a similar manner, proving the claim. □

Lemma B.3. *Let* $\boldsymbol{z} \sim \mathcal{N}_p(c\boldsymbol{v}, \boldsymbol{I}_p)$, *for some* $c \in \mathbb{R}$ *and* $\boldsymbol{v} \in \mathbb{R}^p$. *Then*

$$\mathrm{E}\{(\boldsymbol{v}'\boldsymbol{z})^k\boldsymbol{z}\} = \|\boldsymbol{v}\|^{-2}\mathrm{E}\{(\boldsymbol{v}'\boldsymbol{z})^{k+1}\}\boldsymbol{v}, \quad \textit{and}$$

$$\mathrm{E}\{(\boldsymbol{v}'\boldsymbol{z})^k\boldsymbol{z}\boldsymbol{z}'\} = \mathrm{E}\{(\boldsymbol{v}'\boldsymbol{z})^k\}(\boldsymbol{I}_p - \|\boldsymbol{v}\|^{-2}\boldsymbol{v}\boldsymbol{v}') + \|\boldsymbol{v}\|^{-4}\mathrm{E}\{(\boldsymbol{v}'\boldsymbol{z})^{k+2}\}\boldsymbol{v}\boldsymbol{v}'.$$

Proof of Lemma B.3. The conditional distribution of $\mathbf{z}$ given $\mathbf{v}'\mathbf{z}$ is

$$\mathbf{z} \mid \mathbf{v}'\mathbf{z} = s \sim \mathcal{N}(\|\mathbf{v}\|^{-2}s\mathbf{v}, \mathbf{I}_p - \|\mathbf{v}\|^{-2}\mathbf{v}\mathbf{v}').$$

Thus,

$$\mathrm{E}\{(\mathbf{v}'\mathbf{z})^k\mathbf{z}\} = \mathrm{E}[\mathrm{E}\{(\mathbf{v}'\mathbf{z})^k\mathbf{z} \mid \mathbf{v}'\mathbf{z}\}] = \mathrm{E}[(\mathbf{v}'\mathbf{z})^k\mathrm{E}\{\mathbf{z} \mid \mathbf{v}'\mathbf{z}\}] = \|\mathbf{v}\|^{-2}\mathrm{E}\{(\mathbf{v}'\mathbf{z})^{k+1}\}\mathbf{v}.$$

The second claim is shown analogously and by using the fact that $\mathrm{E}(\mathbf{z}\mathbf{z}' \mid \mathbf{v}'\mathbf{z}) = \mathrm{Cov}(\mathbf{z} \mid \mathbf{v}'\mathbf{z}) + \mathrm{E}(\mathbf{z} \mid \mathbf{v}'\mathbf{z})\mathrm{E}(\mathbf{z}' \mid \mathbf{v}'\mathbf{z})$. □

Proof of Lemma 1. The distribution of the projection $\mathbf{u}'\tilde{\mathbf{x}}$ is

$$\mathbf{u}'\tilde{\mathbf{x}} \sim \alpha_1 \mathcal{N}(-\alpha_2 t, g) + \alpha_2 \mathcal{N}(\alpha_1 t, g),$$

where $t := \mathbf{u}'\mathbf{h}$, $g := \mathbf{u}'\boldsymbol{\Sigma}\mathbf{u}$ and $\mathbf{h} := \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$. By the moment formulas of univariate normal distribution, $\mathrm{E}\{(\mathbf{u}'\tilde{\mathbf{x}})^2\} = g + \beta t^2$, where $\beta := \alpha_1\alpha_2$. Similarly, $\mathrm{E}\{(\mathbf{u}'\tilde{\mathbf{x}})^4\} = \beta(\alpha_1^3 + \alpha_2^3)t^4 + 6\beta t^2 g + 3g^2$ which can be further simplified by noting that $\alpha_1^3 + \alpha_2^3 = 1 - 3\beta$. Hence,

$$\{\kappa(\mathbf{u}) - 3\}^2 = \beta^2(1 - 6\beta)^2 \frac{f^4}{(1 + \beta f)^4}, \tag{B.2}$$

where $f := t^2/g \geq 0$.

If $\alpha_1 \in \{\delta_1, \delta_2\}$, then $1 - 6\beta = 0$ making $\{\kappa(\mathbf{u}) - 3\}^2 = 0$. Assume then that $\alpha_1 \notin \{\delta_1, \delta_2\}$, implying that $(1 - 6\beta)^2 > 0$. The derivative of the map $x \mapsto x^4/(1 + \beta x)^4$ is $4x^3/(1 + \beta x)^5$, showing that the map is strictly increasing in $(0, \infty)$. Hence, $\{\kappa(\mathbf{u}) - 3\}^2$ is maximal when f is at its largest. Now,

$$f = \frac{t^2}{g} = \left\{ \left(\frac{\boldsymbol{\Sigma}^{1/2}\mathbf{u}}{\|\boldsymbol{\Sigma}^{1/2}\mathbf{u}\|} \right)' \boldsymbol{\Sigma}^{-1/2}\mathbf{h} \right\}^2,$$

showing that, by the Cauchy-Schwarz inequality, f is maximal if and only if $\boldsymbol{\Sigma}^{1/2}\mathbf{u} \propto \boldsymbol{\Sigma}^{-1/2}\mathbf{h}$, i.e., when $\mathbf{u} = \pm\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ (where $\boldsymbol{\theta} = \boldsymbol{\Sigma}^{-1}\mathbf{h}$). □

Proof of Theorem 2. The objective functions are translation invariant, meaning that we may, without loss of generality, assume that $\mathrm{E}(\mathbf{x}) = \mathbf{0}$. This makes the marginal distribution of $\mathbf{x}$ be $\mathbf{x} \sim \alpha_1 \mathcal{N}_p(-\alpha_2\mathbf{h}, \boldsymbol{\Sigma}) + \alpha_2 \mathcal{N}_p(\alpha_1\mathbf{h}, \boldsymbol{\Sigma})$, where $\mathbf{h} := \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$.

The strong consistency of the estimator can be shown in the usual way by establishing that the objective function is strongly uniformly convergent in the compact parameter set $\mathbb{S}^{p-1}$ (or, more precisely, in its subset where the sign of the estimator is fixed), that is,

$$\sup_{\mathbf{u} \in \mathbb{S}^{p-1}} |\{\kappa_n(\mathbf{u}) - 3\}^2 - \{\kappa(\mathbf{u}) - 3\}^2| \to 0, \quad \text{a.s.} \tag{B.3}$$

For simplicity, we give the proof of the uniform convergence only in Theorem 3, in the context of skewness (having lower moments than kurtosis), and similar (but lengthier) arguments can be used to show (B.3).

To show the limiting normality, note that the Largrangian corresponding to the optimization problem is $\ell_n(\mathbf{u}) = \{\kappa_n(\mathbf{u}) - 3\}^2 + \lambda_n(\mathbf{u}'\mathbf{u} - 1)$ where λ_n is the Lagrangian multiplier. Using some matrix calculus, the corresponding gradient is seen to be

$$\nabla \ell_n(\mathbf{u}) = \frac{8}{\tilde{s}_{n2}(\mathbf{u})^3} \{\kappa_n(\mathbf{u}) - 3\} \{\tilde{s}_{n2}(\mathbf{u})\tilde{\mathbf{m}}_{n3}(\mathbf{u}) - \tilde{s}_{n4}(\mathbf{u})\tilde{\mathbf{m}}_{n1}(\mathbf{u})\} - 2\lambda_n \mathbf{u},$$

where $\tilde{s}_{nk}(\mathbf{u}) := (1/n)\sum_i (\mathbf{u}'\tilde{\mathbf{x}}_i)^k$ and $\tilde{\mathbf{m}}_{nk}(\mathbf{u}) := (1/n)\sum_i (\mathbf{u}'\tilde{\mathbf{x}}_i)^k \tilde{\mathbf{x}}_i$. The gradient vanishes at the (sign-adjusted) sample maximum $s_n\mathbf{u}_n$ and multiplication

of the gradient from the left with $s_n\mathbf{u}_n'$ thus yields that $0 = s_n\mathbf{u}_n'\nabla\ell_n(s_n\mathbf{u}_n) = -2\lambda_n$, showing that $\lambda_n = 0$.

We next work on the level of individual probability elements $\omega \in \Omega$. By Lemma 1, LLN and the strong consistency of $s_n\mathbf{u}_n$, there exists a probability one set $\mathcal{H}$ such that $s_n\mathbf{u}_n \to \mathbf{u}_0$ and $\kappa_n(\mathbf{u}_n) - 3 \to t \neq 0$ for all $\omega \in \mathcal{H}$. Thus, for each $\omega \in \mathcal{H}$, the maximizer $\mathbf{u}_n$ satisfies, for n large enough, the estimating equation $\tilde{s}_{n2}(s_n\mathbf{u}_n)\tilde{\mathbf{m}}_{n3}(s_n\mathbf{u}_n) - \tilde{s}_{n4}(s_n\mathbf{u}_n)\tilde{\mathbf{m}}_{n1}(s_n\mathbf{u}_n) = \mathbf{0}$. Using Lagrangian multipliers we can similarly show that the population maximizer $\mathbf{u}_0 := \boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ satisfies $s_2(\mathbf{u}_0)\mathbf{m}_3(\mathbf{u}_0) - s_4(\mathbf{u}_0)\mathbf{m}_1(\mathbf{u}_0) = 0$, where $s_k(\mathbf{u}) = \mathrm{E}\{(\mathbf{u}'\mathbf{x})^k\}$ and $\mathbf{m}_k(\mathbf{u}) = \mathrm{E}\{(\mathbf{u}'\mathbf{x})^k\mathbf{x}\}$.

Let $g_{n\kappa} : \mathbb{R}^p \setminus \{\mathbf{0}\} \to \mathbb{R}$ be such that $g_{n\kappa}(\mathbf{u}) = \tilde{s}_{n2}(\mathbf{u})\tilde{\mathbf{m}}_{n3}(\mathbf{u}) - \tilde{s}_{n4}(\mathbf{u})\tilde{\mathbf{m}}_{n1}(\mathbf{u})$. For each $\omega \in \mathcal{H}$, we have, for n large enough, the Taylor expansion

$$\begin{aligned} g_{n\kappa}(s_n\mathbf{u}_n) = g_{n\kappa}(\mathbf{u}_0) + \nabla g_{n\kappa}(\mathbf{u}_0)(s_n\mathbf{u}_n - \mathbf{u}_0) \\ + \{(s_n\mathbf{u}_n - \mathbf{u}_0)' \times \nabla'\nabla g_{n\kappa}(\tilde{\mathbf{u}}_n)\}(s_n\mathbf{u}_n - \mathbf{u}_0), \end{aligned}$$

where $\nabla'\nabla g_{n\kappa}(\tilde{\mathbf{u}}_n)$ is the third order tensor of second derivatives of g, the symbol $\times$ denotes the vector-by-tensor multiplication (producing a matrix) and $\tilde{\mathbf{u}}_n$ satisfies $\|\tilde{\mathbf{u}}_n - \mathbf{u}_0\| \leq \|\mathbf{u}_n - \mathbf{u}_0\|$, implying that $\tilde{\mathbf{u}}_n \to \mathbf{u}_0$. Multiplying the expansion by $\sqrt{n}$ and using the fact that $g_{n\kappa}(s_n\mathbf{u}_n) = \mathbf{0}$ gives that

$$\{(s_n\mathbf{u}_n - \mathbf{u}_0)' \times \nabla'\nabla g_{n\kappa}(\tilde{\mathbf{u}}_n) + \nabla g_{n\kappa}(\mathbf{u}_0)\}\sqrt{n}(s_n\mathbf{u}_n - \mathbf{u}_0) = \sqrt{n}g_{n\kappa}(\mathbf{u}_0). \tag{B.4}$$

Now, the elements of $\nabla'\nabla g_{n\kappa}(\mathbf{u})$ are polynomials of the sample moments of $\tilde{\mathbf{x}}_i$ and the elements of $\mathbf{u}$ implying that, by LLN, $\nabla'\nabla g_{n\kappa}(\tilde{\mathbf{u}}_n)$ converges to a constant and $(s_n\mathbf{u}_n - \mathbf{u}_0)' \times \nabla'\nabla g_{n\kappa}(\tilde{\mathbf{u}}_n)$ converges to zero, for any $\omega \in \mathcal{H}$. Now, by the unit lengths of $s_n\mathbf{u}_n$ and $\mathbf{u}_0$, we have $c_0\mathbf{h}(s_n\mathbf{u}_n + \mathbf{u}_0)'\sqrt{n}(s_n\mathbf{u}_n - \mathbf{u}_0) = 0$, where $c_0 := (1/2)\{3s_2(\mathbf{u}_0)^2 - s_4(\mathbf{u}_0)\}\|\boldsymbol{\theta}\|(\mathbf{h}'\boldsymbol{\Sigma}^{-1}\mathbf{h})^{-1}$ (the inclusion of the constant c_0 simplifies things later on). Summing this with equation (B.4) gives

$$\begin{aligned} &\{(s_n\mathbf{u}_n - \mathbf{u}_0)' \times \nabla'\nabla g_{n\kappa}(\tilde{\mathbf{u}}_n) + \nabla g_{n\kappa}(\mathbf{u}_0) + c_0\mathbf{h}(s_n\mathbf{u}_n + \mathbf{u}_0)'\}\sqrt{n}(s_n\mathbf{u}_n - \mathbf{u}_0) \\ &= \sqrt{n}g_{n\kappa}(\mathbf{u}_0). \end{aligned}$$

Assume now for a moment that $\nabla g_{n\kappa}(\mathbf{u}_0) + c_0\mathbf{h}(s_n\mathbf{u}_n + \mathbf{u}_0)'$ converges to a full-rank matrix $\mathbf{G} \in \mathbb{R}^{p\times p}$. Then, for n large enough, we have,

$$\sqrt{n}(s_n\mathbf{u}_n - \mathbf{u}_0) = \{(s_n\mathbf{u}_n - \mathbf{u}_0)' \times \nabla'\nabla g_{n\kappa}(\tilde{\mathbf{u}}_n) + \nabla g_{n\kappa}(\mathbf{u}_0) \tag{B.5}$$

$$+ c_0\mathbf{h}(s_n\mathbf{u}_n + \mathbf{u}_0)'\}^{-1}\sqrt{n}g_{n\kappa}(\mathbf{u}_0). \tag{B.6}$$

Hence, assuming further that we have $\sqrt{n}g_{n\kappa}(\mathbf{u}_0) \rightsquigarrow \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Pi})$, then the limiting distribution of $s_n\mathbf{u}_n$ is, by Slutsky's theorem,

$$\sqrt{n}(s_n\mathbf{u}_n - \mathbf{u}_0) \rightsquigarrow \mathcal{N}_p\{\mathbf{0}, \mathbf{G}^{-1}\boldsymbol{\Pi}(\mathbf{G}^{-1})'\}. \tag{B.7}$$

Thus, to complete the proof, we next derive expressions for $\mathbf{G}$ and $\boldsymbol{\Pi}$ (and show that the former has indeed full rank).

The Jacobian of $g_{n\kappa}$ is,

$$\begin{aligned}\nabla g_{n\kappa}(\mathbf{u}_0) =& 2\tilde{\mathbf{m}}_{n1}(\mathbf{u}_0)\tilde{\mathbf{m}}_{n3}(\mathbf{u}_0)' + 3\tilde{s}_{n2}(\mathbf{u}_0)\tilde{\mathbf{G}}_{n2}(\mathbf{u}_0) - 4\tilde{\mathbf{m}}_{n3}(\mathbf{u}_0)\tilde{\mathbf{m}}_{n1}(\mathbf{u}_0)' \\ & - \tilde{s}_{n4}(\mathbf{u}_0)\tilde{\mathbf{G}}_{n0}(\mathbf{u}_0),\end{aligned}$$

where $\tilde{\mathbf{G}}_{nk}(\mathbf{u}) := (1/n)\sum_i (\mathbf{u}'\tilde{\mathbf{x}}_i)^k \tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i'$. Thus, by LLN and using the population level estimating equation, $s_2(\mathbf{u}_0)\mathbf{m}_3(\mathbf{u}_0) = s_4(\mathbf{u}_0)\mathbf{m}_1(\mathbf{u}_0)$, we get

$$\nabla g_{n\kappa}(\mathbf{u}_0) \to_p = -2\frac{s_4(\mathbf{u}_0)}{s_2(\mathbf{u}_0)}\mathbf{m}_1(\mathbf{u}_0)\mathbf{m}_1(\mathbf{u}_0)' + 3s_2(\mathbf{u}_0)\mathbf{G}_2(\mathbf{u}_0) - s_4(\mathbf{u}_0)\mathbf{G}_0(\mathbf{u}_0), \tag{B.8}$$

where $\mathbf{G}_k(\mathbf{u}) := \mathrm{E}\{(\mathbf{u}'\mathbf{x})^k\mathbf{x}\mathbf{x}'\}$. Denote next $\tau := \mathbf{h}'\boldsymbol{\Sigma}^{-1}\mathbf{h}$.
To compute the moments $\mathbf{m}_k(\mathbf{u}_0)$ and $\mathbf{G}_k(\mathbf{u}_0)$, we use Lemma B.3. The former satisfies $\boldsymbol{\Sigma}^{-1/2}\mathbf{m}_k(\mathbf{u}_0) = \|\boldsymbol{\theta}\|^{-k}\mathrm{E}\{(\mathbf{v}'\mathbf{z})^k\mathbf{z}\}$, where $\mathbf{v} := \boldsymbol{\Sigma}^{-1/2}\mathbf{h}$ and $\mathbf{z} \sim \alpha_1\mathcal{N}_p(-\alpha_2\mathbf{h}, \mathbf{I}_p) + \alpha_2\mathcal{N}_p(\alpha_1\mathbf{h}, \mathbf{I}_p)$. Denoting the components of the mixture by $\mathbf{z}_1$ and $\mathbf{z}_2$, we have, by the first part of Lemma B.3, for $\mathbf{z}_1$ that $\mathrm{E}\{(\mathbf{v}'\mathbf{z}_1)^k\mathbf{z}_1\} = \|\mathbf{v}\|^{-2}\mathrm{E}\{(\mathbf{v}'\mathbf{z}_1)^{k+1}\}\mathbf{v}$, and similarly for $\mathbf{z}_2$. Hence,

$$\boldsymbol{\Sigma}^{-1/2}\mathbf{m}_k(\mathbf{u}_0) = \|\boldsymbol{\theta}\|^{-k}\|\mathbf{v}\|^{-2}\mathrm{E}\{(\mathbf{v}'\mathbf{z})^{k+1}\}\mathbf{v}.$$

Finally, since $s_k(\mathbf{u}_0) = \|\boldsymbol{\theta}\|^{-k}\mathrm{E}\{(\mathbf{v}'\mathbf{z})^k\}$, we get

$$\mathbf{m}_k(\mathbf{u}_0) = \|\boldsymbol{\theta}\|\|\mathbf{v}\|^{-2}s_{k+1}(\mathbf{u}_0)\boldsymbol{\Sigma}^{1/2}\mathbf{v} = \|\boldsymbol{\theta}\|\tau^{-1}s_{k+1}(\mathbf{u}_0)\mathbf{h}. \tag{B.9}$$

For $\mathbf{G}_k(\mathbf{u}_0)$, we have, using the same notation, that

$$\boldsymbol{\Sigma}^{-1/2}\mathbf{G}_k(\mathbf{u}_0)\boldsymbol{\Sigma}^{-1/2} = \|\boldsymbol{\theta}\|^{-k}\mathrm{E}\{(\mathbf{v}'\mathbf{z})^k\mathbf{z}\mathbf{z}'\}.$$

The second part of Lemma B.3 then shows that

$$\begin{aligned}\mathbf{G}_k(\mathbf{u}_0) &= \|\boldsymbol{\theta}\|^{-k}\boldsymbol{\Sigma}^{1/2}[\mathrm{E}\{(\mathbf{v}'\mathbf{z})^k\}(\mathbf{I}_p - \|\mathbf{v}\|^{-2}\mathbf{v}\mathbf{v}') + \|\mathbf{v}\|^{-4}\mathrm{E}\{(\mathbf{v}'\mathbf{z})^{k+2}\}\mathbf{v}\mathbf{v}']\boldsymbol{\Sigma}^{1/2} \\ &= \|\boldsymbol{\theta}\|^{-k}[\|\boldsymbol{\theta}\|^k s_k(\mathbf{u}_0)(\boldsymbol{\Sigma} - \tau^{-1}\mathbf{h}\mathbf{h}') + \tau^{-2}\|\boldsymbol{\theta}\|^{k+2}s_{k+2}(\mathbf{u}_0)\mathbf{h}\mathbf{h}'] \\ &= s_k(\mathbf{u}_0)\boldsymbol{\Sigma} + \tau^{-1}\{\tau^{-1}\|\boldsymbol{\theta}\|^2 s_{k+2}(\mathbf{u}_0) - s_k(\mathbf{u}_0)\}\mathbf{h}\mathbf{h}'.\end{aligned} \tag{B.10}$$

Plugging in the expressions to (B.8), we get $\nabla g_{n\kappa}(\mathbf{u}_0) \to_p (3s_2^2 - s_4)\boldsymbol{\Sigma}^{1/2}(\mathbf{I}_p - \mathbf{w}\mathbf{w}')\boldsymbol{\Sigma}^{1/2}$, where $\mathbf{w} := \boldsymbol{\Sigma}^{-1/2}\mathbf{h}/\|\boldsymbol{\Sigma}^{-1/2}\mathbf{h}\|$ and $s_k \equiv s_k(\mathbf{u}_0)$. Moreover, we also have $c_0\mathbf{h}(s_n\mathbf{u}_n + \mathbf{u}_0)' \to_p (3s_2^2 - s_4)\boldsymbol{\Sigma}^{1/2}\mathbf{w}\mathbf{w}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{1/2}$. Now $\mathbf{G}$ is the sum of these two, giving,

$$\mathbf{G} = (3s_2^2 - s_4)\boldsymbol{\Sigma}^{1/2}(\mathbf{I}_p - \mathbf{w}\mathbf{w}' + \mathbf{w}\mathbf{w}'\boldsymbol{\Sigma}^{-1})\boldsymbol{\Sigma}^{1/2}.$$

The invertibility of $\mathbf{G}$ now follows from Lemma B.2, which also gives

$$\begin{aligned}(\mathbf{I}_p - \mathbf{w}\mathbf{w}' + \mathbf{w}\mathbf{w}'\boldsymbol{\Sigma}^{-1})^{-1} &= \mathbf{I}_p + (\mathbf{w}'\boldsymbol{\Sigma}^{-1}\mathbf{w})^{-1}\mathbf{w}\mathbf{w}'(\mathbf{I}_p - \boldsymbol{\Sigma}^{-1}) \\ &= \mathbf{I}_p + \|\boldsymbol{\theta}\|^{-2}\boldsymbol{\Sigma}^{-1/2}\mathbf{h}\mathbf{h}'\boldsymbol{\Sigma}^{-1/2}(\mathbf{I}_p - \boldsymbol{\Sigma}^{-1}).\end{aligned}$$

Finally, this makes the inverse of $\mathbf{G}$ be,

$$\mathbf{G}^{-1} = \frac{1}{3s_2^2 - s_4}\left\{\mathbf{\Sigma}^{-1} + \frac{1}{\|\boldsymbol{\theta}\|^2}\boldsymbol{\theta}\boldsymbol{\theta}'(\mathbf{I}_p - \mathbf{\Sigma}^{-1})\right\}.$$

The fact that $3s_2^2 - s_4 \neq 0$ follows from the formulas for s_k given later in the proof.

We next obtain the limiting distribution of

$$\sqrt{n}g_{n\kappa}(\mathbf{u}_0) = \sqrt{n}\{\tilde{s}_{n2}(\mathbf{u}_0)\tilde{\mathbf{m}}_{n3}(\mathbf{u}_0) - \tilde{s}_{n4}(\mathbf{u}_0)\tilde{\mathbf{m}}_{n1}(\mathbf{u}_0)\}.$$

Define non-centered counterparts for the sample moments as $s_{nk}(\mathbf{u}) := (1/n)\sum_i(\mathbf{u}'\mathbf{x}_i)^k$ and $\mathbf{m}_{nk}(\mathbf{u}) := (1/n)\sum_i(\mathbf{u}'\mathbf{x}_i)^k\mathbf{x}_i$. Then, LLN together with the calculus of $o_p(1)$ and $O_p(1)$ sequences shows that $\tilde{s}_{n2}(\mathbf{u}) = s_{n2}(\mathbf{u}) + o_p(1/\sqrt{n})$ and $\tilde{\mathbf{m}}_{n1}(\mathbf{u}) = \mathbf{m}_{n1}(\mathbf{u}) + o_p(1/\sqrt{n})$. However, the same equivalence does not hold for the terms $\tilde{\mathbf{m}}_{n3}(\mathbf{u})$ and $\tilde{s}_{n4}(\mathbf{u})$ but we instead have

$$\tilde{\mathbf{m}}_{n3}(\mathbf{u}) = \mathbf{m}_{n3}(\mathbf{u}) - 3s_{n1}(\mathbf{u})\mathbf{m}_2(\mathbf{u}) - s_3(\mathbf{u})\mathbf{m}_{n0}(\mathbf{u}) + o_p(1/\sqrt{n}),$$

and

$$\tilde{s}_{n4}(\mathbf{u}) = s_{n4}(\mathbf{u}) - 4s_3(\mathbf{u})s_{n1}(\mathbf{u}) + o_p(1/\sqrt{n}).$$

Using these, we expand $\sqrt{n}g_{n\kappa}(\mathbf{u}_0)$ to be (dropping $\mathbf{u}_0$ from the notation),

$$\begin{aligned}\sqrt{n}g_{n\kappa} =& \sqrt{n}(s_{n2} - s_2)\mathbf{m}_3 + s_2\sqrt{n}(\mathbf{m}_{n3} - \mathbf{m}_3) - \sqrt{n}(s_{n4} - s_4)\mathbf{m}_1 \\ &- s_4\sqrt{n}(\mathbf{m}_{n1} - \mathbf{m}_1) + (4s_3\mathbf{m}_1 - 3s_2\mathbf{m}_2)\sqrt{n}s_{n1} - s_2s_3\sqrt{n}\mathbf{m}_{n0}.\end{aligned} \tag{B.11}$$

Hence, by CLT, $\sqrt{n}g_{n\kappa}$ has a limiting normal distribution with the covariance matrix,

$$\begin{aligned}\mathbf{\Pi} = \mathrm{Cov}\{&(\mathbf{u}_0'\mathbf{x})^2\mathbf{m}_3 + s_2(\mathbf{u}_0'\mathbf{x})^3\mathbf{x} - (\mathbf{u}_0'\mathbf{x})^4\mathbf{m}_1 - s_4(\mathbf{u}_0'\mathbf{x})\mathbf{x} \\ &+ (4s_3\mathbf{m}_1 - 3s_2\mathbf{m}_2)(\mathbf{u}_0'\mathbf{x}) - s_2s_3\mathbf{x}\}.\end{aligned}$$

This matrix consists of 36 terms, which we next present and simplify using (B.9) and (B.10). We use the notation $\psi = \|\boldsymbol{\theta}\|\tau^{-1}$. Note that $s_1 = 0$, $\mathbf{m}_0 = \mathbf{0}$ and $\mathbf{f} := 4s_3\mathbf{m}_1 - 3s_2\mathbf{m}_2 = \psi s_2 s_3\mathbf{h}$.

(1, 1): $(s_4 - s_2^2)\mathbf{m}_3\mathbf{m}_3' = \psi^2 s_4^2(s_4 - s_2^2)\mathbf{h}\mathbf{h}'$.
(1, 2): $s_2(\mathbf{m}_3\mathbf{m}_5' - s_2\mathbf{m}_3\mathbf{m}_3') = \psi^2 s_2 s_4(s_6 - s_2 s_4)\mathbf{h}\mathbf{h}'$.
(1, 3): $-(s_6 - s_2s_4)\mathbf{m}_3\mathbf{m}_1' = -\psi^2 s_2 s_4(s_6 - s_2 s_4)\mathbf{h}\mathbf{h}'$.
(1, 4): $-s_4(\mathbf{m}_3\mathbf{m}_3' - s_2\mathbf{m}_3\mathbf{m}_1') = -\psi^2 s_4^2(s_4 - s_2^2)\mathbf{h}\mathbf{h}'$.
(1, 5): $s_3\mathbf{m}_3\mathbf{f}' = \psi^2 s_2 s_3^2 s_4\mathbf{h}\mathbf{h}'$.
(1, 6): $-s_2s_3\mathbf{m}_3\mathbf{m}_2' = -\psi^2 s_2 s_3^2 s_4\mathbf{h}\mathbf{h}'$.
(2, 1): $\psi^2 s_2 s_4(s_6 - s_2 s_4)\mathbf{h}\mathbf{h}'$.
(2, 2): $s_2^2(\mathbf{G}_6 - \mathbf{m}_3\mathbf{m}_3') = s_2^2 s_6\mathbf{\Sigma} + s_2^2\{\psi^2(s_8 - s_4^2) - \tau^{-1}s_6\}\mathbf{h}\mathbf{h}'$.
(2, 3): $-s_2(\mathbf{m}_7\mathbf{m}_1' - s_4\mathbf{m}_3\mathbf{m}_1') = -\psi^2 s_2^2(s_8 - s_4^2)\mathbf{h}\mathbf{h}'$.

(2, 4): $-s_2s_4(\mathbf{G}_4 - \mathbf{m}_3\mathbf{m}_1') = -s_2s_4^2\boldsymbol{\Sigma} - s_2s_4\{\psi^2(s_6 - s_2s_4) - \tau^{-1}s_4\}\mathbf{h}\mathbf{h}'$.
(2, 5): $s_2\mathbf{m}_4\mathbf{f}' = \psi^2 s_2^2 s_3 s_5 \mathbf{h}\mathbf{h}'$.
(2, 6): $-s_2^2 s_3 \mathbf{G}_3 = -s_2^2 s_3^2 \boldsymbol{\Sigma} - s_2^2 s_3\{\psi^2 s_5 - \tau^{-1}s_3\}\mathbf{h}\mathbf{h}'$.
(3, 1): $-\psi^2 s_2 s_4(s_6 - s_2 s_4)\mathbf{h}\mathbf{h}'$.
(3, 2): $-\psi^2 s_2^2(s_8 - s_4^2)\mathbf{h}\mathbf{h}'$.
(3, 3): $(s_8 - s_4^2)\mathbf{m}_1\mathbf{m}_1' = \psi^2 s_2^2(s_8 - s_4^2)\mathbf{h}\mathbf{h}'$.
(3, 4): $s_4(\mathbf{m}_1\mathbf{m}_5' - s_4\mathbf{m}_1\mathbf{m}_1') = \psi^2 s_2 s_4(s_6 - s_2 s_4)\mathbf{h}\mathbf{h}'$.
(3, 5): $-s_5\mathbf{m}_1\mathbf{f}' = -\psi^2 s_2^2 s_3 s_5 \mathbf{h}\mathbf{h}'$.
(3, 6): $s_2 s_3 \mathbf{m}_1\mathbf{m}_4' = \psi^2 s_2^2 s_3 s_5 \mathbf{h}\mathbf{h}'$.
(4, 1): $-\psi^2 s_4^2(s_4 - s_2^2)\mathbf{h}\mathbf{h}'$.
(4, 2): $-s_2 s_4^2\boldsymbol{\Sigma} - s_2 s_4\{\psi^2(s_6 - s_4 s_2) - \tau^{-1}s_4\}\mathbf{h}\mathbf{h}'$.
(4, 3): $\psi^2 s_2 s_4(s_6 - s_2 s_4)\mathbf{h}\mathbf{h}'$.
(4, 4): $s_4^2(\mathbf{G}_2 - \mathbf{m}_1\mathbf{m}_1') = s_2 s_4^2\boldsymbol{\Sigma} + s_4^2\{\psi^2(s_4 - s_2^2) - \tau^{-1}s_2\}\mathbf{h}\mathbf{h}'$.
(4, 5): $-s_4\mathbf{m}_2\mathbf{f}' = -\psi^2 s_2 s_3^2 s_4 \mathbf{h}\mathbf{h}'$.
(4, 6): $s_2 s_3 s_4 \mathbf{G}_1 = \psi^2 s_2 s_3^2 s_4 \mathbf{h}\mathbf{h}'$.
(5, 1): $\psi^2 s_2 s_3^2 s_4 \mathbf{h}\mathbf{h}'$.
(5, 2): $\psi^2 s_2^2 s_3 s_5 \mathbf{h}\mathbf{h}'$.
(5, 3): $-\psi^2 s_2^2 s_3 s_5 \mathbf{h}\mathbf{h}'$.
(5, 4): $-\psi^2 s_2 s_3^2 s_4 \mathbf{h}\mathbf{h}'$.
(5, 5): $s_2\mathbf{f}\mathbf{f}' = \psi^2 s_2^3 s_3^2 \mathbf{h}\mathbf{h}'$.
(5, 6): $-s_2 s_3 \mathbf{f}\mathbf{m}_1' = -\psi^2 s_2^3 s_3^2 \mathbf{h}\mathbf{h}'$.
(6, 1): $-\psi^2 s_2 s_3^2 s_4 \mathbf{h}\mathbf{h}'$.
(6, 2): $-s_2^2 s_3^2\boldsymbol{\Sigma} - s_2^2 s_3\{\psi^2 s_5 - \tau^{-1}s_3\}\mathbf{h}\mathbf{h}'$.
(6, 3): $\psi^2 s_2^2 s_3 s_5 \mathbf{h}\mathbf{h}'$.
(6, 4): $\psi^2 s_2 s_3^2 s_4 \mathbf{h}\mathbf{h}'$.
(6, 5): $-\psi^2 s_2^3 s_3^2 \mathbf{h}\mathbf{h}'$.
(6, 6): $s_2^2 s_3^2 \mathbf{G}_0 = s_2^2 s_3^2\boldsymbol{\Sigma} + s_2^2 s_3^2(\psi^2 s_2 - \tau^{-1})\mathbf{h}\mathbf{h}'$.

Summation of the previous 36 terms results in $\boldsymbol{\Pi} = s_2(s_2 s_6 - s_2 s_3^2 - s_4^2)(\boldsymbol{\Sigma} - \tau^{-1}\mathbf{h}\mathbf{h}')$. Thus, from the reasoning preceding (B.7), we have that $\sqrt{n}(s_n\mathbf{u}_n - \mathbf{u}_0)$ has a limiting normal distribution and with the covariance matrix $\boldsymbol{\Psi}_\kappa = \mathbf{G}^{-1}\boldsymbol{\Pi}(\mathbf{G}^{-1})'$. Plugging now in the values of $\mathbf{G}$ and $\boldsymbol{\Pi}$ and simplifying, we obtain,

$$\boldsymbol{\Psi}_\kappa = \frac{s_2(s_2 s_6 - s_2 s_3^2 - s_4^2)}{(3s_2^2 - s_4)^2}\left(\mathbf{I}_p - \frac{\boldsymbol{\theta}\boldsymbol{\theta}'}{\|\boldsymbol{\theta}\|^2}\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{I}_p - \frac{\boldsymbol{\theta}\boldsymbol{\theta}'}{\|\boldsymbol{\theta}\|^2}\right). \tag{B.12}$$

Now, recall that $s_k \equiv s_k(\mathbf{u}_0) = \mathrm{E}\{(\mathbf{u}_0'\mathbf{x})^k\} = \|\boldsymbol{\theta}\|^{-k}\mathrm{E}\{(\boldsymbol{\theta}'\mathbf{x})^k\}$ where $\boldsymbol{\theta}'\mathbf{x} \sim \alpha_1\mathcal{N}_1(-\alpha_2\tau, \tau) + \alpha_2\mathcal{N}_1(\alpha_1\tau, \tau)$ and $\tau = \boldsymbol{\theta}'\mathbf{h} = \boldsymbol{\theta}'\boldsymbol{\Sigma}\boldsymbol{\theta}$. Using the moment formulas for univariate normal distribution we now obtain that

$$\begin{aligned} s_2 &= \|\boldsymbol{\theta}\|^{-2}\tau(1 + \beta\tau), \quad s_3 = \|\boldsymbol{\theta}\|^{-3}(\alpha_1 - \alpha_2)\beta\tau^3, \\ s_4 &= \|\boldsymbol{\theta}\|^{-4}\tau^2\{\beta\tau^2(1 - 6\beta) + 3(1 + \beta\tau)^2\}, \end{aligned}$$

and

$$s_6 = \|\boldsymbol{\theta}\|^{-6}\tau^3\{\beta(1 - 5\beta + 5\beta^2)\tau^3 + 15\beta(1 - 3\beta)\tau^2 + 45\beta\tau + 15\}$$

where $\beta := \alpha_1\alpha_2$ and we have used the identities $\alpha_1^3+\alpha_2^3 = 1-3\beta$ and $\alpha_1^5+\alpha_2^5 = 1-5\beta+5\beta^2$. Plugging these in to (B.12) and simplifying (using $(\alpha_1-\alpha_2)^2 = 1-4\beta$), shows that the constant in front is

$$\frac{(1+\beta\tau)(6+24\beta\tau+9\beta\tau^2+\beta\tau^3-18\beta^2\tau^2-3\beta^2\tau^3)}{\tau^3\beta^2(6\beta-1)^2\|\boldsymbol{\theta}\|^2}$$

□

Proof of Lemma 2. The distribution of the projection $\mathbf{u}'\tilde{\mathbf{x}}$ is

$$\mathbf{u}'\tilde{\mathbf{x}} \sim \alpha_1\mathcal{N}_p(-\alpha_2 t, g) + \alpha_2\mathcal{N}_p(\alpha_1 t, g),$$

where $t := \mathbf{u}'\mathbf{h}$, $g := \mathbf{u}'\boldsymbol{\Sigma}\mathbf{u}$ and $\mathbf{h} := \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$. By the moment formulas of univariate normal distribution, $\mathrm{E}\{(\mathbf{u}'\tilde{\mathbf{x}})^2\} = g+\beta t^2$, where $\beta := \alpha_1\alpha_2$. Similarly, $\mathrm{E}\{(\mathbf{u}'\tilde{\mathbf{x}})^3\} = (\alpha_1-\alpha_2)\beta t^3$. Hence,

$$\gamma(\mathbf{u}) = \beta^2(1-4\beta)\frac{f^3}{(1+\beta f)^3},$$

where $f := t^2/g \geq 0$. Now, if $\alpha_1 = \alpha_2 = 1/2$, then clearly $\gamma(\mathbf{u})^2 = 0$. If $\alpha_1 \neq 1/2$, the derivative of the map $x \mapsto x^3/(1+\beta x)^3$ is $3x^2/(1+\beta x)^4$, showing that the map is strictly increasing outside of the origin. The conclusion now follows as in the proof of Lemma 1.

□

Proof of Theorem 3. The strong consistency follows as soon as we show the strong uniform consistency,

$$\sup_{\mathbf{u}\in\mathbb{S}^{p-1}} |\gamma_n(\mathbf{u})^2 - \gamma(\mathbf{u})^2| \to 0, \quad \text{a.s..} \tag{B.13}$$

By Theorem 2 and Lemma 1 in [4], (B.13) holds if, 1) the parameter space is compact, 2) we have $\gamma_n(\mathbf{u})^2 \to \gamma(\mathbf{u})^2$, a.s., for all $\mathbf{u} \in \mathbb{S}^{p-1}$ (this holds by LLN and the continuous mapping theorem), 3) γ^2 is uniformly continuous in $\mathbf{u}$ and, 4) γ_n^2 is Lipschitz continuous in the sense that $|\gamma_n(\mathbf{u}_1)^2 - \gamma_n(\mathbf{u}_2)^2| \leq K_n\|\mathbf{u}_1-\mathbf{u}_2\|$ for all $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{S}^{p-1}$ and some random variable K_n converging almost surely to a constant.

We now verify condition 4) above. Using the notation of the proof of Theorem 2, we have

$$\begin{aligned}|\gamma_n(\mathbf{u}_1)^2 - \gamma_n(\mathbf{u}_2)^2| &= \left|\frac{\tilde{s}_{n3}^2(\mathbf{u}_1)}{\tilde{s}_{n2}^3(\mathbf{u}_1)} - \frac{\tilde{s}_{n3}^2(\mathbf{u}_2)}{\tilde{s}_{n2}^3(\mathbf{u}_2)}\right| \\ &\leq \frac{|\tilde{s}_{n3}^2(\mathbf{u}_1)-\tilde{s}_{n3}^2(\mathbf{u}_2)|\tilde{s}_{n2}^3(\mathbf{u}_2) - \tilde{s}_{n3}^2(\mathbf{u}_2)|\tilde{s}_{n2}^3(\mathbf{u}_1)-\tilde{s}_{n2}^3(\mathbf{u}_2)|}{\tilde{s}_{n2}^3(\mathbf{u}_1)\tilde{s}_{n2}^3(\mathbf{u}_2)}.\end{aligned}$$

Now, $\tilde{s}_{n2}(\mathbf{u})$ is, for all $\mathbf{u} \in \mathbb{S}^{p-1}$, lower bounded by the smallest eigenvalue of the sample covariance matrix, which by the continuity of the eigenvalues and

the positive-definiteness of the covariance matrix converges almost surely to a positive constant. Moreover, we have

$$|\tilde{s}_{n3}(\mathbf{u})| \leq \frac{1}{n}\sum_{i=1}^{n}|\mathbf{u}'\tilde{\mathbf{x}}_i|^3 \leq \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i-\bar{\mathbf{x}}\|^3 \leq \frac{1}{n}\sum_{i=1}^{n}(\|\mathbf{x}_i\|+\|\bar{\mathbf{x}}\|)^3,$$

which converges, by LLN, almost surely to a constant, and similar result can be shown for $|\tilde{s}_{n2}(\mathbf{u})|$. Finally,

$$|\tilde{s}^2_{n3}(\mathbf{u}_1)-\tilde{s}^2_{n3}(\mathbf{u}_2)| \leq |\tilde{s}_{n3}(\mathbf{u}_1)-\tilde{s}_{n3}(\mathbf{u}_2)|\frac{2}{n}\sum_{i=1}^{n}\|\mathbf{x}_i-\bar{\mathbf{x}}\|^3,$$

and

$$\begin{aligned}|\tilde{s}_{n3}(\mathbf{u}_1)-\tilde{s}_{n3}(\mathbf{u}_2)| &\leq \frac{1}{n}\sum_{i=1}^{n}|(\mathbf{u}_1-\mathbf{u}_2)'\tilde{\mathbf{x}}_i||(\mathbf{u}_1'\tilde{\mathbf{x}}_i)^2+\mathbf{u}_1'\tilde{\mathbf{x}}_i\mathbf{u}_2'\tilde{\mathbf{x}}_i+(\mathbf{u}_2'\tilde{\mathbf{x}}_i)^2| \\ &\leq \|\mathbf{u}_1-\mathbf{u}_2\|\frac{3}{n}\sum_{i=1}^{n}\|\mathbf{x}_i-\bar{\mathbf{x}}\|^3,\end{aligned}$$

and putting everything above together, we conclude that the Lipschitz continuity 4) holds. What remains to be verified is then condition 3), which can be shown similarly to 4) after recalling that Lipschitz continuity implies uniform continuity. Hence, the strong consistency of the estimator follows.

Also, the proof of the limiting normality has exactly the same steps as in the proof of Theorem 2 and we only provide the key steps and expressions, using the same notation as in the proof of Theorem 2. The gradient of γ_n is

$$\nabla\gamma_n(\mathbf{u}) = \frac{6}{\tilde{s}_{n2}(\mathbf{u})^{5/2}}\gamma_n(\mathbf{u})\{\tilde{s}_{n2}(\mathbf{u})\tilde{\mathbf{m}}_{n2}(\mathbf{u})-\tilde{s}_{n3}(\mathbf{u})\tilde{\mathbf{m}}_{n1}(\mathbf{u})\},$$

leading to the estimating equation $g_{n\gamma}(\mathbf{u}_n)=\mathbf{0}$, for $g_{n\gamma}(\mathbf{u}) := \tilde{s}_{n2}(\mathbf{u})\tilde{\mathbf{m}}_{n2}(\mathbf{u}) - \tilde{s}_{n3}(\mathbf{u})\tilde{\mathbf{m}}_{n1}(\mathbf{u})$. The Jacobian of $g_{n\gamma}$ at $\mathbf{u}_0$ satisfies (after simplification via the estimating equation)

$$\nabla g_{n\gamma}(\mathbf{u}_0) \to_p -\frac{s_3(\mathbf{u}_0)}{s_2(\mathbf{u}_0)}\mathbf{m}_1(\mathbf{u}_0)\mathbf{m}_1(\mathbf{u}_0)' + 2s_2(\mathbf{u}_0)\mathbf{G}_1(\mathbf{u}_0) - s_3(\mathbf{u}_0)\mathbf{G}_0(\mathbf{u}_0).$$

By formulas for $\mathbf{m}_k(\mathbf{u}_0)$ and $\mathbf{G}_k(\mathbf{u}_0)$, the limit equals $-s_3\boldsymbol{\Sigma}^{1/2}(\mathbf{I}_p-\mathbf{w}\mathbf{w}')\boldsymbol{\Sigma}^{1/2}$. Using the same trick as in the proof of Theorem 2 to make the Jacobian full rank (addition of $c_0\mathbf{h}(s_n\mathbf{u}_n+\mathbf{u}_0)'\sqrt{n}(s_n\mathbf{u}_n-\mathbf{u}_0)=0$ for suitably chosen c_0 to the Taylor expansion), we obtain the corresponding matrix $\mathbf{G}$ to be

$$\mathbf{G} = -s_3\boldsymbol{\Sigma}^{1/2}(\mathbf{I}_p-\mathbf{w}\mathbf{w}'+\mathbf{w}\mathbf{w}'\boldsymbol{\Sigma}^{-1})\boldsymbol{\Sigma}^{1/2},$$

with the inverse,

$$\mathbf{G}^{-1} = -\frac{1}{s_3}\left\{\boldsymbol{\Sigma}^{-1}+\frac{1}{\|\boldsymbol{\theta}\|^2}\boldsymbol{\theta}\boldsymbol{\theta}'(\mathbf{I}_p-\boldsymbol{\Sigma}^{-1})\right\}.$$

Moving to study the limiting distribution of $\sqrt{n}g_{n\gamma}$, we note that

$$\tilde{\mathbf{m}}_{n2}(\mathbf{u}) = \mathbf{m}_{n2}(\mathbf{u}) - 2s_{n1}(\mathbf{u})\mathbf{m}_1(\mathbf{u}) - s_2(\mathbf{u})\mathbf{m}_{n0}(\mathbf{u}) + o_p(1/\sqrt{n}),$$

and

$$\tilde{s}_{n3}(\mathbf{u}) = s_{n3}(\mathbf{u}) - 3s_2(\mathbf{u})s_{n1}(\mathbf{u}) + o_p(1/\sqrt{n}).$$

With these, we expand $\sqrt{n}g_{n\gamma}(\mathbf{u}_0)$ to be,

$$\begin{aligned}\sqrt{n}g_{n\gamma} =& \sqrt{n}(s_{n2} - s_2)\mathbf{m}_2 + s_2\sqrt{n}(\mathbf{m}_{n2} - \mathbf{m}_2) - \sqrt{n}(s_{n3} - s_3)\mathbf{m}_1 \\ & - s_3\sqrt{n}(\mathbf{m}_{n1} - \mathbf{m}_1) + s_2\mathbf{m}_1\sqrt{n}s_{n1} - s_2^2\sqrt{n}\mathbf{m}_{n0}. \qquad \text{(B.14)}\end{aligned}$$

Hence, by CLT, $\sqrt{n}g_{n\gamma}$ has a limiting normal distribution with the covariance matrix,

$$\mathbf{\Pi} = \mathrm{Cov}\{(\mathbf{u}_0'\mathbf{x})^2\mathbf{m}_2 + s_2(\mathbf{u}_0'\mathbf{x})^2\mathbf{x} - (\mathbf{u}_0'\mathbf{x})^3\mathbf{m}_1 - s_3(\mathbf{u}_0'\mathbf{x})\mathbf{x} + s_2\mathbf{m}_1(\mathbf{u}_0'\mathbf{x}) - s_2^2\mathbf{x}\}.$$

The covariance matrix has the following 36 terms.

(1, 1): $(s_4 - s_2^2)\mathbf{m}_2\mathbf{m}_2' = \psi^2 s_3^2(s_4 - s_2^2)\mathbf{h}\mathbf{h}'$.
(1, 2): $s_2(\mathbf{m}_2\mathbf{m}_4' - s_2\mathbf{m}_2\mathbf{m}_2') = \psi^2 s_2 s_3(s_5 - s_2 s_3)\mathbf{h}\mathbf{h}'$.
(1, 3): $-(s_5 - s_2 s_3)\mathbf{m}_2\mathbf{m}_1' = -\psi^2 s_2 s_3(s_5 - s_2 s_3)\mathbf{h}\mathbf{h}'$.
(1, 4): $-s_3(\mathbf{m}_2\mathbf{m}_3' - s_2\mathbf{m}_2\mathbf{m}_1') = -\psi^2 s_3^2(s_4 - s_2^2)\mathbf{h}\mathbf{h}'$.
(1, 5): $s_2 s_3\mathbf{m}_2\mathbf{m}_1' = \psi^2 s_2^2 s_3^2\mathbf{h}\mathbf{h}'$.
(1, 6): $-s_2^2\mathbf{m}_2\mathbf{m}_2' = -\psi^2 s_2^2 s_3^2\mathbf{h}\mathbf{h}'$.
(2, 1): $\psi^2 s_2 s_3(s_5 - s_2 s_3)\mathbf{h}\mathbf{h}'$.
(2, 2): $s_2^2(\mathbf{G}_4 - \mathbf{m}_2\mathbf{m}_2') = s_2^2 s_4(\mathbf{\Sigma} - \tau^{-1}\mathbf{h}\mathbf{h}') + \psi^2 s_2^2(s_6 - s_3^2)\mathbf{h}\mathbf{h}'$.
(2, 3): $-s_2(\mathbf{m}_5\mathbf{m}_1' - s_3\mathbf{m}_2\mathbf{m}_1') = -\psi^2 s_2^2(s_6 - s_3^2)\mathbf{h}\mathbf{h}'$.
(2, 4): $-s_2 s_3(\mathbf{G}_3 - \mathbf{m}_2\mathbf{m}_1') = -s_2 s_3^2(\mathbf{\Sigma} - \tau^{-1}\mathbf{h}\mathbf{h}') - \psi^2 s_2 s_3(s_5 - s_2 s_3)\mathbf{h}\mathbf{h}'$.
(2, 5): $s_2^2\mathbf{m}_3\mathbf{m}_1' = \psi^2 s_2^3 s_4\mathbf{h}\mathbf{h}'$.
(2, 6): $-s_2^3\mathbf{G}_2 = -s_2^4(\mathbf{\Sigma} - \tau^{-1}\mathbf{h}\mathbf{h}') - \psi^2 s_2^3 s_4\mathbf{h}\mathbf{h}'$.
(3, 1): $-\psi^2 s_2 s_3(s_5 - s_2 s_3)\mathbf{h}\mathbf{h}'$.
(3, 2): $-\psi^2 s_2^2(s_6 - s_3^2)\mathbf{h}\mathbf{h}'$.
(3, 3): $(s_6 - s_3^2)\mathbf{m}_1\mathbf{m}_1' = \psi^2 s_2^2(s_6 - s_3^2)\mathbf{h}\mathbf{h}'$.
(3, 4): $s_3(\mathbf{m}_1\mathbf{m}_4' - s_3\mathbf{m}_1\mathbf{m}_1') = \psi^2 s_2 s_3(s_5 - s_2 s_3)\mathbf{h}\mathbf{h}'$.
(3, 5): $-s_2 s_4\mathbf{m}_1\mathbf{m}_1' = -\psi^2 s_2^3 s_4\mathbf{h}\mathbf{h}'$.
(3, 6): $s_2^2\mathbf{m}_1\mathbf{m}_3' = \psi^2 s_2^3 s_4\mathbf{h}\mathbf{h}'$.
(4, 1): $-\psi^2 s_3^2(s_4 - s_2^2)\mathbf{h}\mathbf{h}'$.
(4, 2): $-s_2 s_3^2(\mathbf{\Sigma} - \tau^{-1}\mathbf{h}\mathbf{h}') - \psi^2 s_2 s_3(s_5 - s_2 s_3)\mathbf{h}\mathbf{h}'$.
(4, 3): $\psi^2 s_2 s_3(s_5 - s_2 s_3)\mathbf{h}\mathbf{h}'$.
(4, 4): $s_3^2(\mathbf{G}_2 - \mathbf{m}_1\mathbf{m}_1') = s_2 s_3^2(\mathbf{\Sigma} - \tau^{-1}\mathbf{h}\mathbf{h}') + \psi^2 s_3^2(s_4 - s_2^2)\mathbf{h}\mathbf{h}'$.
(4, 5): $-s_2 s_3\mathbf{m}_2\mathbf{m}_1' = -\psi^2 s_2^2 s_3^2\mathbf{h}\mathbf{h}'$.
(4, 6): $s_2^2 s_3\mathbf{G}_1 = \psi^2 s_2^2 s_3^2\mathbf{h}\mathbf{h}'$.
(5, 1): $\psi^2 s_2^2 s_3^2\mathbf{h}\mathbf{h}'$.
(5, 2): $\psi^2 s_2^3 s_4\mathbf{h}\mathbf{h}'$.
(5, 3): $-\psi^2 s_2^3 s_4\mathbf{h}\mathbf{h}'$.
(5, 4): $-\psi^2 s_2^2 s_3^2\mathbf{h}\mathbf{h}'$.

(5, 5): $s_2^3\mathbf{m}_1\mathbf{m}_1' = \psi^2 s_2^5\mathbf{h}\mathbf{h}'$.
(5, 6): $-s_2^3\mathbf{m}_1\mathbf{m}_1' = -\psi^2 s_2^5\mathbf{h}\mathbf{h}'$.
(6, 1): $-\psi^2 s_2^2 s_3^2\mathbf{h}\mathbf{h}'$.
(6, 2): $-s_2^4(\mathbf{\Sigma} - \tau^{-1}\mathbf{h}\mathbf{h}') - \psi^2 s_2^3 s_4\mathbf{h}\mathbf{h}'$.
(6, 3): $\psi^2 s_2^3 s_4\mathbf{h}\mathbf{h}'$.
(6, 4): $\psi^2 s_2^2 s_3^2\mathbf{h}\mathbf{h}'$.
(6, 5): $-\psi^2 s_2^5\mathbf{h}\mathbf{h}'$.
(6, 6): $s_2^4\mathbf{G}_0 = s_2^4(\mathbf{\Sigma} - \tau^{-1}\mathbf{h}\mathbf{h}') + \psi^2 s_2^5\mathbf{h}\mathbf{h}'$.

Summing the terms gives $\mathbf{\Pi} = s_2(s_2 s_4 - s_2^3 - s_3^2)(\mathbf{\Sigma} - \tau^{-1}\mathbf{h}\mathbf{h}')$. This yields the limiting covariance,

$$\mathbf{\Psi}_\kappa = \mathbf{G}^{-1}\mathbf{\Pi}(\mathbf{G}^{-1})' = \frac{s_2(s_2 s_4 - s_2^3 - s_3^2)}{s_3^2}\left(\mathbf{I}_p - \frac{\boldsymbol{\theta}\boldsymbol{\theta}'}{\|\boldsymbol{\theta}\|^2}\right)\mathbf{\Sigma}^{-1}\left(\mathbf{I}_p - \frac{\boldsymbol{\theta}\boldsymbol{\theta}'}{\|\boldsymbol{\theta}\|^2}\right). \tag{B.15}$$

Finally, simplifying the constant in front shows that it equals

$$\frac{(1+\beta\tau)(2+6\beta\tau+\beta\tau^2)}{\tau^2\beta^2(1-4\beta)\|\boldsymbol{\theta}\|^2}.$$

□

Proof of Theorem 4. Again, the strong consistency follows as in Theorem 2 and we omit its proof. For the limiting distribution, we give in the following the key steps of the proof (and use the same notation as in the proofs of Theorems 2 and 3).

The gradient of η_n is

$$\nabla\eta_n(\mathbf{u}) = \frac{1}{\tilde{s}_{n2}^3}[6w_1\gamma_n(\mathbf{u})\tilde{s}_{n2}^{1/2}(\mathbf{u})g_{n\gamma}(\mathbf{u}) + 8w_2\{\kappa_n(\mathbf{u}) - 3\}g_{n\kappa}(\mathbf{u})],$$

where $g_{n\gamma}(\mathbf{u}) = \tilde{s}_{n2}(\mathbf{u})\tilde{\mathbf{m}}_{n2}(\mathbf{u}) - \tilde{s}_{n3}(\mathbf{u})\tilde{\mathbf{m}}_{n1}(\mathbf{u})$ and $g_{n\kappa}(\mathbf{u}) = \tilde{s}_{n2}(\mathbf{u})\tilde{\mathbf{m}}_{n3}(\mathbf{u}) - \tilde{s}_{n4}(\mathbf{u})\tilde{\mathbf{m}}_{n1}(\mathbf{u})$ were used in the proofs of Theorems 3 and 2, respectively. Thus, $\mathbf{u}_n$ solves the estimating equation $g_{n\eta}(\mathbf{u}_n) = \mathbf{0}$, where

$$g_{n\eta}(\mathbf{u}) := 3w_1\gamma_n(\mathbf{u})\tilde{s}_{n2}^{1/2}(\mathbf{u})g_{n\gamma}(\mathbf{u}) + 4w_2\{\kappa_n(\mathbf{u}) - 3\}g_{n\kappa}(\mathbf{u}).$$

The Jacobian of $g_{n\eta}$ at $\mathbf{u}_0$ satisfies

$$\begin{aligned}\nabla g_{n\eta}(\mathbf{u}_0) =& 3w_1[\nabla\{\gamma_n(\mathbf{u}_0)\tilde{s}_{n2}^{1/2}(\mathbf{u}_0)\}g_{n\gamma}(\mathbf{u}_0)' + \gamma_n(\mathbf{u}_0)\tilde{s}_{n2}^{1/2}(\mathbf{u}_0)\nabla g_{n\gamma}(\mathbf{u}_0)] \\ &+ 4w_2[\nabla\{\kappa_n(\mathbf{u}_0) - 3\}g_{n\kappa}(\mathbf{u}_0)' + \{\kappa_n(\mathbf{u}_0) - 3\}\nabla g_{n\kappa}(\mathbf{u}_0)].\end{aligned}$$

Recalling that $\mathbf{m}_k(\mathbf{u}_0) = \psi s_{k+1}\mathbf{h}$ and $\mathbf{G}_k(\mathbf{u}_0) = s_k(\mathbf{\Sigma} - \tau^{-1}\mathbf{h}\mathbf{h}') + \psi^2 s_{k+2}\mathbf{h}\mathbf{h}'$, where $\psi = \|\boldsymbol{\theta}\|\tau^{-1}$, LLN now gives that $g_{n\gamma}(\mathbf{u}_0) \to_p \mathbf{0}$ and $g_{n\kappa}(\mathbf{u}_0) \to_p \mathbf{0}$, implying that

$$\nabla g_{n\eta}(\mathbf{u}_0) \to_p -s_2^{-2}\{3w_1 s_2 s_3^2 + 4w_2(s_4 - 3s_2^2)^2\}\mathbf{\Sigma}^{1/2}(\mathbf{I}_p - \mathbf{w}\mathbf{w}')\mathbf{\Sigma}^{1/2}.$$

Completing now this matrix to full rank through the unit length constraint on $\mathbf{u}_n$ (as in the proofs of Theorems 3 and 2), we now obtain that,

$$\mathbf{G}^{-1} = \frac{-s_2^2}{3w_1 s_2 s_3^2 + 4w_2(s_4 - 3s_2^2)^2} \left\{ \mathbf{\Sigma}^{-1} + \frac{1}{\|\boldsymbol{\theta}\|^2} \boldsymbol{\theta}\boldsymbol{\theta}'(\mathbf{I}_p - \mathbf{\Sigma}^{-1}) \right\}.$$

We then derive the limiting distribution of

$$\sqrt{n} g_{n\eta}(\mathbf{u}_0) = 3w_1 \sqrt{n} \tilde{s}_{n2}^{1/2}(\mathbf{u}_0) g_{n\gamma}(\mathbf{u}_0) + 4w_2 \sqrt{n} g_{n\kappa}(\mathbf{u}_0).$$

Recalling that the population version satisfies

$$3w_1 \gamma(\mathbf{u}_0) s_2^{1/2}(\mathbf{u}_0) g_\gamma(\mathbf{u}_0) + 4w_2 \{\kappa(\mathbf{u}_0) - 3\} g_\kappa(\mathbf{u}_0) = 0,$$

we get the expansion,

$$\begin{aligned}\sqrt{n} g_{n\eta} =& 3w_1 \{\sqrt{n}(\gamma_n \tilde{s}_{n2}^{1/2} - \gamma s_2^{1/2})(s_2 \mathbf{m}_2 - s_3 \mathbf{m}_1) + s_2^{-1} s_3 \sqrt{n} g_{n\gamma}\} \\ &+ 4w_2 \{\sqrt{n}(\kappa_n - \kappa)(s_2 \mathbf{m}_3 - s_4 \mathbf{m}_1) + s_2^{-2}(s_4 - 3s_2^2)\sqrt{n} g_{n\kappa}\} + o_p(1),\end{aligned}$$

where $\sqrt{n} g_{n\kappa}$ has the expansion given in (B.11). Now, $s_2 \mathbf{m}_2 - s_3 \mathbf{m}_1 = \mathbf{0}$ and $s_2 \mathbf{m}_3 - s_4 \mathbf{m}_1 = \mathbf{0}$, implying that

$$\sqrt{n} g_{n\eta} = 3w_1 s_2^{-1} s_3 \sqrt{n} g_{n\gamma} + 4w_2 s_2^{-2}(s_4 - 3s_2^2)\sqrt{n} g_{n\kappa} + o_p(1),$$

where $\sqrt{n} g_{n\gamma}$ has the expansion given in (B.14). Consequently, by CLT, $\sqrt{n} g_{n\eta}$ has a limiting normal distribution. By the proof of Theorem 2, the limiting covariance matrix of $4w_2 s_2^{-2}(s_4 - 3s_2^2)\sqrt{n} g_{n\kappa}$ is $16w_2^2 s_2^{-3}(s_4 - 3s_2^2)^2(s_2 s_6 - s_2 s_3^2 - s_4^2)(\mathbf{\Sigma} - \tau^{-1}\mathbf{h}\mathbf{h}')$ and, by the proof of Theorem 3, the limiting covariance matrix of $3w_1 s_2^{-1} s_3 \sqrt{n} g_{n\gamma}$ is $9w_1^2 s_2^{-1} s_3^2 (s_2 s_4 - s_2^3 - s_3^2)(\mathbf{\Sigma} - \tau^{-1}\mathbf{h}\mathbf{h}')$. Thus, the limiting covariance matrix of $\sqrt{n} g_{n\eta}$ is $\{9w_1^2 s_2^{-1} s_3^2 (s_2 s_4 - s_2^3 - s_3^2) + 16w_2^2 s_2^{-3}(s_4 - 3s_2^2)^2(s_2 s_6 - s_2 s_3^2 - s_4^2)\}(\mathbf{\Sigma} - \tau^{-1}\mathbf{h}\mathbf{h}') + 24w_1 w_2 s_2^{-3} s_3 (s_4 - 3s_2^2)\mathrm{Cov}(\mathbf{y}_1, \mathbf{y}_2)$, where

$$\begin{aligned}\mathbf{y}_1 &:= (\mathbf{u}_0'\mathbf{x})^2 \mathbf{m}_2 + s_2(\mathbf{u}_0'\mathbf{x})^2 \mathbf{x} - (\mathbf{u}_0'\mathbf{x})^3 \mathbf{m}_1 - s_3(\mathbf{u}_0'\mathbf{x})\mathbf{x} + s_2(\mathbf{u}_0'\mathbf{x})\mathbf{m}_1 - s_2^2 \mathbf{x}, \\ \mathbf{y}_2 &:= (\mathbf{u}_0'\mathbf{x})^2 \mathbf{m}_3 + s_2(\mathbf{u}_0'\mathbf{x})^3 \mathbf{x} - (\mathbf{u}_0'\mathbf{x})^4 \mathbf{m}_1 - s_4(\mathbf{u}_0'\mathbf{x})\mathbf{x} + s_3(\mathbf{u}_0'\mathbf{x})\mathbf{m}_1 - s_2 s_3 \mathbf{x}.\end{aligned}$$

The matrix $\mathrm{Cov}(\mathbf{y}_1, \mathbf{y}_2)$ consists of the following 36 terms:

(1, 1): $\psi^2 s_3 s_4 (s_4 - s_2^2)\mathbf{h}\mathbf{h}'$.
(1, 2): $\psi^2 s_2 s_3 (s_6 - s_2 s_4)\mathbf{h}\mathbf{h}'$.
(1, 3): $-\psi^2 s_2 s_3 (s_6 - s_2 s_4)\mathbf{h}\mathbf{h}'$.
(1, 4): $-\psi^2 s_3 s_4 (s_4 - s_2^2)\mathbf{h}\mathbf{h}'$.
(1, 5): $\psi^2 s_2 s_3^3 \mathbf{h}\mathbf{h}'$.
(1, 6): $-\psi^2 s_2 s_3^3 \mathbf{h}\mathbf{h}'$.
(2, 1): $\psi^2 s_2 s_4 (s_5 - s_2 s_3)\mathbf{h}\mathbf{h}'$.
(2, 2): $s_2^2 s_5 (\mathbf{\Sigma} - \tau^{-1}\mathbf{h}\mathbf{h}') + \psi^2 s_2^2 (s_7 - s_3 s_4)\mathbf{h}\mathbf{h}'$.
(2, 3): $-\psi^2 s_2^2 (s_7 - s_3 s_4)\mathbf{h}\mathbf{h}'$.
(2, 4): $-s_2 s_3 s_4 (\mathbf{\Sigma} - \tau^{-1}\mathbf{h}\mathbf{h}') - \psi^2 s_2 s_4 (s_5 - s_2 s_3)\mathbf{h}\mathbf{h}'$.
(2, 5): $\psi^2 s_2^2 s_3 s_4 \mathbf{h}\mathbf{h}'$.

(2, 6): $-s_2^3 s_3(\mathbf{\Sigma} - \tau^{-1}\mathbf{h}\mathbf{h}') - \psi^2 s_2^2 s_3 s_4 \mathbf{h}\mathbf{h}'$.
(3, 1): $-\psi^2 s_2 s_4 (s_5 - s_2 s_3)\mathbf{h}\mathbf{h}'$.
(3, 2): $-\psi^2 s_2^2 (s_7 - s_3 s_4)\mathbf{h}\mathbf{h}'$.
(3, 3): $\psi^2 s_2^2 (s_7 - s_3 s_4)\mathbf{h}\mathbf{h}'$.
(3, 4): $\psi^2 s_2 s_4 (s_5 - s_2 s_3)\mathbf{h}\mathbf{h}'$.
(3, 5): $-\psi^2 s_2^2 s_3 s_4 \mathbf{h}\mathbf{h}'$.
(3, 6): $\psi^2 s_2^2 s_3 s_4 \mathbf{h}\mathbf{h}'$.
(4, 1): $-\psi^2 s_3 s_4 (s_4 - s_2^2)\mathbf{h}\mathbf{h}'$.
(4, 2): $-s_2 s_3 s_4(\mathbf{\Sigma} - \tau^{-1}\mathbf{h}\mathbf{h}') - \psi^2 s_2 s_3 (s_6 - s_2 s_4)\mathbf{h}\mathbf{h}'$.
(4, 3): $\psi^2 s_2 s_3 (s_6 - s_2 s_4)\mathbf{h}\mathbf{h}'$.
(4, 4): $s_2 s_3 s_4(\mathbf{\Sigma} - \tau^{-1}\mathbf{h}\mathbf{h}') + \psi^2 s_3 s_4 (s_4 - s_2^2)\mathbf{h}\mathbf{h}'$.
(4, 5): $-\psi^2 s_2 s_3^3 \mathbf{h}\mathbf{h}'$.
(4, 6): $\psi^2 s_2 s_3^3 \mathbf{h}\mathbf{h}'$.
(5, 1): $\psi^2 s_2^2 s_3 s_4 \mathbf{h}\mathbf{h}'$.
(5, 2): $\psi^2 s_2^3 s_5 \mathbf{h}\mathbf{h}'$.
(5, 3): $-\psi^2 s_2^3 s_5 \mathbf{h}\mathbf{h}'$.
(5, 4): $-\psi^2 s_2^2 s_3 s_4 \mathbf{h}\mathbf{h}'$.
(5, 5): $\psi^2 s_2^4 s_3 \mathbf{h}\mathbf{h}'$.
(5, 6): $-\psi^2 s_2^4 s_3 \mathbf{h}\mathbf{h}'$.
(6, 1): $-\psi^2 s_2^2 s_3 s_4 \mathbf{h}\mathbf{h}'$.
(6, 2): $-s_2^3 s_3(\mathbf{\Sigma} - \tau^{-1}\mathbf{h}\mathbf{h}') - \psi^2 s_2^3 s_5 \mathbf{h}\mathbf{h}'$.
(6, 3): $\psi^2 s_2^3 s_5 \mathbf{h}\mathbf{h}'$.
(6, 4): $\psi^2 s_2^2 s_3 s_4 \mathbf{h}\mathbf{h}'$.
(6, 5): $-\psi^2 s_2^4 s_3 \mathbf{h}\mathbf{h}'$.
(6, 6): $s_2^3 s_3(\mathbf{\Sigma} - \tau^{-1}\mathbf{h}\mathbf{h}') + \psi^2 s_2^4 s_3 \mathbf{h}\mathbf{h}'$.

The sum of the 36 terms is $s_2(s_2 s_5 - s_2^2 s_3 - s_3 s_4)(\mathbf{\Sigma} - \tau^{-1}\mathbf{h}\mathbf{h}')$. Hence, the limiting covariance matrix of $\sqrt{n} g_{n\eta}$ is $\{9w_1^2 s_2^{-1} s_3^2 (s_2 s_4 - s_2^3 - s_3^2) + 24 w_1 w_2 s_2^{-2} s_3 (s_4 - 3s_2^2)(s_2 s_5 - s_2^2 s_3 - s_3 s_4) + 16 w_2^2 s_2^{-3}(s_4 - 3s_2^2)^2 (s_2 s_6 - s_2 s_3^2 - s_4^2)\}(\mathbf{\Sigma} - \tau^{-1}\mathbf{h}\mathbf{h}')$. Consequently, the limiting covariance of $\sqrt{n}(\mathbf{u}_n - \mathbf{u}_0)$ is

$$\mathbf{\Psi}_\eta = \left\{ \frac{9w_1^2 s_2^3 s_3^2 (s_2 s_4 - s_2^3 - s_3^2) + 24 w_1 w_2 s_2^2 s_3 (s_4 - 3s_2^2)(s_2 s_5 - s_2^2 s_3 - s_3 s_4)}{\{3w_1 s_2 s_3^2 + 4w_2 (s_4 - 3s_2^2)^2\}^2} \right.$$
$$\left. + \frac{16 w_2^2 s_2 (s_4 - 3s_2^2)^2 (s_2 s_6 - s_2 s_3^2 - s_4^2)}{\{3w_1 s_2 s_3^2 + 4w_2 (s_4 - 3s_2^2)^2\}^2} \right\} \left(\mathbf{I}_p - \frac{\boldsymbol{\theta}\boldsymbol{\theta}'}{\|\boldsymbol{\theta}\|^2} \right) \mathbf{\Sigma}^{-1} \left(\mathbf{I}_p - \frac{\boldsymbol{\theta}\boldsymbol{\theta}'}{\|\boldsymbol{\theta}\|^2} \right).$$

Using now the expressions for s_2, s_3, s_4, s_6 in the proof of Theorem 2 and the analogously obtainable formula $s_5 = \|\boldsymbol{\theta}\|^{-5}(\alpha_1 - \alpha_2)\beta\tau^4\{(1-2\beta)\tau + 10\}$, the factor in front of the covariance matrix simplifies to $C_\eta(1+\beta\tau)/(\|\boldsymbol{\theta}\|^2\beta)$, where

$$C_\eta = \frac{9w_1^2(1+\beta\tau)^2(1-4\beta)(2+6\beta\tau+\beta\tau^2)}{\beta\tau^2\{3w_1(1+\beta\tau)(1-4\beta) + 4w_2\tau(1-6\beta)^2\}^2}$$
$$+ \frac{24 w_1 w_2 \tau^2 (1+\beta\tau)(1-4\beta)\beta(1-6\beta)(6+\tau) + 16 w_2^2 \tau (1-6\beta)^2 \Delta}{\beta\tau^2\{3w_1(1+\beta\tau)(1-4\beta) + 4w_2\tau(1-6\beta)^2\}^2}.$$

and $\Delta := 6 + 24\beta\tau + 9\beta(1-2\beta)\tau^2 + \beta(1-3\beta)\tau^3$.

□

Proof of Lemma 4. The limiting distribution of $\{\sqrt{n}(s_n\mathbf{u}_n - \boldsymbol{\theta}/\|\boldsymbol{\theta}\|)\}'\{\sqrt{n}(s_n\mathbf{u}_n - \boldsymbol{\theta}/\|\boldsymbol{\theta}\|)\}$ is the same as that of $\mathbf{z}'\boldsymbol{\Psi}\mathbf{z}$. The first claim now follows by observing that,

$$\{\sqrt{n}(s_n\mathbf{u}_n - \boldsymbol{\theta}/\|\boldsymbol{\theta}\|)\}'\{\sqrt{n}(s_n\mathbf{u}_n - \boldsymbol{\theta}/\|\boldsymbol{\theta}\|)\} = 2n(1 - s_n\mathbf{u}_n'\boldsymbol{\theta}/\|\boldsymbol{\theta}\|).$$

Finally, $\mathrm{E}(\mathbf{z}'\boldsymbol{\Psi}\mathbf{z}) = \mathrm{tr}\{\boldsymbol{\Psi}\mathrm{E}(\mathbf{z}\mathbf{z}')\} = \mathrm{tr}(\boldsymbol{\Psi})$. □

Proof of Lemma A.1. We first show that i) implies ii). The positive-definiteness of $\boldsymbol{\Sigma}$ in conjunction with the relation $\boldsymbol{\Sigma}\mathbf{h} = \phi\mathbf{h}$ gives that $\boldsymbol{\Sigma}^{-1}\mathbf{h} = \phi^{-1}\mathbf{h}$. Consequently,

$$\mathrm{Cov}(\mathbf{x})\frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|} = (\boldsymbol{\Sigma} + \beta\mathbf{h}\mathbf{h}')\frac{\boldsymbol{\Sigma}^{-1}\mathbf{h}}{\|\boldsymbol{\Sigma}^{-1}\mathbf{h}\|} = \frac{\phi}{\|\mathbf{h}\|}(1+\beta\tau)\mathbf{h} = \phi(1+\beta\tau)\frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|}.$$

Hence, $\pm\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ are the unique leading unit-length eigenvectors of $\mathrm{Cov}(\mathbf{x})$ if the second-to-largest eigenvalue of $\mathrm{Cov}(\mathbf{x})$ is smaller than $\phi(1+\beta\tau)$, the eigenvalue corresponding to $\pm\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$.

To see that ii) implies i), denote the eigenvalue of $\mathrm{Cov}(\mathbf{x})$ corresponding to $\pm\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ by ρ. Then, $(\boldsymbol{\Sigma} + \beta\mathbf{h}\mathbf{h}')\boldsymbol{\theta} = \rho\boldsymbol{\theta}$ or, equivalently,

$$\boldsymbol{\Sigma}\mathbf{h} = \frac{\rho}{1+\beta\tau}\mathbf{h},$$

showing that $\mathbf{h}$ is indeed an eigenvector of $\boldsymbol{\Sigma}$ corresponding to the eigenvalue $\phi := \rho/(1+\beta\tau)$. Finally, since $\pm\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ are the unique leading unit length eigenvectors of $\mathrm{Cov}(\mathbf{x})$, we have $\phi_2\{\mathrm{Cov}(\mathbf{x})\} < \rho = \phi(1+\beta\tau)$, concluding the proof. □

Proof of Theorem A.1. The proof of the strong consistency is done similarly as in Theorem 2 and we omit it. For the limiting normality we again, without loss of generality, assume that $\mathbf{x}$ has zero mean, implying that $\mathbf{x} \sim \alpha_1\mathcal{N}_p(-\alpha_2\mathbf{h}, \boldsymbol{\Sigma}) + \alpha_2\mathcal{N}_p(\alpha_1\mathbf{h}, \boldsymbol{\Sigma})$, where $\mathbf{h} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$.

Let $\mathbf{u}_1 := \boldsymbol{\theta}/\|\boldsymbol{\theta}\|$, where $\boldsymbol{\theta} = \boldsymbol{\Sigma}^{-1}\mathbf{h}$, and recall that it is a leading eigenvector of $\mathrm{Cov}(\mathbf{x}) = \boldsymbol{\Sigma} + \beta\mathbf{h}\mathbf{h}'$. Denote any set of the remaining $p-1$ eigenvectors by $\mathbf{u}_2, \ldots, \mathbf{u}_p$ and the corresponding eigenvalues by $\phi =: \phi_1 > \phi_2 \geq \cdots \geq \phi_p > 0$.

The gradient of the Lagrangian corresponding to the extraction of the leading unit length eigenvector of $\mathbf{C}_n$ is

$$2\mathbf{C}_n\mathbf{u} - 2\lambda_n\mathbf{u},$$

where λ_n is the Lagrangian multiplier. The gradient vanishes at $s_n\mathbf{u}_n$, allowing us to solve the value of $\lambda_n = \mathbf{u}_n'\mathbf{C}_n\mathbf{u}_n$ by multiplying the gradient equation from left with $\mathbf{u}_n$. Plugging the multiplier back in gives $(\mathbf{I}_p - \mathbf{P}_n)\mathbf{C}_n s_n\mathbf{u}_n = \mathbf{0}$, where $\mathbf{P}_n := \mathbf{u}_n\mathbf{u}_n'$. This equation is equivalent to the following equality,

$$\mathbf{A}_n\sqrt{n}(s_n\mathbf{u}_n - \mathbf{u}_1) = -(\mathbf{I}_p - \mathbf{P}_n)\sqrt{n}\{\mathbf{C}_n - \mathrm{Cov}(\mathbf{x})\}\mathbf{u}_1, \tag{B.16}$$

where

$$\mathbf{A}_n := (\mathbf{I}_p - \mathbf{P}_n)\mathbf{C}_n - s_n\phi_1(\mathbf{u}_n'\mathbf{u}_1)\mathbf{I}_p + \phi_1 s_n\mathbf{u}_1\mathbf{u}_n',$$

$\phi_1 = \mathbf{u}_1'\mathrm{Cov}(\mathbf{x})\mathbf{u}_1 = \|\boldsymbol{\theta}\|^{-2}\tau(1+\beta\tau)$ and $\tau = \mathbf{h}'\boldsymbol{\Sigma}^{-1}\mathbf{h}$. Now, by the strong consistency $s_n\mathbf{u}_n \to \mathbf{u}_1$, we have $\mathbf{P}_n \to_p \mathbf{u}_1\mathbf{u}_1' =: \mathbf{P}$. Consequently,

$$\mathbf{A}_n \to_p (\mathbf{I}_p - \mathbf{P})\mathrm{Cov}(\mathbf{x}) - \phi_1\mathbf{I}_p + \phi_1\mathbf{u}_1\mathbf{u}_1' = \mathrm{Cov}(\mathbf{x}) - \phi_1\mathbf{I}_p.$$

Observe then that we have the identity $\mathbf{B}_n\sqrt{n}(s_n\mathbf{u}_n - \mathbf{u}_1) = \mathbf{0}$ where $\mathbf{B}_n := \mathbf{u}_1\mathbf{u}_1' + s_n\mathbf{u}_1\mathbf{u}_n' \to_p 2\mathbf{u}_1\mathbf{u}_1'$. As $\sqrt{n}\{\mathbf{C}_n - \mathrm{Cov}(\mathbf{x})\} = \mathcal{O}_p(1)$, summing the previous equation and (B.16), yields,

$$(\mathbf{A}_n + \mathbf{B}_n)\sqrt{n}(s_n\mathbf{u}_n - \mathbf{u}_1) = -(\mathbf{I}_p - \mathbf{P})\sqrt{n}\{\mathbf{C}_n - \mathrm{Cov}(\mathbf{x})\}\mathbf{u}_1 + o_p(1),$$

where $\mathbf{A}_n + \mathbf{B}_n \to_p 2\mathbf{u}_1\mathbf{u}_1' + \sum_{j=2}^p(\phi_j - \phi_1)\mathbf{u}_j\mathbf{u}_j'$ and $\phi_j - \phi_1 < 0$ for all $j = 2, \ldots, p$. Hence, by Slutsky's theorem, the limiting distribution of $\sqrt{n}(s_n\mathbf{u}_n - \mathbf{u}_1)$ is that of

$$\begin{aligned} &-\left(\frac{1}{2}\mathbf{u}_1\mathbf{u}_1' + \sum_{j=2}^p \frac{1}{\phi_j - \phi_1}\mathbf{u}_j\mathbf{u}_j'\right)(\mathbf{I}_p - \mathbf{P})\sqrt{n}\{\mathbf{C}_n - \mathrm{Cov}(\mathbf{x})\}\mathbf{u}_1 \\ &= -\left(\sum_{j=2}^p \frac{1}{\phi_j - \phi_1}\mathbf{u}_j\mathbf{u}_j'\right)\sqrt{n}\{\mathbf{C}_n - \mathrm{Cov}(\mathbf{x})\}\mathbf{u}_1. \end{aligned}$$

Observing that $(1/n)\sum_i(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' = (1/n)\sum_i \mathbf{x}_i\mathbf{x}_i' + o_p(1/\sqrt{n})$, the limiting covariance matrix of $\sqrt{n}(s_n\mathbf{u}_n - \mathbf{u}_1)$ is hence

$$\boldsymbol{\Psi}_{\mathrm{PCA}} = \left(\sum_{j=2}^p \frac{1}{\phi_j - \phi_1}\mathbf{u}_j\mathbf{u}_j'\right)\mathrm{Cov}\{(\mathbf{u}_1'\mathbf{x})\mathbf{x}\}\left(\sum_{j=2}^p \frac{1}{\phi_j - \phi_1}\mathbf{u}_j\mathbf{u}_j'\right).$$

In the notation of Theorem 2, we have $\mathrm{Cov}\{(\mathbf{u}_1'\mathbf{x})\mathbf{x}\} = \mathbf{G}_2 - \mathbf{m}_1\mathbf{m}_1' = s_2(\boldsymbol{\Sigma} - \tau^{-1}\mathbf{h}\mathbf{h}') + \psi^2(s_4 - s_2^2)\mathbf{h}\mathbf{h}'$, where $s_2 = \|\boldsymbol{\theta}\|^{-2}\tau(1+\beta\tau)$ and $s_4/s_2^2 = \kappa(\mathbf{u}_1) = \kappa(\boldsymbol{\theta})$, yielding the result.

□

To prove Theorem 5, we first establish two sets of auxiliary results. The first one has to do with bounding a specific Orlicz norm of the third power of a Gaussian mixture and the second one is devoted to the subresults that are required in the M-estimator argument (uniform convergence of the sample objective function and uniform identifiability of the maximizer).

Auxiliary results for the proof of Theorem 5, part 1

Recall that the ψ_λ-Orlicz "norm" of a random variable X is defined as $\|X\|_{\psi_\lambda} := \inf\{\theta > 0 \mid \mathrm{E}\exp[(|X|/\theta)^\lambda] \le 2\}$. We write "norm" as $\|\cdot\|_{\psi_\lambda}$ fails to satisfy the triangle inequality for $\lambda < 1$, see, e.g., [33]. As our first task, we show that the random variable $X^3 - \mathrm{E}(X^3)$, where X is a centered normal mixture, has $\|X^3 - \mathrm{E}(X^3)\|_{\psi_{2/3}} < \infty$.

Lemma B.4. *Let* $X \sim \alpha_1\mathcal{N}(-\alpha_2 h, \sigma^2) + \alpha_2\mathcal{N}(\alpha_1 h, \sigma^2)$ *where* $h \in \mathbb{R}$, $\sigma^2 > 0$. *Then,*

$$\|X^3 - \mathrm{E}(X^3)\|_{\psi_{2/3}} \leq 32(\sigma^2 + h^2)^{3/2}.$$

Proof of Lemma B.4. We write $X = b_1X_1 + (1 - b_1)X_2$ where X_1, X_2, b_1 are independent and $X_1 \sim \mathcal{N}(-\alpha_2 h, \sigma^2)$, $X_2 \sim \mathcal{N}(\alpha_1 h, \sigma^2)$ and $b_1 \sim \mathrm{Ber}(\alpha_1)$. Consequently, the moment-generating function of X has,

$$\begin{aligned}\mathrm{E}\exp(tX) =& \alpha_1 \exp(-\alpha_2 ht + \sigma^2 t^2/2) + \alpha_2 \exp(\alpha_1 ht + \sigma^2 t^2/2) \\ =& \exp(\sigma^2 t^2/2)\{\alpha_1 \exp(-\alpha_2 ht) + \alpha_2 \exp(\alpha_1 ht)\}.\end{aligned} \tag{B.17}$$

By the Kearns-Saul inequality [31, Lemma 1], the second multiplicand on the right-hand side of (B.17) has the upper bound

$$\exp\{(1/4)h^2t^2(\alpha_1 - \alpha_2)/\log(\alpha_1/\alpha_2)\},$$

for all $h, \sigma^2, \alpha_1, \alpha_2$, where $(\alpha_1 - \alpha_2)/\log(\alpha_1/\alpha_2)$ is interpreted as its limit $1/2$ when $\alpha_1 = 1/2$, see [7, Theorem 2.3]. This bound can further be verified to be (very crudely) upper bounded by $\exp\{(1/2)h^2t^2\}$. Consequently, plugging in to (B.17), we get

$$\mathrm{E}\exp(tX) \leq \exp\{(\sigma^2 + h^2)t^2/2\}, \tag{B.18}$$

for all $t \in \mathbb{R}$. We next use this bound for the MGF to derive a tail bound for X, using the typical approach via Markov's inequality. That is, for each $x \geq 0$, $\lambda > 0$, we have,

$$\begin{aligned}\mathrm{P}(X \geq x) &= \mathrm{P}\{\exp(\lambda X) \geq \exp(\lambda x)\} \\ &\leq \exp(-\lambda x)\mathrm{E}\exp(\lambda X) \\ &\leq \exp\{-\lambda x + (\sigma^2 + h^2)\lambda^2/2\}.\end{aligned}$$

Optimizing the bound over $\lambda > 0$, we find that the minimum is reached at $\lambda = x/(\sigma^2 + h^2)$, giving,

$$\mathrm{P}(X \geq x) \leq \exp(-bx^2),$$

for all $x \geq 0$, where we use the notation $b = b(\sigma^2, h^2) := (\sigma^2 + h^2)^{-1}/2$. Repeating the exercise with $-X$ in place of X, we find that it obeys the same tail bound, implying that

$$\mathrm{P}(|X| \geq x) \leq 2\exp(-bx^2),$$

for all $x \geq 0$. And, consequently, we get a sub-Weibull tail bound for X^3:

$$\mathrm{P}(|X^3| \geq x) \leq 2\exp(-bx^{2/3}), \tag{B.19}$$

for all $x \geq 0$.

Next, we incorporate the mean $\mu := \mathrm{E}(X^3) = \alpha_1\alpha_2(\alpha_1 - \alpha_2)h^3$ into (B.19):

$$\mathrm{P}(|X^3 - \mu| \geq x) \leq \mathrm{P}(|X^3| \geq x - |\mu|) \leq 2\exp\{-b(x - |\mu|)^{2/3}\},$$

for all $x \geq |\mu|$. The inequality $|r + s|^{2/3} \leq |r|^{2/3} + |s|^{2/3}$, valid for all $r, s \in \mathbb{R}$, then gives, with $r = |\mu|, s = x - |\mu|$,

$$\begin{aligned} \mathrm{P}(|X^3 - \mu| \geq x) &\leq 2\exp\{-bx^{2/3} + b|\mu|^{2/3}\} \\ &\leq 2\exp\{-bx^{2/3} + bh^2\} \\ &\leq 4\exp\{-bx^{2/3}\}, \end{aligned} \tag{B.20}$$

for all $x \geq |\mu|$, where we have used $|\alpha_1\alpha_2(\alpha_1 - \alpha_2)h^3| \leq |h^3|$, $bh^2 \leq 1/2$ and $\exp(1/2) \leq 2$. As $b|\mu|^{2/3} < 1/2$, we have that the final upper bound in (B.20) takes a value greater than one when $x = |\mu|$. As probabilities are trivially upper bounded by one and since the final upper bound in (B.20) is a decreasing function of x, we observe that this bound actually holds for all $x \geq 0$.

Denoting $|Z| := |X^3 - \mu|$, the inequality (B.20) now lets us write, for all $t \geq 1$, $\theta > 0$,

$$\mathrm{P}[\exp\{(|Z|/\theta)^{2/3}\} \geq t] = \mathrm{P}[|Z| \geq \theta\{\log(t)\}^{3/2}] \leq 4t^{-b\theta^{2/3}}.$$

This, in turn, finally allows us to bound the desired Orlicz norm $\|Z\|_{\psi_{2/3}}$ as follows: Letting $\theta > 0$, we have,

$$\begin{aligned} \mathrm{E}\exp\{(|Z|/\theta)^{2/3}\} &= \int_0^\infty \mathrm{P}[\exp\{(|Z|/\theta)^{2/3}\} \geq t]dt \\ &\leq 1 + 4\int_1^\infty t^{-b\theta^{2/3}}dt \\ &= 1 - \frac{4}{1 - b\theta^{2/3}}, \end{aligned}$$

for all $\theta = ab^{-3/2}$, where $a > 1$. Choosing now $a = 5^{3/2}$, makes this upper bound equal to 2, finally giving,

$$\|Z\|_{\psi_{2/3}} \leq 5^{3/2}b^{-3/2} \leq 32(\sigma^2 + h^2)^{3/2}.$$

□

Having established the finiteness of the norm allows us to next apply the general tail bound for averages of centered sub-Weibull variables given in [33].

Lemma B.5. *Let $X_1, \ldots, X_n$ be a random sample from the distribution of $X \sim \alpha_1\mathcal{N}(-\alpha_2 h, \sigma^2) + \alpha_2\mathcal{N}(\alpha_1 h, \sigma^2)$ where $h \in \mathbb{R}$, $\sigma^2 > 0$. Then,*

$$\mathrm{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n X_i^3 - \mathrm{E}(X^3)\right| \geq \varepsilon\right\} \leq 2\exp\left\{-\left(\frac{\sqrt{n}\varepsilon}{K\{1 + (\sigma^2 + h^2)^{3/2}\}}\right)^{2/3}\right\},$$

for all $\varepsilon \geq K\{1 + (\sigma^2 + h^2)^{3/2}\}n^{-1/2}$ where $K > 0$ is a constant not depending on any of the parameters.

Proof of Lemma B.5. The result is a consequence of Theorem 3.1 in [33] and we give the details below. We first note that the result itself can be applied as the random variables $X_i^3 - E(X^3)$ are i.i.d., have zero means and satisfy $\|X_i^3 - \mathrm{E}(X^3)\|_{\psi_{2/3}} < \infty$ by Lemma B.4.

Now, using the notation of [33], we have $\|b\|_2 \leq 32n^{-1/2}(\sigma^2 + h^2)^{3/2}$ (by Lemma B.4) and $\|b\|_\infty / \|b\|_2 = n^{-1/2}$. The quantity on the RHS inside the probability statement in (3.1) in [33] then has the upper bound,

$$64eC(2/3)n^{-1/2}(\sigma^2 + h^2)^{3/2}\sqrt{t} + 2eC(2/3)4^{3/2}2^{-1/2}n^{-1/2}t^{3/2}, \tag{B.21}$$

where $C(2/3)$ can be checked to satisfy $C(2/3) < 1500$ and we also have $4^{3/2}2^{-1/2} < 6$. Assuming further that $t \geq 1$, we also have $t^{1/2} \leq t^{3/2}$, implying that (B.21) has the upper bound,

$$\{96000e(\sigma^2 + h^2)^{3/2} + 18000e\}n^{-1/2}t^{3/2} \leq K\{(\sigma^2 + h^2)^{3/2} + 1\}n^{-1/2}t^{3/2},$$

where $K := 96000e$. Consequently, Theorem 3.1 in [33] gives us the bound,

$$\mathrm{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i^3 - E(X^3)\right| \geq K\{(\sigma^2 + h^2)^{3/2} + 1\}n^{-1/2}t^{3/2}\right] \leq 2\exp(-t),$$

for all $t \geq 1$. Setting now $\varepsilon := K\{(\sigma^2 + h^2)^{3/2} + 1\}n^{-1/2}t^{3/2}$ gives the claim. □

In the low-dimensional case (i.e., constant p), Lemma B.5 simply recovers the law of large numbers for X_i^3 (with explicit tail bound), but unlike the standard LLN, Lemma B.5 retains its usefulness also in the high-dimensional case. In particular, it allows us to establish a uniform law of large numbers for the projections of a high-dimensional X_i.

Theorem B.1. *Let $\boldsymbol{x}_{n1}, \ldots, \boldsymbol{x}_{nn}$ be a random sample from the model (2) and assume that*

$$p_n \to \infty \qquad \text{and} \qquad p_n(\|\boldsymbol{\Sigma}_n\|_2 + \|\boldsymbol{h}_n\|^2) = o(n^{1/3}).$$

Assume further that for some $C > 0$, we have $\|\boldsymbol{\Sigma}_n\|_2 + \|\boldsymbol{h}_n\|^2 > C$ for all n large enough. Then,

$$\sup_{\boldsymbol{u} \in \mathbb{S}^{p_n - 1}} \left|\frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{u}'\boldsymbol{x}_{ni})^3 - \mathrm{E}\{(\boldsymbol{u}'\boldsymbol{x}_n)^3\}\right| \to_p 0,$$

as $n \to \infty$.

Proof of Theorem B.1. Let

$$\mathcal{R}_n := (h_{jk\ell}) = (1/n)\sum_{i=1}^{n}\{x_{nij}x_{nik}x_{ni\ell} - \mathrm{E}(x_{nj}x_{nk}x_{n\ell})\}$$

be the $p_n \times p_n \times p_n$ third-order symmetric tensor containing all centered sample third moments. Given $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3 \in \mathbb{R}^{p_n}$ we denote by $\mathcal{R}_n \times (\mathbf{u}_1 \otimes \mathbf{u}_2 \otimes \mathbf{u}_3)$

the scalar $\sum_{j=1}^{p_n}\sum_{k=1}^{p_n}\sum_{\ell=1}^{p_n} u_{1j}u_{2k}u_{3\ell}h_{jk\ell}$. Using this notation, our quantity of interest is,

$$\|\mathcal{R}_n\|_2 := \sup_{\mathbf{u}\in\mathbb{S}^{p_n-1}} |\mathcal{R}_n \times (\mathbf{u}\otimes\mathbf{u}\otimes\mathbf{u})|\,,$$

i.e., the spectral norm of the tensor $\mathcal{R}_n$, see, e.g., [18, Lemma 1].

Fix next $\varepsilon > 0$ and let $\mathcal{N}_{n\varepsilon} \subseteq \mathbb{S}^{p_n-1}$ be an ε-net of $\mathbb{S}^{p_n-1}$. That is, for each $\mathbf{u} \in \mathbb{S}^{p_n-1}$ there exists $\mathbf{v} \in \mathcal{N}_{n\varepsilon}$ such that $\|\mathbf{u}-\mathbf{v}\| \le \varepsilon$. From ([61], Lemma 5.2) we know that $\mathcal{N}_{n\varepsilon}$ can be chosen such that its cardinality $|\mathcal{N}_{n\varepsilon}|$ is at most $(1+2/\varepsilon)^{p_n}$. Fix now $\mathbf{u} \in \mathbb{S}^{p_n-1}$ and let $\mathbf{v} \equiv \mathbf{v}_u$ be its ε-neighbour in $\mathcal{N}_{n\varepsilon}$. Then, we have,

$$\begin{aligned}
&|\mathcal{R}_n \times (\mathbf{u}\otimes\mathbf{u}\otimes\mathbf{u}) - \mathcal{R}_n \times (\mathbf{v}\otimes\mathbf{v}\otimes\mathbf{v})| \\
\le&|\mathcal{R}_n \times (\mathbf{u}\otimes\mathbf{u}\otimes\{\mathbf{u}-\mathbf{v}\})| \\
+&|\mathcal{R}_n \times (\mathbf{u}\otimes\{\mathbf{u}-\mathbf{v}\}\otimes\mathbf{v})| \\
+&|\mathcal{R}_n \times (\{\mathbf{u}-\mathbf{v}\}\otimes\mathbf{v}\otimes\mathbf{v})| \\
\le&3\varepsilon\|\mathcal{R}_n\|_2,
\end{aligned} \tag{B.22}$$

where the final inequality follows as $|\mathcal{R}_n \times (\mathbf{a}_1\otimes\mathbf{a}_2\otimes\mathbf{a}_3)| \le \|\mathcal{R}_n\|_2\|\mathbf{a}_1\|\|\mathbf{a}_2\|\|\mathbf{a}_3\|$ for any $\mathbf{a}_1,\mathbf{a}_2,\mathbf{a}_3 \in \mathbb{R}^{p_n}$, see [18]. By the reverse triangle inequality, (B.22) gives,

$$\begin{aligned}
&|\mathcal{R}_n \times (\mathbf{u}\otimes\mathbf{u}\otimes\mathbf{u})| \le |\mathcal{R}_n \times (\mathbf{v}\otimes\mathbf{v}\otimes\mathbf{v})| + 3\varepsilon\|\mathcal{R}_n\|_2 \\
\le& \max_{\mathbf{w}\in\mathcal{N}_{n\varepsilon}} |\mathcal{R}_n \times (\mathbf{w}\otimes\mathbf{w}\otimes\mathbf{w})| + 3\varepsilon\|\mathcal{R}_n\|_2.
\end{aligned}$$

Since the above holds for all $\mathbf{u} \in \mathcal{S}^{p_n-1}$, we further get,

$$\|\mathcal{R}_n\|_2 = \sup_{\mathbf{u}\in\mathbb{S}^{p_n-1}} |\mathcal{R}_n \times (\mathbf{u}\otimes\mathbf{u}\otimes\mathbf{u})| \le \max_{\mathbf{w}\in\mathcal{N}_{n\varepsilon}} |\mathcal{R}_n \times (\mathbf{w}\otimes\mathbf{w}\otimes\mathbf{w})| + 3\varepsilon\|\mathcal{R}_n\|_2,$$

that is,

$$\|\mathcal{R}_n\|_2 \le \frac{1}{1-3\varepsilon} \max_{\mathbf{u}\in\mathcal{N}_{n\varepsilon}} |\mathcal{R}_n \times (\mathbf{u}\otimes\mathbf{u}\otimes\mathbf{u})|,$$

whenever $\varepsilon < 1/3$. We next apply this bound with the choice $\varepsilon = 2/9$ to get

$$\begin{aligned}
\mathrm{P}(\|\mathcal{R}_n\|_2 \ge t) \le& \mathrm{P}\Big(\max_{\mathbf{u}\in\mathcal{N}_{n(2/9)}} |\mathcal{R}_n \times (\mathbf{u}\otimes\mathbf{u}\otimes\mathbf{u})| \ge t/3\Big) \\
\le& \sum_{\mathbf{u}\in\mathcal{N}_{n(2/9)}} \mathrm{P}(|\mathcal{R}_n \times (\mathbf{u}\otimes\mathbf{u}\otimes\mathbf{u})| \ge t/3).
\end{aligned} \tag{B.23}$$

Now, for a fixed $\mathbf{u} \in \mathcal{N}_{n(2/9)}$, the quantity $\mathcal{R}_n \times (\mathbf{u}\otimes\mathbf{u}\otimes\mathbf{u})$ is distributed as $\frac{1}{n}\sum_{i=1}^n X_i^3 - E(X^3)$ where $X_1,\ldots,X_n$ is a random sample from the distribution of $X \sim \alpha_1\mathcal{N}(-\alpha_2\mathbf{u}'\mathbf{h}_n, \mathbf{u}'\boldsymbol{\Sigma}_n\mathbf{u}) + \alpha_2\mathcal{N}(\alpha_1\mathbf{u}'\mathbf{h}_n, \mathbf{u}'\boldsymbol{\Sigma}_n\mathbf{u})$. Consequently, Lemma B.5 implies that

$$\mathrm{P}(|\mathcal{R}_n \times (\mathbf{u}\otimes\mathbf{u}\otimes\mathbf{u})| \ge t/3) \le 2\exp\left[-\left(\frac{\sqrt{n}t}{3K\{1+(\|\boldsymbol{\Sigma}_n\|_2 + \|\mathbf{h}_n\|^2)^{3/2}\}}\right)^{2/3}\right]$$

whenever $t \geq 3K[1 + \{\mathbf{u}'\boldsymbol{\Sigma}_n\mathbf{u} + (\mathbf{u}'\mathbf{h}_n)^2\}^{3/2}]n^{-1/2}$. I.e., in particular when $t \geq 3K\{1 + (\|\boldsymbol{\Sigma}_n\|_2 + \|\mathbf{h}_n\|^2)^{3/2}\}n^{-1/2}$. As $|\mathcal{N}_{n\varepsilon}| \leq 10^{p_n}$, plugging in to (B.23), we finally get,

$$\mathrm{P}(\|\mathcal{R}_n\|_2 \geq t) \leq 2\exp\left\{-\left(\frac{\sqrt{n}t}{3K\{1 + (\|\boldsymbol{\Sigma}_n\|_2 + \|\mathbf{h}_n\|^2)^{3/2}\}}\right)^{2/3} + p_n \log 10\right\},$$

for all $t \geq 3K\{1+(\|\boldsymbol{\Sigma}_n\|_2+\|\mathbf{h}_n\|^2)^{3/2}\}n^{-1/2}$, which is, for a fixed $t > 0$, satisfied for all large enough n, thanks to our assumptions that $p_n \to \infty$ and $p_n(\|\boldsymbol{\Sigma}_n\|_2 + \|\mathbf{h}_n\|^2) = o(n^{1/3})$. Denoting $H_n := \|\boldsymbol{\Sigma}_n\|_2 + \|\mathbf{h}_n\|^2$, the same assumptions, in conjunction with the requirement that $H_n \geq C$ for all n large enough, guarantee that we have, for each fixed $t > 0$, that

$$-\frac{n^{1/3}}{H_n}\left(\frac{t^{2/3}}{\{3K\}^{2/3}\{H_n^{-3/2} + 1\}^{2/3}} + \frac{p_n H_n}{n^{1/3}}\log 10\right) \to -\infty,$$

as $n \to \infty$. Thus the claim follows. □

We next show equivalent results for the second moment instead of the third, beginning with the finiteness of the ψ_1-Orlicz "norm" of $X^2 - E(X^2)$ for univariate normal mixtures. The proof of this is exactly analogous to that of Lemma B.4 and we omit it.

Lemma B.6. *Let $X \sim \alpha_1\mathcal{N}(-\alpha_2 h, \sigma^2) + \alpha_2\mathcal{N}(\alpha_1 h, \sigma^2)$ where $h \in \mathbb{R}$, $\sigma^2 > 0$. Then,*

$$\|X^2 - \mathrm{E}(X^2)\|_{\psi_1} \leq 10(\sigma^2 + h^2).$$

Lemma B.6 allows us to derive a concentration bound for the sample-centered second moment. Again, the proof exactly follows Lemma B.5, causing us to leave it out.

Lemma B.7. *Let $X_1, \ldots, X_n$ be a random sample from the distribution of $X \sim \alpha_1\mathcal{N}(-\alpha_2 h, \sigma^2) + \alpha_2\mathcal{N}(\alpha_1 h, \sigma^2)$ where $h \in \mathbb{R}$, $\sigma^2 > 0$. Then,*

$$\mathrm{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i^2 - E(X^2)\right| \geq \varepsilon\right\} \leq 2\exp\left\{-\frac{\sqrt{n}\varepsilon}{K(1 + \sigma^2 + h^2)}\right\},$$

for all $\varepsilon \geq K(1 + \sigma^2 + h^2)n^{-1/2}$ where $K > 0$ is a constant not depending on any of the parameters.

Finally, Lemmas B.6 and B.7 give us the uniform law of large numbers for the second sample moment over all projections.

Theorem B.2. *Let $\boldsymbol{x}_{n1}, \ldots, \boldsymbol{x}_{nn}$ be a random sample from the model (2) and assume that*

$$p_n \to \infty \qquad \text{and} \qquad p_n(\|\boldsymbol{\Sigma}_n\|_2 + \|\boldsymbol{h}_n\|^2) = o(n^{1/2}).$$

Assume further that for some $C > 0$, we have $\|\boldsymbol{\Sigma}_n\|_2 + \|\boldsymbol{h}_n\|^2 > C$ for all n large enough. Then,

$$\sup_{\boldsymbol{u}\in\mathbb{S}^{p_n-1}} \left| \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{u}'\boldsymbol{x}_{ni})^2 - \mathrm{E}\{(\boldsymbol{u}'\boldsymbol{x}_n)^2\} \right| \to_p 0,$$

as $n \to \infty$.

Proof of Theorem B.2. We provide only the main steps of the proof (which is very similar to that of Theorem B.1). Letting $\mathbf{S}_n := (1/n)\sum_{i=1}^{n} \mathbf{x}_{ni}\mathbf{x}'_{ni} - \mathrm{E}\{\mathbf{x}_n\mathbf{x}'_n\}$, our claim is that $\|\mathbf{S}_n\|_2 \to_p 0$.

Arguing as in Theorem B.1 we obtain,

$$\|\mathbf{S}_n\|_2 \le \frac{1}{1-2\varepsilon} \max_{\mathbf{u}\in\mathcal{N}_{n\varepsilon}} |\mathbf{u}'\mathbf{S}_n\mathbf{u}|,$$

whenever $\varepsilon < 1/2$, where $\mathcal{N}_{n\varepsilon}$ is an ε-net of $\mathbb{S}^{p_n-1}$. Thus,

$$\mathrm{P}(\|\mathbf{S}_n\|_2 \ge t) \le \sum_{\mathbf{u}\in\mathcal{N}_{n(1/4)}} \mathrm{P}(|\mathbf{u}'\mathbf{S}_n\mathbf{u}| \ge t/2), \tag{B.24}$$

where the right-hand side probabilities each satisfy, by Lemma B.7,

$$\mathrm{P}(|\mathbf{u}'\mathbf{S}_n\mathbf{u}| \ge t/2) \le 2\exp\left\{ -\frac{\sqrt{n}t}{2K(1+\|\boldsymbol{\Sigma}_n\|_2 + \|\mathbf{h}_n\|^2)} \right\},$$

when $t \ge 2K(1+\|\boldsymbol{\Sigma}_n\|_2 + \|\mathbf{h}_n\|^2)n^{-1/2}$. Plugging in to (B.24), we get,

$$\mathrm{P}(\|\mathbf{S}_n\|_2 \ge t) \le 2\exp\left\{ -\frac{\sqrt{n}t}{2K(1+\|\boldsymbol{\Sigma}_n\|_2 + \|\mathbf{h}_n\|^2)} + p_n \log 9 \right\},$$

for all $t \ge 2K(1+\|\boldsymbol{\Sigma}_n\|_2 + \|\mathbf{h}_n\|^2)n^{-1/2}$, a condition which is, for a fixed $t > 0$, satisfied for all large enough n by our assumptions. The desired result now follows. □

Auxiliary results for the proof of Theorem 5, part 2

In this subsection, we prove the two main parts required in the M-estimator argument, i.e., uniform identifiability and uniform law of large numbers. However, before them, we first give a lemma that quantifies the extent to which invertible linear transformations preserve the non-parallelity of vectors.

Lemma B.8. *Assume that $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^p$, $\boldsymbol{a}, \boldsymbol{b} \neq \boldsymbol{0}$, are such that*

$$\left| \frac{\boldsymbol{a}'\boldsymbol{b}}{\|\boldsymbol{a}\|\|\boldsymbol{b}\|} \right| \le 1 - \varepsilon,$$

for some $\varepsilon > 0$. Moreover, let $\boldsymbol{M} \in \mathbb{R}^{p\times p}$ satisfy $\max\{\|\boldsymbol{M}\|_2, \|\boldsymbol{M}^{-1}\|_2\} \le C$ for some $C > 0$. Then,

$$\left| \frac{\boldsymbol{a}'\boldsymbol{M}^2\boldsymbol{b}}{\|\boldsymbol{M}\boldsymbol{a}\|\|\boldsymbol{M}\boldsymbol{b}\|} \right| \le 1 - \varepsilon C^{-4}.$$

Proof of Lemma B.8. We have

$$\begin{aligned}\frac{\mathbf{a}'\mathbf{M}^2\mathbf{b}}{\|\mathbf{Ma}\|\|\mathbf{Mb}\|} &= 1 - \frac{1}{2}\left\|\frac{\mathbf{Ma}}{\|\mathbf{Ma}\|} - \frac{\mathbf{Mb}}{\|\mathbf{Mb}\|}\right\|^2 \\ &\leq 1 - \frac{1}{2\|\mathbf{M}^{-1}\|_2^2}\left\|\frac{\mathbf{a}}{\|\mathbf{Ma}\|} - \frac{\mathbf{b}}{\|\mathbf{Mb}\|}\right\|^2 \\ &\leq 1 - \frac{1}{2C^2}\left\{\left(\frac{\|\mathbf{a}\|}{\|\mathbf{Ma}\|} - \frac{\|\mathbf{b}\|}{\|\mathbf{Mb}\|}\right)^2 + 2\varepsilon\frac{\|\mathbf{a}\|\|\mathbf{b}\|}{\|\mathbf{Ma}\|\|\mathbf{Mb}\|}\right\} \\ &\leq 1 - \varepsilon C^{-4},\end{aligned}$$

where the last line uses $\|\mathbf{Ma}\| \leq \|\mathbf{M}\|_2\|\mathbf{a}\| \leq C\|\mathbf{a}\|$ and similarly for $\|\mathbf{Mb}\|$. The desired bound for the other direction can be obtained by starting from the equation,

$$-\frac{\mathbf{a}'\mathbf{M}^2\mathbf{b}}{\|\mathbf{Ma}\|\|\mathbf{Mb}\|} = 1 - \frac{1}{2}\left\|\frac{\mathbf{Ma}}{\|\mathbf{Ma}\|} + \frac{\mathbf{Mb}}{\|\mathbf{Mb}\|}\right\|^2,$$

and proceeding analogously. □

Lemma B.9 then next gives sufficient conditions for the model parameters under which the sequence of population level maximizers is identifiable uniformly in n.

Lemma B.9. *Let $\boldsymbol{x}_{n1}, \ldots, \boldsymbol{x}_{nn}$ be a random sample from the model (2) with $\alpha_1 \neq \alpha_2$ and assume that there exists $C_1, C_2 > 0$ such that, for all n,*

$$1/C_1 \leq \|\boldsymbol{h}_n\| \leq C_1, \quad \|\boldsymbol{\Sigma}_n\|_2 \leq C_2, \quad \|\boldsymbol{\Sigma}_n^{-1}\|_2 \leq C_2.$$

Then, for all fixed $\varepsilon \in (0, C_2^2)$, there exists $\delta \equiv \delta(\varepsilon, C_1, C_2, \alpha_1) > 0$ such that (a) below implies (b) for all n.

(a) $\boldsymbol{u} \in \mathbb{S}^{p_n-1}$ is such that $|\boldsymbol{u}'\boldsymbol{\theta}_n| \leq 1 - \varepsilon$.
(b) $\gamma_{n0}^2(\boldsymbol{\theta}_n) - \gamma_{n0}^2(\boldsymbol{u}) \geq \delta$.

Proof of Lemma B.9. By the proof of Lemma 2,

$$\gamma_{n0}^2(\mathbf{u}) = (\alpha_1\alpha_2)^4(1-4\alpha_1\alpha_2)^2\left(\frac{f_n(\mathbf{u})}{1+\alpha_1\alpha_2 f_n(\mathbf{u})}\right)^6 =: (\alpha_1\alpha_2)^4(1-4\alpha_1\alpha_2)^2 g_n^6(\mathbf{u}),$$

where $f_n(\mathbf{u}) := (\mathbf{u}'\mathbf{h}_n)^2/\mathbf{u}'\boldsymbol{\Sigma}_n\mathbf{u}$ and the constant $\iota := (\alpha_1\alpha_2)^4(1-4\alpha_1\alpha_2)^2$ is strictly positive. Moreover, $g_n(\boldsymbol{\theta}_n) \geq g_n(\mathbf{u}) \geq 0$, implying that we have

$$\begin{aligned}\gamma_{n0}^2(\boldsymbol{\theta}_n) - \gamma_{n0}^2(\mathbf{u}) &\geq \iota\{g_n(\boldsymbol{\theta}_n)^3 - g_n(\mathbf{u})^3\}\{g_n(\boldsymbol{\theta}_n)^3 + g_n(\mathbf{u})^3\} \\ &\geq \iota\{g_n(\boldsymbol{\theta}_n) - g_n(\mathbf{u})\}g_n^5(\boldsymbol{\theta}_n).\end{aligned}$$

Now, $f_n(\boldsymbol{\theta}_n) = \mathbf{h}_n'\boldsymbol{\Sigma}_n^{-1}\mathbf{h}_n \geq \|\mathbf{h}_n\|^2\|\boldsymbol{\Sigma}_n\|_2^{-1} \geq C_1^{-2}C_2^{-1}$ for all n, implying that there exists $C_3 > 0$ such that $g_n(\boldsymbol{\theta}_n) \geq C_3$ for all n. Thus, the claim of the

lemma holds once we show that $g_n(\boldsymbol{\theta}_n) - g_n(\mathbf{u}) \geq C_4$ for some $C_4 > 0$ not depending on n.

Now,

$$\begin{aligned} g_n(\boldsymbol{\theta}_n) - g_n(\mathbf{u}) =& \frac{f_n(\boldsymbol{\theta}_n) - f_n(\mathbf{u})}{\{1+\alpha_1\alpha_2 f_n(\boldsymbol{\theta}_n)\}\{1+\alpha_1\alpha_2 f_n(\mathbf{u})\}} \\ \geq& \frac{f_n(\boldsymbol{\theta}_n) - f_n(\mathbf{u})}{\{1+\alpha_1\alpha_2 f_n(\boldsymbol{\theta}_n)\}^2} \\ \geq& \frac{f_n(\boldsymbol{\theta}_n) - f_n(\mathbf{u})}{\{1+\alpha_1\alpha_2 \|\mathbf{h}_n\|^2 \|\boldsymbol{\Sigma}_n^{-1}\|_2\}^2} \\ \geq& \frac{f_n(\boldsymbol{\theta}_n) - f_n(\mathbf{u})}{\{1+\alpha_1\alpha_2 C_1^2 C_2\}^2}, \end{aligned}$$

showing that it remains to lower bound $f_n(\boldsymbol{\theta}_n) - f_n(\mathbf{u})$ is a manner not depending on n. Some algebra reveals that,

$$f_n(\mathbf{u}) = \left(\frac{\mathbf{u}'\boldsymbol{\Sigma}_n\boldsymbol{\theta}_n}{\|\boldsymbol{\Sigma}_n^{1/2}\mathbf{u}\|\|\boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\theta}_n\|} \right)^2 \mathbf{h}_n'\boldsymbol{\Sigma}_n^{-1}\mathbf{h}_n \leq (1-\varepsilon C_2^{-2})^2 \mathbf{h}_n'\boldsymbol{\Sigma}_n^{-1}\mathbf{h}_n,$$

where the inequality follows from applying Lemma B.8 with $\mathbf{a} = \mathbf{u}$, $\mathbf{b} = \boldsymbol{\theta}_n$ and $\mathbf{M} = \boldsymbol{\Sigma}_n^{1/2}$. Consequently,

$$f_n(\boldsymbol{\theta}_n) - f_n(\mathbf{u}) \geq C_1^{-2}C_2^{-3}\varepsilon(2-\varepsilon C_2^{-2}) \geq C_1^{-2}C_2^{-3}\varepsilon.$$

The lower bound does not depend on n, finally implying the desired claim. □

Before showing the uniform law of large numbers for our objective function, we first establish an auxiliary result on some basic properties of $o_p(1)$ and $\mathcal{O}_p(1)$ sequences of random variables.

Lemma B.10. *Assume that a sequence of random variables $Y_n > 0$ has $Y_n \geq \varepsilon + R_n$ for some $\varepsilon > 0$ and some sequence of random variables $R_n = o_p(1)$. Then,*

$$\frac{1}{Y_n} \leq \frac{1}{\varepsilon} + T_n,$$

for some sequence of random variables $T_n = o_p(1)$.

Proof of Lemma B.10. We first show that $1/Y_n = \mathcal{O}_p(1)$. To see this, we write,

$$\mathrm{P}(1/Y_n > 2/\varepsilon) = \mathrm{P}(Y_n < \varepsilon/2) \leq \mathrm{P}(\varepsilon + R_n < \varepsilon/2) = \mathrm{P}(R_n < -\varepsilon/2),$$

where the final probability goes to zero as $n \to \infty$. Hence $1/Y_n = \mathcal{O}_p(1)$, and we can write,

$$\frac{1}{Y_n} - \frac{1}{\varepsilon} = \frac{\varepsilon - Y_n}{Y_n\varepsilon} \leq \frac{-R_n}{Y_n\varepsilon} = o_p(1)\mathcal{O}_p(1) = o_p(1),$$

concluding the proof. □

Lemma B.11. *Let $\boldsymbol{x}_{n1}, \ldots, \boldsymbol{x}_{nn}$ be a random sample from the model (2) and assume that there exists $C_1, C_2 > 0$ such that, for all n,*

$$\|\boldsymbol{h}_n\| \leq C_1, \quad \|\boldsymbol{\Sigma}_n\|_2 \leq C_2, \quad \|\boldsymbol{\Sigma}_n^{-1}\|_2 \leq C_2.$$

Assume further that,

$$p_n \to \infty \qquad \text{and} \qquad p_n = o(n^{1/3}).$$

Then,

$$\sup_{\boldsymbol{u} \in \mathbb{S}^{p_n - 1}} \left| \gamma_n^2(\boldsymbol{u}) - \gamma_{n0}^2(\boldsymbol{u}) \right| \to_p 0,$$

as $n \to \infty$.

Proof of Lemma B.11. We use the notation $r_{n0}(\mathbf{u}) := \mathrm{E}\{(\mathbf{u}'\mathbf{x}_n)^3\}$, $r_n(\mathbf{u}) := (1/n)\sum_{i=1}^n (\mathbf{u}'\mathbf{x}_{ni})^3$, $s_{n0}(\mathbf{u}) := \mathrm{E}\{(\mathbf{u}'\mathbf{x}_n)^2\}$ and $s_n(\mathbf{u}) := (1/n)\sum_{i=1}^n (\mathbf{u}'\mathbf{x}_{ni})^2$.

Now, $r_{n0}(\mathbf{u})$ is uniformly upper bounded in $\mathbf{u}$ and n, as can be seen by writing $|r_{n0}(\mathbf{u})| = |\alpha_1\alpha_2(\alpha_1 - \alpha_2)(\mathbf{u}'\mathbf{h}_n)^3| \leq C_1^3$. Similarly, we have $|s_{n0}(\mathbf{u})| = \mathbf{u}'\boldsymbol{\Sigma}_n\mathbf{u} + \alpha_1\alpha_2(\mathbf{u}'\mathbf{h}_n)^2 \geq \|\boldsymbol{\Sigma}_n^{-1}\|_2^{-1} \geq C_2^{-1}$ and $|s_{n0}(\mathbf{u})| \leq \|\boldsymbol{\Sigma}_n\|_2 + \|\mathbf{h}_n\|^2 \leq C_2 + C_1^2$. And, by writing,

$$|s_{n0}(\mathbf{u})| \leq |s_{n0}(\mathbf{u}) - s_n(\mathbf{u})| + |s_n(\mathbf{u})| \leq \sup_{\mathbf{u} \in \mathbb{S}^{p_n - 1}} |s_{n0}(\mathbf{u}) - s_n(\mathbf{u})| + |s_n(\mathbf{u})|,$$

Theorem B.2 further gives that $|s_n(\mathbf{u})| \geq |s_{n0}(\mathbf{u})| + q_n \geq C_2^{-1} + q_n$, where $q_n = o_p(1)$. We also note that this implies that $|s_n(\mathbf{u})|^3 \geq (C_2^{-1} + q_n)^3 = C_2^{-3} + o_p(1)$, as taking the third power is an increasing mapping. This further gives $1/|s_n(\mathbf{u})|^3 \leq C_2^3 + o_p(1)$, almost surely, by Lemma B.10, as the second moment $s_n(\mathbf{u})$ is almost surely positive for all large enough n (in the sequel, we implicitly restrict to this almost sure set).

We now write,

$$\begin{aligned} &\sup_{\mathbf{u} \in \mathbb{S}^{p_n - 1}} \left| \gamma_n^2(\mathbf{u}) - \gamma_{n0}^2(\mathbf{u}) \right| \\ &\leq \sup_{\mathbf{u} \in \mathbb{S}^{p_n - 1}} \frac{1}{s_{n0}^3(\mathbf{u})} |r_n(\mathbf{u}) - r_{n0}(\mathbf{u})||r_n(\mathbf{u}) + r_{n0}(\mathbf{u})| \\ &+ \sup_{\mathbf{u} \in \mathbb{S}^{p_n - 1}} r_n^2(\mathbf{u}) \left| \frac{1}{s_n^3(\mathbf{u})} - \frac{1}{s_{n0}^3(\mathbf{u})} \right|, \end{aligned} \tag{B.25}$$

and treat the two supremums on the right-hand side separately. Denoting $B_n := \sup_{\mathbf{u} \in \mathbb{S}^{p_n - 1}} |r_n(\mathbf{u}) - r_{n0}(\mathbf{u})|$, the first one has the upper bound,

$$C_2^3 B_n (B_n + 2C_1^3),$$

which is of the order $o_p(1)$ by Theorem B.1. Whereas, the second supremum has

the upper bound,

$$\begin{aligned}
&\{B_n(B_n+2C_1^3)+C_1^6\}\sup_{\mathbf{u}\in\mathbb{S}^{p_n-1}}\left|\frac{s_{n0}^3(\mathbf{u})-s_n^3(\mathbf{u})}{s_n^3(\mathbf{u})s_{n0}^3(\mathbf{u})}\right| \\
\leq&\{B_n(B_n+2C_1^3)+C_1^6\}C_2^3\sup_{\mathbf{u}\in\mathbb{S}^{p_n-1}}\left|\frac{s_{n0}^3(\mathbf{u})-s_n^3(\mathbf{u})}{s_n^3(\mathbf{u})}\right| \\
\leq&\{B_n(B_n+2C_1^3)+C_1^6\}C_2^3\{C_2^3+o_p(1)\}\sup_{\mathbf{u}\in\mathbb{S}^{p_n-1}}|s_{n0}^3(\mathbf{u})-s_n^3(\mathbf{u})|.
\end{aligned} \tag{B.26}$$

Denoting,

$$E_n:=\sup_{\mathbf{u}\in\mathbb{S}^{p_n-1}}|s_n(\mathbf{u})-s_{n0}(\mathbf{u})|,$$

the left-over supremum in the final expression of (B.26) has the upper bound,

$$\begin{aligned}
&E_n\sup_{\mathbf{u}\in\mathbb{S}^{p_n-1}}|s_{n0}^2(\mathbf{u})+s_{n0}(\mathbf{u})s_n(\mathbf{u})+s_n^2(\mathbf{u})| \\
\leq&E_n\{3(C_2+C_1^2)^2+3(C_2+C_1^2)E_n+E_n^2\},
\end{aligned}$$

which converges in probability to zero by Theorem B.2. Hence, both terms on the right-hand side of (B.25) are upper bounded by sequences of random variables converging in probability to zero, and we finally obtain the desired claim. □

Proof of Theorem 5

Proof of Theorem 5. We first note that a sequence of maximizers exists almost surely for all n large enough as γ_n^2 is a rational function whose denominator is a positive power of a quantity of the form $\mathbf{u}'\{(1/n)\sum_{i=1}^n\mathbf{x}_{ni}\mathbf{x}_{ni}'\}\mathbf{u}$ where the matrix $\sum_{i=1}^n\mathbf{x}_{ni}\mathbf{x}_{ni}'$ is almost surely positive-definite when $p_n\leq n$ (again, we restrict implicitly to the corresponding almost sure set).

Fix now $\varepsilon\in(0,C_2^2)$. Then, by Lemma B.9, we have,

$$\mathrm{P}(1-|\mathbf{u}_n'\boldsymbol{\theta}_n|\geq\varepsilon)\leq\mathrm{P}\{\gamma_{n0}^2(\boldsymbol{\theta}_n)-\gamma_{n0}^2(\mathbf{u}_n)\geq\delta\},$$

for some $\delta\equiv\delta(\varepsilon,C_1,C_2,\alpha_1)>0$. Consequently, by the triangle inequality,

$$\begin{aligned}
\mathrm{P}(1-|\mathbf{u}_n'\boldsymbol{\theta}_n|\geq\varepsilon)\leq&\mathrm{P}\{|\gamma_{n0}^2(\boldsymbol{\theta}_n)-\gamma_n^2(\boldsymbol{\theta}_n)|\geq\delta/3\} \\
&+\mathrm{P}\{\gamma_n^2(\boldsymbol{\theta}_n)-\gamma_n^2(\mathbf{u}_n)\geq\delta/3\} \\
&+\mathrm{P}\{|\gamma_n^2(\mathbf{u}_n)-\gamma_{n0}^2(\mathbf{u}_n)|\geq\delta/3\}.
\end{aligned} \tag{B.27}$$

The first and the third term on the right-hand side in B.27 are by Lemma B.11 $o(1)$, while the second term equals zero as $\mathbf{u}_n$ is a maximizer of $\mathbf{u}\mapsto\gamma_n^2(\mathbf{u})$. Hence, the claim follows after noting that we always have $|\mathbf{u}_n'\boldsymbol{\theta}_n|\leq 1$ due to the unit lengths of $\mathbf{u}_n$ and $\boldsymbol{\theta}_n$. □

Appendix C: Additional simulation results

In this section, we give supporting plots as supplementary material to claims made and plots presented in the article. Simulations and the corresponding plots are done using R 3.6.1 [58] together with R packages ICtest [48], mvtnorm [21], MASS [60], GGally [55], ggpubr [30], dplyr [66], tidyr [65] and RColorBrewer [46].

Figures C.1 and C.2 show the standard deviation of maximal similarity index $s_n \mathbf{u}_n' \boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ where $\mathbf{u}_n$ is one of PP estimators discussed in the article, as a function of the Mahalanobis distance between the group means τ and mixing proportion α_1, for sample sizes $n \in \{500, 1000, 2000, 4000\}$ and $n \in \{8000, 16000, 32000\}$ respectively.

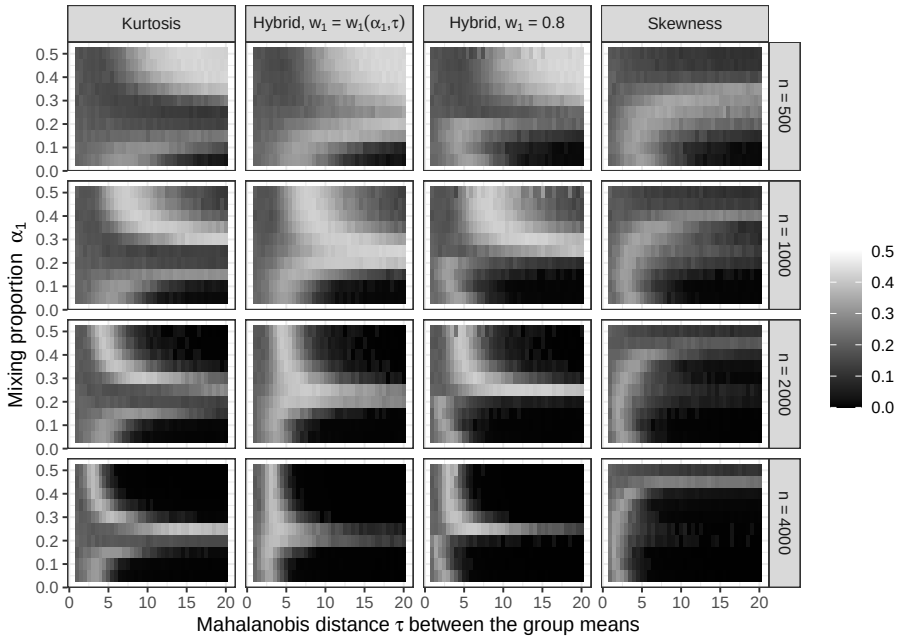

FIG C.1. *The heatmaps show standard deviation of the MSI* $s_n \boldsymbol{u}_n' \boldsymbol{\theta}/||\boldsymbol{\theta}||$ *as a function of Mahalanobis distance between the group means* τ *and mixing proportion* α_1*, where* $\boldsymbol{u}_n$ *is one of estimators of* $\boldsymbol{\theta}/||\boldsymbol{\theta}||$ *obtained by maximizing* $(\kappa_n - 3)^2$*,* γ_n^2*,* $\eta_n(\cdot; 0.8)$ *and* $\eta_n(\cdot; w_1)$*, where in the latter case,* $w_1 = w_1(\alpha_1, \tau)$ *maximizes asymptotic relative efficiency of hybrid estimator w.r.t. LDA, for sample sizes* $n \in \{500, 1000, 2000, 4000\}$*. Sign* s_n *is chosen such that* $s_n \boldsymbol{u}_n' \boldsymbol{\theta}/||\boldsymbol{\theta}|| \geq 0$*. In each setting, mean is calculated using* $m = 1000$ *replicates and the data is randomly generated from* 10*-dimensional normal mixtures with covariance matrix* $\boldsymbol{\Sigma}$ *having AR(1) structure with* $\rho = 0.6$*., while* $\boldsymbol{\mu}_1 = \boldsymbol{0}$ *and* $\boldsymbol{\mu}_2$ *chosen in each setting such that* $\boldsymbol{\mu}_2' \boldsymbol{\Sigma} \boldsymbol{\mu}_2 = \tau$*.*

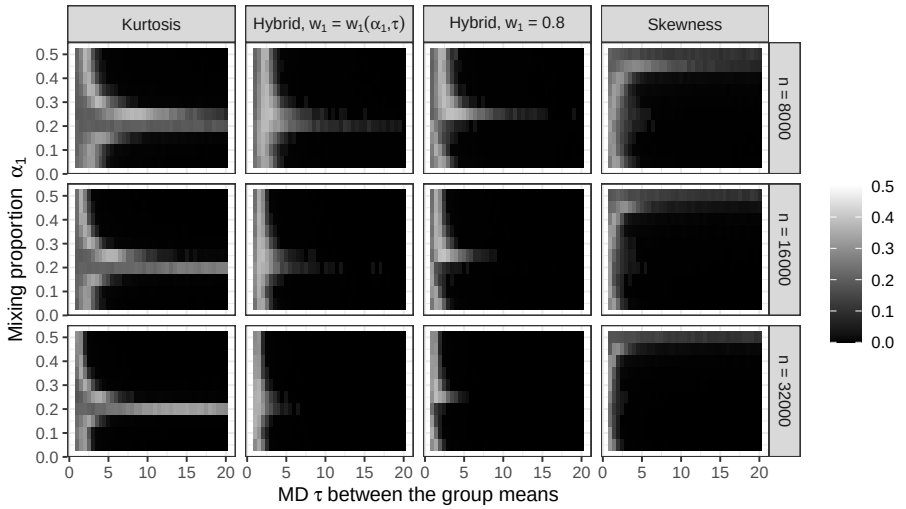

FIG C.2. *The heatmaps show standard deviation of the MSI* $s_n \boldsymbol{u}_n' \boldsymbol{\theta}/||\boldsymbol{\theta}||$ *as a function of Mahalanobis distance between the group means* τ *and mixing proportion* α_1*, where* $\boldsymbol{u}_n$ *is one of estimators of* $\boldsymbol{\theta}/||\boldsymbol{\theta}||$ *obtained by maximizing* $(\kappa_n - 3)^2$*,* γ_n^2*,* $\eta_n(\cdot; 0.8)$ *and* $\eta_n(\cdot; w_1)$*, where in the latter case,* $w_1 = w_1(\alpha_1, \tau)$ *maximizes asymptotic relative efficiency of hybrid estimator w.r.t. LDA, for sample sizes* $n \in \{8000, 16000, 32000\}$*. Sign* s_n *is chosen such that* $s_n \boldsymbol{u}_n' \boldsymbol{\theta}/||\boldsymbol{\theta}|| \geq 0$*. In each setting, mean is calculated using* $m = 1000$ *replicates and the data is randomly generated from* 10*-dimensional normal mixtures with covariance matrix* $\boldsymbol{\Sigma}$ *having AR(1) structure with* $\rho = 0.6$*, while* $\boldsymbol{\mu}_1 = \boldsymbol{0}$ *and* $\boldsymbol{\mu}_2$ *chosen in each setting such that* $\boldsymbol{\mu}_2' \boldsymbol{\Sigma} \boldsymbol{\mu}_2 = \tau$*.*

Figure C.3 shows mean of the maximal similarity index $s_n \mathbf{u}_n' \boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ where $\mathbf{u}_n$ is one of PP estimators discussed in the article, as a function of the Mahalanobis distance between the group means τ and mixing proportion α_1, for large sample sizes, $n \in \{8000, 16000, 32000\}$.

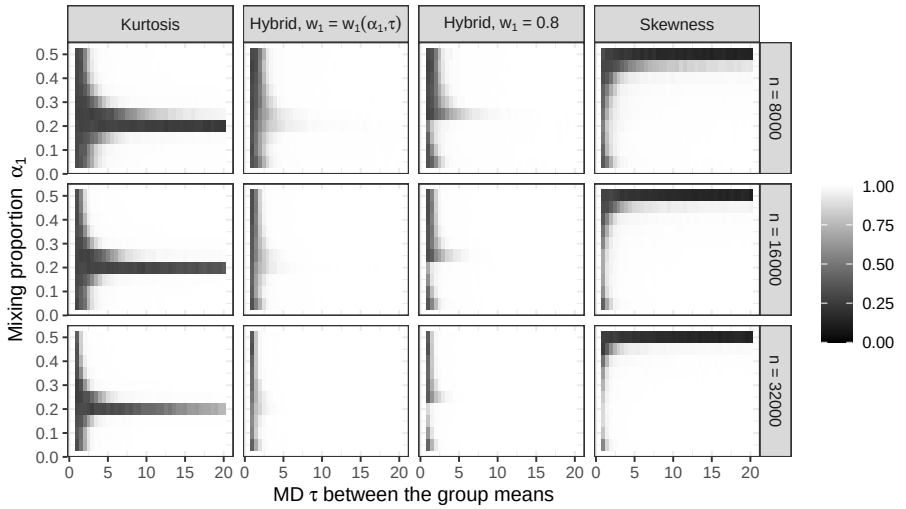

FIG C.3. *The heatmaps show means of the MSI* $s_n \boldsymbol{u}_n' \boldsymbol{\theta}/||\boldsymbol{\theta}||$ *as a function of Mahalanobis distance between the group means* τ *and mixing proportion* α_1*, where* $\boldsymbol{u}_n$ *is one of estimators of* $\boldsymbol{\theta}/||\boldsymbol{\theta}||$ *obtained by maximizing* $(\kappa_n - 3)^2$*,* γ_n^2*,* $\eta_n(\cdot; 0.8)$ *and* $\eta_n(\cdot; w_1)$*, where in the latter case,* $w_1 = w_1(\alpha_1, \tau)$ *maximizes asymptotic relative efficiency of hybrid estimator w.r.t. LDA, for sample sizes* $n \in \{500, 1000, 2000, 4000\}$*. Sign* s_n *is chosen such that* $s_n \boldsymbol{u}_n' \boldsymbol{\theta}/||\boldsymbol{\theta}|| \geq 0$*. In each setting, mean is calculated using* $m = 1000$ *replicates and the data is randomly generated from* 10*-dimensional normal mixtures with covariance matrix* $\boldsymbol{\Sigma}$ *having AR(1) structure with* $\rho = 0.6$*, while* $\boldsymbol{\mu}_1 = \boldsymbol{0}$ *and* $\boldsymbol{\mu}_2$ *chosen in each setting such that* $\boldsymbol{\mu}_2' \boldsymbol{\Sigma} \boldsymbol{\mu}_2 = \tau$*.*

Figure C.4 shows a scatter matrix plot of the *finance* data set from the R-package *Rmodmix*, where the point in the plot is being colored red if the company is being bankrupt, and blue otherwise, as well as the marginal densities for both groups which are given at the diagonal.

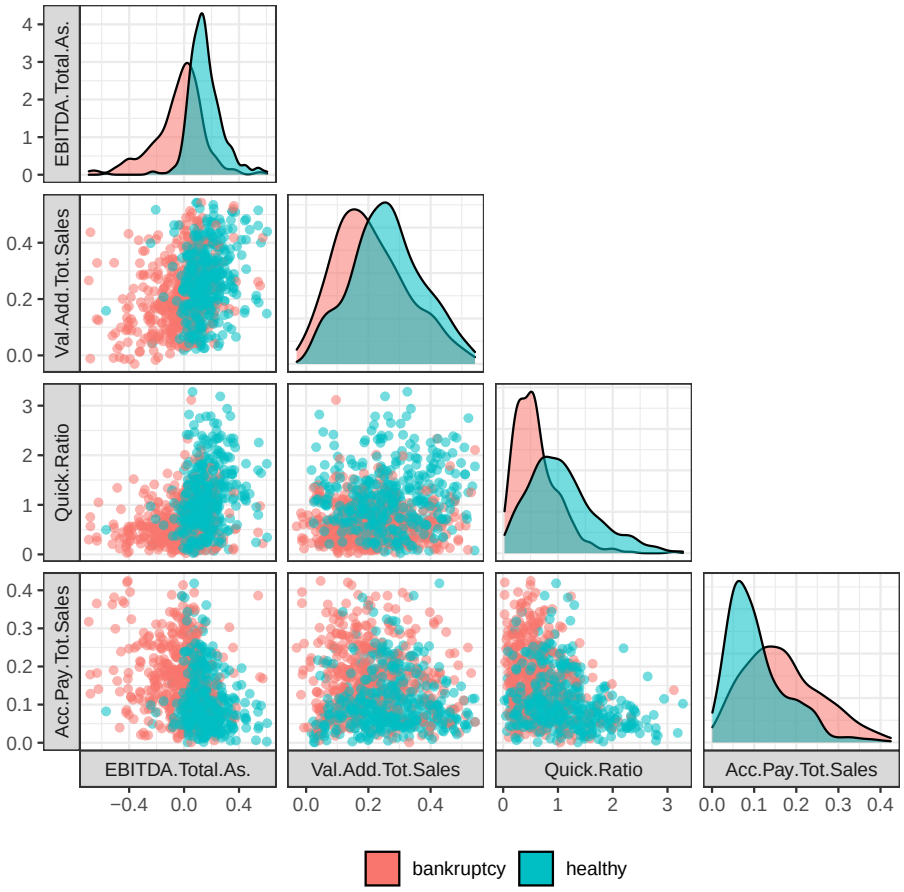

FIG C.4. *Scatter matrix plot of the finance data set, where the point is colored red if the company is being bankrupt, and blue otherwise. Marginal densities for both groups are given at the diagonal.*

Figure C.5 shows boxplots of the projection scores of the *finance* data set from the R-package *Rmodmix* along the PP directions based on mclust, PCA, LDA, kurtosis, skewness, and hybrid estimator $\eta_n(\cdot, w_1)$, for $w_1 = 0.1, 0.2, \ldots, 0.9$ for healthy and bankrupted companies.

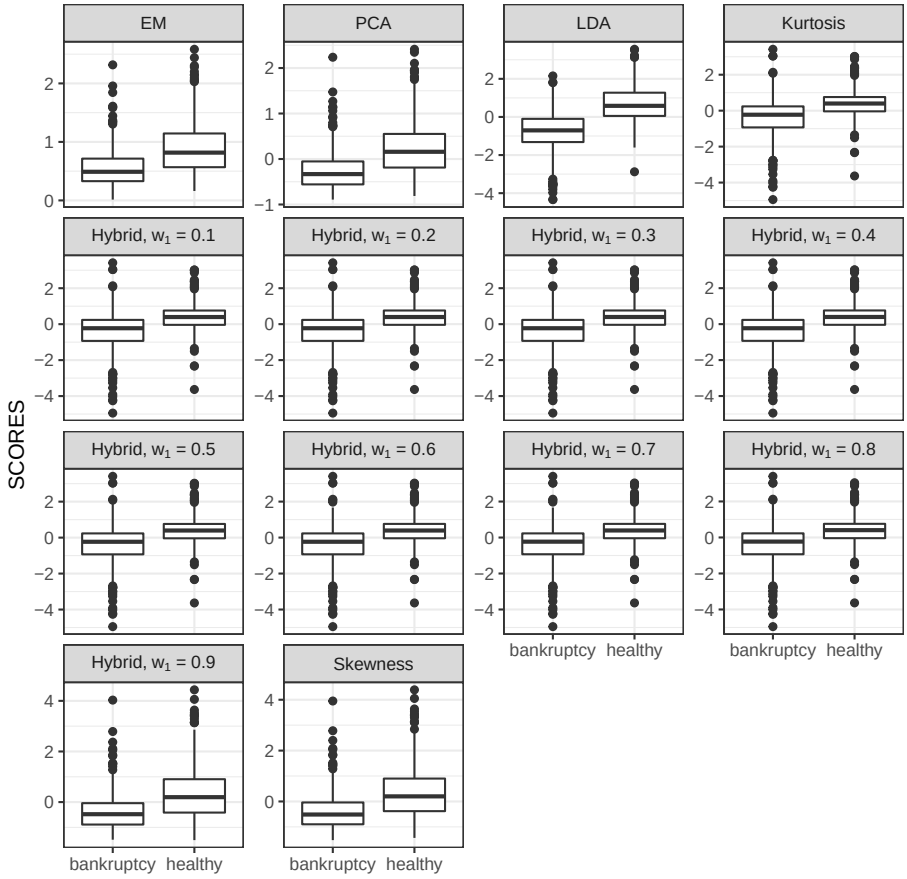

FIG C.5. *Plot shows boxplots of the projection scores of the finance data along the directions based on mclust, PCA, LDA, as well as the PP directions obtained by maximizing* $(\kappa_n - 3)^2$, γ_n^2 *and* $\eta_n(\cdot, w_1)$, *for* $w_1 = 0.1, 0.2, \ldots, 0.9$ *for healthy and bankrupted companies.*

Acknowledgements

The authors thank the Editors and two referees for helpful and insightful comments and suggestions that greatly improved the manuscript.
The work of Joni Virta was supported by the Academy of Finland (Grant 335077).

References

[1] ALASHWALI, F. and KENT, J. T. (2016). The use of a common location measure in the invariant coordinate selection and projection pursuit. *Journal of Multivariate Analysis* **152** 145–161.

[2] Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics* **34** 122–148.
[3] Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York. Second edition.
[4] Andrews, D. W. (1992). Generic uniform convergence. *Econometric Theory* 241–257.
[5] Balakrishnan, S., Wainwright, M. J. and Yu, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Annals of Statistics* **45** 77 – 120.
[6] Baringhaus, L. and Henze, N. (1991). Limit distributions for measures of multivariate skewness and kurtosis based on projections. *Journal of Multivariate Analysis* **38** 51-69.
[7] Berend, D. and Kontorovich, A. (2013). On the concentration of the missing mass. *Electronic Communications in Probability* **18** 1–7.
[8] Bickel, P. J., Kur, G. and Nadler, B. (2018). Projection pursuit in high dimensions. *Proceedings of the National Academy of Sciences* **115** 9151–9156.
[9] Bickel, P. J. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* **10** 989 – 1010.
[10] Bolton, R. J. and Krzanowski, W. J. (2003). Projection pursuit clustering for exploratory data analysis. *Journal of Computational and Graphical Statistics* **12** 121–142.
[11] Cai, T. T., Ma, J. and Zhang, L. (2019). CHIME: Clustering of high-dimensional Gaussian mixtures with EM algorithm and its optimality. *Annals of Statistics* **47** 1234-1267.
[12] Cardoso, J.-F. (1989). Source separation using higher order moments. In *Proceedings of IEEE International Conference on Accoustics, Speech and Signal Processing* 2109–2112.
[13] Cook, D., Buja, A. and Cabrera, J. (1993). Projection pursuit indexes based on orthonormal function expansions. *Journal of Computational and Graphical Statistics* **2** 225–250.
[14] Davis, A. (1977). Asymptotic theory for principal component analysis: Non-normal case. *Australian Journal of Statistics* **19** 206–212.
[15] Dermoune, A. and Wei, T. (2013). FastICA algorithm: five criteria for the optimal choice of the nonlinearity function. *IEEE Transactions on Signal Processing* **61** 2078–2087.
[16] Fischer, D., Berro, A., Nordhausen, K. and Ruiz-Gazen, A. (2019). REPPlab: An R package for detecting clusters and outliers using exploratory projection pursuit. *Communications in Statistics-Simulation and Computation* 1–23.
[17] Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97** 611–631.
[18] Friedland, S. and Wang, L. (2020). Spectral norm of a symmetric tensor and its computation. *Mathematics of Computation* **89** 2175–2215.

[19] Friedman, J., Hastie, T. and Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer, New York.
[20] Friedman, J. H. (1987). Exploratory projection pursuit. *Journal of the American Statistical Association* **82** 249–266.
[21] Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F. and Hothorn, T. (2020). mvtnorm: Multivariate Normal and t Distributions R package version 1.1-1.
[22] Hardt, M. (2015). Tight bounds for learning a mixture of two Gaussians. In *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing* 753–760.
[23] Hosseini, R. and Sra, S. (2020). An alternative to EM for Gaussian mixture models: Batch and stochastic Riemannian optimization. *Mathematical Programming* **181** 187-223.
[24] Hou, S. and Wentzell, P. D. (2014). Re-centered kurtosis as a projection pursuit index for multivariate data analysis. *Journal of Chemometrics* **28** 370–384.
[25] Huber, P. J. (1985). Projection pursuit. *Annals of Statistics* **13** 435–475.
[26] Jarque, C. M. and Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters* **6** 255–259.
[27] Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer Second edition.
[28] Jones, M. C. and Sibson, R. (1987). What is projection pursuit? *Journal of the Royal Statistical Society: Series A (General)* **150** 1–18.
[29] Kalai, A. T., Moitra, A. and Valiant, G. (2010). Efficiently learning mixtures of two Gaussians. In *Proceedings of the Forty-Second ACM Symposium on Theory of Computing* 553–562. Association for Computing Machinery, New York, NY, USA.
[30] Kassambara, A. (2020). ggpubr: 'ggplot2' based publication ready plots R package version 0.4.0.
[31] Kearns, M. and Saul, L. (1998). Large deviation methods for approximate probabilistic inference. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI)* 311–319.
[32] Kerenidis, I., Luongo, A. and Prakash, A. (2020). Quantum Expectation-Maximization for Gaussian mixture models. In *Proceedings of the 37th International Conference on Machine Learning* (H. D. III and A. Singh, eds.). *Proceedings of Machine Learning Research* **119** 5187–5197. PMLR.
[33] Kuchibhotla, A. K. and Chakrabortty, A. (2018). Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *arXiv preprint arXiv:1804.02605*.
[34] Kumar, N. S. L. P., Satoor, S. and Buck, I. (2009). Fast parallel Expectation Maximization for Gaussian mixture models on GPUs Using CUDA. In *2009 11th IEEE International Conference on High Performance Computing and Communications* 103-109.
[35] Kuriki, S. and Takemura, A. (2008). The tube method for the moment

index in projection pursuit. *Journal of Statistical Planning and Inference* **138** 2749-2762.

[36] Langrognet, F., Lebret, R., Poli, C., Iovleff, S., Auder, B. and Iovleff, S. (2020). Rmixmod: Classification with Mixture Modelling R package version 2.1.5.

[37] Loperfido, N. (2013). Skewness and the linear discriminant function. *Statistics & Probability Letters* **83** 93–99.

[38] Loperfido, N. (2015). Vector-valued skewness for model-based clustering. *Statistics & Probability Letters* **99**.

[39] Loperfido, N. (2018). Skewness-based projection pursuit: A computational approach. *Computational Statistics & Data Analysis* **120** 42–57.

[40] Loperfido, N. (2020). Kurtosis-based projection pursuit for outlier detection in financial time series. *The European Journal of Finance* **26** 142-164.

[41] Machado, S. G. (1983). Two statistics for testing for multivariate normality. *Biometrika* **70** 713-718.

[42] Malkovich, J. F. and Afifi, A. A. (1973). On tests for multivariate normality. *Journal of the American Statistical Association* **68** 176–179.

[43] Mardia, K., Kent, J. and Bibby, J. (1995). *Multivariate Analysis*. Academic Press.

[44] Miettinen, J., Taskinen, S., Nordhausen, K. and Oja, H. (2015). Fourth moments and independent component analysis. *Statistical Science* **30** 372–390.

[45] Naito, K. (1997). A generalized projection pursuit procedure and its significance level. *Hiroshima Mathematical Journal* **27** 513 – 554.

[46] Neuwirth, E. (2014). RColorBrewer: ColorBrewer Palettes R package version 1.1-2.

[47] Nielsen, S. F. (2000). The stochastic EM algorithm: Estimation and asymptotic results. *Bernoulli* **6** 457–489.

[48] Nordhausen, K., Oja, H., Tyler, D. E. and Virta, J. (2021). ICtest: Estimating and Testing the Number of Interesting Components in Linear Dimension Reduction R package version 0.3-3.

[49] Ollila, E. (2009). The deflation-based FastICA estimator: Statistical analysis revisited. *IEEE transactions on Signal Processing* **58** 1527–1541.

[50] Peña, D. and Prieto, F. J. (2001). Cluster identification using projections. *Journal of the American Statistical Association* **96** 1433–1445.

[51] Peña, D., Prieto, F. J. and Viladomat, J. (2010). Eigenvectors of a kurtosis matrix as interesting directions to reveal cluster structure. *Journal of Multivariate Analysis* **101** 1995–2007.

[52] Peña, D. and Prieto, F. J. (2001). Multivariate outlier detection and robust covariance matrix estimation. *Technometrics* **43** 286–300.

[53] Radojičić, U., Nordhausen, K. and Oja, H. (2020). Notion of information and independent component analysis. *Applied Mathematics* **65** 311–330.

[54] Rocke, D. M. and Woodruff, D. L. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association* **91** 1047–1061.

[55] Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A. and Crowley, J. (2021). GGally: Extension to 'ggplot2' R package version 2.1.0.
[56] Scrucca, L., Fop, M., Murphy, T. B. and Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal* **8** 289–317.
[57] Sun, J. (1991). Significance levels in exploratory projection pursuit. *Biometrika* **78** 759-769.
[58] R Core Team (2020). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.
[59] Tyler, D. E., Critchley, F., Dümbgen, L. and Oja, H. (2009). Invariant co-ordinate selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71** 549–592.
[60] Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*, Fourth ed. Springer, New York.
[61] Vershynin, R. (2012). *Introduction to the non-asymptotic analysis of random matrices* In *Compressed Sensing: Theory and Applications* 210–268. Cambridge University Press.
[62] Virta, J., Nordhausen, K. and Oja, H. (2016). Projection pursuit for non-Gaussian independent components. *arXiv preprint arXiv:1612.05445.*
[63] Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing* **17** 395–416.
[64] Wang, Z., Gu, Q., Ning, Y. and Liu, H. (2015). High Dimensional EM Algorithm: Statistical Optimization and Asymptotic Normality. In *Advances in Neural Information Processing Systems* (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama and R. Garnett, eds.) **28** 2521–2529. Curran Associates, Inc.
[65] Wickham, H. (2020). tidyr: Tidy Messy Data R package version 1.1.2.
[66] Wickham, H., François, R., Henry, L. and Müller, K. (2021). dplyr: A Grammar of Data Manipulation R package version 1.0.3.
[67] Xu, J. and Mareček, J. (2018). Parameter estimation in Gaussian mixture models with malicious noise, without balanced mixing coefficients. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)* 446-453.
[68] Yi, X. and Caramanis, C. (2015). Regularized EM algorithms: A unified framework and statistical guarantees. In *Advances in Neural Information Processing Systems* (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama and R. Garnett, eds.) **28** 1567–1575. Curran Associates, Inc.
[69] Zhao, Y., Shrivastava, A. K. and Tsui, K.-L. (2019). Regularized Gaussian mixture model for high-dimensional clustering. *IEEE Transactions on Cybernetics* **49** 3677-3688.
[70] Zhu, R., Wang, L., Zhai, C. and Gu, Q. (2017). High-dimensional variance-reduced stochastic gradient Expectation-Maximization algorithm. In *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.). *Proceedings of Machine Learning Research* **70** 4180–4188. PMLR.

[71] Çağlar Ari, Aksoy, S. and Arikan, O. (2012). Maximum likelihood estimation of Gaussian mixture models using stochastic search. *Pattern Recognition* **45** 2804-2816.