Journal of Multivariate Analysis 211 (2026) 105524

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Unsupervised linear discrimination using skewness ☆

Una Radojičić [a,*], Klaus Nordhausen [b,c], Joni Virta [d]

[a] Vienna University of Technology, Austria
[b] University of Helsinki, Finland
[c] University of Jyväskylä, Finland
[d] University of Turku, Finland

ARTICLE INFO

AMS 2020 subject classifications:
primary 62H12
secondary 62F12

Keywords:
Asymptotic normality
Fisher's linear discriminant
Gaussian mixture
Limiting efficiency
Third moment

ABSTRACT

It is well-known that, in Gaussian two-group separation, the optimally discriminating projection direction can be estimated without any knowledge on the group labels. In this work, we gather several such unsupervised estimators based on skewness and derive their limiting distributions. As one of our main results, we show that all affine equivariant estimators of the optimal direction have proportional asymptotic covariance matrices, making their comparison straightforward. Two of our four estimators are novel and two have been proposed already earlier. We use simulations to verify our results and to inspect the finite-sample behaviors of the estimators.

1. Introduction

Assume that our observed sample $\mathbf{x}_1, \dots, \mathbf{x}_n$ is i.i.d. from the p-variate normal location mixture,

$$\mathbf{x} \sim \alpha_1 \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + \alpha_2 \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}), \tag{1}$$

where the mixture weights $\alpha_1, \alpha_2 > 0$, $\alpha_1 + \alpha_2 = 1$, are fixed, the means are distinct, $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ is positive definite.

The objective of this work is to study the estimation of a vector $\mathbf{u}$, $\|\mathbf{u}\| = 1$, such that the univariate projection $\mathbf{u}^\top \mathbf{x}$ offers the best possible separation between the two mixture components/classes. In case we had observed also the group labels $y_1, \dots, y_n \in \{-1, 1\}$, this problem would be trivially solvable by the classical linear discriminant analysis which says that the Bayes optimal projection direction is $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$, where $\boldsymbol{\theta} := \boldsymbol{\Sigma}^{-1}\mathbf{h}$, for $\mathbf{h} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$. However, we approach this problem in an unsupervised ("blind") fashion where the class labels are not known to us, meaning that the estimation is carried out solely based on $\mathbf{x}_1, \dots, \mathbf{x}_n$ and utilizing the usual class-specific estimators of $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}$ is not possible.

Interestingly, the optimal direction $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ is still estimable even in the unsupervised context in several different ways, as described in the earlier literature: [1] showed that if $\min(\alpha_1, \alpha_2) < (3 - \sqrt{3})/6$, the projection direction attaining maximal kurtosis among all projections exactly corresponds to $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ (up to sign) while if $\min(\alpha_1, \alpha_2) > (3 - \sqrt{3})/6$ it is the projection direction attaining minimal kurtosis. [2] proved that the eigenvectors of a specific fourth-moment matrix have the same property. [3] estimated the optimal direction as a singular vector of a matrix of third standardized cumulants and [4] achieved the same using the skewness vector defined in [5]. Most recently, [6] compared several projection pursuit-based estimators from an asymptotic viewpoint, through their limiting efficiencies. In this context, limiting efficiency refers to the "ratio" between the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}/\|\hat{\boldsymbol{\theta}}\|$ and

☆ This article is part of a Special issue entitled: 'JMVA_Dimension Reduction in Multivariate Analysis' published in Journal of Multivariate Analysis.

* Corresponding author.
E-mail address: una.radojicic@tuwien.ac.at (U. Radojičić).

https://doi.org/10.1016/j.jmva.2025.105524
Received 26 June 2024; Received in revised form 16 January 2025; Accepted 7 November 2025
Available online 8 November 2025

the asymptotic covariance matrix of the supervised LDA-based estimator. In this work, we continue this line of research, by deriving the limiting efficiencies of a total of four skewness-based unsupervised estimators of $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$: a novel moment-based estimator, the estimators proposed by Loperfido in [3,4], and a novel joint diagonalization -type estimator, called 3-JADE. We note that some of these earlier works have considered more general models than (1), e.g., elliptical location mixtures [1] or location mixtures of weakly symmetric distributions with proportional covariance matrices [4]. However, while narrower, the normal mixture has the advantage of being analytically tractable enough to allow the comprehensive study of the deeper asymptotic properties of the methods. In fact, up to our best knowledge, limiting distributions of the unsupervised estimators of $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ have earlier been considered only by [6], and even this was in the context of model (1).

We next provide a summary of the main findings of this work: (i) Three of the four estimators we consider are affine equivariant, meaning that the projections given by them are essentially unaffected by the coordinate system of the original data, see Section 3 for the precise definition. This property (affine equivariance) turns out to be such strong that it almost completely determines the asymptotic behavior of an estimator, and we show that the asymptotic covariance matrices of all affine equivariant estimators of $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ are proportional to each other. This unified form makes it easy to compare two affine equivariant estimators through their corresponding constant factors. (ii) As a sort of complement to the previous point, we show that the non-affine equivariant method of moments estimator does not have an asymptotic covariance matrix of the described form. This goes to show that the requirement of affine equivariance cannot be dropped in the corresponding result. (iii) We show that the estimator proposed in [3] and the novel 3-JADE are equally efficient not only to each other, but also to a skewness-based projection pursuit estimator proposed earlier in [6]. (iv) We establish that the fourth considered estimator, proposed in [4], is strictly less efficient than the estimators mentioned in the previous point. (v) In a simulation study we confirm the limiting distribution results of the affine equivariant estimators and observe that, from the estimators discussed in this paper, the novel 3-JADE approach seems to be the best unsupervised estimator from a practical point of view. Finally, up to our best knowledge, out of the seven theorems and five lemmas included in the main text, only Lemmas 2, 3 and 5 have been included in earlier literature (in one form or another), see the corresponding parts of this manuscript for details.

The paper is organized as follows: the four considered estimators are treated individually in Sections 2, 4, 5, 6. Section 3 is devoted to studying the implications of affine equivariance to the estimation. In Section 7 we present our simulation studies and Section 8 contains discussion about the results.

As the symmetric mixture with $\alpha_1 = \alpha_2$ has skewness zero, we exclude this case and make throughout the paper the assumption that $\alpha_1 > \alpha_2$. For a non-zero vector $\mathbf{v} \in \mathbb{R}^p$, we use $\mathbf{P}_{\mathbf{v}} := \mathbf{v}\mathbf{v}^\top/\|\mathbf{v}\|^2$ and $\mathbf{Q}_{\mathbf{v}} := \mathbf{I}_p - \mathbf{P}_{\mathbf{v}}$ to denote the orthogonal projections to the subspace spanned by $\mathbf{v}$ and to its orthogonal complement, respectively. The standard basis vectors of $\mathbb{R}^p$ are denoted as $\mathbf{e}_k$, $k \in \{1, \ldots, p\}$. The following quantities are encountered often enough for them to warrant their own notation $\tau := \mathbf{h}^\top \boldsymbol{\Sigma}^{-1}\mathbf{h}$, $\beta := \alpha_1\alpha_2$, $\gamma := \alpha_1 - \alpha_2$. Finally, we note that most of the methods we consider estimate a population quantity that is only proportional to $\boldsymbol{\theta}$ and the normalization by $\|\boldsymbol{\theta}\|$ is thus done to facilitate a comparison between the different methods. From a practical point of view, the normalization only affects the scale of the projection $\mathbf{x}^\top\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ and not its direction, and is, as such, without loss of generality.

2. Method of moments estimator

We first consider a simple method of moments estimator, based on the following second and third moments of the observed mixture,

$$\mathbf{C}_2(\mathbf{x}) \equiv \mathbf{C}_2 := \mathrm{E}[\{\mathbf{x} - \mathrm{E}(\mathbf{x})\}\{\mathbf{x} - \mathrm{E}(\mathbf{x})\}^\top], \qquad \mathbf{c}_3(\mathbf{x}) \equiv \mathbf{c}_3 := \mathrm{E}[\{\mathbf{x} - \mathrm{E}(\mathbf{x})\}\{\mathbf{x} - \mathrm{E}(\mathbf{x})\}^\top\{\mathbf{x} - \mathrm{E}(\mathbf{x})\}].$$

It turns out that these two moments together contain enough information to estimate the discriminating direction $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$, as shown in the next lemma.

Lemma 1. *We have*

$$\boldsymbol{\theta} = (\mathbf{C}_2 - \beta^{1/3}\gamma^{-2/3}\|\mathbf{c}_3\|^{-4/3}\mathbf{c}_3\mathbf{c}_3^\top)^{-1}\beta^{-1/3}\gamma^{-1/3}\|\mathbf{c}_3\|^{-2/3}\mathbf{c}_3.$$

Lemma 1 states that, if one knows the mixing weights α_1, α_2, the moments $\mathbf{C}_2, \mathbf{c}_3$ can be used to construct $\boldsymbol{\theta}$. As such, a natural sample method of moments estimator of $\boldsymbol{\theta}$ is then obtained as

$$\hat{\boldsymbol{\theta}}_{\mathrm{M}} = (\hat{\mathbf{C}}_2 - \beta^{1/3}\gamma^{-2/3}\|\hat{\mathbf{c}}_3\|^{-4/3}\hat{\mathbf{c}}_3\hat{\mathbf{c}}_3^\top)^{-1}\beta^{-1/3}\gamma^{-1/3}\|\hat{\mathbf{c}}_3\|^{-2/3}\hat{\mathbf{c}}_3,$$

where the sample second and third moments are

$$\hat{\mathbf{C}}_2 := \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top, \quad \hat{\mathbf{c}}_3 := \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top(\mathbf{x}_i - \bar{\mathbf{x}}).$$

As the main result of this section, we give the limiting distribution of the normalized estimator $\hat{\boldsymbol{\theta}}_{\mathrm{M}}$.

Theorem 1. *We have, as* $n \to \infty$,

$$\sqrt{n}\left(\frac{\hat{\boldsymbol{\theta}}_{\mathrm{M}}}{\|\hat{\boldsymbol{\theta}}_{\mathrm{M}}\|} - \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|}\right) \rightsquigarrow \mathcal{N}_p\left(\mathbf{0}, \left\{\omega_1\omega_2 - \frac{\tau(1+\beta\tau)}{\|\boldsymbol{\theta}\|^2}\right\}\mathbf{Q}_{\boldsymbol{\theta}}\boldsymbol{\Sigma}^{-1}\mathbf{Q}_{\boldsymbol{\theta}} + 4\omega_1\mathbf{Q}_{\boldsymbol{\theta}}(\boldsymbol{\Sigma} + \beta\mathbf{h}\mathbf{h}^\top)\mathbf{Q}_{\boldsymbol{\theta}}\right),$$

where $\omega_1 := (1+\beta\tau)^2/(\|\mathbf{h}\|^4\beta^2(1-4\beta)\|\boldsymbol{\theta}^2\|)$ *and* $\omega_2 := 2\mathrm{tr}(\boldsymbol{\Sigma}^2) + 4\beta\mathbf{h}^\top\boldsymbol{\Sigma}\mathbf{h} + \beta(1-4\beta)\|\mathbf{h}\|^4$.

The method of moments estimator is not entirely satisfactory. The main drawback is that its use requires knowing the true mixture weights α_1, α_2, making it very impractical. Moreover, its limiting distribution in Theorem 1 is rather cumbersome and difficult to interpret. In the next section, we show that better-behaving estimators are obtained by restricting one's attention to affine equivariant functionals in a specific sense.

3. Affine equivariant estimators

Let now $g(\mathbf{x}) \in \mathbb{R}^p$ be a functional of (the distribution of) the random vector $\mathbf{x}$ that is affine equivariant (AE) in the sense that $g(\mathbf{A}^\top \mathbf{x} + \mathbf{b}) = \mathbf{A}^{-1} g(\mathbf{x})$ for all $\mathbf{b} \in \mathbb{R}^p$ and all invertible $\mathbf{A} \in \mathbb{R}^{p \times p}$. The above form of affine equivariance guarantees that the projection yielded by an AE functional is (up to location) unaffected by the coordinate system of the data,

$$g(\mathbf{A}^\top \mathbf{x} + \mathbf{b})^\top (\mathbf{A}^\top \mathbf{x}_0 + \mathbf{b}) = g(\mathbf{x})^\top \mathbf{x}_0 + g(\mathbf{x})^\top (\mathbf{A}^{-1})^\top \mathbf{b},$$

where $g(\mathbf{x})^\top \mathbf{x}_0$ is the projection in the original basis and $g(\mathbf{x})^\top (\mathbf{A}^{-1})^\top \mathbf{b}$ is a location artifact that does not depend on the projected point $\mathbf{x}_0$.

We next show that the limiting distributions of all affine equivariant estimators of $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ are identical apart from a single degree of freedom. Note that, even though we consider only skewness-based estimators in this work, this result is wider and indeed applies to all AE estimators of the optimal direction. Below, $\hat{\boldsymbol{\theta}}(\mathbf{x}_i)$ denotes a functional (statistic) of the sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Theorem 2. *Assume that (i) $\hat{\boldsymbol{\theta}}(\mathbf{x}_i)$ is affine equivariant, and (ii) for every $\mathbf{h}$ and $\boldsymbol{\Sigma}$, there exists a non-zero constant B such that $\sqrt{n}\{\hat{\boldsymbol{\theta}}(\mathbf{x}_i) - B\boldsymbol{\theta}\}$ admits a limiting normal distribution. Then, there exists $C \equiv C(\mathbf{h}, \boldsymbol{\Sigma}) > 0$ such that, as $n \to \infty$,*

$$\sqrt{n}\left\{ \frac{\hat{\boldsymbol{\theta}}(\mathbf{x}_i)}{\|\hat{\boldsymbol{\theta}}(\mathbf{x}_i)\|} - \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|} \right\} \rightsquigarrow \mathcal{N}_p\left(\mathbf{0}, C \frac{\tau}{\|\boldsymbol{\theta}\|^2} \mathbf{Q}_{\boldsymbol{\theta}} \boldsymbol{\Sigma}^{-1} \mathbf{Q}_{\boldsymbol{\theta}} \right).$$

We note that $\boldsymbol{\theta} = \boldsymbol{\Sigma}^{-1}\mathbf{h}$ in Theorem 2 is technically also a function of $\mathbf{h}$ and $\boldsymbol{\Sigma}$, but to keep the exposition more readable, we have not made this explicit. Theorem 2 essentially states that every affine equivariant estimator of $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ that admits a limiting distribution has the same limiting covariance matrix up to a constant. This result makes comparing different AE estimators of $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ considerably easier as it is sufficient to compare the corresponding factors C only. Recall that we assumed in Section 1 that the mixture weights α_1, α_2 are fixed. As such, while our notation does not explicitly show it, the constant C depends also on the mixture weights for any given estimator. Note also that the method of moments estimator in Section 2 is not affine equivariant and thus its limiting covariance matrix in Theorem 1 reveals that, if the assumption of affine equivariance is dropped, then the form postulated in Theorem 2 might no longer hold.

Recall from Section 1 that if, in addition to the data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$, we also knew the labels $y_1, \ldots, y_n \in \{-1, 1\}$ indicating the group memberships, then supervised methods could be used to estimate $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$. In such a scenario, the Bayes optimal estimator is given by the classical linear discriminant analysis (LDA) estimator, and in [6, Theorem 1] it was shown that this estimator also has a limiting covariance proportional to $\mathbf{Q}_{\boldsymbol{\theta}} \boldsymbol{\Sigma}^{-1} \mathbf{Q}_{\boldsymbol{\theta}}$, with the coefficient C equal to $C = (1 + \beta\tau)/(\beta\tau)$. As LDA is indeed supervised, this value thus serves as a lower limit for the constant C in the current unsupervised estimation scenario, since one cannot really expect to surpass the performance of LDA in the absence of label information.

As our second result of this section, we derive a "shortcut" for finding the constant C for affine equivariant estimators of a particular form. Namely, we assume for the remainder of this section that

$$\hat{\boldsymbol{\theta}}(\mathbf{x}_i) = \hat{\mathbf{C}}_2^{-1/2}(\mathbf{x}_i) \hat{\mathbf{u}}(\hat{\mathbf{C}}_2^{-1/2}(\mathbf{x}_i)(\mathbf{x}_i - \bar{\mathbf{x}})), \tag{2}$$

where $\hat{\mathbf{u}}$ is a unit-length estimator/functional that transforms as $\hat{\mathbf{u}}(\mathbf{O}\mathbf{x}_i) = \mathbf{O}\hat{\mathbf{u}}(\mathbf{x}_i)$ for any orthogonal $p \times p$ matrix $\mathbf{O}$. Theorem 2.1 in [7] can be used to show that any such estimator $\hat{\boldsymbol{\theta}}$ is indeed affine equivariant, see also the proof of Lemma 2.

Theorem 3. *In addition to the assumptions of Theorem 2, assume that (iii) $\hat{\boldsymbol{\theta}}(\mathbf{x}_i)$ is of the form (2), and (iv) for every $\mathbf{m}$, we have*

$$\hat{\mathbf{r}} := \sqrt{n}\left\{ \hat{\mathbf{u}}(\hat{\mathbf{C}}_2^{-1/2}(\mathbf{z}_i)(\mathbf{z}_i - \bar{\mathbf{z}})) - \frac{\mathbf{m}}{\|\mathbf{m}\|} \right\} = \mathcal{O}_P(1),$$

where $\mathbf{z}_i \sim \alpha_1 \mathcal{N}_p(-\alpha_2 \mathbf{m}, \mathbf{I}_p) + \alpha_2 \mathcal{N}_p(\alpha_1 \mathbf{m}, \mathbf{I}_p)$. Then, as $n \to \infty$, we have

$$-\frac{1}{1 + \sqrt{1 + \beta\tau}} \left(\frac{\mathbf{m}}{\|\mathbf{m}\|} \otimes \mathbf{t} \right)^\top \sqrt{n}\mathrm{vec}(\hat{\mathbf{C}}_2(\mathbf{z}_i) - \mathbf{C}_2(\mathbf{z})) + \sqrt{1 + \beta\tau} \cdot \mathbf{t}^\top \hat{\mathbf{r}} \rightsquigarrow \mathcal{N}(0, C),$$

where $\mathbf{t} \in \mathbb{R}^p$ is any unit-length vector satisfying $\mathbf{m}^\top \mathbf{t} = 0$.

Theorem 3 states that if one has for $\hat{\mathbf{r}}$ a linearization of the form $\hat{\mathbf{r}} = (1\sqrt{n}) \sum_{i=1}^n [g(\mathbf{z}_i) - \mathrm{E}\{g(\mathbf{z})\}] + o_P(1)$ for some $g : \mathbb{R}^p \to \mathbb{R}^p$, then the univariate central limit theorem can be used to find C. We use this result (or its suitable variant) to find the constants C for all the methods considered in the subsequent sections.

We conclude the section by examining more closely the transformation $\mathbf{x} \mapsto \mathbf{C}_2^{-1/2}(\mathbf{x})\{\mathbf{x} - \mathrm{E}(\mathbf{x})\}$ that was used also in (2). This mapping, which is known as standardization or "whitening" is typically used as preprocessing in many multivariate methods [8]. One of its benefits is that if some orthogonally equivariant methodology is applied to whitened data, then the full procedure is affine equivariant, see the proof of Lemma 2 for an example of this. From a heuristic viewpoint, the role of the standardization is

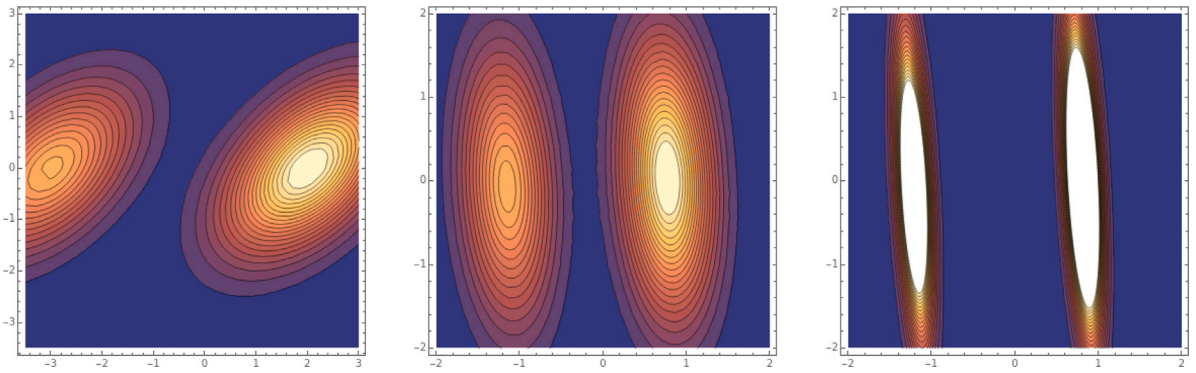

Fig. 1. The three panels contain, from left to right: (a) The contour plot of a normal mixture $\mathbf{x}$ in (1) for some specific choices of $\alpha_1, \alpha_2, \mathbf{h}, \boldsymbol{\Sigma}$. (b) The contour plot of the standardized $\mathbf{C}_2^{-1/2}(\mathbf{x})\{\mathbf{x} - \mathrm{E}(\mathbf{x})\}$, which demonstrates the result of Theorem 4 that the direction $\mathbf{w}$ joining the two groups centers is indeed the one with the least variation. (c) The same as in the previous panel but for a mixture with larger value of τ, yielding an even smaller variation in the direction $\mathbf{w}$.

to remove from the data any effects that are artifacts of the used coordinate system, to allow better focusing on the deeper features of the data. Our next result demonstrated this fact in the context of the normal mixture (1). The result is rather simple but, as far as we are aware, previously unknown, most likely due to the non-identifiability of the result up to orthogonal transformations.

Theorem 4. *We have* $\mathbf{C}_2^{-1/2}(\mathbf{x})\{\mathbf{x} - \mathrm{E}(\mathbf{x})\} = \mathbf{O}\mathbf{z}$ *for some orthogonal* $p \times p$ *matrix* $\mathbf{O}$ *and*

$$\mathbf{z} \sim \alpha_1 \mathcal{N}_p\left(-\alpha_2\sqrt{\frac{\tau}{1+\beta\tau}}\mathbf{w}, \mathbf{I}_p - \frac{\beta\tau}{1+\beta\tau}\mathbf{w}\mathbf{w}^\top\right) + \alpha_2 \mathcal{N}_p\left(\alpha_1\sqrt{\frac{\tau}{1+\beta\tau}}\mathbf{w}, \mathbf{I}_p - \frac{\beta\tau}{1+\beta\tau}\mathbf{w}\mathbf{w}^\top\right),$$

where $\mathbf{w} = \boldsymbol{\Sigma}^{-1/2}\mathbf{h}/\|\boldsymbol{\Sigma}^{-1/2}\mathbf{h}\|$.

Theorem 4 shows that, after the standardization, the normal mixture is such that (a) the two groups means are separated along the direction $\mathbf{Ow}$, and (b) the variation of the data is one in every direction orthogonal to $\mathbf{Ow}$ and strictly smaller in the direction $\mathbf{Ow}$. This effect has been demonstrated in Fig. 1. Note that the whitening does not change the standardized distance between the group means. That is, $\{\tau/(1+\beta\tau)\}\mathbf{w}^\top[\mathbf{I}_p - \{\beta\tau/(1+\beta\tau)\}\mathbf{w}\mathbf{w}^\top]^{-1}\mathbf{w} = \mathbf{h}^\top\boldsymbol{\Sigma}^{-1}\mathbf{h}$. However, what the whitening does is to essentially position the data in an optimal coordinate system for the detection of the linear discriminant direction $\mathbf{Ow}$.

4. Affine equivariant method of moments

Motivated by the previous section, we next improve the method of moments estimator by obtaining an affine equivariant version of it through the whitened observation described in the previous section. That is, we first define $\mathbf{x}_w = \mathbf{C}_2^{-1/2}(\mathbf{x})\{\mathbf{x} - \mathrm{E}(\mathbf{x})\}$, and then take $\boldsymbol{\theta}_R(\mathbf{x}) := \mathbf{C}_2^{-1/2}(\mathbf{x})\mathbf{c}_3(\mathbf{x}_w)$. The resulting estimator is then the same one that was already studied in [4]. Additionally, [9] investigated the closely related quantity $\mathbf{c}_3(\mathbf{x}_w)$, known as the canonical skewness vector, in a model-free context.

It is not immediately obvious that this results in an affine equivariant estimator, so, for completeness, we provide a proof. See also [9, Theorem 3] for an equivalent result for the canonical skewness vector.

Lemma 2. *The functional* $\boldsymbol{\theta}_R(\mathbf{x})$ *is affine equivariant.*

Having established the affine equivariance, we next show that $\boldsymbol{\theta}_R(\mathbf{x})$ indeed estimates $\boldsymbol{\theta}$, up to scale, obtaining a quantitative version of Theorem 1 in [4]. Lemma 2 simplifies this task by essentially allowing us to consider only the case $\boldsymbol{\Sigma} = \mathbf{I}_p$.

Lemma 3. *We have*

$$\boldsymbol{\theta}_{\mathrm{R}}(\mathbf{x}) = \frac{\beta\gamma\tau}{(1+\beta\tau)^2}\boldsymbol{\theta}.$$

The corresponding sample estimator is obtained by simply replacing the population moments with their empirical counterparts and we denote it by $\hat{\boldsymbol{\theta}}_{\mathrm{R}}$. Its affine equivariance follows similarly as for $\boldsymbol{\theta}_R(\mathbf{x})$ in Lemma 2. Loperfido did not consider the asymptotic properties of the estimator in [4] and, to complement his work, we next derive the limiting distribution of $\hat{\boldsymbol{\theta}}_R$. In presenting this and the remaining limiting covariance matrices, we use the notation that

$$C_0 := (1+\beta\tau)\frac{\beta\tau^2 + 6\beta\tau + 2}{\beta^2(1-4\beta)\tau^3}. \tag{3}$$

The significance behind the quantity C_0 is that it is the asymptotic C-constant of the estimators discussed later in Sections 5 and 6 (see Theorems 6 and 7). As such, Theorem 5 below reveals that the estimator $\hat{\boldsymbol{\theta}}_{\mathrm{R}}$ is strictly less efficient than either of these. We

note that, interestingly, the C-constant corresponding to the skewness-based projection pursuit studied in [6, Theorem 3] is also equal to C_0.

Theorem 5. *We have, as $n \to \infty$,*

$$\sqrt{n}\left(\frac{\hat{\theta}_{\mathrm{R}}}{\|\hat{\theta}_{\mathrm{R}}\|} - \frac{\theta}{\|\theta\|}\right) \rightsquigarrow \mathcal{N}_p\left(\mathbf{0}, C\frac{\tau}{\|\theta\|^2}\mathbf{Q}_\theta \boldsymbol{\Sigma}^{-1}\mathbf{Q}_\theta\right),$$

where

$$C = C_0 + \frac{2(p+1)(1+\beta\tau)^4}{\beta^2(1-4\beta)\tau^3}.$$

5. Third order blind identification

As our second affine equivariant estimator, we consider the third-moment based estimator proposed originally by Loperfido in [3]. Denoting again the standardized observation by $\mathbf{x}_w = \mathbf{C}_2^{-1/2}(\mathbf{x})\{\mathbf{x} - \mathrm{E}(\mathbf{x})\}$, Loperfido defines the estimator $\theta_{\mathrm{L}}(\mathbf{x}) = \mathbf{C}_2^{-1/2}(\mathbf{x})\mathbf{u}(\mathbf{x})$, where $\mathbf{u}(\mathbf{x})$ is the leading unit-length eigenvector of the matrix

$$\mathbf{T}(\mathbf{x}_w) := [\mathrm{E}\{(\mathbf{x}_w \otimes \mathbf{x}_w)\mathbf{x}_w^\top\}]^\top [\mathrm{E}\{(\mathbf{x}_w \otimes \mathbf{x}_w)\mathbf{x}_w^\top\}].$$

The vector $\theta_{\mathrm{L}}(\mathbf{x})$ was further studied under skew-normal scale mixtures in [10] and under a model-free context in [11]. The latter connected $\theta_{\mathrm{L}}(\mathbf{x})$ to several classical measures of multivariate skewness, with the particular consequence that the vector $\theta_{\mathrm{L}}(\mathbf{x})$ does not, in general, coincide with the direction yielding maximal univariate skewness. However, this equivalence holds in some special, structured cases, such as in the current normal location mixture [3,6], independent component models [11], and skew-normal scale mixtures [10].

The original work [3] did not consider the limiting distribution of the estimator $\theta_{\mathrm{L}}(\mathbf{x})$, and the purpose of this section is to derive this result under the normal mixture. But first, we will show an alternative form for the matrix $\mathbf{B}(\mathbf{x}_w)$ that connects the estimator to fourth-order blind identification (FOBI) [12], a seminal method of independent component analysis.

Lemma 4. *Defining $\mathbf{T}_k(\mathbf{x}) := \mathrm{E}(\mathbf{x}\mathbf{x}^\top \mathbf{e}_k \mathbf{x}^\top)$, we have*

$$\mathbf{T}(\mathbf{x}_w) = \sum_{k=1}^{p} \mathbf{T}_k(\mathbf{x}_w)^2.$$

The matrices $\mathbf{T}_k(\mathbf{x}_w)$ can be seen as the third-order counterparts of the matrices $\mathbf{B}^{ij}$ [13, page 379], which are summed over i, j to obtain the matrix used in FOBI. As such, since Loperfido did not name his estimator, we propose calling $\theta_{\mathrm{L}}(\mathbf{x})$, due to this connection, as the third-order blind identification (TOBI) estimator. See also [14] for a similar skewness-based estimator that achieves affine equivariance by first applying FOBI to the data.

The next result shows the Fisher consistency of $\theta_{\mathrm{L}}(\mathbf{x})$, serving as a quantitative, population counterpart to [3, Proposition 3].

Lemma 5. *We have,*

$$\theta_L(\mathbf{x}) = s\frac{1}{\{\tau(1+\beta\tau)\}^{1/2}}\theta,$$

for some sign $s \in \{-1, 1\}$.

The arbitrariness of the sign in Lemma 5 is caused by the fact that eigenvectors are unique (at most) only up to sign. This is taken into account later in our limiting results by considering a sign-corrected version of the sample estimator.

Denoting $\hat{\mathbf{C}}_2 \equiv \hat{\mathbf{C}}_2(\mathbf{x}_i)$, the sample TOBI-estimator is then defined as $\hat{\theta}_{\mathrm{L}}(\mathbf{x}_i) := \hat{\mathbf{C}}_2^{-1/2}\hat{\mathbf{u}}$, where $\hat{\mathbf{u}}$ is any leading eigenvector of the matrix

$$\hat{\mathbf{T}}(\hat{\mathbf{C}}_2^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}})) := \sum_{k=1}^{p} \hat{\mathbf{T}}_k(\hat{\mathbf{C}}_2^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}}))^2$$

where $\hat{\mathbf{T}}_k(\mathbf{x}_i) := (1/n)\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top \mathbf{e}_k \mathbf{x}_i^\top$. That the TOBI-estimator is affine equivariant is proven in the next result.

Lemma 6. *For any invertible $\mathbf{A} \in \mathbb{R}^{p\times p}$ and any $\mathbf{b} \in \mathbb{R}^p$, we have, almost surely $\hat{\theta}_{\mathrm{L}}(\mathbf{A}^\top\mathbf{x}_i + \mathbf{b}) = s\mathbf{A}^{-1}\hat{\theta}_{\mathrm{L}}(\mathbf{x}_i)$, for some sign $s \in \{-1, 1\}$.*

Note that the affine equivariance in Lemma 6 holds only almost surely as it requires the leading eigenvalue of the matrix $\hat{\mathbf{T}}(\hat{\mathbf{C}}_2^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}}))$ to be simple. Proceeding as in the proof of Lemma 6, it is straightforwardly checked that also the population-level TOBI-estimator enjoys the analogous form of affine equivariance (without the "almost surely"-part).

The main result of this section, the limiting distribution of TOBI, is given next.

Theorem 6. *We have, as $n \to \infty$,*

$$\sqrt{n}\left(\frac{\hat{\theta}_{\mathrm{L}}}{\|\hat{\theta}_{\mathrm{L}}\|} - \frac{\theta}{\|\theta\|}\right) \rightsquigarrow \mathcal{N}_p\left(\mathbf{0}, C_0\frac{\tau}{\|\theta\|^2}\mathbf{Q}_\theta \boldsymbol{\Sigma}^{-1}\mathbf{Q}_\theta\right),$$

where C_0 is defined in (3).

6. Novel 3-JADE estimator

To conclude our collection of estimators, we next propose a novel alternative to θ_M, θ_R and θ_L. Using the matrices $\mathbf{T}_k$, we define this estimator as $\theta_{\mathrm{J}} := \mathbf{C}_2^{-1/2}(\mathbf{x})\mathbf{u}$ where

$$\mathbf{u} = \underset{\mathbf{v}\in\mathbb{R}^p, \|\mathbf{v}\|=1}{\arg\max} \sum_{k=1}^{p} \left\{\mathbf{v}^\top \mathbf{T}_k(\mathbf{x}_w)\mathbf{v}\right\}^2. \tag{4}$$

The proof of Lemma 7 later reveals that the maximizer in (4) is indeed unique, up to a sign change. The underlying idea behind θ_{J} is that, essentially, it is to TOBI what the classical method known as joint approximate diagonalization of eigenmatrices (JADE) [15] is to FOBI. Whereas FOBI finds the eigendecomposition of a matrix formed by summing fourth cumulant matrices, JADE jointly diagonalizes these cumulant matrices, see [13]. The relationship between θ_{J} and TOBI is the same, with the exception that we search in (4) for a single projection only (JADE finds p simultaneously). As such, we call the estimator θ_{J} the (one-step) 3-JADE in the following. The following lemma shows the Fisher consistency of 3-JADE, i.e., that it is capable of estimating the linear discriminant direction.

Lemma 7. *We have $\theta_{\mathrm{J}} = s\{\tau(1+\beta\tau)\}^{-1/2}\theta$ for some sign $s \in \{-1, 1\}$.*

The sample estimator $\hat{\theta}_{\mathrm{J}}$ is defined analogously, using the sample moments instead of population ones. The next result then shows that the estimator is affine equivariant, and is given without proof as it can be derived using the techniques presented in the Proof of Lemma 6.

Lemma 8. *For any invertible $\mathbf{A} \in \mathbb{R}^{p\times p}$ and any $\mathbf{b} \in \mathbb{R}^p$, we have, $\hat{\theta}_{\mathrm{J}}(\mathbf{A}^\top\mathbf{x}_i + \mathbf{b}) = \mathbf{A}^{-1}\hat{\theta}_{\mathrm{J}}(\mathbf{x}_i)$, where*

$$\hat{\theta}_{\mathrm{J}}(\mathbf{x}_i) \in \underset{\mathbf{v}\in\mathbb{R}^p, \|\mathbf{v}\|=1}{\arg\max} \sum_{k=1}^{p} \left(\mathbf{v}^\top \hat{\mathbf{T}}_k \mathbf{v}\right)^2$$

and $\hat{\mathbf{T}}_k$ is the sample version of $\mathbf{T}_k(\mathbf{x}_w)$, $k \in \{1, \ldots, p\}$.

JADE-3 being a novel method, we next present an optimization algorithm for solving the optimization problem (4). Then, after obtaining the estimate $\hat{\mathbf{u}}$ as described below, the final 3-JADE estimate is found as $\hat{\theta}_{\mathrm{J}} = \hat{\mathbf{C}}_2^{-1/2}\hat{\mathbf{u}}$. Denoting the sample versions of the matrices $\mathbf{T}_k(\mathbf{x}_w)$ as $\hat{\mathbf{T}}_k$, the Lagrangian of the objective function in (4) and its gradient are

$$\ell_n(\mathbf{u}) = \sum_{k=1}^{p} (\mathbf{u}^\top \hat{\mathbf{T}}_k \mathbf{u})^2 - \lambda(\mathbf{u}^\top\mathbf{u} - 1), \qquad \nabla\ell_n(\mathbf{u}) = 4\sum_{k=1}^{p} (\mathbf{u}^\top \hat{\mathbf{T}}_k \mathbf{u})\hat{\mathbf{T}}_k\mathbf{u} - 2\lambda\mathbf{u},$$

respectively. Letting $\mathbf{u}_n$ denote a unit-length null point of the gradient, and multiplying by $\mathbf{u}_n^\top$ from the left shows that $\lambda = 2\sum_{k=1}^{p}(\mathbf{u}_n^\top\hat{\mathbf{T}}_k\mathbf{u}_n)^2$. Hence, any solution $\mathbf{u}_n$ needs to satisfy

$$\sum_{k=1}^{p} (\mathbf{u}_n^\top \hat{\mathbf{T}}_k \mathbf{u}_n)\hat{\mathbf{T}}_k\mathbf{u}_n = \sum_{k=1}^{p} (\mathbf{u}_n^\top \hat{\mathbf{T}}_k \mathbf{u}_n)^2\mathbf{u}_n,$$

implying that

$$\mathbf{u}_n \propto \sum_{k=1}^{p} (\mathbf{u}_n^\top \hat{\mathbf{T}}_k \mathbf{u}_n)\hat{\mathbf{T}}_k\mathbf{u}_n. \tag{5}$$

Motivated by the fixed point Eq. (5), we propose next the Algorithm 1 for estimating $\hat{\mathbf{u}}$.

Algorithm 1 3-JADE

1: Initialize $\mathbf{u}_n$;
2: **while** not converged **do**
3: $\mathbf{u}_n \leftarrow \sum_{k=1}^{p} (\mathbf{u}_n{}^\top \hat{\mathbf{T}}_k \mathbf{u}_n)\hat{\mathbf{T}}_k\mathbf{u}_n$;
4: $\mathbf{u}_n \leftarrow \frac{\mathbf{u}_n}{\|\mathbf{u}_n\|}$;
5: **end while**

Finally, we conclude the section with the limiting normality of 3-JADE, showing that it, like TOBI, has a limiting efficiency exactly equal to the skewness-based projection pursuit studied in [6, Theorem 3].

Theorem 7. *We have, as $n \to \infty$,*

$$\sqrt{n}\left(\frac{\hat{\theta}_{\mathrm{J}}}{\|\hat{\theta}_{\mathrm{J}}\|} - \frac{\theta}{\|\theta\|}\right) \rightsquigarrow \mathcal{N}_p\left(\mathbf{0}, C_0\frac{\tau}{\|\theta\|^2}\mathbf{Q}_\theta\boldsymbol{\Sigma}^{-1}\mathbf{Q}_\theta\right),$$

where C_0 is defined in (3).

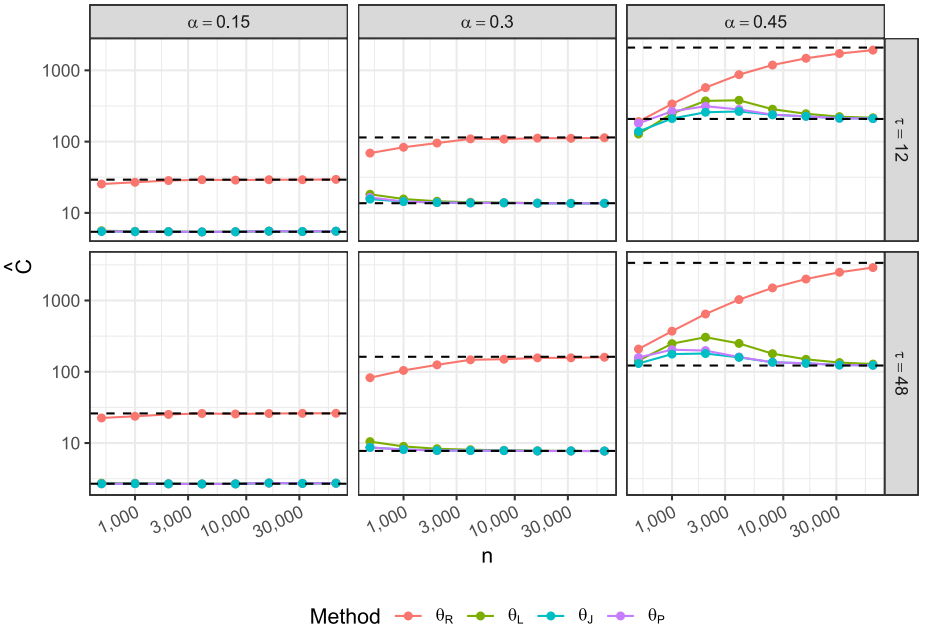

Fig. 2. Convergence of $\hat{C}$ to C (dashed black lines) for the estimators $\boldsymbol{\theta}_R$, $\boldsymbol{\theta}_L$, $\boldsymbol{\theta}_J$ and $\boldsymbol{\theta}_P$ based on $M = 10000$ data sets of size n. Note that both axes have a log scale.

As a summary of our results concerning the limiting distributions of the affine equivariant methods (Theorems 5, 6 and 7), the superior methods are TOBI and 3-JADE, both having exactly the same limiting covariance matrix. As routines for the computation of eigendecompositions are efficient and widely available, from a purely asymptotic viewpoint using TOBI is preferable over the other methods. However, since the finite-sample properties of the methods can still differ, we next compare their behaviors under simulated data.

7. Simulations

The simulations involve a total of six estimators, $\boldsymbol{\theta}_{LDA}, \boldsymbol{\theta}_M, \boldsymbol{\theta}_R, \boldsymbol{\theta}_L, \boldsymbol{\theta}_J, \boldsymbol{\theta}_P$, corresponding to linear discriminant analysis, the estimators discussed in Sections 2, 4, 5, 6 and the skewness-based projection pursuit estimator studied in [6], respectively. The first goal of the simulation study is to verify the constants C for the affine equivariant estimators $\boldsymbol{\theta}_R, \boldsymbol{\theta}_L, \boldsymbol{\theta}_J, \boldsymbol{\theta}_P$. Of these, the final three all share the same constant $C = C_0$ defined in (3). Due to the methods' affine equivariance we consider in the first simulation only the case with $\boldsymbol{\Sigma} = \mathbf{I}_p$ and $\mathrm{E}(\mathbf{x}) = \mathbf{0}$ in model (1).

Let $\mathbf{X}_n^m$, $m \in \{1, \ldots, M\}$ correspond then to M realized data matrices consisting of i.i.d. samples of size n from this distribution for a given vector $\mathbf{h}$. Then it can be shown that for any affine equivariant estimator $\hat{\boldsymbol{\theta}}(\mathbf{X}_n^m)$ and for any unit-length vector $\mathbf{t} \in \mathbb{R}^p$ such that $\mathbf{t}^\top \mathbf{h} = 0$, we have, as $n \to \infty$,

$$\hat{C} = n\mathrm{Var}\left[\frac{\mathbf{t}^\top \hat{\boldsymbol{\theta}}(\mathbf{X}_n^m)}{\|\hat{\boldsymbol{\theta}}(\mathbf{X}_n^m)\|}\right] \to C.$$

Using this result, we compute for a wide range of sample sizes for $\alpha = \alpha_1 \in \{0.15, 0.30, 0.45\}$ and $\mathbf{h}$ such that $\tau \in \{12, 48\}$ with $p = 3$ for $\boldsymbol{\theta}_R$, $\boldsymbol{\theta}_L$, $\boldsymbol{\theta}_J$ and $\boldsymbol{\theta}_P$ the corresponding $\hat{C}$ with $M = 10000$. The results are shown in Fig. 2 and illustrate clearly the correctness of the derived constants. The convergence rate to the true value seems however slower when the distribution is more "symmetric", which is as expected, since in the perfectly symmetric case $\alpha = 0.50$, the optimal direction is no longer estimable using skewness. In cases under consideration, $\boldsymbol{\theta}_R$ is the least preferred estimator while for the asymptotically equivalent estimators, it seems that $\boldsymbol{\theta}_L$ exhibits more variation for small sample sizes than expected compared to $\boldsymbol{\theta}_J$ and $\boldsymbol{\theta}_P$ which behave quite similarly. The results also demonstrate that a larger τ does not make the task for all methods easier.

While this simulation shows that the methods have the expected variation in the limit, we are also interested in how well they actually estimate $\boldsymbol{\theta}$ for finite samples. For that purpose, we follow [6] and use as a performance measure the maximal similarity index (MSI) which corresponds to the inner product between the normalized true and estimated directions. MSI takes values in $[0, 1]$ where 1 indicates that the two vectors point in the same direction and the estimation works perfectly.

In this second simulation, besides LDA, we included also the non-affine equivariant $\boldsymbol{\theta}_M$ and for a fair comparison therefore chose for each data set a random matrix $\boldsymbol{\Sigma}$ which is of the form $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^\top$ where $\mathbf{A}$ is a $p \times p$ matrix where all elements are independently drawn from $\mathcal{N}(0, 1)$. For all combinations of $p = \{3, 10\}$, $n \in \{500, 1000, 2000, 4000\}$, $\alpha = \alpha_1 \in \{0.05, 0.07, \ldots, 0.49\}$ and $\mathbf{h}$ such that $\tau \in \{1, \ldots, 20\}$ we sampled $M = 1000$ data sets and Figs. 3 and 4 provide the obtained average performances.

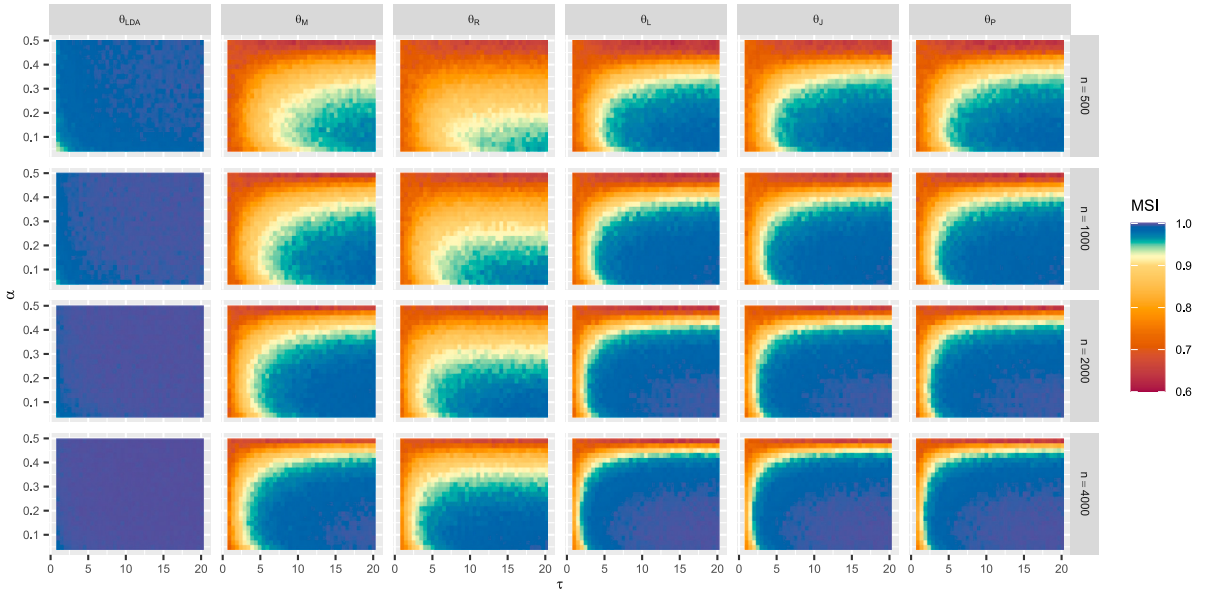

Fig. 3. Average MSI values for the different estimators for a range of α's and τ's and different sample sizes based on $M = 1000$ when $p = 3$.

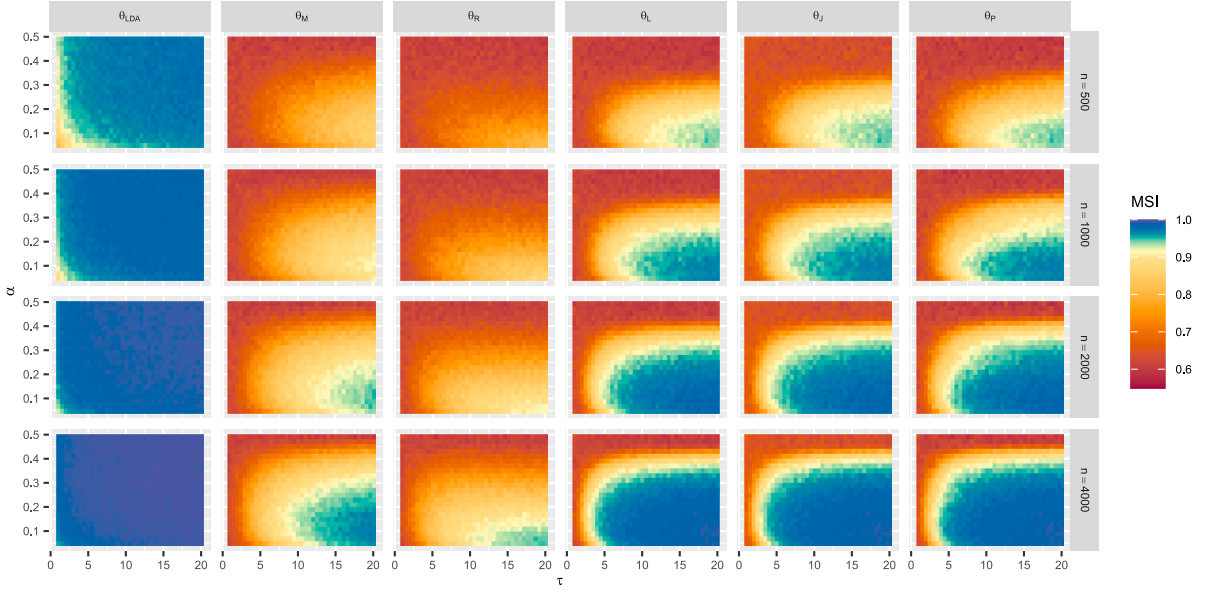

Fig. 4. Average MSI values for the different estimators for a range of α's and τ's and different sample sizes based on $M = 1000$ when $p = 10$.

The figures demonstrate that LDA is much better than all unsupervised methods. The moment estimator $\boldsymbol{\theta}_M$ which requires the knowledge of α seems also to make use of this additional knowledge and clearly outperforms $\boldsymbol{\theta}_R$. Maybe a surprise is then that this extra knowledge seems insufficient to outperform $\boldsymbol{\theta}_L$, $\boldsymbol{\theta}_J$, and $\boldsymbol{\theta}_P$, which all perform very similarly. When comparing $\boldsymbol{\theta}_L$, $\boldsymbol{\theta}_J$, and $\boldsymbol{\theta}_P$ an important feature to point out is that all methods except $\boldsymbol{\theta}_J$ and $\boldsymbol{\theta}_P$ always produced estimators. $\boldsymbol{\theta}_J$, using $\boldsymbol{\theta}_L$ as an initial value, had 0.11% and 1.65% convergence failures when $p = 3$ and $p = 10$ respectively while $\boldsymbol{\theta}_P$, as implemented in the R package ICtest [16], had 4.36% and 7.52% convergence failures respectively. For more details of the convergence problems of $\boldsymbol{\theta}_P$ see also Fig S1 in the Supplement. The code for reproducing the results from the simulation study is available at https://github.com/uradojic/Unsupervised-linear-discrimination-using-skewness.

Thus, to conclude the simulations, it seems that the novel estimator $\boldsymbol{\theta}_J$ is the unsupervised estimator based on skewness, which can be recommended for the estimation of the linear discriminant in practice. It belongs to the best estimators under consideration, has hardly any convergence issues, and converges quickest to the limiting distribution.

Note that both the simulations and theoretical aspects addressed in this study focus on scenarios where the dimension remains fixed while the sample size increases. This approach is adopted because LDA is well-known to underperform in high-dimensional settings [17]. Consequently, there is little rationale to expect that unsupervised methods, like the ones considered here, would be effective under these conditions. For readers interested in exploring this further, an additional simulation in the Supplement confirms these limitations. Although the results are suboptimal, they still surpass however random guessing.

8. Discussion

We briefly comment on possible avenues for future research. While [6] considered the asymptotics of kurtosis-based projection pursuit for estimating the discriminating direction, the limiting behaviors of other kurtosis-based estimators, such as [2], are still unknown. As such, a natural continuation of this work would be to conduct an equivalent study of fourth moments instead of third, see [13] for such a study under the independent component model. Another possible extension would be to allow for elliptical mixtures instead of normal ones. As elliptical distributions have a similar joint moment structure as the multivariate normal, it is reasonable to expect that analogous results could be derived for them. A third option would be to assume that the observed mixture contains $k > 2$ components and estimate a $k-1$ dimensional subspace that best separates them; see [18, Theorem 4] for a related Fisher consistency result. [19] considered the latter two problems under the finite mixtures of weakly symmetric distributions only differing in their means and showed that directions maximizing skewness of the projection can indeed be used in these situations. Interestingly, eigenvectors of particular, symmetric third-order tensors are shown to be consistent for the separating direction. As the asymptotic distribution of such vectors is beyond the scope of this paper, we find it to be an interesting venue for future research. Finally, a natural continuation is to study the high-dimensional asymptotics of the methods in a scenario where $p \equiv p_n \to \infty$ as $n \to \infty$. Such a study was conducted for the projection pursuit estimator in [6] and the tools involved there could prove useful also for the methods considered in the current work.

CRediT authorship contribution statement

Una Radojičić: Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing. **Klaus Nordhausen:** Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing. **Joni Virta:** Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing.

Acknowledgments

The work of JV was supported by the Research Council of Finland (Grants 335077, 347501, 353769). KN was supported by the HiTEc COST Action (CA21163) and by the Research Council of Finland (363261). The work of UR was supported by the Austrian Science Fund (FWF), [10.55776/I5799]. The authors are grateful to the editors and the two anonymous reviewers whose comments were of great help in improving the manuscript.

Appendix A. Proofs of technical results

We present the proofs grouped by section. Throughout the proofs, the various estimators such as $\hat{\theta}_{\mathrm{M}}$ will be presented without a subscript. It will always be clear from the context which estimator we are working with.

A.1. Proofs of the results in Section 2

In the proofs, we denote the estimator, for simplicity, by $\hat{\theta}$ (instead of $\hat{\theta}_{\mathrm{M}}$). Additionally, to simplify the notation and without loss of generality (since all our procedures involve centering), we also impose the condition $\mathrm{E}(\mathbf{x}) = \mathbf{0}$, implying that $\boldsymbol{\mu}_1 = -\alpha_2\mathbf{h}$ and $\boldsymbol{\mu}_2 = \alpha_1\mathbf{h}$ for some non-zero $\mathbf{h} \in \mathbb{R}^p$. The Bayes optimal projection direction is then as in the main text, $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$, where $\boldsymbol{\theta} := \boldsymbol{\Sigma}^{-1}\mathbf{h}$.

Proof of Lemma 1. Straightforward computation reveals that $\mathbf{C}_2 = \boldsymbol{\Sigma} + \beta\mathbf{h}\mathbf{h}^\top$ and for the third moment $\mathbf{c}_3$ one can use, e.g., [20], to see that $\mathbf{c}_3 = \beta\gamma\|\mathbf{h}\|^2\mathbf{h}$. Consequently, $\beta^{-1/3}\gamma^{-1/3}\|\mathbf{c}_3\|^{-2/3}\mathbf{c}_3 = \mathbf{h}$, from which the claim follows. □

Before proving Theorem 1, we first present three auxiliary lemmas: Lemma 9 obtains the joint limiting distribution of $\hat{\mathbf{C}}_2$ and $\hat{\mathbf{c}}_3$, in terms of which the limiting distribution of $\hat{\theta}$ is expressed in Lemma 10. Lemma 11 collects different moments of order up to the sixth that are required in finding the limiting covariance matrix of the method of moments estimator.

Lemma 9. *As $n \to \infty$,*

$$\sqrt{n}\begin{pmatrix}\mathrm{vec}(\hat{\mathbf{C}}_2) - \mathrm{vec}(\mathbf{C}_2)\\ \hat{\mathbf{c}}_3 - \mathbf{c}_3\end{pmatrix} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix}(\mathbf{x}_i \otimes \mathbf{x}_i) - \mathrm{E}(\mathbf{x}\otimes\mathbf{x})\\ (\mathbf{x}_i\mathbf{x}_i{}^\top\mathbf{x}_i - \mathrm{E}(\mathbf{x}\mathbf{x}^\top\mathbf{x})) - (2\mathrm{E}(\mathbf{x}\mathbf{x}^\top) + \mathrm{E}(\mathbf{x}^\top\mathbf{x})\mathbf{I}_p)\mathbf{x}_i\end{pmatrix} + o_P(1).$$

converges in distribution to the normal distribution with the covariance matrix

$$\boldsymbol{\Theta} = \begin{pmatrix}\boldsymbol{\Theta}_{1,1} & \boldsymbol{\Theta}_{1,2}\\ \boldsymbol{\Theta}_{2,1} & \boldsymbol{\Theta}_{2,2}\end{pmatrix},$$

the blocks of which are given in the proof of the lemma.

Proof of Lemma 9. We denote the non-centered counterparts of $\hat{\mathbf{C}}_2$ and $\hat{\mathbf{c}}_3$ by

$$\hat{\mathbf{C}}_{02} := \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top \quad \text{and} \quad \hat{\mathbf{c}}_{03} := \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{x}_i.$$

As $\mathrm{E}(\mathbf{x}) = \mathbf{0}$, linearization coupled with the fact that $\sqrt{n}\bar{\mathbf{x}} = \mathcal{O}_P(1)$, reveals that $\sqrt{n}(\hat{\mathbf{C}}_2 - \mathbf{C}_2) = \sqrt{n}(\hat{\mathbf{C}}_{02} - \mathbf{C}_2) + o_P(1)$, allowing us to ignore the centering for the second moment. For the third moment, we have that,

$$\sqrt{n}(\hat{\mathbf{c}}_3 - \mathbf{c}_3) = \sqrt{n}(\hat{\mathbf{c}}_{03} - \mathbf{c}_3) - 2\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\right)\sqrt{n}\bar{\mathbf{x}} - \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^\top\mathbf{x}_i\right)\sqrt{n}\bar{\mathbf{x}} + o_P(1) = \sqrt{n}(\hat{\mathbf{c}}_{03} - \mathbf{c}_3) - \mathbf{B}\sqrt{n}\bar{\mathbf{x}} + o_P(1),$$

where $\mathbf{B} := 2\boldsymbol{\Sigma} + 2\beta\mathbf{h}\mathbf{h}^\top + \mathrm{tr}(\boldsymbol{\Sigma})\mathbf{I}_p + \beta\|\mathbf{h}\|^2\mathbf{I}_p$. Note that all terms in the expansion that involve more than one instance of $\bar{\mathbf{x}}$ are $o_P(1)$ as $\sqrt{n}\bar{\mathbf{x}} = \mathcal{O}_P(1)$.

Consequently, the limiting covariance matrix of $(\mathrm{vec}(\hat{\mathbf{C}}_2), \hat{\mathbf{c}}_3)$ is

$$\begin{pmatrix} \mathbf{0} & \mathbf{I}_{p^2} & \mathbf{0} \\ -\mathbf{B} & \mathbf{0} & \mathbf{I}_p \end{pmatrix} \mathbf{V} \begin{pmatrix} \mathbf{0} & \mathbf{I}_{p^2} & \mathbf{0} \\ -\mathbf{B} & \mathbf{0} & \mathbf{I}_p \end{pmatrix}^\top,$$

where the $(p + p^2 + p) \times (p + p^2 + p)$ matrix $\mathbf{V}$ is the covariance matrix of the random vector $(\mathbf{x}, \mathbf{x}\otimes\mathbf{x}, \mathbf{x}\mathbf{x}^\top\mathbf{x})$, a block matrix with blocks

$$\begin{aligned}
&\mathbf{V}_{1,1} = \mathrm{Cov}(\mathbf{x}) = \mathrm{E}(\mathbf{x}\mathbf{x}^\top) = \mathbf{C}_2, \quad \mathbf{V}_{2,2} = \mathrm{Cov}(\mathbf{x}\otimes\mathbf{x}) = \mathrm{E}\{(\mathbf{x}\otimes\mathbf{x})(\mathbf{x}\otimes\mathbf{x})^\top\} - \mathrm{E}(\mathbf{x}\otimes\mathbf{x})\mathrm{E}(\mathbf{x}\otimes\mathbf{x})^\top,\\
&\mathbf{V}_{1,2} = \mathrm{Cov}(\mathbf{x}, \mathbf{x}\otimes\mathbf{x}) = \mathrm{E}\{\mathbf{x}(\mathbf{x}\otimes\mathbf{x})^\top\}, \quad \mathbf{V}_{2,3} = \mathrm{Cov}\{(\mathbf{x}\otimes\mathbf{x}), \mathbf{x}\mathbf{x}^\top\mathbf{x}\} = \mathrm{E}\{(\mathbf{x}\otimes\mathbf{x})\mathbf{x}^\top\mathbf{x}\mathbf{x}^\top\} - \mathrm{E}(\mathbf{x}\otimes\mathbf{x})\mathrm{E}(\mathbf{x}\mathbf{x}^\top\mathbf{x})^\top,\\
&\mathbf{V}_{1,3} = \mathrm{Cov}(\mathbf{x}, \mathbf{x}\mathbf{x}^\top\mathbf{x}) = \mathrm{E}(\mathbf{x}\mathbf{x}^\top\mathbf{x}\mathbf{x}^\top), \quad \mathbf{V}_{3,3} = \mathrm{Cov}(\mathbf{x}\mathbf{x}^\top\mathbf{x}) = \mathrm{E}(\mathbf{x}\mathbf{x}^\top\mathbf{x}\mathbf{x}^\top\mathbf{x}\mathbf{x}^\top) - \mathrm{E}(\mathbf{x}\mathbf{x}^\top\mathbf{x})\mathrm{E}(\mathbf{x}\mathbf{x}^\top\mathbf{x})^\top.
\end{aligned}$$

The formulas for these six blocks are computed in Lemma 11, and they can be substituted to

$$\boldsymbol{\Theta} = \begin{pmatrix} \mathbf{V}_{2,2} & -\mathbf{V}_{2,1}\mathbf{B}^\top + \mathbf{V}_{2,3} \\ -\mathbf{B}\mathbf{V}_{1,2} + \mathbf{V}_{3,2} & \mathbf{B}\mathbf{V}_{1,1}\mathbf{B}^\top - \mathbf{V}_{3,1}\mathbf{B}^\top - \mathbf{B}\mathbf{V}_{1,3} + \mathbf{V}_{3,3} \end{pmatrix},$$

to obtain the final limiting covariance matrix. □

Lemma 10. *We have,*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = -(\boldsymbol{\theta}^\top \otimes \boldsymbol{\Sigma}^{-1})\sqrt{n}\{\mathrm{vec}(\hat{\mathbf{C}}_2) - \mathrm{vec}(\mathbf{C}_2)\} + \{(1+\beta\tau)\boldsymbol{\Sigma}^{-1} + \beta\boldsymbol{\theta}\boldsymbol{\theta}^\top\}\mathbf{A}\sqrt{n}(\hat{\mathbf{c}}_3 - \mathbf{c}_3) + o_P(1),$$

where

$$\mathbf{A} := \beta^{-1/3}\gamma^{-1/3}\|\mathbf{c}_3\|^{-2/3}\mathbf{I}_p - \frac{2}{3}\beta^{-1/3}\gamma^{-1/3}\|\mathbf{c}_3\|^{-2/3}\frac{\mathbf{h}\mathbf{h}^\top}{\|\mathbf{h}\|^2}.$$

Proof of Lemma 10. We denote by $\hat{\mathbf{h}} := \beta^{-1/3}\gamma^{-1/3}\|\hat{\mathbf{c}}_3\|^{-2/3}\hat{\mathbf{c}}_3$ the sample third-moment estimator of $\mathbf{h}$. Standard asymptotic linearization shows that $\sqrt{n}(\|\hat{\mathbf{c}}_3\|^2 - \|\mathbf{c}_3\|^2) = 2\mathbf{c}_3^\top\sqrt{n}(\hat{\mathbf{c}}_3 - \mathbf{c}_3) + o_P(1)$, and, consequently, by the delta method, we get

$$\sqrt{n}(\|\hat{\mathbf{c}}_3\|^{-2/3} - \|\mathbf{c}_3\|^{-2/3}) = -\frac{2}{3}\|\mathbf{c}_3\|^{-8/3}\mathbf{c}_3^\top\sqrt{n}(\hat{\mathbf{c}}_3 - \mathbf{c}_3) + o_P(1).$$

Using this, the asymptotic linearization of $\hat{\mathbf{h}}$ is

$$\sqrt{n}(\hat{\mathbf{h}} - \mathbf{h}) = \beta^{-1/3}\gamma^{-1/3}\{\sqrt{n}(\|\hat{\mathbf{c}}_3\|^{-2/3} - \|\mathbf{c}_3\|^{-2/3})\mathbf{c}_3 + \|\mathbf{c}_3\|^{-2/3}\sqrt{n}(\hat{\mathbf{c}}_3 - \mathbf{c}_3)\} + o_P(1) = \mathbf{A}\sqrt{n}(\hat{\mathbf{c}}_3 - \mathbf{c}_3) + o_P(1),$$

where $\mathbf{A}$ is as in the statement of the lemma.

Denote next $\hat{\boldsymbol{\Sigma}} := \hat{\mathbf{C}}_2 - \beta\hat{\mathbf{h}}\hat{\mathbf{h}}^\top$. Then, we have $\sqrt{n}(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}) = \sqrt{n}(\hat{\mathbf{C}}_2 - \mathbf{C}_2) - \beta\mathbf{A}\sqrt{n}(\hat{\mathbf{c}}_3 - \mathbf{c}_3)\mathbf{h}^\top - \beta\mathbf{h}\sqrt{n}(\hat{\mathbf{c}}_3 - \mathbf{c}_3)^\top\mathbf{A}^\top$. To obtain a linearization for the inverse $\hat{\boldsymbol{\Sigma}}^{-1}$, we observe that $\mathbf{0} = \sqrt{n}(\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\Sigma}}^{-1} - \mathbf{I}_p) = \sqrt{n}(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\hat{\boldsymbol{\Sigma}}^{-1} + \boldsymbol{\Sigma}\sqrt{n}(\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1})$, giving $\sqrt{n}(\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}) = -\boldsymbol{\Sigma}^{-1}\sqrt{n}(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1} + o_P(1)$. Hence, $\hat{\boldsymbol{\theta}}$ has the linearization,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = -\boldsymbol{\Sigma}^{-1}\sqrt{n}(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\mathbf{h} + \boldsymbol{\Sigma}^{-1}\sqrt{n}(\hat{\mathbf{h}} - \mathbf{h}) + o_P(1),$$

and plugging in our earlier expressions for $\sqrt{n}(\hat{\mathbf{h}} - \mathbf{h})$ and $\sqrt{n}(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})$ now yields the claim. □

Lemma 11. *Let* $\mathbf{x}$ *follow the mixture distribution (1) with* $\boldsymbol{\mu}_1 = -\alpha_2\mathbf{h}$ *and* $\boldsymbol{\mu}_2 = \alpha_1\mathbf{h}$ *for some* $\mathbf{h} \in \mathbb{R}^p$. *Then we have the following moments.*

$$\begin{aligned}
\mathrm{Cov}(\mathbf{x}) &= \boldsymbol{\Sigma} + \beta\mathbf{h}\mathbf{h}^\top, \quad \mathrm{Cov}(\mathbf{x}, \mathbf{x}\otimes\mathbf{x}) = \beta\gamma\mathbf{h}(\mathbf{h}\otimes\mathbf{h})^\top,\\
\mathrm{Cov}(\mathbf{x}, \mathbf{x}\mathbf{x}^\top\mathbf{x}) &= \mathrm{tr}(\boldsymbol{\Sigma})\boldsymbol{\Sigma} + 2\boldsymbol{\Sigma}^2 + 2\beta\mathbf{h}\mathbf{h}^\top\boldsymbol{\Sigma} + 2\beta\boldsymbol{\Sigma}\mathbf{h}\mathbf{h}^\top + \beta\|\mathbf{h}\|^2\boldsymbol{\Sigma} + \beta\mathrm{tr}(\boldsymbol{\Sigma})\mathbf{h}\mathbf{h}^\top + \beta(1-3\beta)\|\mathbf{h}\|^2\mathbf{h}\mathbf{h}^\top,\\
\mathrm{Cov}(\mathbf{x}\otimes\mathbf{x}) &= (\mathbf{I}_{p^2} + \mathbf{K}_{p,p})(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma} + \beta\mathbf{h}\mathbf{h}^\top\otimes\boldsymbol{\Sigma} + \beta\boldsymbol{\Sigma}\otimes\mathbf{h}\mathbf{h}^\top) + \beta(1-4\beta)\mathbf{h}\mathbf{h}^\top\otimes\mathbf{h}\mathbf{h}^\top\\
\mathrm{Cov}\{(\mathbf{x}\otimes\mathbf{x}), \mathbf{x}\mathbf{x}^\top\mathbf{x}\} &= \beta\gamma\{\mathrm{tr}(\boldsymbol{\Sigma}) + (1-3\beta)\|\mathbf{h}\|^2\}(\mathbf{h}\otimes\mathbf{h})\mathbf{h}^\top + 2\beta\gamma\{(\mathbf{I}_p\otimes\boldsymbol{\Sigma}) + (\boldsymbol{\Sigma}\otimes\mathbf{I}_p)\}(\mathbf{h}\otimes\mathbf{h})\mathbf{h}^\top\\
&\quad + 2\beta\gamma(\mathbf{h}\otimes\mathbf{h})\mathbf{h}^\top\boldsymbol{\Sigma} + \beta\gamma\|\mathbf{h}\|^2(\mathbf{h}\otimes\boldsymbol{\Sigma}) + \beta\gamma\|\mathbf{h}\|^2(\boldsymbol{\Sigma}\otimes\mathbf{h}),\\
\mathrm{Cov}(\mathbf{x}\mathbf{x}^\top\mathbf{x}) &= 4\beta\mathrm{tr}(\boldsymbol{\Sigma})\boldsymbol{\Sigma}\mathbf{h}\mathbf{h}^\top + 8\beta\boldsymbol{\Sigma}^2\mathbf{h}\mathbf{h}^\top + 4\beta\gamma\|\mathbf{h}\|^2\boldsymbol{\Sigma}\mathbf{h}\mathbf{h}^\top + [2\mathrm{tr}(\boldsymbol{\Sigma}^2) + \{\mathrm{tr}(\boldsymbol{\Sigma})\}^2](\boldsymbol{\Sigma} + \beta\mathbf{h}\mathbf{h}^\top)\\
&\quad + 4\left\{\mathrm{tr}(\boldsymbol{\Sigma})\boldsymbol{\Sigma} + 2\boldsymbol{\Sigma}^2 + 2\beta\mathbf{h}\mathbf{h}^\top\boldsymbol{\Sigma} + \beta\mathrm{tr}(\boldsymbol{\Sigma})\mathbf{h}\mathbf{h}^\top + 2\beta\boldsymbol{\Sigma}\mathbf{h}\mathbf{h}^\top + \beta\gamma\|\mathbf{h}\|^2\mathbf{h}\mathbf{h}^\top + \beta\|\mathbf{h}\|^2\boldsymbol{\Sigma}\right\}\boldsymbol{\Sigma}
\end{aligned}$$

$$+\beta\{2\mathrm{tr}(\boldsymbol{\Sigma})\|\mathbf{h}\|^2+4\mathbf{h}^\top\boldsymbol{\Sigma}\mathbf{h}\}\{\boldsymbol{\Sigma}+(1-3\beta)\mathbf{h}\mathbf{h}^\top\}+\beta(1-3\beta)\|\mathbf{h}\|^4\{\boldsymbol{\Sigma}+(1-3\beta)\mathbf{h}\mathbf{h}^\top\},$$

where $\mathbf{K}_{p,p}$ *is the* (p,p)*-commutation matrix.*

Proof of Lemma 11. Let $\mathbf{y}_1\sim\mathcal{N}_p(-\alpha_2\mathbf{h},\boldsymbol{\Sigma})$ and $\mathbf{y}_2\sim\mathcal{N}_p(\alpha_1\mathbf{h},\boldsymbol{\Sigma})$ be independent random vectors such that $\mathbf{x}=B\mathbf{y}_1+(1-B)\mathbf{y}_2$, where $B\sim\mathrm{Ber}(\alpha_1)$ is independent of $\mathbf{y}_1$ and $\mathbf{y}_2$. Then, using the law of total expectation, we get $\mathrm{E}\{f(\mathbf{x})\}=\alpha_1\mathrm{E}\{f(\mathbf{y}_1)\}+\alpha_2\mathrm{E}\{f(\mathbf{y}_2)\}$. We now treat each of the six claims one-by-one:

1. $\mathrm{Cov}(\mathbf{x})=\boldsymbol{\Sigma}+\beta\mathbf{h}\mathbf{h}^\top$ follows straightforwardly from the basic moment formulas for multivariate normal distribution.
2. The (i,j)-element of $\mathrm{E}\{(\mathbf{x}\otimes\mathbf{x})\mathbf{x}^\top\}$ is $\alpha_1\mathrm{E}\{(\mathbf{y}_1\otimes\mathbf{y}_1)_i(\mathbf{y}_1)_j\}+\alpha_2\mathrm{E}\{(\mathbf{y}_2\otimes\mathbf{y}_2)_i(\mathbf{y}_2)_j\}$, where we take $i=i(k,l)=(k-1)p+l$, for $k,l=1,\dots,p$. In that case, $(\mathbf{x}\otimes\mathbf{x})_i=x_kx_l$. Focus now on the first expectation.

$$E_{1,i,j}:=\mathrm{E}\{(\mathbf{y}_1\otimes\mathbf{y}_1)_i(\mathbf{y}_1)_j\}=\mathrm{E}(y_{1,k}y_{1,l}y_{1,j})=\mathrm{E}(y_{1,k}y_{1,l})\mu_{1,j}+\sum_{s=1}^{p}\Sigma_{js}\mathrm{E}\left\{\frac{\partial}{\partial y_{1,s}}g(y_{1,k},y_{1,l})\right\},$$

   where the last equality holds due to multivariate Stein's identity [21, Lemma 1] with $g(y_{1,k},y_{1,l})=y_{1,k}y_{1,l}$. Applying simple algebra we obtain that $E_{1,i,j}=\Sigma_{kl}\mu_{1,j}+\mu_{1,k}\mu_{1,l}\mu_{1,j}+\Sigma_{jk}\mu_{1,l}+\Sigma_{jl}\mu_{1,k}$. Finally, since $\alpha_1\boldsymbol{\mu}_1+\alpha_2\boldsymbol{\mu}_2=0$, we obtain that $\mathrm{E}\{(\mathbf{x}\otimes\mathbf{x})\mathbf{x}^\top\}=\beta(\alpha_1-\alpha_2)(\mathbf{h}\otimes\mathbf{h})\mathbf{h}^\top$.
3. Using again the law of total expectation and by Theorem 2.3.8(vi) in [22], $\mathrm{Cov}(\mathbf{x}\mathbf{x}^\top\mathbf{x},\mathbf{x})=\mathrm{E}(\mathbf{x}\mathbf{x}^\top\mathbf{x}\mathbf{x}^\top)$ equals

$$\mathrm{tr}(\boldsymbol{\Sigma})\boldsymbol{\Sigma}+2\boldsymbol{\Sigma}^2+2\beta\mathbf{h}\mathbf{h}^\top\boldsymbol{\Sigma}+2\beta\boldsymbol{\Sigma}\mathbf{h}\mathbf{h}^\top+\beta\|\mathbf{h}\|^2\boldsymbol{\Sigma}+\beta\mathrm{tr}(\boldsymbol{\Sigma})\mathbf{h}\mathbf{h}^\top+\beta(1-3\beta)\|\mathbf{h}\|^2\mathbf{h}\mathbf{h}^\top,$$

   where we have used the fact that $\alpha_1^3+\alpha_2^3=1-3\beta$.
4. Observe first that $\mathrm{E}\{(\mathbf{x}\otimes\mathbf{x})(\mathbf{x}\otimes\mathbf{x})^\top\}=\mathrm{E}\{(\mathbf{x}\mathbf{x}^\top)\otimes(\mathbf{x}\mathbf{x}^\top)\}$. Then using again the law of total expectation and Theorem 3.1(v) in von Rosen [23], we obtain that

$$\begin{aligned}\mathrm{E}\{(\mathbf{x}\otimes\mathbf{x})(\mathbf{x}\otimes\mathbf{x})^\top\}=\;&\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma}+\mathrm{vec}(\boldsymbol{\Sigma})\mathrm{vec}(\boldsymbol{\Sigma})^\top+\mathbf{K}_{p,p}(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})+\beta\boldsymbol{\Sigma}\otimes\mathbf{h}\mathbf{h}^\top+\beta(\mathbf{h}\otimes\mathbf{h})\mathrm{vec}(\boldsymbol{\Sigma})^\top\\&+\beta\mathbf{K}_{p,p}(\mathbf{h}\mathbf{h}^\top\otimes\boldsymbol{\Sigma}+\boldsymbol{\Sigma}\otimes\mathbf{h}\mathbf{h}^\top)+\beta\mathrm{vec}(\boldsymbol{\Sigma})(\mathbf{h}\otimes\mathbf{h})^\top+\beta\mathbf{h}\mathbf{h}^\top\otimes\boldsymbol{\Sigma}+\beta(1-3\beta)\mathbf{h}\mathbf{h}^\top\otimes\mathbf{h}\mathbf{h}^\top,\end{aligned}$$

   where $\mathbf{K}_{p,p}$ is the (p,p)-commutation matrix. Furthermore, $\mathrm{E}(\mathbf{x}\otimes\mathbf{x})=\mathrm{vec}\{\mathrm{E}(\mathbf{x}\mathbf{x}^\top)\}=\mathrm{vec}(\boldsymbol{\Sigma})+\beta(\mathbf{h}\otimes\mathbf{h})$, giving

$$\mathrm{E}(\mathbf{x}\otimes\mathbf{x})\mathrm{E}(\mathbf{x}\otimes\mathbf{x})^\top=\mathrm{vec}(\boldsymbol{\Sigma})\mathrm{vec}(\boldsymbol{\Sigma})^\top+\beta\mathrm{vec}(\boldsymbol{\Sigma})(\mathbf{h}\otimes\mathbf{h})^\top+\beta(\mathbf{h}\otimes\mathbf{h})\mathrm{vec}(\boldsymbol{\Sigma})^\top+\beta^2(\mathbf{h}\otimes\mathbf{h})(\mathbf{h}\otimes\mathbf{h})^\top.$$

   Finally, $\mathrm{Cov}(\mathbf{x}\otimes\mathbf{x})$ takes the form $(\mathbf{I}_{p^2}+\mathbf{K}_{p,p})(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma}+\beta\mathbf{h}\mathbf{h}^\top\otimes\boldsymbol{\Sigma}+\beta\boldsymbol{\Sigma}\otimes\mathbf{h}\mathbf{h}^\top)+\beta(1-4\beta)\mathbf{h}\mathbf{h}^\top\otimes\mathbf{h}\mathbf{h}^\top$.
5. For $i=i(k,l)=(k-1)p+l$, $\mathrm{E}\{(\mathbf{x}\otimes\mathbf{x})\mathbf{x}^\top\mathbf{x}\mathbf{x}^\top\}_{i,j}=\mathrm{E}(\|\mathbf{x}\|^2x_kx_lx_j)$. Due to the law of total expectation, we have $\mathrm{E}\{(\mathbf{x}\otimes\mathbf{x})\mathbf{x}^\top\mathbf{x}\mathbf{x}^\top\}=\alpha_1\mathrm{E}\{(\mathbf{y}_1\otimes\mathbf{y}_1)\mathbf{y}_1^\top\mathbf{y}_1\mathbf{y}_1^\top\}+\alpha_2\mathrm{E}\{(\mathbf{y}_2\otimes\mathbf{y}_2)\mathbf{y}_2^\top\mathbf{y}_2\mathbf{y}_2^\top\}$. We focus now on the first expression in this sum, denoting it by $E_{i,j}(\mathbf{y})$ and omitting the subscript "1" in the following.
   By multivariate Stein's lemma, with $g(\mathbf{y})=\|\mathbf{y}\|^2y_ky_l$, the quantity $E_{i,j}(\mathbf{y})$ equals,

$$\begin{aligned}&\mathrm{E}(\|\mathbf{y}\|^2y_ky_l)\mu_j+2\sum_{s=1}^{p}\Sigma_{js}\mathrm{E}(y_sy_ky_l)+\Sigma_{jk}\mathrm{E}(y_l\|\mathbf{y}\|^2)+\Sigma_{jl}\mathrm{E}(y_k\|\mathbf{y}\|^2)\\&\quad=\mathrm{E}(\|\mathbf{y}\|^2(\mathbf{y}\otimes\mathbf{y})_i\mu_j)+2\mathbf{e}_j^\top\boldsymbol{\Sigma}\mathrm{E}\{\mathbf{y}(\mathbf{y}\otimes\mathbf{y})_i\}+\mathrm{E}\{(\mathbf{e}_j^\top\boldsymbol{\Sigma})_ky_l\|\mathbf{y}\|^2\}+\mathrm{E}\{y_k(\boldsymbol{\Sigma}\mathbf{e}_j)_l\|\mathbf{y}\|^2)\\&\quad=[\mathrm{E}\{\|\mathbf{y}\|^2(\mathbf{y}\otimes\mathbf{y})\}\boldsymbol{\mu}^\top]_{i,j}+2[\mathrm{E}\{(\mathbf{y}\otimes\mathbf{y})\mathbf{y}^\top\}\boldsymbol{\Sigma}]_{i,j}+[\boldsymbol{\Sigma}\otimes\mathrm{E}(\mathbf{y}\mathbf{y}^\top\mathbf{y})]_{i,j}+[\mathrm{E}(\mathbf{y}\mathbf{y}^\top\mathbf{y})\otimes\boldsymbol{\Sigma}]_{i,j}.\end{aligned}\tag{A.1}$$

   Furthermore, again applying multivariate Stein's lemma with $g(\mathbf{y})=\|\mathbf{y}\|^2y_l$ and $i=i(k,l)$, we obtain

$$\mathrm{E}\{\|\mathbf{y}\|^2(\mathbf{y}\otimes\mathbf{y})_i\}=\mathrm{E}(\|\mathbf{y}\|^2y_l)\mu_k+2\sum_{s=1}^{p}\Sigma_{ks}\mathrm{E}(y_sy_l)+\Sigma_{kl}\mathrm{E}(\|\mathbf{y}\|^2),$$

   giving

$$\begin{aligned}\mathrm{E}\{||\mathbf{y}||^2(\mathbf{y}\otimes\mathbf{y})\}\boldsymbol{\mu}^\top=\;&\mathrm{E}\{\|\mathbf{y}\|^2(\boldsymbol{\mu}\otimes\mathbf{y})\}\boldsymbol{\mu}^\top+2\mathrm{E}(\boldsymbol{\Sigma}\mathbf{y}\otimes\mathbf{y})\boldsymbol{\mu}^\top+\mathrm{E}(\|\mathbf{y}\|^2)\mathrm{vec}(\boldsymbol{\Sigma})\boldsymbol{\mu}^\top\\=\;&\{\mathrm{tr}(\boldsymbol{\Sigma})\mathbf{I}_{p^2}+\|\boldsymbol{\mu}\|^2\mathbf{I}_{p^2}+2(\boldsymbol{\Sigma}\otimes\mathbf{I}_p)+2(\mathbf{I}_p\otimes\boldsymbol{\Sigma})\}(\boldsymbol{\mu}\otimes\boldsymbol{\mu})\boldsymbol{\mu}^\top+\{\mathrm{tr}(\boldsymbol{\Sigma})\mathbf{I}_{p^2}+\|\boldsymbol{\mu}\|^2\mathbf{I}_{p^2}+2(\boldsymbol{\Sigma}\otimes\mathbf{I}_p)\}\mathrm{vec}(\boldsymbol{\Sigma})\boldsymbol{\mu}^\top,\end{aligned}\tag{A.2}$$

   where we have used the fact that $\mathrm{E}(\|\mathbf{y}\|^2\mathbf{y})=2\boldsymbol{\Sigma}\boldsymbol{\mu}+\|\boldsymbol{\mu}\|^2\boldsymbol{\mu}+\mathrm{tr}(\boldsymbol{\Sigma})\boldsymbol{\mu}$ by Stein's lemma. The three final terms in (A.1) contribute the following quantities to the final sum, respectively,

$$2\mathrm{E}\{(\mathbf{x}\otimes\mathbf{x})\mathbf{x}^\top\}\boldsymbol{\Sigma}=2\beta\gamma(\mathbf{h}\otimes\mathbf{h})\mathbf{h}^\top\boldsymbol{\Sigma},\quad\boldsymbol{\Sigma}\otimes\mathrm{E}(\mathbf{x}\mathbf{x}^\top\mathbf{x})=\beta\gamma\|\mathbf{h}\|^2(\boldsymbol{\Sigma}\otimes\mathbf{h}),\quad\mathrm{E}(\mathbf{x}\mathbf{x}^\top\mathbf{x})\otimes\boldsymbol{\Sigma}=\beta\gamma\|\mathbf{h}\|^2(\mathbf{h}\otimes\boldsymbol{\Sigma}).$$

   The contribution of the term corresponding to (A.2) can be obtained with the help of the formula $\alpha_1^2+\alpha_2^2=1-2\beta$, and putting now all four terms together and subtracting $\mathrm{E}(\mathbf{x}\otimes\mathbf{x})\mathrm{E}(\mathbf{x}\mathbf{x}^\top\mathbf{x})^\top$ gives the claim.
6. We focus first on $\mathrm{E}(\mathbf{x})=\mathrm{E}\{(\mathbf{x}^\top\mathbf{x})^2\mathbf{x}\mathbf{x}^\top\}=\alpha_1\mathbb{E}(\|\mathbf{y}_1\|^4\mathbf{y}_1\mathbf{y}_1^\top)+\alpha_2\mathbb{E}(\|\mathbf{y}_2\|^4\mathbf{y}_2\mathbf{y}_2^\top)$. Then, dropping the subscript "1" for convenience, the (i,j)-element of the $\mathbf{y}_1$-part equals, by Stein's lemma with $g(\mathbf{y})=\|\mathbf{y}\|^4y_i$,

$$\mathrm{E}\{(\mathbf{y}^\top\mathbf{y})^2y_iy_j\}=\mathrm{E}\{(\mathbf{y}^\top\mathbf{y})^2y_i\}\mu_j+4\sum_{s=1}^{p}\Sigma_{js}\mathrm{E}(\|\mathbf{y}\|^2y_iy_s)+\Sigma_{ij}\mathrm{E}(\|\mathbf{y}\|^4),$$

implying $\mathrm{E}(\|\mathbf{y}\|^4\mathbf{y}\mathbf{y}^\top) = \mathrm{E}(\|\mathbf{y}\|^4\mathbf{y})\boldsymbol{\mu}^\top + 4\mathrm{E}(\|\mathbf{y}\|^2\mathbf{y}\mathbf{y}^\top)\boldsymbol{\Sigma} + \mathrm{E}(\|\mathbf{y}\|^4)\boldsymbol{\Sigma}$. Using Stein's lemma with $g(\mathbf{y}) = \|\mathbf{y}\|^4$ we get

$$\mathrm{E}(\|\mathbf{y}\|^4 y_i) = \mathrm{E}(\|\mathbf{y}\|^4)\mu_i + 4\sum_{s=1}^{p}\Sigma_{is}\mathrm{E}(\|\mathbf{y}\|^2 y_s),$$

further implying that $\mathrm{E}(\|\mathbf{y}\|^4\mathbf{y}) = \mathrm{E}(\|\mathbf{y}\|^4)\boldsymbol{\mu} + 4\boldsymbol{\Sigma}\mathrm{E}(\mathbf{y}\mathbf{y}^\top\mathbf{y})$. Hence, we have the identity $\mathrm{E}(\|\mathbf{y}\|^4\mathbf{y}\mathbf{y}^\top) = 4\boldsymbol{\Sigma}\mathrm{E}(\mathbf{y}\mathbf{y}^\top\mathbf{y})\boldsymbol{\mu}^\top + 4\mathrm{E}(\|\mathbf{y}\|^2\mathbf{y}\mathbf{y}^\top)\boldsymbol{\Sigma} + \mathrm{E}(\|\mathbf{y}\|^4)(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top)$. We next inspect the above three terms one-by-one (using the moment formulas derived earlier in the proof):

- The first one equals,

$$4\boldsymbol{\Sigma}\mathrm{E}(\mathbf{y}\mathbf{y}^\top\mathbf{y})\boldsymbol{\mu}^\top = 4\mathrm{tr}(\boldsymbol{\Sigma})\boldsymbol{\Sigma}\boldsymbol{\mu}\boldsymbol{\mu}^\top + 8\boldsymbol{\Sigma}^2\boldsymbol{\mu}\boldsymbol{\mu}^\top + 4\|\boldsymbol{\mu}\|^2\boldsymbol{\Sigma}\boldsymbol{\mu}\boldsymbol{\mu}^\top = 4\{\mathrm{tr}(\boldsymbol{\Sigma})\mathbf{I}_p + 2\boldsymbol{\Sigma} + \|\boldsymbol{\mu}\|^2\mathbf{I}_p\}\boldsymbol{\Sigma}\boldsymbol{\mu}\boldsymbol{\mu}^\top.$$

- The second one equals,

$$\begin{aligned}4\mathrm{E}(\|\mathbf{y}\|^2\mathbf{y}\mathbf{y}^\top)\boldsymbol{\Sigma} &= 4\mathrm{E}(\|\mathbf{y}\|^2\boldsymbol{\mu}\mathbf{y}^\top)\boldsymbol{\Sigma} + 8\boldsymbol{\Sigma}\mathrm{E}(\mathbf{y}\mathbf{y}^\top)\boldsymbol{\Sigma} + 4\mathrm{E}(\|\mathbf{y}\|^2)\boldsymbol{\Sigma}^2\\ &= 8\boldsymbol{\mu}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^2 + 4\|\boldsymbol{\mu}\|^2\boldsymbol{\mu}\boldsymbol{\mu}^\top\boldsymbol{\Sigma} + 4\mathrm{tr}(\boldsymbol{\Sigma})\boldsymbol{\mu}\boldsymbol{\mu}^\top\boldsymbol{\Sigma} + 8\boldsymbol{\Sigma}^3 + 8\boldsymbol{\Sigma}\boldsymbol{\mu}\boldsymbol{\mu}^\top\boldsymbol{\Sigma} + 4\mathrm{tr}(\boldsymbol{\Sigma})\boldsymbol{\Sigma}^2 + 4\|\boldsymbol{\mu}\|^2\boldsymbol{\Sigma}^2\\ &= 4\{\mathrm{tr}(\boldsymbol{\Sigma})\boldsymbol{\Sigma} + 2\boldsymbol{\Sigma}^2 + 2\boldsymbol{\mu}\boldsymbol{\mu}^\top\boldsymbol{\Sigma} + \mathrm{tr}(\boldsymbol{\Sigma})\boldsymbol{\mu}\boldsymbol{\mu}^\top + 2\boldsymbol{\Sigma}\boldsymbol{\mu}\boldsymbol{\mu}^\top + \|\boldsymbol{\mu}\|^2\boldsymbol{\mu}\boldsymbol{\mu}^\top + \|\boldsymbol{\mu}\|^2\boldsymbol{\Sigma}\}\boldsymbol{\Sigma}.\end{aligned}$$

- The third one equals,

$$\mathrm{E}(\|\mathbf{y}\|^4)(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top) = [2\mathrm{tr}(\boldsymbol{\Sigma}^2) + \{\mathrm{tr}(\boldsymbol{\Sigma})\}^2 + 2\mathrm{tr}(\boldsymbol{\Sigma})\|\boldsymbol{\mu}\|^2 + 4\boldsymbol{\mu}^\top\boldsymbol{\Sigma}\boldsymbol{\mu} + \|\boldsymbol{\mu}\|^4](\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top).$$

Therefore,

$$\begin{aligned}\mathrm{E}(\|\mathbf{y}\|^4\mathbf{y}\mathbf{y}^\top) &= 4\{\mathrm{tr}(\boldsymbol{\Sigma})\mathbf{I}_p + 2\boldsymbol{\Sigma} + \|\boldsymbol{\mu}\|^2\mathbf{I}_p\}\boldsymbol{\Sigma}\boldsymbol{\mu}\boldsymbol{\mu}^\top + 4\{\mathrm{tr}(\boldsymbol{\Sigma})\boldsymbol{\Sigma} + 2\boldsymbol{\Sigma}^2 + 2\boldsymbol{\mu}\boldsymbol{\mu}^\top\boldsymbol{\Sigma} + \mathrm{tr}(\boldsymbol{\Sigma})\boldsymbol{\mu}\boldsymbol{\mu}^\top + 2\boldsymbol{\Sigma}\boldsymbol{\mu}\boldsymbol{\mu}^\top\\ &\quad + \|\boldsymbol{\mu}\|^2\boldsymbol{\mu}\boldsymbol{\mu}^\top + \|\boldsymbol{\mu}\|^2\boldsymbol{\Sigma}\}\boldsymbol{\Sigma} + [2\mathrm{tr}(\boldsymbol{\Sigma}^2) + \{\mathrm{tr}(\boldsymbol{\Sigma})\}^2 + 2\mathrm{tr}(\boldsymbol{\Sigma})\|\boldsymbol{\mu}\|^2 + 4\boldsymbol{\mu}^\top\boldsymbol{\Sigma}\boldsymbol{\mu} + \|\boldsymbol{\mu}\|^4](\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top).\end{aligned}$$

Thus, finally, using $\alpha_1^5 + \alpha_2^5 = 1 - 5\beta + 5\beta^2$ and $\alpha_1^3 + \alpha_2^3 = 1 - 3\beta$,

$$\begin{aligned}\mathrm{E}(\|\mathbf{x}\|^4\mathbf{x}\mathbf{x}^\top) &= 4\beta\mathrm{tr}(\boldsymbol{\Sigma})\boldsymbol{\Sigma}\mathbf{h}\mathbf{h}^\top + 8\beta\boldsymbol{\Sigma}^2\mathbf{h}\mathbf{h}^\top + 4\beta(1-3\beta)\|\mathbf{h}\|^2\boldsymbol{\Sigma}\mathbf{h}\mathbf{h}^\top + [2\mathrm{tr}(\boldsymbol{\Sigma}^2) + \{\mathrm{tr}(\boldsymbol{\Sigma})\}^2](\boldsymbol{\Sigma} + \beta\mathbf{h}\mathbf{h}^\top)\\ &\quad + 4\left\{\mathrm{tr}(\boldsymbol{\Sigma})\boldsymbol{\Sigma} + 2\boldsymbol{\Sigma}^2 + 2\beta\mathbf{h}\mathbf{h}^\top\boldsymbol{\Sigma} + \beta\mathrm{tr}(\boldsymbol{\Sigma})\mathbf{h}\mathbf{h}^\top + 2\beta\boldsymbol{\Sigma}\mathbf{h}\mathbf{h} + \beta(1-3\beta)\|\mathbf{h}\|^2\mathbf{h}\mathbf{h}^\top + \beta\|\mathbf{h}\|^2\boldsymbol{\Sigma}\right\}\boldsymbol{\Sigma}\\ &\quad + \beta\{2\mathrm{tr}(\boldsymbol{\Sigma})\|\mathbf{h}\|^2 + 4\mathbf{h}^\top\boldsymbol{\Sigma}\mathbf{h}\}\{\boldsymbol{\Sigma} + (1-3\beta)\mathbf{h}\mathbf{h}^\top\} + \beta\|\mathbf{h}\|^4\{(1-3\beta)\boldsymbol{\Sigma} + (1-5\beta+5\beta^2)\mathbf{h}\mathbf{h}^\top\}.\end{aligned}$$

Subtracting now $\beta^2\gamma^2\|\mathbf{h}\|^4\mathbf{h}\mathbf{h}^\top$ and using $\gamma^2 = 1 - 4\beta$ and $1 - 6\beta + 9\beta^2 = (1-3\beta)^2$ yields the claim. □

Proof of Theorem 1. By the delta method, the normalized estimator has

$$\sqrt{n}\left(\frac{\hat{\boldsymbol{\theta}}}{\|\hat{\boldsymbol{\theta}}\|} - \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|}\right) = \frac{1}{\|\boldsymbol{\theta}\|}\left(\mathbf{I}_p - \frac{\boldsymbol{\theta}\boldsymbol{\theta}^\top}{\|\boldsymbol{\theta}\|^2}\right)\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + o_P(1).$$

Denoting the projection matrix onto the orthogonal complement of $\boldsymbol{\theta}$ by $\mathbf{Q}_{\boldsymbol{\theta}}$, we thus have, by Lemmas 9 and 10 (using their notation) that

$$\begin{aligned}\boldsymbol{\Upsilon}_{\mathrm{M}} =& \frac{1}{\|\boldsymbol{\theta}\|^2}\mathbf{Q}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^\top\otimes\boldsymbol{\Sigma}^{-1})\boldsymbol{\Theta}_{1,1}(\boldsymbol{\theta}\otimes\boldsymbol{\Sigma}^{-1})\mathbf{Q}_{\boldsymbol{\theta}} - \frac{1}{\|\boldsymbol{\theta}\|^2}\mathbf{Q}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^\top\otimes\boldsymbol{\Sigma}^{-1})\boldsymbol{\Theta}_{1,2}\mathbf{A}^\top\{(1+\beta\tau)\boldsymbol{\Sigma}^{-1} + \beta\boldsymbol{\theta}\boldsymbol{\theta}^\top\}\mathbf{Q}_{\boldsymbol{\theta}}\\ &- \frac{1}{\|\boldsymbol{\theta}\|^2}\mathbf{Q}_{\boldsymbol{\theta}}\{(1+\beta\tau)\boldsymbol{\Sigma}^{-1} + \beta\boldsymbol{\theta}\boldsymbol{\theta}^\top\}\mathbf{A}\boldsymbol{\Theta}_{2,1}(\boldsymbol{\theta}\otimes\boldsymbol{\Sigma}^{-1})\mathbf{Q}_{\boldsymbol{\theta}} + \frac{1}{\|\boldsymbol{\theta}\|^2}\mathbf{Q}_{\boldsymbol{\theta}}\{(1+\beta\tau)\boldsymbol{\Sigma}^{-1} + \beta\boldsymbol{\theta}\boldsymbol{\theta}^\top\}\mathbf{A}\boldsymbol{\Theta}_{2,2}\mathbf{A}^\top\{(1+\beta\tau)\boldsymbol{\Sigma}^{-1} + \beta\boldsymbol{\theta}\boldsymbol{\theta}^\top\}\mathbf{Q}_{\boldsymbol{\theta}}.\end{aligned}$$

The expressions for the matrices $\boldsymbol{\Theta}_{k,\ell}$ are given in Lemma 9 as functions of the blocks $\mathbf{V}_{k',\ell'}$, which themselves are computed in Lemma 11. Plugging everything in above and simplifying (using the fact that $\mathbf{Q}_{\boldsymbol{\theta}}\boldsymbol{\theta} = \mathbf{0}$ whenever possible) gives then the desired expression. □

A.2. Proofs of the results in Section 3

Lemma 12. *Let the assumptions of Theorem 2 hold and let*

$$\mathbf{z} \sim \alpha_1\mathcal{N}_p(-\alpha_2\boldsymbol{\Sigma}^{-1/2}\mathbf{h}, \mathbf{I}_p) + \alpha_2\mathcal{N}_p(\alpha_1\boldsymbol{\Sigma}^{-1/2}\mathbf{h}, \mathbf{I}_p).$$

Using the notation $\mathbf{m} := \boldsymbol{\Sigma}^{-1/2}\mathbf{h}$*, we then have*

$$\sqrt{n}\left\{\frac{\hat{\boldsymbol{\theta}}(\boldsymbol{\Sigma}^{1/2}\mathbf{z}_i)}{\|\hat{\boldsymbol{\theta}}(\boldsymbol{\Sigma}^{1/2}\mathbf{z}_i)\|} - \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|}\right\} = \frac{\|\mathbf{m}\|}{\|\boldsymbol{\theta}\|}\mathbf{Q}_{\boldsymbol{\theta}}\boldsymbol{\Sigma}^{-1/2}\sqrt{n}\left\{\frac{\hat{\boldsymbol{\theta}}(\mathbf{z}_i)}{\|\hat{\boldsymbol{\theta}}(\mathbf{z}_i)\|} - \frac{\mathbf{m}}{\|\mathbf{m}\|}\right\} + o_P(1).$$

Proof of Lemma 12. By affine equivariance, the left-hand side can be expanded as

$$\sqrt{n}\left\{\frac{\hat{\boldsymbol{\theta}}(\boldsymbol{\Sigma}^{1/2}\mathbf{z}_i)}{\|\hat{\boldsymbol{\theta}}(\boldsymbol{\Sigma}^{1/2}\mathbf{z}_i)\|} - \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|}\right\} = \boldsymbol{\Sigma}^{-1/2}\sqrt{n}\left\{\frac{\hat{\boldsymbol{\theta}}(\mathbf{z}_i)}{\|\hat{\boldsymbol{\theta}}(\mathbf{z}_i)\|} - \frac{\mathbf{m}}{\|\mathbf{m}\|}\right\}\frac{\|\hat{\boldsymbol{\theta}}(\mathbf{z}_i)\|}{\|\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\theta}}(\mathbf{z}_i)\|} + \frac{\boldsymbol{\theta}}{\|\mathbf{m}\|}\sqrt{n}\left\{\frac{\|\hat{\boldsymbol{\theta}}(\mathbf{z}_i)\|}{\|\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\theta}}(\mathbf{z}_i)\|} - \frac{\|\mathbf{m}\|}{\|\boldsymbol{\theta}\|}\right\}. \tag{A.3}$$

By Slutsky's lemma, the first term on the RHS of (A.3) equals

$$\frac{\|\mathbf{m}\|}{\|\theta\|}\boldsymbol{\Sigma}^{-1/2}\sqrt{n}\left\{\frac{\hat{\theta}(\mathbf{z}_i)}{\|\hat{\theta}(\mathbf{z}_i)\|}-\frac{\mathbf{m}}{\|\mathbf{m}\|}\right\}+o_P(1).$$

Let now H be a constant such that $\hat{\theta}(\mathbf{z}_i)\to_P H\mathbf{m}$. Then, the $\sqrt{n}$-part of the second term on the RHS of (A.3) can be written as

$$\frac{1}{H\|\theta\|}\sqrt{n}\{\|\hat{\theta}(\mathbf{z}_i)\|-H\|\mathbf{m}\|\}+H\|\mathbf{m}\|\sqrt{n}\left\{\frac{1}{\|\boldsymbol{\Sigma}^{-1/2}\hat{\theta}(\mathbf{z}_i)\|}-\frac{1}{H\|\theta\|}\right\}+o_P(1). \tag{A.4}$$

The delta method shows that the second term in (A.4) has the form

$$-\frac{\|\mathbf{m}\|}{2H^2\|\theta\|^3}\sqrt{n}\{\|\boldsymbol{\Sigma}^{-1/2}\hat{\theta}(\mathbf{z}_i)\|^2-H^2\|\theta\|^2\}+o_P(1)=-\frac{\|\mathbf{m}\|}{H\|\theta\|^3}\sqrt{n}\{\hat{\theta}(\mathbf{z}_i)-H\mathbf{m}\}^\top\boldsymbol{\Sigma}^{-1}\mathbf{m}+o_P(1).$$

Finally, $\sqrt{n}\{\hat{\theta}(\mathbf{z}_i)-H\mathbf{m}\}$ can be written as

$$\sqrt{n}\{\hat{\theta}(\mathbf{z}_i)-H\mathbf{m}\}=\sqrt{n}\left\{\frac{\hat{\theta}(\mathbf{z}_i)}{\|\hat{\theta}(\mathbf{z}_i)\|}-\frac{\mathbf{m}}{\|\mathbf{m}\|}\right\}H\|\mathbf{m}\|+\frac{\mathbf{m}}{\|\mathbf{m}\|}\sqrt{n}\{\|\hat{\theta}(\mathbf{z}_i)\|-H\|\mathbf{m}\|\}+o_P(1).$$

Collecting everything to (A.4), we see that the second term on the RHS of (A.3) is

$$-\frac{\|\mathbf{m}\|}{\|\theta\|^3}\theta\theta^\top\boldsymbol{\Sigma}^{-1/2}\sqrt{n}\left\{\frac{\hat{\theta}(\mathbf{z}_i)}{\|\hat{\theta}(\mathbf{z}_i)\|}-\frac{\mathbf{m}}{\|\mathbf{m}\|}\right\}+o_P(1).$$

Putting now everything together to (A.3), we obtain the claim. □

Lemma 13. *Let $\mathbf{u}_1\in\mathbb{R}^p$, $\|\mathbf{u}_1\|=1$, be fixed and asumme that that $\mathbf{B}\in\mathbb{R}^{p\times p}$ is a symmetric matrix satsfying*

$$\mathbf{B}=\mathbf{OBO}^\top,$$

for all $p\times p$ orthogonal matrices $\mathbf{O}$ for which $\mathbf{Ou}_1=\mathbf{u}_1$. Then, there exists $a,b\in\mathbb{R}$ such that

$$\mathbf{B}=a\mathbf{u}_1\mathbf{u}_1+b(\mathbf{I}_p-\mathbf{u}_1\mathbf{u}_1^\top).$$

Proof of Lemma 12. Choose the unit-length vectors $\mathbf{u}_2,\dots,\mathbf{u}_p$ such that $\mathbf{U}=(\mathbf{u}_1,\dots,\mathbf{u}_p)\in\mathbb{R}^{p\times p}$ is an orthogonal matrix. Parametrize then $\mathbf{B}:=\mathbf{URU}^\top$. Choosing now $\mathbf{O}$ such that $\mathbf{OU}=(\mathbf{u}_1,-\mathbf{u}_2,\dots,\mathbf{u}_p)$, we obtain the equation $\mathbf{U}^\top(\mathbf{OU})\mathbf{R}(\mathbf{OU})^\top\mathbf{U}=\mathbf{R}$, which shows that all elements expect $(2,2)$ in the second row and column of $\mathbf{R}$ must be zero. Similarly we can show that all other off-diagonal elements of $\mathbf{R}$ must be zero as well. Taking then $\mathbf{O}$ to be such that it permutes two of the columns $\mathbf{u}_2,\dots,\mathbf{u}_p$ of $\mathbf{U}$, we find that the final $p-1$ diagonal elements of $\mathbf{R}$ must be equal. Hence, $\mathbf{R}=\mathrm{diag}(a,b,\dots,b)$ for some $a,b\in\mathbb{R}$, yielding the claim. □

Proof of Theorem 2. Under our assumptions, Lemma 12 implies that

$$\sqrt{n}\left\{\frac{\hat{\theta}(\mathbf{x}_i)}{\|\hat{\theta}(\mathbf{x}_i)\|}-\frac{\theta}{\|\theta\|}\right\}=\frac{\|\mathbf{m}\|}{\|\theta\|}\mathbf{Q}_\theta\boldsymbol{\Sigma}^{-1/2}\sqrt{n}\left\{\frac{\hat{\theta}(\mathbf{z}_i)}{\|\hat{\theta}(\mathbf{z}_i)\|}-\frac{\mathbf{m}}{\|\mathbf{m}\|}\right\}+o_P(1), \tag{A.5}$$

where $\mathbf{m}$ and $\mathbf{z}_i$ are as in Lemma 12. Now, the distribution of $\mathbf{z}_i$ is invariant to all orthogonal transformations $\mathbf{Oz}_i$ where $\mathbf{O}$ is such that $\mathbf{Om}=\mathbf{m}$. Lemma 13 then states that the limiting covariance matrix of

$$\sqrt{n}\left\{\frac{\hat{\theta}(\mathbf{z}_i)}{\|\hat{\theta}(\mathbf{z}_i)\|}-\frac{\mathbf{m}}{\|\mathbf{m}\|}\right\}$$

has the form

$$a\frac{\mathbf{mm}^\top}{\|\mathbf{m}\|^2}+C\left(\mathbf{I}_p-\frac{\mathbf{mm}^\top}{\|\mathbf{m}\|^2}\right),$$

for some $a,C\in\mathbb{R}$. Hence the limiting covariance matrix of (A.5) is

$$\frac{\|\mathbf{m}\|^2}{\|\theta\|^2}\mathbf{Q}_\theta\boldsymbol{\Sigma}^{-1/2}\left\{a\frac{\mathbf{mm}^\top}{\|\mathbf{m}\|^2}+C\left(\mathbf{I}_p-\frac{\mathbf{mm}^\top}{\|\mathbf{m}\|^2}\right)\right\}\boldsymbol{\Sigma}^{-1/2}\mathbf{Q}_\theta=C\frac{\|\mathbf{m}\|^2}{\|\theta\|^2}\mathbf{Q}_\theta\boldsymbol{\Sigma}^{-1}\mathbf{Q}_\theta,$$

concluding the proof. □

For the proof of Theorem 3, we need the following lemma.

Lemma 14. *For any $\alpha\geq 0$ and $\mathbf{u}\in\mathbb{R}^p$, $\|\mathbf{u}\|=1$,*

$$\left[\{\mathbf{I}_p\otimes(\mathbf{I}_p+\alpha\mathbf{uu}^\top)\}+\{(\mathbf{I}_p+\alpha\mathbf{uu}^\top)\otimes\mathbf{I}_p\}\right]^{-1}=\frac{1}{2(\alpha+2)}\left[\mathbf{I}_p\otimes\mathbf{I}_p+(\alpha+1)\left\{\left(\mathbf{I}_p-\frac{\alpha}{\alpha+1}\mathbf{uu}^\top\right)\otimes\left(\mathbf{I}_p-\frac{\alpha}{\alpha+1}\mathbf{uu}^\top\right)\right\}\right].$$

Proof of Lemma 14. The statement is proven by direct multiplication. The Sherman–Morrison formula implies that $\mathbf{I}_p - \{\alpha/(\alpha+1)\}\mathbf{u}\mathbf{u}^\top = (\mathbf{I}_p + \alpha\mathbf{u}\mathbf{u}^\top)^{-1}$. If we denote now

$$\mathbf{A} = \left\{\mathbf{I}_p \otimes \left(\mathbf{I}_p + \alpha\mathbf{u}\mathbf{u}^\top\right)\right\} + \left\{\left(\mathbf{I}_p + \alpha\mathbf{u}\mathbf{u}^\top\right) \otimes \mathbf{I}_p\right\},$$

$$\mathbf{B} = \frac{1}{2(\alpha+2)}\left[\mathbf{I}_p \otimes \mathbf{I}_p + (\alpha+1)\left\{\left(\mathbf{I}_p - \frac{\alpha}{\alpha+1}\mathbf{u}\mathbf{u}^\top\right) \otimes \left(\mathbf{I}_p - \frac{\alpha}{\alpha+1}\mathbf{u}\mathbf{u}^\top\right)\right\}\right],$$

then

$$\begin{aligned}2(\alpha+2)\mathbf{A}\mathbf{B} &= \mathbf{I}_p \otimes (\mathbf{I}_p + \alpha\mathbf{u}\mathbf{u}^\top) + (\alpha+1)\left\{\left(\mathbf{I}_p - \frac{\alpha}{\alpha+1}\mathbf{u}\mathbf{u}^\top\right) \otimes \mathbf{I}_p\right\} + (\mathbf{I}_p + \alpha\mathbf{u}\mathbf{u}^\top) \otimes \mathbf{I}_p + (\alpha+1)\left\{\mathbf{I}_p \otimes \left(\mathbf{I}_p - \frac{\alpha}{\alpha+1}\mathbf{u}\mathbf{u}^\top\right)\right\} \\ &= \mathbf{I}_{p^2} + \alpha\mathbf{I}_p \otimes (\mathbf{u}\mathbf{u}^\top) + (\alpha+1)\mathbf{I}_{p^2} - \alpha(\mathbf{u}\mathbf{u}^\top) \otimes \mathbf{I}_p + \mathbf{I}_{p^2} + \alpha(\mathbf{u}\mathbf{u}^\top) \otimes \mathbf{I}_p + (\alpha+1)\mathbf{I}_{p^2} - \alpha\mathbf{I}_p \otimes (\mathbf{u}\mathbf{u}^\top) \\ &= 2(\alpha+2)\mathbf{I}_{p^2}.\end{aligned}$$

Thus, $\mathbf{A}\mathbf{B} = \mathbf{I}_p \otimes \mathbf{I}_p = \mathbf{I}_{p^2}$. As $\mathbf{A}$ and $\mathbf{B}$ are square matrices, every left inverse is also a right inverse, concluding the proof. □

Proof of Theorem 3. We have $\mathbf{C}_2(\mathbf{z}) = \mathbf{I}_p + \beta\|\mathbf{m}\|^2\mathbf{w}\mathbf{w}^\top$ where $\mathbf{w} := \mathbf{m}/\|\mathbf{m}\|$. Denoting $d = (1+\beta\|\mathbf{m}\|^2)^{-1} \in (0,1)$, we then have $\mathbf{C}_2(\mathbf{z})^{-1} = \mathbf{I}_p - (1-d)\mathbf{w}\mathbf{w}^\top$ and $\mathbf{C}_2(\mathbf{z})^{-1/2} = \mathbf{I}_p - (1-d^{1/2})\mathbf{w}\mathbf{w}^\top$. Consequently, dropping the parenthetical references to the data, $\|\hat{\boldsymbol{\theta}}(\mathbf{z}_i)\|^2 = \hat{\mathbf{u}}^\top\hat{\mathbf{C}}_2^{-1}\hat{\mathbf{u}} \to_p d$. Let now $\mathbf{t} \in \mathbb{R}^p$ be any unit-length vector satisfying $\mathbf{m}^\top\mathbf{t} = 0$. Then, by the proof of Theorem 2, we have

$$\mathbf{t}^\top\sqrt{n}\left\{\frac{\hat{\boldsymbol{\theta}}(\mathbf{z}_i)}{\|\hat{\boldsymbol{\theta}}(\mathbf{z}_i)\|} - \frac{\mathbf{m}}{\|\mathbf{m}\|}\right\} \rightsquigarrow \mathcal{N}(0, C).$$

This further gives $\mathbf{t}^\top\sqrt{n}\{\hat{\boldsymbol{\theta}}(\mathbf{z}_i) - d^{1/2}\mathbf{w}\} \rightsquigarrow \mathcal{N}(0, dC)$. The left-hand side of this can be expanded, using Slutsky's lemma, to obtain

$$\mathbf{t}^\top\sqrt{n}(\hat{\mathbf{C}}_2^{-1/2} - \mathbf{C}_2^{-1/2})\mathbf{w} + \mathbf{t}^\top\sqrt{n}(\hat{\mathbf{u}} - \mathbf{w}) + o_P(1). \tag{A.6}$$

Linearizing the equation $\sqrt{n}(\hat{\mathbf{C}}_2^{-1/2}\hat{\mathbf{C}}_2\hat{\mathbf{C}}_2^{-1/2} - \mathbf{I}_p) = 0$ and using the formula $\mathrm{vec}(\mathbf{A}\mathbf{X}\mathbf{B}^\top) = (\mathbf{B}\otimes\mathbf{A})\mathrm{vec}(\mathbf{X})$, we get

$$\sqrt{n}\mathrm{vec}(\hat{\mathbf{C}}_2^{-1/2} - \mathbf{C}_2^{-1/2}) = -\{(\mathbf{C}_2^{1/2}\otimes\mathbf{I}_p) + (\mathbf{I}_p\otimes\mathbf{C}_2^{1/2})\}^{-1}(\mathbf{C}_2^{-1/2}\otimes\mathbf{C}_2^{-1/2})\sqrt{n}\mathrm{vec}(\hat{\mathbf{C}}_2 - \mathbf{C}_2) + o_P(1).$$

Now, $\mathbf{C}_2^{1/2} = \mathbf{I}_d + (d^{-1/2}-1)\mathbf{w}\mathbf{w}^\top$ and, using Lemma 14, we get

$$\mathbf{t}^\top\sqrt{n}(\hat{\mathbf{C}}_2^{-1/2} - \mathbf{C}_2^{-1/2})\mathbf{w} = -\frac{d^{1/2}}{1+d^{-1/2}}\mathbf{t}^\top\sqrt{n}(\hat{\mathbf{C}}_2 - \mathbf{C}_2)\mathbf{w} + o_P(1).$$

Plugging in to (A.6) now yields the claim. □

Proof of Theorem 4. Let $\mathbf{R} := \mathbf{C}_2^{-1/2}(\mathbf{x})\boldsymbol{\Sigma}^{1/2}$. Then $\mathbf{R}^\top\mathbf{R} = \mathbf{I}_p - \{(\beta\tau)/(1+\beta\tau)\}\mathbf{w}\mathbf{w}^\top$ where $\mathbf{w} = \boldsymbol{\Sigma}^{-1/2}\mathbf{h}/\|\boldsymbol{\Sigma}^{-1/2}\mathbf{h}\|$. Consequently, the unique symmetric positive definite square root of $\mathbf{R}^\top\mathbf{R}$ is $\mathbf{P} := \mathbf{I}_p - \{1-(1+\beta\tau)^{-1/2}\}\mathbf{w}\mathbf{w}^\top$. Consequently $\mathbf{R} = \mathbf{O}\mathbf{P}$ for some orthogonal matrix $\mathbf{O}$. This implies that $\mathbf{C}_2^{-1/2}(\mathbf{x})\{\mathbf{x} - \mathrm{E}(\mathbf{x})\}$ can be written as

$$\mathbf{C}_2^{-1/2}(\mathbf{x})\{\mathbf{x} - \mathrm{E}(\mathbf{x})\} = \mathbf{R}\boldsymbol{\Sigma}^{-1/2}\mathbf{x} = \mathbf{O}\mathbf{z},$$

where $\mathbf{z}$ is as described in the theorem statement. □

A.3. Proofs of the results in Section 4

Proof of Lemma 2. Let $\mathbf{A}$ and $\mathbf{b}$ be an invertible $p\times p$ -matrix and p-vector, respectively. Then

$$\boldsymbol{\theta}(\mathbf{A}^\top\mathbf{x}+\mathbf{b}) = \mathbf{C}_2^{-1/2}(\mathbf{A}^\top\mathbf{x}+\mathbf{b})\mathbf{c}_3(\mathbf{C}_2^{-1/2}(\mathbf{A}^\top\mathbf{x}+\mathbf{b})\mathbf{A}^\top\{\mathbf{x}-\mathrm{E}(\mathbf{x})\}).$$

As shown in the proof of Lemma 6, $\mathbf{C}_2^{-1/2}(\mathbf{A}^\top\mathbf{x}+\mathbf{b}) = \mathbf{U}\mathbf{C}_2^{-1/2}(\mathbf{x})(\mathbf{A}^{-1})^\top$, for some orthogonal matrix $\mathbf{U}$ that makes $\mathbf{C}_2^{-1/2}(\mathbf{A}^\top\mathbf{x}+\mathbf{b})$ symmetric. Therefore, $(\mathbf{A}^\top\mathbf{x}+\mathbf{b})_{\mathrm{w}} = \mathbf{U}\mathbf{x}_{\mathrm{w}}$. Furthermore, it is straightforward to show that for any orthogonal matrix $\mathbf{O}$, $\mathbf{c}_3(\mathbf{O}\mathbf{x}) = \mathbf{O}\mathbf{c}_3(\mathbf{x})$. And finally, since $\mathbf{U}$ is such that $\mathbf{C}_2^{-1/2}(\mathbf{A}^\top\mathbf{x}+\mathbf{b})$ is symmetric, we can write $\mathbf{C}_2^{-1/2}(\mathbf{A}^\top\mathbf{x}+\mathbf{b}) = \mathbf{A}^{-1}\mathbf{C}_2^{-1/2}(\mathbf{x})\mathbf{U}^\top$, which finally gives

$$\boldsymbol{\theta}(\mathbf{A}^\top\mathbf{x}+\mathbf{b}) = \mathbf{C}_2^{-1/2}(\mathbf{A}^\top\mathbf{x}+\mathbf{b})\mathbf{c}_3((\mathbf{A}^\top\mathbf{x}+\mathbf{b})_{\mathrm{w}}) = \mathbf{A}^{-1}\mathbf{C}_2^{-1/2}(\mathbf{x})\mathbf{c}_3(\mathbf{x}_{\mathrm{w}}) = \mathbf{A}^{-1}\boldsymbol{\theta}(\mathbf{x}). \quad \square \quad \square$$

Proof of Lemma 3. By AE, it is sufficient to assume that

$$\mathbf{z} \sim \alpha_1\mathcal{N}_p(-\alpha_2\mathbf{m}, \mathbf{I}_p) + \alpha_2\mathcal{N}_p(\alpha_1\mathbf{m}, \mathbf{I}_p),$$

and show that $\boldsymbol{\theta}_R(\mathbf{z})$ is proportional to $\mathbf{m} := \boldsymbol{\Sigma}^{-1/2}\mathbf{h}$. Now, by the proof of Theorem 3, we have $\mathbf{C}_2(\mathbf{z})^{-1/2} = \mathbf{I} - (1-d^{1/2})\mathbf{w}\mathbf{w}^\top$ where $d = (1+\beta\|\mathbf{m}\|^2)^{-1}$ and $\mathbf{w} := \mathbf{m}/\|\mathbf{m}\|$. Hence,

$$\mathbf{z}_w \sim \alpha_1\mathcal{N}_p(-\alpha_2 d^{1/2}\mathbf{m}, \mathbf{C}_2(\mathbf{z})^{-1}) + \alpha_2\mathcal{N}_p(\alpha_1 d^{1/2}\mathbf{m}, \mathbf{C}_2(\mathbf{z})^{-1}),$$

and, by Lemma 11, we have $\mathbf{c}_3(\mathbf{z}_w) = \beta\gamma d^{3/2}\mathbf{m}\mathbf{m}^\top\mathbf{m}$. The claim now follows, after recalling that $\tau = \mathbf{h}^\top\boldsymbol{\Sigma}^{-1}\mathbf{h} = \|\mathbf{m}\|^2$. □

Throughout the remainder of this section, without loss of generality, we assume $\mathbf{x} \sim \alpha_1\mathcal{N}(-\alpha_2\mathbf{m}, \mathbf{I}_p) + \alpha_2\mathcal{N}(\alpha_1\mathbf{m}, \mathbf{I}_p)$, where $\mathbf{m} := \boldsymbol{\Sigma}^{-1/2}\mathbf{h}$. For simplicity of the notation, we denote $\mathbf{A} := \mathbf{C}_2{}^{-1/2}(\mathbf{x})$ and $\hat{\mathbf{A}} := \hat{\mathbf{C}}_2{}^{-1/2}(\mathbf{x}_i) := \left(\frac{1}{n}\sum_{i=1}^n \tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top\right)^{-1/2}$, where $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \bar{\mathbf{x}}$. We also use the notation $\Delta^2 = (1+\beta\tau)^{-1}$ where we recall that $\tau = \|\mathbf{m}\|^2$.

Lemma 15. *Under the above assumptions* $\mathrm{E}(\mathbf{x}\mathbf{x}^\top\mathbf{A}^2\mathbf{x}) = \beta\gamma\Delta^2\tau\mathbf{m}$.

Proof of Lemma 15. To prove the statement we use the fact that the moments of the multivariate normal mixture are convex combinations of the moments of its normal components, as well as the multivariate Stein's lemma [21] with $g(\mathbf{x}) = \mathbf{x}^\top\mathbf{A}^2\mathbf{x}$. By further observing that $\mathrm{tr}(\mathbf{A}^2\mathbf{m}\mathbf{m}^\top) = \mathbf{m}^\top\mathbf{A}^2\mathbf{m} = \Delta^2\tau$, we obtain the claim. □

Proof of Theorem 5. We start by the linearization of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$;

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \sqrt{n}(\hat{\mathbf{A}}\hat{\mathbf{c}}_3 - \mathbf{A}\mathbf{c}_3) = \sqrt{n}(\hat{\mathbf{A}} - \mathbf{A})\hat{\mathbf{c}}_3 + \mathbf{A}\sqrt{n}(\hat{\mathbf{c}}_3 - \mathbf{c}_3) = \sqrt{n}(\hat{\mathbf{A}} - \mathbf{A})\mathbf{c}_3 + \mathbf{A}\sqrt{n}(\hat{\mathbf{c}}_3 - \mathbf{c}_3) + o_P(1). \tag{A.7}$$

Since $\mathbf{c}_3 = \mathbf{A}\mathrm{E}(\mathbf{x}\mathbf{x}^\top\mathbf{A}^2\mathbf{x})$, Lemma 15 implies that $\mathbf{c}_3 = \beta\gamma\Delta^3\tau\mathbf{m}$. Thus,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \beta\gamma\Delta^3\tau\sqrt{n}(\hat{\mathbf{A}}\mathbf{m} - \mathbf{A}\mathbf{m}) + \mathbf{A}\sqrt{n}(\hat{\mathbf{c}}_3 - \mathbf{c}_3) + o_P(1). \tag{A.8}$$

Denote now $\mathrm{I} = \beta\gamma\Delta^3\tau\sqrt{n}(\hat{\mathbf{A}}\mathbf{m} - \mathbf{A}\mathbf{m})$ and $\mathrm{II} = \mathbf{A}\sqrt{n}(\hat{\mathbf{c}}_3 - \mathbf{c}_3)$, and let us focus first on the expression II. We again begin by linearization, this time of $\hat{\mathbf{c}}_3$:

$$\begin{aligned}\sqrt{n}(\hat{\mathbf{c}}_3 - \mathbf{c}_3) &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \hat{\mathbf{A}}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i{}^\top\hat{\mathbf{A}}^\top\hat{\mathbf{A}}\tilde{\mathbf{x}}_i - \mathbf{c}_3\right)\\ &= \sqrt{n}(\hat{\mathbf{A}} - \mathbf{A})\frac{1}{n}\sum_{i=1}^n \tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top\hat{\mathbf{A}}^2\tilde{\mathbf{x}}_i + \mathbf{A}\frac{1}{n}\sum_{i=1}^n \tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top\sqrt{n}(\hat{\mathbf{A}}^2 - \mathbf{A}^2)\tilde{\mathbf{x}}_i + \mathbf{A}\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top\mathbf{A}^2\tilde{\mathbf{x}}_i - \mathbf{c}_3\right).\end{aligned} \tag{A.9}$$

We examine now all three expressions in (A.9) separately.

$$\sqrt{n}(\hat{\mathbf{A}} - \mathbf{A})\frac{1}{n}\sum_{i=1}^n \tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top\hat{\mathbf{A}}^2\tilde{\mathbf{x}}_i = \sqrt{n}(\hat{\mathbf{A}} - \mathbf{A})\mathrm{E}\left(\mathbf{x}\mathbf{x}^\top\mathbf{A}^2\mathbf{x}\right) + o_P(1) = \beta\gamma\Delta^2\tau\sqrt{n}(\hat{\mathbf{A}}\mathbf{m} - \mathbf{A}\mathbf{m}) + o_P(1). \tag{A.10}$$

Furthermore,

$$\begin{aligned}\mathbf{A}\frac{1}{n}\sum_{i=1}^n \tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top\sqrt{n}(\hat{\mathbf{A}}^2 - \mathbf{A}^2)\tilde{\mathbf{x}}_i &= \mathbf{A}\frac{1}{n}\sum_{i=1}^n \tilde{\mathbf{x}}_i\mathrm{vec}\left\{\tilde{\mathbf{x}}_i^\top\sqrt{n}(\hat{\mathbf{A}}^2 - \mathbf{A}^2)\tilde{\mathbf{x}}_i\right\} = \mathbf{A}\mathrm{E}\{\mathbf{x}(\mathbf{x}\otimes\mathbf{x})^\top\}\sqrt{n}\mathrm{vec}\left(\hat{\mathbf{A}}^2 - \mathbf{A}^2\right) + o_P(1)\\ &= \beta\gamma\mathbf{A}\mathbf{m}(\mathbf{m}\otimes\mathbf{m})\sqrt{n}\mathrm{vec}\left(\hat{\mathbf{A}}^2 - \mathbf{A}^2\right) + o_P(1),\end{aligned}$$

which turns out to be a negligible term in the expression for the limiting covariance of the estimator standardized to unit norm as $\mathbf{A}\mathbf{m} \propto \mathbf{m}$ and since the final linearization will in the end by multiplied from the left by $\mathbf{Q}_{\boldsymbol{\theta}}\boldsymbol{\Sigma}^{-1/2}$, giving $\mathbf{Q}_{\boldsymbol{\theta}}\boldsymbol{\Sigma}^{-1/2}\mathbf{m} = \mathbf{0}$. Thus, we ignore this term in the sequel. Finally, we focus on a final term in (A.9).

$$\begin{aligned}\mathbf{A}\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top\mathbf{A}^2\tilde{\mathbf{x}}_i - \mathbf{c}_3\right) =& \sqrt{n}(\hat{\mathbf{c}}_{0,3} - \mathbf{c}_3) - \mathbf{A}\sqrt{n}\bar{\mathbf{x}}\frac{1}{n}\sum_{i=1}^n \tilde{\mathbf{x}}_i^\top\mathbf{A}^2\tilde{\mathbf{x}}_i - \mathbf{A}\sqrt{n}\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\bar{\mathbf{x}}^\top\mathbf{A}^2\tilde{\mathbf{x}}_i - \mathbf{A}\sqrt{n}\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\mathbf{A}^2\bar{\mathbf{x}}\\ &+ o_P(1),\end{aligned} \tag{A.11}$$

where $\hat{\mathbf{c}}_{0,3} := \frac{1}{n}\sum_{i=1}^n \mathbf{A}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{A}^2\mathbf{x}_i$. Since, $\sqrt{n}\bar{\mathbf{x}} = O_P(1)$ and $\bar{\mathbf{x}} = o_P(1)$,

$$\mathbf{A}\sqrt{n}\bar{\mathbf{x}}\frac{1}{n}\sum_{i=1}^n \tilde{\mathbf{x}}_i^\top\mathbf{A}^2\tilde{\mathbf{x}}_i = \mathbf{A}\sqrt{n}\bar{\mathbf{x}}\mathrm{E}(\mathbf{x}^\top\mathbf{A}^2\mathbf{x}) + o_P(1) = p\mathbf{A}\sqrt{n}\bar{\mathbf{x}},$$

since $\mathbf{A}\mathbf{x}$ is standardized to have a unit covariance matrix. Furthermore,

$$\mathbf{A}\sqrt{n}\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\mathbf{A}^2\bar{\mathbf{x}} = \mathbf{A}\sqrt{n}\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\bar{\mathbf{x}}^\top\mathbf{A}^2\mathbf{x}_i = \mathbf{A}\mathrm{E}(\mathbf{x}\mathbf{x}^\top)\mathbf{A}^2\sqrt{n}\bar{\mathbf{x}} + o_P(1).$$

Observe first that $\mathrm{E}(\mathbf{x}\mathbf{x}^\top) = \mathbf{I}_p + \beta\mathbf{m}\mathbf{m}^\top$. As above, since we are interested in the limiting covariance of the normalized estimator $\hat{\boldsymbol{\theta}}/\|\hat{\boldsymbol{\theta}}\|$, we can neglect any term starting with $\mathbf{m}$, or $\mathbf{A}\mathbf{m}$ for that matter, as any such term will vanish in the end. For the same reason, multiplication with matrix $\mathbf{A}$ from the very left, has the same impact to the final limiting covariance as multiplication by an identity, and can therefore be omitted. Hence, as before, we can write

$$\mathbf{A}\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top\mathbf{A}^2\tilde{\mathbf{x}}_i - \mathbf{c}_3\right) = \sqrt{n}(\hat{\mathbf{c}}_{0,3} - \mathbf{c}_3) - (p+2)\sqrt{n}\bar{\mathbf{x}} + o_P(1). \tag{A.12}$$

Therefore, we obtain

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \beta\gamma\Delta^2(\Delta+1)\tau\sqrt{n}(\hat{\mathbf{A}}\mathbf{m} - \mathbf{A}\mathbf{m}) - (p+2)\sqrt{n}\bar{\mathbf{x}} + \sqrt{n}(\hat{\mathbf{c}}_{0,3} - \mathbf{c}_3) + o_P(1). \tag{A.13}$$

As for the final part of the linearization, we focus on the first term in the sum in (A.13) and express it as a linear function of $\sqrt{n}(\hat{\mathbf{C}}_{0,2} - \mathbf{C}_2)\mathbf{m}$, where $\hat{\mathbf{C}}_{0,2} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$. As in the proof of Theorem 3,

$$\sqrt{n}\mathrm{vec}(\hat{\mathbf{A}} - \mathbf{A}) = -(\mathbf{I}_p \otimes \mathbf{A}^{-1} + \mathbf{A}^{-1} \otimes \mathbf{I}_p)^{-1}(\mathbf{A} \otimes \mathbf{A})\sqrt{n}\mathrm{vec}(\hat{\mathbf{C}}_{0,2} - \mathbf{C}_2) + o_P(1).$$

Now, since $\mathbf{C}_2 = \mathbf{I}_p + \beta\mathbf{m}\mathbf{m}^\top$ is a rank-1 perturbation of identity, and $\mathbf{C}_2\mathbf{m} = \Delta^{-2}\mathbf{m}$, we obtain $\mathbf{A}^{-1}\mathbf{m} = \mathbf{C}_2^{1/2}\mathbf{m} = \Delta^{-1}\mathbf{m}$, giving $\mathbf{A}^{-1} = \mathbf{I}_p + (\Delta^{-1} - 1)(\mathbf{m}\mathbf{m}^\top)/\tau$. Lemma 14 then implies that

$$(\mathbf{I}_p \otimes \mathbf{A}^{-1} + \mathbf{A}^{-1} \otimes \mathbf{I}_p)^{-1} = \frac{1}{2(\Delta^{-1} + 1)}(\mathbf{I}_p \otimes \mathbf{I}_p + \Delta^{-1}\mathbf{A} \otimes \mathbf{A}),$$

further giving

$$\sqrt{n}\mathrm{vec}(\hat{\mathbf{A}} - \mathbf{A}) = \frac{-1}{2(\Delta^{-1} + 1)}\left\{\mathbf{A} \otimes \mathbf{A} + \Delta^{-1}(\mathbf{A}^2 \otimes \mathbf{A}^2)\right\}\sqrt{n}\mathrm{vec}(\hat{\mathbf{C}}_{0,2} - \mathbf{C}_2). \tag{A.14}$$

Unvectorizing (A.14) gives

$$\sqrt{n}(\hat{\mathbf{A}} - \mathbf{A}) = \frac{-1}{2(\Delta^{-1} + 1)}\left\{\mathbf{A}\sqrt{n}(\hat{\mathbf{C}}_{0,2} - \mathbf{C}_2)\mathbf{A} + \Delta^{-1}\mathbf{A}^2\sqrt{n}(\hat{\mathbf{C}}_{0,2} - \mathbf{C}_2)\mathbf{A}^2\right\} + o_P(1).$$

Same as above, we can ignore the multiplications by $\mathbf{A}$ and $\mathbf{A}^2$ from the left, and finally obtain

$$\sqrt{n}(\hat{\mathbf{A}} - \mathbf{A})\mathbf{m} = \frac{-\Delta^2}{\Delta + 1}\sqrt{n}(\hat{\mathbf{C}}_{0,2} - \mathbf{C}_2)\mathbf{m} + o_P(1). \tag{A.15}$$

Plugging (A.15) into (A.13) gives

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = -(p+2)\sqrt{n}\bar{\mathbf{x}} - \beta\gamma\Delta^4\tau\sqrt{n}(\hat{\mathbf{C}}_{0,2} - \mathbf{C}_2)\mathbf{m} + \sqrt{n}(\hat{\mathbf{c}}_{0,3} - \mathbf{c}_3) + o_P(1), \tag{A.16}$$

showing that $\hat{\boldsymbol{\theta}}$, and consequently $\hat{\boldsymbol{\theta}}/\|\hat{\boldsymbol{\theta}}\|$, has a limiting normal distribution.

To conclude the proof, we still compute the constant C in the limiting covariance matrix. We have $\boldsymbol{\theta} = \beta\gamma\tau\Delta^4\mathbf{m}$ and $\|\boldsymbol{\theta}\|^2 = \beta^2\gamma^2\tau^3\Delta^8$. Consequently, the proof of Theorem 3 reveals that $\mathbf{t}^\top\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \rightsquigarrow \mathcal{N}(0, \beta^2\gamma^2\tau^3\Delta^8 C)$, for any unit-length vector $\mathbf{t}$ satisfying $\mathbf{t}^\top\mathbf{m} = 0$. By working in an orthogonal basis given by $(\mathbf{m}/\|\mathbf{m}\|, \mathbf{t}, \mathbf{u}_3, \ldots, \mathbf{u}_p)$, where the final $p-2$ vectors are arbitrary, we thus have, by the CLT and (A.15) that

$$\beta^2\gamma^2\tau^3\Delta^8 C = \mathrm{Var}\{-(p+2)X_2 - \beta\gamma\Delta^4\tau^{3/2}X_1X_2 + X_2(\Delta^2X_1^2 + X_2^2 + X_3^2 + \cdots + X_p^2)\},$$

where $X_1 \sim \alpha_1\mathcal{N}(-\alpha_2\tau^{1/2}, 1) + \alpha_2\mathcal{N}(\alpha_1\tau^{1/2}, 1)$, $X_2, \ldots, X_p \sim \mathcal{N}(0,1)$ and all these variables are independent of each other. We have the moments $\mathrm{E}(X_1) = 0$, $\mathrm{E}(X_1^2) = 1 + \beta\tau$, $\mathrm{E}(X_1^3) = \beta\gamma\tau^{3/2}$ and $\mathrm{E}(X_1^4) = \beta(1-3\beta)\tau^2 + 6\beta\tau + 3$, allowing us to simplify the above equation to $\beta^2\gamma^2\tau^3\Delta^8 C = 2p + 1 - \beta^2\gamma^2\Delta^6\tau^3 + \Delta^4\mathrm{E}(X_1^4)$. Observing now that $\gamma^2 = 1 - 4\beta$, the desired limiting normal distribution follows. □

A.4. Proofs of the results in Section 5

Proof of Lemma 4. Letting $\mathbf{y}, \mathbf{z}$ denote independent copies of $\mathbf{x}_w$, the right-hand side writes as $\mathrm{E}\{\mathbf{y}(\mathbf{y}\otimes\mathbf{y})^\top(\mathbf{z}\otimes\mathbf{z})\mathbf{z}^\top\} = \mathrm{E}\{(\mathbf{y}^\top\mathbf{z})^2\mathbf{y}\mathbf{z}^\top\}$. Whereas, the left-hand side takes the form $\sum_{k=1}^p \mathrm{E}(\mathbf{y}\mathbf{y}^\top\mathbf{e}_k\mathbf{y}^\top\mathbf{z}\mathbf{e}_k^\top\mathbf{z}\mathbf{z}^\top) = \mathrm{E}\{(\mathbf{y}^\top\mathbf{z})^2\mathbf{y}\mathbf{z}^\top\}$, proving the claim. □

Proof of Lemma 5. Letting $\mathbf{A} := \mathbf{C}_2^{-1/2}$, we have

$$\mathbf{T}_k(\mathbf{A}(\mathbf{x} - \mathrm{E}(\mathbf{x}))) = \mathbf{A}\mathrm{E}\{(\mathbf{e}_k^\top\mathbf{A}\mathbf{x})\cdot\mathbf{x}\mathbf{x}^\top\}\mathbf{A} = \beta\gamma(\mathbf{e}_k^\top\mathbf{A}\mathbf{h})\mathbf{A}\mathbf{h}\mathbf{h}^\top\mathbf{A},$$

by the moment formulas used also in deriving Lemma 11. Consequently,

$$\mathbf{T}(\mathbf{A}(\mathbf{x} - \mathrm{E}(\mathbf{x}))) = \sum_{k=1}^p \mathbf{T}_k(\mathbf{A}\mathbf{x})^2 = \beta^2\gamma^2(\mathbf{h}^\top\mathbf{A}^2\mathbf{h})^2\mathbf{A}\mathbf{h}\mathbf{h}^\top\mathbf{A}.$$

By the Sherman–Morrison formula,

$$\mathbf{A}^2\mathbf{h} = (\boldsymbol{\Sigma} + \beta\mathbf{h}\mathbf{h}^\top)^{-1}\mathbf{h} = \frac{1}{1+\beta\tau}\boldsymbol{\theta}, \quad \mathbf{h}^\top\mathbf{A}^2\mathbf{h} = \frac{\tau}{1+\beta\tau},$$

where $\tau = \mathbf{h}^\top\boldsymbol{\Sigma}^{-1}\mathbf{h}$. Thus, $\mathbf{u} = s\mathbf{A}\mathbf{h}/\|\mathbf{A}\mathbf{h}\|$ for some sign s, and

$$\mathbf{C}_2^{-1/2}\mathbf{u} = s\frac{\mathbf{A}^2\mathbf{h}}{\|\mathbf{A}\mathbf{h}\|} = s\{\tau(1+\beta\tau)\}^{-1/2}\boldsymbol{\theta},$$

completing the proof. □

Proof of Lemma 6. We have $\hat{\mathbf{C}}_2(\mathbf{A}^\top\mathbf{x}_i + \mathbf{b}) = \mathbf{A}^\top\hat{\mathbf{C}}_2(\mathbf{x}_i)\mathbf{A}$. Consequently, by Theorem 2.1 in [7],

$$\hat{\mathbf{C}}_2(\mathbf{A}^\top\mathbf{x}_i + \mathbf{b})^{-1/2} = \mathbf{V}\hat{\mathbf{C}}_2(\mathbf{x}_i)^{-1/2}(\mathbf{A}^{-1})^\top, \tag{A.17}$$

where $\mathbf{V}$ is the unique orthogonal matrix that makes $\mathbf{V}\hat{\mathbf{C}}_2(\mathbf{x}_i)^{-1/2}(\mathbf{A}^{-1})^\top$ symmetric. Hence, writing $\tilde{\mathbf{x}}_i := \mathbf{x}_i - \bar{\mathbf{x}}$,

$$\hat{\mathbf{T}}_k(\hat{\mathbf{C}}_2(\mathbf{A}^\top\mathbf{x}_i + \mathbf{b})^{-1/2}\mathbf{A}^\top(\mathbf{x}_i - \bar{\mathbf{x}})) = \frac{1}{n}\sum_{i=1}^n \mathbf{V}\hat{\mathbf{C}}_2(\mathbf{x}_i)^{-1/2}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top\hat{\mathbf{C}}_2(\mathbf{x}_i)^{-1/2}\mathbf{V}^\top\mathbf{e}_k\tilde{\mathbf{x}}_i^\top\hat{\mathbf{C}}_2(\mathbf{x}_i)^{-1/2}\mathbf{V}^\top.$$

Consequently,

$$\hat{\mathbf{T}}(\hat{\mathbf{C}}_2(\mathbf{A}^\top\mathbf{x}_i+\mathbf{b})^{-1/2}\mathbf{A}^\top(\mathbf{x}_i-\bar{\mathbf{x}})) = \frac{1}{n^2}\sum_{k=1}^n\sum_{j=1}^n \mathbf{V}\hat{\mathbf{C}}_2(\mathbf{x}_i)^{-1/2}\tilde{\mathbf{x}}_k\tilde{\mathbf{x}}_k^\top\hat{\mathbf{C}}_2(\mathbf{x}_i)^{-1}\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_k^\top\hat{\mathbf{C}}_2(\mathbf{x}_i)^{-1}\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_j^\top\hat{\mathbf{C}}_2(\mathbf{x}_i)^{-1/2}\mathbf{V}^\top$$
$$= \mathbf{V}\hat{\mathbf{T}}(\hat{\mathbf{C}}_2(\mathbf{x}_i)^{-1/2}(\mathbf{x}_i-\bar{\mathbf{x}}))\mathbf{V}^\top.$$

As the normal mixture is absolutely continuous w.r.t. the Lebesgue measure, the matrix $\hat{\mathbf{T}}(\hat{\mathbf{C}}_2(\mathbf{x}_i)^{-1/2}(\mathbf{x}_i-\bar{\mathbf{x}}))$ has almost surely distinct eigenvalues. Thus, almost surely,

$$\hat{\mathbf{u}}(\mathbf{A}^\top\mathbf{x}_i+\mathbf{b}) = s\mathbf{V}\hat{\mathbf{u}}(\mathbf{x}_i),$$

for some sign s. The desired result now follows by combining the above with (A.17). □

We begin with an auxiliary result that shows how to obtain the limiting distribution of the leading eigenvector of an estimator that converges to a rank-one matrix. Its proof is exactly analogous to the proof of Theorem A.1.2 in [6] and we refrain from including it here.

Lemma 16. *Assume that the n-indexed sequence $\hat{\mathbf{H}}$ of estimators, taking values in the set of symmetric $p\times p$ matrices, satisfies $\sqrt{n}(\hat{\mathbf{H}}-\psi\mathbf{h}\mathbf{h}^\top)=\mathcal{O}_P(1)$ for some $\psi>0$ and $\mathbf{h}\in\mathbb{R}^p$, $\|\mathbf{h}\|=1$. Then, letting $\hat{\mathbf{u}}$ denote any leading eigenvector of $\hat{\mathbf{H}}$, we have,*

$$\sqrt{n}(\hat{s}\hat{\mathbf{u}}-\mathbf{h}) = \frac{1}{\psi}\left(\mathbf{I}_p-\mathbf{h}\mathbf{h}^\top\right)\sqrt{n}(\hat{\mathbf{H}}-\psi\mathbf{h}\mathbf{h}^\top)\mathbf{h}+o_P(1),$$

for some n-indexed sequence of signs $\hat{s}$.

Let now $\mathbf{z}\sim\alpha_1\mathcal{N}(-\alpha_2\mathbf{m},\mathbf{I}_p)+\alpha_2\mathcal{N}(\alpha_1\mathbf{m},\mathbf{I}_p)$, where $\mathbf{m}:=\boldsymbol{\Sigma}^{-1/2}\mathbf{h}$. We next provide a linearization of the estimator $\hat{s}_z\hat{\boldsymbol{\theta}}_{\mathrm{L}}(\mathbf{z}_i)$, omitting, for convenience, both the signs $\hat{s}_z$ and the parenthesis notation specifying the sample we use to be $\mathbf{z}_i$.

Lemma 17. *We have*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{\mathrm{L}}-\{\tau(1+\beta\tau)\}^{-1/2}\mathbf{m}\right) = -\frac{\Delta}{\|\mathbf{m}\|}\left(\mathbf{Q}_{\mathbf{m}}+\frac{1}{2}\Delta^2\frac{\mathbf{m}\mathbf{m}^\top}{\|\mathbf{m}\|^2}\right)\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^n(\mathbf{m}^\top\mathbf{z}_i)\mathbf{z}_i-(1+\beta\tau)\mathbf{m}\right\}$$
$$-\frac{1}{\beta\gamma\tau\Delta^3\|\mathbf{m}\|}\mathbf{Q}_{\mathbf{m}}\sqrt{n}\bar{\mathbf{z}}+\frac{1}{\beta\gamma\tau^2\Delta\|\mathbf{m}\|}\mathbf{Q}_{\mathbf{m}}\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^n(\mathbf{m}^\top\mathbf{z}_i)^2\mathbf{z}_i-\beta\gamma\tau^2\mathbf{m}\right\}+o_P(1),$$

where $\Delta^2:=1/(1+\beta\tau)$ and $\mathbf{Q}_{\mathbf{m}}$ is the projection onto the orthogonal complement of $\mathbf{m}$.

Proof of Lemma 17. Linearization and Lemma 16 together give,

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{\mathrm{L}}-\{\tau(1+\beta\tau)\}^{-1/2}\mathbf{m}\right) = \sqrt{n}(\hat{\mathbf{A}}-\mathbf{A})\frac{\mathbf{m}}{\|\mathbf{m}\|}+\mathbf{A}\sqrt{n}\left(\hat{\mathbf{u}}-\frac{\mathbf{m}}{\|\mathbf{m}\|}\right)+o_P(1)$$
$$=\sqrt{n}(\hat{\mathbf{A}}-\mathbf{A})\frac{\mathbf{m}}{\|\mathbf{m}\|}+\frac{1}{\lambda}\mathbf{Q}_{\mathbf{m}}\sqrt{n}\left(\hat{\mathbf{T}}-\lambda\frac{\mathbf{m}\mathbf{m}^\top}{\|\mathbf{m}\|^2}\right)\frac{\mathbf{m}}{\|\mathbf{m}\|}+o_P(1), \tag{A.18}$$

where $\hat{\mathbf{A}}:=\hat{\mathbf{C}}_2^{-1/2}\to_p\mathbf{C}_2^{-1/2}=:\mathbf{A}$, $\lambda=\beta^2\gamma^2\tau^3\Delta^6$ is the leading (and only non-zero) eigenvalue of $\mathbf{T}$, $\Delta=(1+\beta\tau)^{-1/2}$ and $\mathbf{Q}_{\mathbf{m}}$ is the projection onto the orthogonal complement of $\mathbf{m}$.

As our first task, we obtain an expression for the second part of (A.18) by linearizing $\hat{\mathbf{T}}$: $\sqrt{n}(\hat{\mathbf{T}}-\mathbf{T})=\sum_{k=1}^p\{\sqrt{n}(\hat{\mathbf{T}}_k-\mathbf{T}_k)\mathbf{T}_k+\mathbf{T}_k\sqrt{n}(\hat{\mathbf{T}}_k-\mathbf{T}_k)\}+o_P(1)$. Now, $\mathbf{T}_k$ is proportional to $\mathbf{m}\mathbf{m}^\top$, meaning that $\mathbf{Q}_{\mathbf{m}}\mathbf{T}_k=\mathbf{0}$, allowing us to ignore the second term inside the sum above (as it later gets plugged in into (A.18) and canceled by $\mathbf{Q}_{\mathbf{m}}$ there). We thus next linearize $\sqrt{n}(\hat{\mathbf{T}}_k-\mathbf{T}_k)$, which essentially amounts to replacing each $\hat{\mathbf{A}}$ with $(\hat{\mathbf{A}}-\mathbf{A})+\mathbf{A}$ and splitting each $\mathbf{x}_i-\bar{\mathbf{x}}$. In this, we use the fact that $\mathbf{A}\mathbf{m}=\Delta\mathbf{m}$, obtained by simplifying $\mathbf{A}(\mathbf{I}+\beta\mathbf{m}\mathbf{m}^\top)\mathbf{A}-\mathbf{A}^2$. Thus, any resulting terms in the linearization that are proportional to identity or that start with either "$\mathbf{m}$" or "$\mathbf{A}\mathbf{m}$" get later canceled by the projection $\mathbf{Q}_{\mathbf{m}}$ and, ignoring them, we obtain,

$$\sqrt{n}(\hat{\mathbf{T}}_k-\mathbf{T}_k) = \frac{1}{\sqrt{n}}\sum_{i=1}^n\hat{\mathbf{A}}(\mathbf{z}_i-\bar{\mathbf{z}})(\mathbf{z}_i-\bar{\mathbf{z}})^\top\hat{\mathbf{A}}\mathbf{e}_k(\mathbf{z}_i-\bar{\mathbf{z}})^\top\hat{\mathbf{A}}-\sqrt{n}\mathbf{T}_k$$
$$=\sqrt{n}(\hat{\mathbf{A}}-\mathbf{A})\mathrm{E}(\mathbf{z}\mathbf{z}^\top\mathbf{A}\mathbf{e}_k\mathbf{z}^\top)\mathbf{A}-\mathbf{A}\sqrt{n}\bar{\mathbf{z}}\mathbf{e}_k^\top-\mathbf{e}_k\sqrt{n}\bar{\mathbf{z}}^\top\mathbf{A}+\frac{1}{\sqrt{n}}\sum_{i=1}^n\mathbf{A}\mathbf{z}_i\mathbf{z}_i^\top\mathbf{A}\mathbf{e}_k\mathbf{z}_i^\top\mathbf{A}-\sqrt{n}\mathbf{T}_k+o_P(1).$$

Using the facts that $\mathrm{E}(\mathbf{z}\mathbf{z}^\top\mathbf{A}\mathbf{e}_k\mathbf{z}^\top)=\beta\gamma\Delta\mathbf{m}\mathbf{m}^\top\mathbf{e}_k\mathbf{m}^\top$ and $\mathbf{T}_k=\beta\gamma\Delta^3\mathbf{m}\mathbf{m}^\top\mathbf{e}_k\mathbf{m}^\top$, we thus get,

$$\sum_{k=1}^p\sqrt{n}(\hat{\mathbf{T}}_k-\mathbf{T}_k)\mathbf{T}_k=\sqrt{n}(\hat{\mathbf{A}}-\mathbf{A})\beta^2\gamma^2\tau^2\Delta^5\mathbf{m}\mathbf{m}^\top-\beta\gamma\tau\Delta^3\mathbf{A}\sqrt{n}\bar{\mathbf{z}}\mathbf{m}^\top+\sqrt{n}\left(\beta\gamma\Delta^5\frac{1}{n}\sum_{i=1}^n\mathbf{A}\mathbf{z}_i\mathbf{z}_i^\top\mathbf{m}\mathbf{z}_i^\top\mathbf{m}\mathbf{m}^\top-\mathbf{T}\right)+o_P(1).$$

Plugging now in to (A.18), and using $\mathbf{Q}_{\mathbf{m}}\mathbf{A}=\mathbf{Q}_{\mathbf{m}}$, we get

$$\sqrt{n}(\hat{\mathbf{A}}-\mathbf{A})\frac{\mathbf{m}}{\|\mathbf{m}\|}+\frac{1}{\lambda}\mathbf{Q}_{\mathbf{m}}\sqrt{n}\left(\hat{\mathbf{T}}-\lambda\frac{\mathbf{m}\mathbf{m}^\top}{\|\mathbf{m}\|^2}\right)\frac{\mathbf{m}}{\|\mathbf{m}\|}$$

$$=\sqrt{n}(\hat{\mathbf{A}}-\mathbf{A})\frac{\mathbf{m}}{\|\mathbf{m}\|}+\frac{1}{\Delta}\mathbf{Q}_{\mathbf{m}}\sqrt{n}(\hat{\mathbf{A}}-\mathbf{A})\frac{\mathbf{m}}{\|\mathbf{m}\|}-\frac{1}{\beta\gamma\tau\Delta^3\|\mathbf{m}\|}\mathbf{Q}_{\mathbf{m}}\sqrt{n}\bar{\mathbf{z}}+\frac{1}{\beta\gamma\tau^2\Delta\|\mathbf{m}\|}\mathbf{Q}_{\mathbf{m}}\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbf{m}^\top\mathbf{z}_i)^2\mathbf{z}_i-\beta\gamma\tau^2\mathbf{m}\right).$$

What remains now is the simplification of the terms involving $\sqrt{n}(\hat{\mathbf{A}}-\mathbf{A})$. For these, linearizing the relation $\sqrt{n}(\hat{\mathbf{A}}\hat{\mathbf{C}}_2\hat{\mathbf{A}}-\mathbf{I}_p)=\mathbf{0}$ yields

$$\mathbf{A}\sqrt{n}(\hat{\mathbf{C}}_2-\mathbf{C}_2)\mathbf{A}=-\sqrt{n}(\hat{\mathbf{A}}-\mathbf{A})\mathbf{C}_2^{1/2}-\mathbf{C}_2^{1/2}\sqrt{n}(\hat{\mathbf{A}}-\mathbf{A})+o_P(1).$$

Using $\sqrt{n}(\hat{\mathbf{C}}_2-\mathbf{C}_2)=\sqrt{n}(\hat{\mathbf{C}}_{02}-\mathbf{C}_2)+o_P(1)$ from the proof of Lemma 9 and vectorizing gives

$$(\mathbf{I}_p\otimes\mathbf{C}_2^{1/2}+\mathbf{C}_2^{1/2}\otimes\mathbf{I}_p)\sqrt{n}\mathrm{vec}(\hat{\mathbf{A}}-\mathbf{A})=-(\mathbf{A}\otimes\mathbf{A})\sqrt{n}\mathrm{vec}(\hat{\mathbf{C}}_{02}-\mathbf{C}_2).$$

Now, $\mathbf{C}_2^{1/2}=(\mathbf{I}_p+\beta\mathbf{m}\mathbf{m}^\top)^{1/2}=\mathbf{I}_p+\alpha\mathbf{m}\mathbf{m}^\top/\|\mathbf{m}\|^2$ where $\alpha=\sqrt{1+\beta\tau}-1$, meaning that Lemma 14 gives

$$(\mathbf{I}_p\otimes\mathbf{C}_2^{1/2}+\mathbf{C}_2^{1/2}\otimes\mathbf{I}_p)^{-1}=\frac{1}{2(\alpha+2)}\left[\mathbf{I}_p\otimes\mathbf{I}_p+(\alpha+1)\left\{\left(\mathbf{I}_p-\frac{\alpha}{\alpha+1}\frac{\mathbf{m}\mathbf{m}^\top}{\|\mathbf{m}\|^2}\right)\otimes\left(\mathbf{I}_p-\frac{\alpha}{\alpha+1}\frac{\mathbf{m}\mathbf{m}^\top}{\|\mathbf{m}\|^2}\right)\right\}\right].$$

Now, as $\mathbf{A}=(\mathbf{I}_p+\beta\mathbf{m}\mathbf{m}^\top)^{-1/2}=\mathbf{I}_p+\{(1+\beta\tau)^{-1/2}-1\}\mathbf{m}\mathbf{m}^\top/\|\mathbf{m}\|^2$, we get

$$(\mathbf{I}_p\otimes\mathbf{C}_2^{1/2}+\mathbf{C}_2^{1/2}\otimes\mathbf{I}_p)^{-1}(\mathbf{A}\otimes\mathbf{A})=\frac{1}{2(\alpha+2)}\left[\mathbf{A}\otimes\mathbf{A}+(\alpha+1)\left\{\left(\mathbf{I}_p-(1-\Delta^2)\frac{\mathbf{m}\mathbf{m}^\top}{\|\mathbf{m}\|^2}\right)\otimes\left(\mathbf{I}_p-(1-\Delta^2)\frac{\mathbf{m}\mathbf{m}^\top}{\|\mathbf{m}\|^2}\right)\right\}\right].$$

Thus, by vectorizing and carrying out the Kronecker-multiplications, we get

$$\left\{\frac{\mathbf{m}}{\|\mathbf{m}\|}^\top\otimes\left(\mathbf{I}_p+\frac{1}{\Delta}\mathbf{Q}_{\mathbf{m}}\right)\right\}\sqrt{n}\mathrm{vec}(\hat{\mathbf{A}}-\mathbf{A})=-\frac{1}{(1+\beta\tau)^{1/2}}\left\{\frac{\mathbf{m}}{\|\mathbf{m}\|}^\top\otimes\left(\mathbf{Q}_{\mathbf{m}}+\frac{1}{2(1+\beta\tau)}\frac{\mathbf{m}\mathbf{m}^\top}{\|\mathbf{m}\|^2}\right)\right\}\sqrt{n}\mathrm{vec}(\hat{\mathbf{C}}_{02}-\mathbf{C}_2)+o_P(1),$$

after which unvectorizing yields the claim. □

Proof of Theorem 6. We now obtain the limiting normality and constant C in exactly the same way as in Theorem 5. As such, we point out only the main steps here. The "**t**-argument" together with the linearization in Lemma 17 shows that the constant C satisfies

$$\Delta^2 C=\mathrm{Var}\left\{-\Delta X_1X_2-\frac{1}{\beta\gamma\tau\Delta^3\|\mathbf{m}\|}X_2+\frac{1}{\beta\gamma\tau\Delta\|\mathbf{m}\|}X_1^2X_2\right\},$$

where X_1, X_2 are as in Theorem 5. □

A.5. Proofs of the results in Section 6

Proof of Lemma 7. By affine equivariance, without loss of generality we can assume that the common group covariance is $\boldsymbol{\Sigma}=\mathbf{I}_p$ and the group means are $-\alpha_2\mathbf{m}$ and $\alpha_1\mathbf{m}$, with $\mathbf{m}=\boldsymbol{\Sigma}^{-1/2}\mathbf{h}$. As shown in Lemma 11, we have $\mathbf{T}_k(\mathbf{x}_w)=\beta\gamma\Delta^3(\mathbf{e}_k^\top\mathbf{m})\mathbf{m}\mathbf{m}^\top$, where $\Delta^2=1/(1+\beta\tau)$ and we have used $\mathbf{C}_2^{-1/2}\mathbf{m}=\Delta\mathbf{m}$. Using this notation

$$\sum_{k=1}^{p}\{\mathbf{v}^\top\mathbf{T}_k(\mathbf{x}_w)\mathbf{v}\}^2=\beta^2\gamma^2\Delta^6\tau(\mathbf{v}^\top\mathbf{m})^4,$$

implying that this sum is maximized at $\mathbf{v}=\pm\mathbf{m}/\|\mathbf{m}\|$. Consequently, $\theta_{\mathrm{J}}=\mathbf{C}_2^{-1/2}\mathbf{m}/\|\mathbf{m}\|=s\Delta\mathbf{m}/\|\mathbf{m}\|$. The result now follows via affine equivariance. □

The proof of Theorem 7 is based on a suitably linearized Taylor expansion of the gradient of the Lagrangian and the remainder of this section focuses on that, under the assumption of a standardized mixture. Even though 3-JADE satisfies only a weaker version of affine equivariance (Lemma 8), this assumption (standardized mixture) is justified by the following: (i) Arguing as in Theorem 2, we see that the limiting behavior of $\hat{\theta}_{\mathrm{J}}(\mathbf{x}_i)/\|\hat{\theta}_{\mathrm{J}}(\mathbf{x}_i)\|$ is determined by the limiting behavior of $\hat{\theta}_{\mathrm{J}}(\mathbf{z}_i)/\|\hat{\theta}_{\mathrm{J}}(\mathbf{z}_i)\|$ where $\hat{\theta}_{\mathrm{J}}(\mathbf{z}_i)$ is *some* sequence of 3-JADE estimators for the standardized mixture $\mathbf{z}_i$. (ii) Below we show that *all* sequences $\hat{\theta}_{\mathrm{J}}(\mathbf{z}_i)$ of 3-JADE estimators for $\mathbf{z}_i$ have the same limiting distribution. Hence, the weaker form of affine equivariance is not an issue and we may use Theorem 3 to compute the limiting constant C.

The gradient $\nabla\ell_n(\mathbf{u}_n)$ (with the Lagrangian multiplier plugged in) has the following first-order Taylor expansion around the population optimum $\mathbf{u}=\mathbf{m}/\|\mathbf{m}\|$: $0=\nabla\ell_n(\mathbf{u})+\nabla^\top\nabla\ell_n(\mathbf{u})(\mathbf{u}_n-\mathbf{u})+O(\|\mathbf{u}_n-\mathbf{u}\|^2)$, further allowing us to write

$$-\sqrt{n}\nabla\ell_n(\mathbf{u})=\mathbf{H}\sqrt{n}(\mathbf{u}_n-\mathbf{u})+o_p(1),\tag{A.19}$$

where $\mathbf{H}$ is an a.s. limit of $\nabla^\top\nabla\ell_n(\mathbf{u})$; for more technical details on the expansion see e.g. [6]. That $\mathbf{u}_n$ converges in probability to $\mathbf{u}$ (up to sign) follows from the standard M-estimator argument over compact spaces. Lemma 19 gives a closed-form expression for $\mathbf{H}$, while Lemma 18 comprises the auxiliary results needed for the calculation of $\mathbf{H}$. The proof of the latter is omitted as it follows the same argument as that of Lemma 11.

Lemma 18. *Letting* $\mathbf{u}:=\mathbf{m}/\tau^{1/2}=\mathbf{m}/\|\mathbf{m}\|$, *we have*

$$\mathbf{T}_k(\mathbf{x}_w)=\beta\gamma\Delta^3(\mathbf{e}_k^\top\mathbf{m})\mathbf{m}\mathbf{m}^\top,\qquad\mathbf{T}_k(\mathbf{x}_w)\mathbf{u}=\beta\gamma\Delta^3\tau^{1/2}(\mathbf{e}_k^\top\mathbf{m})\mathbf{m}\quad\textit{and}\quad\mathbf{u}^\top\mathbf{T}_k(\mathbf{x}_w)\mathbf{u}=\beta\gamma\Delta^3\tau(\mathbf{e}_k^\top\mathbf{m}).$$

Lemma 19. *Let $\nabla^\top \nabla \ell_n(\mathbf{u})$ be the gradient of $\nabla \ell_n(\mathbf{u})$ that is given in (5). Then the a.s. limit $\mathbf{H}$ of $\nabla^\top \nabla \ell_n(\mathbf{u})$ evaluated at $\mathbf{u} = \mathbf{m}/\tau^{1/2}$ exists and is given by*

$$\mathbf{H} = -4\beta^2\gamma^2\Delta^6\tau^3\left(\mathbf{I}_p + \frac{\mathbf{m}\mathbf{m}^\top}{\|\mathbf{m}\|^2}\right).$$

Proof of Lemma 19. It is straightforward to verify that

$$\frac{1}{4}\nabla^\top \nabla \ell_n(\mathbf{u}) = \sum_{k=1}^{p}\left(2\hat{\mathbf{T}}_k\mathbf{u}\mathbf{u}^\top\hat{\mathbf{T}}_k + (\mathbf{u}^\top\hat{\mathbf{T}}_k\mathbf{u})\hat{\mathbf{T}}_k - 4(\mathbf{u}^\top\hat{\mathbf{T}}_k\mathbf{u})\hat{\mathbf{T}}_k\mathbf{u}\mathbf{u}^\top - (\mathbf{u}^\top\hat{\mathbf{T}}_k\mathbf{u})^2\mathbf{I}_p\right).$$

The law of large numbers implies that $\nabla^\top \nabla \ell_n(\mathbf{u}) \to_{a.s.} \mathbf{H}$, where

$$\frac{1}{4}\mathbf{H} = \sum_{k=1}^{p}\left(2\mathbf{T}_k\mathbf{u}\mathbf{u}^\top\mathbf{T}_k + (\mathbf{u}^\top\mathbf{T}_k\mathbf{u})\mathbf{T}_k - 4(\mathbf{u}^\top\mathbf{T}_k\mathbf{u})\mathbf{T}_k\mathbf{u}\mathbf{u}^\top - (\mathbf{u}^\top\mathbf{T}_k\mathbf{u})^2\mathbf{I}_p\right).$$

Using auxiliary Lemma 18, let us now simplify the four sums in $\mathbf{H}$:

$$\sum_{k=1}^{p}\mathbf{T}_k\mathbf{u}\mathbf{u}^\top\mathbf{T}_k = \beta^2\gamma^2\Delta^6\tau\mathbf{m}\mathbf{m}^\top\sum_{k=1}^{p}\mathbf{e}_k\mathbf{e}_k^\top\mathbf{m}\mathbf{m}^\top = \beta^2\gamma^2\Delta^6\tau^2\mathbf{m}\mathbf{m}^\top,$$

$$\sum_{k=1}^{p}(\mathbf{u}^\top\mathbf{T}_k\mathbf{u})\mathbf{T}_k = \beta^2\gamma^2\Delta^6\tau\left(\mathbf{m}^\top\sum_{k=1}^{p}\mathbf{e}_k\mathbf{e}_k^\top\mathbf{m}\right)\mathbf{m}\mathbf{m}^\top = \beta^2\gamma^2\Delta^6\tau^2\mathbf{m}\mathbf{m}^\top,$$

$$\sum_{k=1}^{p}(\mathbf{u}^\top\mathbf{T}_k\mathbf{u})\mathbf{T}_k\mathbf{u}\mathbf{u}^\top = \beta^2\gamma^2\Delta^6\tau^2\mathbf{m}\mathbf{m}^\top,$$

$$\sum_{k=1}^{p}(\mathbf{u}^\top\mathbf{T}_k\mathbf{u})^2\mathbf{I}_p = \beta^2\gamma^2\Delta^6\tau^2\left(\mathbf{m}^\top\sum_{k=1}^{p}\mathbf{e}_k\mathbf{e}_k^\top\mathbf{m}\right)\mathbf{I}_p = \beta^2\gamma^2\Delta^6\tau^3\mathbf{I}_p.$$

Plugging these back into the expression for $\mathbf{H}$ gives

$$\mathbf{H} = -4\beta^2\gamma^2\Delta^6\tau^3\left(\mathbf{I}_p + \frac{\mathbf{m}\mathbf{m}^\top}{\tau}\right). \quad \square$$

Lemma 19 additionally implies that $\mathbf{H}$ has only two distinct eigenvalues: $-8\beta^2\gamma^2\Delta^6\tau^3$ belonging to $\mathbf{m}/\|\mathbf{m}\|$, and $-4\beta^2\gamma^2\Delta^6\tau^3$ with multiplicity $p-1$. This further implies that $\mathbf{H}$ is a regular matrix and that for any unit-length vector $\mathbf{t}$ such that $\mathbf{t}^\top\mathbf{m} = 0$, we have $\mathbf{t}^\top\mathbf{H}^{-1} = (-4\beta^2\gamma^2\Delta^6\tau^3)^{-1}\mathbf{t}^\top$.

Proof of Theorem 7. Eq. (A.19) together with Lemma 19 imply that $\sqrt{n}(\mathbf{u}_n - \mathbf{u}) = -\mathbf{H}^{-1}\sqrt{n}\nabla\ell_n(\mathbf{u}) + o_p(1)$. To show that $\sqrt{n}(\mathbf{u}_n - \mathbf{u})$ indeed has a limiting normal distribution, we start by linearizing the terms in $\sqrt{n}\nabla\ell_n(\mathbf{u})$.

$$\begin{aligned}
\frac{1}{4}\nabla\ell_n(\mathbf{u}) &= \sum_{k=1}^{p}(\mathbf{u}^\top\hat{\mathbf{T}}_k\mathbf{u})\hat{\mathbf{T}}_k\mathbf{u} - \sum_{k=1}^{p}(\mathbf{u}^\top\hat{\mathbf{T}}_k\mathbf{u})^2\mathbf{u} = \sum_{k=1}^{p}(\mathbf{u}^\top\mathbf{T}_k\mathbf{u})\hat{\mathbf{T}}_k\mathbf{u} + \sum_{k=1}^{p}\{\mathbf{u}^\top(\hat{\mathbf{T}}_k - \mathbf{T}_k)\mathbf{u}\}\hat{\mathbf{T}}_k\mathbf{u} \\
&\quad - \sum_{k=1}^{p}(\mathbf{u}^\top\mathbf{T}_k\mathbf{u})^2\mathbf{u} - \sum_{k=1}^{p}\{\mathbf{u}^\top(\hat{\mathbf{T}}_k - \mathbf{T}_k)\mathbf{u}\}^2\mathbf{u} - 2\sum_{k=1}^{p}\{\mathbf{u}^\top(\hat{\mathbf{T}}_k - \mathbf{T}_k)\mathbf{u}\}(\mathbf{u}^\top\mathbf{T}_k\mathbf{u})\mathbf{u} \\
&= \sum_{k=1}^{p}(\mathbf{u}^\top\mathbf{T}_k\mathbf{u})(\hat{\mathbf{T}}_k - \mathbf{T}_k)\mathbf{u} + \sum_{k=1}^{p}(\mathbf{u}^\top\mathbf{T}_k\mathbf{u})\mathbf{T}_k\mathbf{u} + \sum_{k=1}^{p}\{\mathbf{u}^\top(\hat{\mathbf{T}}_k - \mathbf{T}_k)\mathbf{u}\}\mathbf{T}_k\mathbf{u} + \sum_{k=1}^{p}\{\mathbf{u}^\top(\hat{\mathbf{T}}_k - \mathbf{T}_k)\mathbf{u}\}(\hat{\mathbf{T}}_k - \mathbf{T}_k)\mathbf{u} \\
&\quad - \sum_{k=1}^{p}(\mathbf{u}^\top\mathbf{T}_k\mathbf{u})^2\mathbf{u} - \sum_{k=1}^{p}\{\mathbf{u}^\top(\hat{\mathbf{T}}_k - \mathbf{T}_k)\mathbf{u}\}^2\mathbf{u} - 2\sum_{k=1}^{p}\{\mathbf{u}^\top(\hat{\mathbf{T}}_k - \mathbf{T}_k)\mathbf{u}\}(\mathbf{u}^\top\mathbf{T}_k\mathbf{u})\mathbf{u}.
\end{aligned}$$

Since $\mathbf{u}$ is a solution to optimization problem (4), we have $\sum_{k=1}^{p}(\mathbf{u}^\top\mathbf{T}_k\mathbf{u})\mathbf{T}_k\mathbf{u} - \sum_{k=1}^{p}(\mathbf{u}^\top\mathbf{T}_k\mathbf{u})^2\mathbf{u} = \mathbf{0}$, further giving

$$\begin{aligned}
\frac{1}{4}\sqrt{n}\nabla\ell_n(\mathbf{u}) &= \sum_{k=1}^{p}(\mathbf{u}^\top\mathbf{T}_k\mathbf{u})\sqrt{n}(\hat{\mathbf{T}}_k - \mathbf{T}_k)\mathbf{u} + \sum_{k=1}^{p}\{\mathbf{u}^\top\sqrt{n}(\hat{\mathbf{T}}_k - \mathbf{T}_k)\mathbf{u}\}\mathbf{T}_k\mathbf{u} + \sum_{k=1}^{p}\{\mathbf{u}^\top\sqrt{n}(\hat{\mathbf{T}}_k - \mathbf{T}_k)\mathbf{u}\}(\hat{\mathbf{T}}_k - \mathbf{T}_k)\mathbf{u} \\
&\quad - \sum_{k=1}^{p}\{\mathbf{u}^\top\sqrt{n}(\hat{\mathbf{T}}_k - \mathbf{T}_k)\mathbf{u}\mathbf{u}^\top(\hat{\mathbf{T}}_k - \mathbf{T}_k)\mathbf{u}\}\mathbf{u} - 2\sum_{k=1}^{p}\{\mathbf{u}^\top\sqrt{n}(\hat{\mathbf{T}}_k - \mathbf{T}_k)\mathbf{u}\}(\mathbf{u}^\top\mathbf{T}_k\mathbf{u})\mathbf{u}.
\end{aligned}$$

Additional linearization and the law of large numbers imply that $\hat{\mathbf{T}}_k - \mathbf{T}_k = o_P(1)$, allowing for simplification in the upper linearization, i.e.,

$$\sqrt{n}\nabla\ell_n(\mathbf{u}) = 4\sum_{k=1}^{p}(\mathbf{u}^\top\mathbf{T}_k\mathbf{u})\sqrt{n}(\hat{\mathbf{T}}_k - \mathbf{T}_k)\mathbf{u} + 4\sum_{k=1}^{p}\{\mathbf{u}^\top\sqrt{n}(\hat{\mathbf{T}}_k - \mathbf{T}_k)\mathbf{u}\}\mathbf{T}_k\mathbf{u} - 8\sum_{k=1}^{p}\{\mathbf{u}^\top\sqrt{n}(\hat{\mathbf{T}}_k - \mathbf{T}_k)\mathbf{u}\}(\mathbf{u}^\top\mathbf{T}_k\mathbf{u})\mathbf{u} + o_P(1).$$

The asymptotic normality of $\sqrt{n}(\hat{\mathbf{T}}_k - \mathbf{T}_k)$ discussed in the proof of Lemma 17 now implies that $\sqrt{n}(\mathbf{u}_n - \mathbf{u})$ is also asymptotically normal. Consequently, Theorems 2 and 3 can be used to obtain the precise form of the asymptotic covariance matrix of the 3-JADE

estimator. The former gives the matrix up to proportionality (due to AE) and the latter allows finding the method-specific constant C. This constant requires obtaining an expansion of $\mathbf{t}^\top\sqrt{n}(\mathbf{u}_n-\mathbf{u}) = -\mathbf{t}^\top\mathbf{H}^{-1}\sqrt{n}\nabla\ell_n(\mathbf{u}) + o_P(1)$, which we will do next.

As shown after Lemma 19, or any unit-length vector $\mathbf{t}$ such that $\mathbf{t}^\top\mathbf{m}=0$, we have $\mathbf{t}^\top\mathbf{H}^{-1} = (-4\beta^2\gamma^2\Delta^6\tau^3)^{-1}\mathbf{t}^\top$, further implying that $\mathbf{t}^\top\sqrt{n}(\mathbf{u}_n-\mathbf{u})$ can be written as

$$(\beta^2\gamma^2\Delta^6\tau^3)^{-1}\mathbf{t}^\top\left[\sum_{k=1}^p(\mathbf{u}^\top\mathbf{T}_k\mathbf{u})\sqrt{n}(\hat{\mathbf{T}}_k-\mathbf{T}_k)\mathbf{u} + \sum_{k=1}^p\{\mathbf{u}^\top\sqrt{n}(\hat{\mathbf{T}}_k-\mathbf{T}_k)\mathbf{u}\}\mathbf{T}_k\mathbf{u} - 2\sum_{k=1}^p\{\mathbf{u}^\top\sqrt{n}(\hat{\mathbf{T}}_k-\mathbf{T}_k)\mathbf{u}\}(\mathbf{u}^\top\mathbf{T}_k\mathbf{u})\mathbf{u}\right] + o_P(1).$$

As $\mathbf{t}^\top\mathbf{u}=0$, and $\mathbf{T}_k\mathbf{t}=\mathbf{0}$, the final two terms in the upper expansion vanish, leaving $\mathbf{t}^\top\sqrt{n}(\mathbf{u}_n-\mathbf{u}) = \frac{1}{\beta\gamma\Delta^3\tau^2}\sum_{k=1}^p\{\mathbf{t}^\top\sqrt{n}(\hat{\mathbf{T}}_k-\mathbf{T}_k)\mathbf{u}\}(\mathbf{e}_k^\top\mathbf{m}) + o_P(1)$. Plugging in the expression for $\sqrt{n}(\hat{\mathbf{T}}_k-\mathbf{T}_k)$ from the proof of Lemma 17, we get

$$\mathbf{t}^\top\sqrt{n}(\mathbf{u}_n-\mathbf{u}) = \frac{1}{\beta\gamma\Delta^3\tau^2}\left\{\beta\gamma\Delta^2\tau^2\mathbf{t}^\top\sqrt{n}(\hat{\mathbf{A}}-\mathbf{A})\mathbf{u} - \tau^{1/2}\mathbf{t}^\top\sqrt{n}\bar{\mathbf{z}} + \Delta^2\tau^{1/2}\mathbf{t}^\top\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n\mathbf{z}_i(\mathbf{z}_i^\top\mathbf{u})^2 - \beta\gamma\tau\mathbf{m}\right)\right\} + o_P(1).$$

Invoking now Theorem 3 and its proof, we have

$$-\mathbf{t}^\top\sqrt{n}(\hat{\mathbf{C}}_2-\mathbf{C}_2)\mathbf{u} - \frac{1}{\beta\gamma\Delta^4\tau^{3/2}}\sqrt{n}\mathbf{t}^\top\bar{\mathbf{z}} + \frac{1}{\beta\gamma\Delta^2\tau^{3/2}}\mathbf{t}^\top\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n\mathbf{z}_i(\mathbf{z}_i^\top\mathbf{u})^2 - \beta\gamma\tau\mathbf{m}\right) \rightsquigarrow \mathcal{N}(0,C).$$

Consequently, the asymptotic variance constant C satisfies, by the CLT, that

$$C = \mathrm{Var}\left(-X_1X_2 - \frac{1}{\beta\gamma\Delta^4\tau^{3/2}}X_2 + \frac{1}{\beta\gamma\Delta^2\tau^{3/2}}X_1^2X_2\right),$$

where X_1, X_2 are as in the proof of Theorem 5. Computing the variance now yields the claim. □

Appendix B. Supplementary data

Supplementary material contains additional numerical experiments and simulation results including Fig. S1–S3.
Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2025.105524.

References

[1] D. Peña, F.J. Prieto, Cluster identification using projections, J. Amer. Statist. Assoc. 96 (2001) 1433–1445.
[2] D. Peña, F.J. Prieto, J. Viladomat, Eigenvectors of a kurtosis matrix as interesting directions to reveal cluster structure, J. Multivariate Anal. 101 (2010) 1995–2007.
[3] N. Loperfido, Skewness and the linear discriminant function, Statist. Probab. Lett. 83 (1) (2013) 93–99.
[4] N. Loperfido, Vector-valued skewness for model-based clustering, Statist. Probab. Lett. 99 (2015) 230–237.
[5] T.F. Móri, V.K. Rohatgi, G. Székely, On multivariate skewness and kurtosis, Theory Probab. Appl. 38 (3) (1994) 547–551.
[6] U. Radojičić, K. Nordhausen, J. Virta, Large-sample properties of blind estimation of the linear discriminant using projection pursuit, Electron. J. Stat. 15 (2) (2021).
[7] P. Ilmonen, H. Oja, R. Serfling, On invariant coordinate system (ICS) functionals, Int. Stat. Rev. 80 (1) (2012) 93–110.
[8] K. Nordhausen, A. Ruiz-Gazen, On the usage of joint diagonalization in multivariate statistics, J. Multivariate Anal. 188 (2022) 104844.
[9] N. Loperfido, On a vector-valued measure of multivariate skewness, Symmetry 13 (10) (2021) 1817.
[10] J.M. Arevalillo, H. Navarro, Skewness-based projection pursuit as an eigenvector problem in scale mixtures of skew-normal distributions, Symmetry 13 (6) (2021) 1056.
[11] N. Loperfido, Singular value decomposition of the third multivariate moment, Linear Algebra Appl. 473 (2015) 202–216.
[12] J.-F. Cardoso, Source separation using higher order moments, in: International Conference on Acoustics, Speech, and Signal Processing, IEEE, 1989, pp. 2109–2112.
[13] J. Miettinen, S. Taskinen, K. Nordhausen, H. Oja, Fourth moments and independent component analysis, Statist. Sci. 30 (3) (2015) 372–390.
[14] J. Virta, K. Nordhausen, H. Oja, Joint use of third and fourth cumulants in independent component analysis, 2015, arXiv preprint arXiv:1505.02613.
[15] J.-F. Cardoso, A. Souloumiac, Blind beamforming for non-Gaussian signals, IEE Proc. F 140 (6) (1993) 362–370.
[16] K. Nordhausen, H. Oja, D. Tyler, J. Virta, ICtest: Estimating and testing the number of interesting components in linear dimension reduction, 2022, R package version 0.3-5.
[17] T. Tony Cai, L. Zhang, High Dimensional Linear Discriminant Analysis: Optimality, Adaptive Algorithm and Missing Data, J. R. Stat. Soc. Ser. B 81 (4) (2019) 675–705.
[18] D.E. Tyler, F. Critchley, L. Dümbgen, H. Oja, Invariant co-ordinate selection, J. R. Stat. Soc. Ser. B 71 (3) (2009) 549–592.
[19] N. Loperfido, Tensor eigenvectors for projection pursuit, TEST 33 (2) (2024) 453–472.
[20] T. Kollo, D. von Rosen, Advanced Multivariate Statistics With Matrices, Springer Dordrecht, 2005.
[21] J.S. Liu, Siegel's formula via Stein's identities, Statist. Probab. Lett. 21 (3) (1994) 247–251.
[22] A.K. Gupta, D.K. Nagar, Matrix Variate Distributions, Chapman and Hall/CRC New York, 2018.
[23] D. von Rosen, Moments for matrix normal variables, Stat.: A J. Theor. Appl. Stat. 19 (4) (1988) 575–583.