International Statistical Review

International Statistical Review (2023), **91**, 1, 140–161 doi: 10.1111/insr.12517

A Computational Perspective on Projection Pursuit in High Dimensions: Feasible or Infeasible Feature Extraction

Chunming Zhang[1], Jimin Ye[2] and Xiaomei Wang[3]

[1]*Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, USA*
[2]*School of Mathematics and Statistics, Xidian University, Xi'an, Shaanxi 710071, China*
[3]*School of Management, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China*
Correspondence *Chunming Zhang, Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, USA. Email: cmzhang@stat.wisc.edu*

Summary

Finding a suitable representation of multivariate data is fundamental in many scientific disciplines. Projection pursuit (PP) aims to extract interesting 'non-Gaussian' features from multivariate data, and tends to be computationally intensive even when applied to data of low dimension. In high-dimensional settings, a recent work (Bickel et al., 2018) on PP addresses asymptotic characterization and conjectures of the feasible projections as the dimension grows with sample size. To gain practical utility of and learn theoretical insights into PP in an integral way, data analytic tools needed to evaluate the behaviour of PP in high dimensions become increasingly desirable but are less explored in the literature. This paper focuses on developing computationally fast and effective approaches central to finite sample studies for (i) visualizing the feasibility of PP in extracting features from high-dimensional data, as compared with alternative methods like PCA and ICA, and (ii) assessing the plausibility of PP in cases where asymptotic studies are lacking or unavailable, with the goal of better understanding the practicality, limitation and challenge of PP in the analysis of large data sets.

Key words: density estimation; empirical distribution function; exploratory data analysis; Gaussian mixture; ICA; PCA.

1 Introduction

The projection pursuit (PP), proposed by Kruskal (1969) and first implemented in Friedman & Tukey (1974), is a technique for finding 'interesting' structures from multivariate data. It is based on the idea of pursuing low-dimensional projections of multivariate data to display highly 'non-Gaussian' features. To be explicit, consider n independent data vectors $\boldsymbol{X}_{\cdot,i} = (X_{1,i}, \dots, X_{p,i})^T$, $i = 1, \dots, n$, from the distribution of a p-variate random vector $\boldsymbol{X} = (X_1, \dots, X_p)^T$, namely,

$$\boldsymbol{X}_{\cdot,1}, \dots, \boldsymbol{X}_{\cdot,n} \overset{\text{i.i.d.}}{\sim} \boldsymbol{X}. \tag{1.1}$$

© 2022 The Authors. International Statistical Review published by John Wiley & Sons Ltd on behalf of International Statistical Institute.
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

Along each projection vector z of unit length ($z^T z = 1$), the n original data vectors are transformed to n projected data scalars,

$$z^T X_{\cdot,1}, \ldots, z^T X_{\cdot,n}, \tag{1.2}$$

with the empirical distribution function (E.D.F.),

$$\hat{G}_z(s) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}(z^T X_{\cdot,i} \leq s), \tag{1.3}$$

where $\mathrm{I}(\cdot)$ denotes the indicator operator. It is anticipated that

$$\hat{G}_z(s) \text{ tends to a target distribution } G(s) \text{ as } n \text{ tends to infinity}, \tag{1.4}$$

where $G(\cdot)$ is a cumulative distribution function (C.D.F.). Loosely speaking, a target distribution G is 'feasible' if there exist projections whose empirical distribution could possibly approximate G. Numerically, the one-dimensional PP seeks a projection onto a unit length vector z, such that

$$\text{the target distribution } G \text{ in } (1.4) \text{ is as '}\textit{non-Gaussian}\text{' as possible}, \tag{1.5}$$

associated with interesting structural features, such as clusters, inhomogeneity, modes, trends and skewness in the data. The search for the desired direction z solves an optimization problem which maximizes a suitable projection index, measuring the degree of interestingness or non-Gaussianity of the projected data. In this regard, projection pursuit methods tend to be computationally intensive (Tyler *et al.*, 2009) even for comparatively low-dimensional datasets. Refer to Huber (1985), Friedman (1987), Jones & Sibson (1987), Sun (2006), Daszykowski (2007), Jee (2009), and Loperfido (2018) and references therein, for more comprehensive reviews and discussions of PP on a range of topics including the dimension of the projection space, choices of the projection index and optimization approaches to the search for interesting projections.

In recent years, high-dimensional data arise in many contemporary problems, where the sample size is small relative to the dimensionality, posing substantial challenges to data analysis and motivating the development of new statistical methods. Yet relatively few works relevant to performing PP on high-dimensional data are available in the statistics literature. For classification of large p small n data, Lee & Cook (2010) devised and implemented a new projection pursuit index. Blanchard *et al.* (2006), De Bie *et al.* (2016), Sasaki *et al.* (2016), Virta *et al.* (2016), and Loperfido (2018) considered algorithms to find non-Gaussian components of a high-dimensional distribution. Bickel *et al.* (2018) gave a more recent theoretically-focused research on PP.

For data points in high-dimensional spaces, two important and challenging issues arise from the feature extraction task by PP: issue (i) regarding asymptotic studies conducted on the number p of variables, the number n of observations and the joint distribution of variables $X_1, \ldots, X_p$ to guarantee the 'existence' of a plausible direction z such that a target distribution G in (1.4) and (1.5) is feasible; issue (ii) concerning data analytic tools for evaluating the behaviour of PP in high dimensions and enabling the asymptotic feasibility results in issue (i) to be accessible to users and computers.

As for the asymptotic issue (i), there have been many discussions (Geman, 1980; Diaconis & Freedman, 1984; Huber, 1985) on the choice of projection vectors z. Particularly, for high-dimensional data, where $p = p_n$ tends to infinity as sample size n tends to infinity, a recent work (Bickel *et al.*, 2018) building on results (Rudelson & Vershynin, 2009) characterizes and

© 2022 The Authors. International Statistical Review published by John Wiley & Sons Ltd on behalf of International Statistical Institute.

17515823, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/insr.12517 by Chunming Zhang - University Of Wisconsin - Madison , Wiley Online Library on [04/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

conjectures impacts of the dimension p growing with n on the target distribution G in a variety of settings (to be reviewed in Section 2.1 and summarized in Table 1), and identifies types of target distributions G to be 'feasible', 'uniquely feasible', or 'infeasible' (formally defined in Section 2.1) for PP. These results are confined to the scenario where variables $X_1, \ldots, X_p$ are independent, and standard Gaussian or zero-mean sub-Gaussian. Affirmative results are not available yet for cases where variables $X_1, \ldots, X_p$ either are non-independent, non-identically distributed, or have other types of non-Gaussian distributions.

In responding to the practical issue (ii), computational strategies and data analytic tools needed to evaluate the behaviour of PP for finite data samples in high dimensions and learn asymptotic feasibility results in Bickel *et al.* (2018) become increasingly desirable but are less explored in the literature. This paper focuses on developing computationally fast and effective approaches and making practical suggestions, central to empirical studies and numerical experiments.

(a) Section 2.2 devises numerical schemes to be used in Section 3 for visualizing the feasibility of PP in feature extraction. First, we integrate a closed-form expression of the Kolmogorov–Smirnov distance into quantifying the maximum discrepancy between the empirical distribution $\hat{G}_z$ and the target distribution G in one-dimensional PPs, while exemplify the demand for an approximate evaluation in the multiple-dimensional analogue. Second, the developed non-parametric 'empirical probability density function' $\hat{g}_z$ in (2.12) enjoys numerical simplicity and better uncovers the multi-modal and/or non-Gaussian types of features in the exploratory PP, from correlated projected data $\{z^T \boldsymbol{X}_{\cdot,i}\}_{i=1}^n$ onto data-dependent directions z, than the kernel smoothing method which is typically -suited for estimating unimodal or Gaussian-type of densities underlying i.i.d. data. These

Table 1. *Summary of asymptotic results in Bickel et al. (2018) of projection pursuit as $n\to\infty$, $p\to\infty$, $p/n\to\gamma$ for different values of γ on mean-centred data.*

$\gamma = \lim_{n\to\infty} \frac{p}{n}$	*zero-mean target distributions in (1.4)*	*result in Bickel* et al., *(2018)*
$\gamma = \infty$	any G is feasible;	Thm. 1
	any G is feasible under (2.2);	Remark 1
	any multivariate G is feasible.	Remark 2
$\gamma \in (1, \infty)$	G is feasible if $\mu_2(G) < \gamma - 1$,	Thm. 2(i)
	where $\mu_2(G) = \int x^2 \mathrm{d}\, G\cdot(x)$;	
	G is feasible under (2.2) if $\mu_2(G) < (\sqrt{\gamma} - 1)^2$;	Remark 3
	G is infeasible if $\mu_2(G) > (\sqrt{\gamma} + 1)^2$;	Thm. 2(ii)
	$G(\frac{s - u_0}{\sigma_0})$ of 'Type-G' is feasible if $\mu_2(G) < \infty$.	Corollary 1
$\gamma \in (0, 1)$	$\frac{\gamma}{L}G + (1 - \frac{\gamma}{L})\Phi$ is feasible if $\mu_2(G) < L - 1$, $L \in (1, \infty)$,	Thm. 3,
	where Φ is the standard Gaussian distribution;	
	G is infeasible if	Thm. 4
	$\max_s \lvert G(s) - \Phi(s)\rvert > C\sqrt{\gamma \log(1/\gamma)}$.	
$\gamma = 1$	$\frac{1}{L}G + (1 - \frac{1}{L})\Phi$ is feasible if $\mu_2(G) < L - 1$, $L \in (1, \infty)$;	Thm. 3,
	G (not a mixture of Gaussian) may be feasible,	Remark 4
	relying on convergence rate to $\gamma = 1$.	
$\gamma = 0$	Φ is uniquely feasible.	Thm. 5
$\gamma \in (0, \infty)$, $\frac{\lVert z\rVert_0}{n} \to 0$	Φ is uniquely feasible.	Thm. 6

See Appendix A for a complete list of notations.

© 2022 The Authors. International Statistical Review published by John Wiley & Sons Ltd on behalf of International Statistical Institute.

17515823, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/insr.12517 by Chunming Zhang - University Of Wisconsin - Madison , Wiley Online Library on [04/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

numerical schemes facilitate comparison with other commonly used feature extraction methods, such as the principal component analysis (PCA) in Jolliffe (2002) and Jolliffe & Cadima (2016) and the independent component analysis (ICA) in Hyvärinen & Oja (2000) and Hyvärinen *et al.* (2001), both of which search for projection directions using criterions with relevance to that of PP; see Section 3.2.1 and Appendix B.1, with an emphasis on the two-dimensional PP.

(b) The plausibility of PP has yet to be evaluated in high-dimensional settings where theoretical justifications are lacking, unsolved or unavailable. For instance, previous work (Bickel *et al.*, 2018) focused on one-dimensional projections, with the exception of its Remark 2 extended to two or more-dimensional orthonormal projections where p/n diverges to infinity; implications and existence of feasible multivariate projections remain less clear for other cases, for example, the ratio p/n tending to a constant $\gamma \in (0, \infty)$. Appendix A.1 presents some extended results to Lemmas E.1–E.2 and Results E.1–E.2 which allow bivariate and multivariate orthonormal projections and also serve to validate numerical schemes in Sections 2.2.1 and 3.2.1. Again, the developed data analytic tools not only enable the extended feasibility results on PP to be visually inspected in Section 3.2.2, but also inspire an empirical assessment of scenarios not yet covered by Bickel *et al.*, (2018) in Appendix B.2, where PP is performed on multivariate t data with an identity covariance matrix.

The rest of the paper is organized as follows. Section 2 incorporates an explicit expression and an approximation scheme in quantifying the proximity between $\hat{G}_z$ and G, and devises a non-parametric density estimation method better suited for visualizing structural features extracted by PP from high-dimensional data vectors. Section 3 utilizes the computational tool in Section 2 to graphically illustrate the feasibility of PP in feature extraction on a case-by-case basis, in contrast to PCA and ICA methods. Section 4 briefly concludes. Notations, expanded results and additional illustrations are collected in Appendices A and B. All numerical results are implemented in Matlab, with codes available at Github https://github.com/ChunmingZhangUW/PP_high_dim.

2 Numerically Assessing Feasibility of PP in Feature Extraction

2.1 The High-dimensional Setup in Bickel et al. (2018)

To facilitate the numerical evaluation of feasibility of PP in high dimensions, we adopt the high-dimensional settings consistent with those in Bickel *et al.* (2018) for PP applied to realizations $\boldsymbol{X}_{\cdot,1}, \ldots, \boldsymbol{X}_{\cdot,n}$ of a random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^T$, and start with a brief review of assumptions on the distribution of $\boldsymbol{X}$, in combination with dimension p and sample size n.

The assumption A1 on

$$\text{variables } X_1, \ldots, X_p \overset{\text{i.i.d.}}{\sim} \mathbb{N}(0,1) \tag{2.1}$$

is made in all feasibility results except Remark 1 and Remark 3, where $\mathbb{N}(0,1)$ denotes the standard Gaussian distribution.

The assumption A2 on

$$\text{variables} X_1, \ldots, X_p \overset{\text{i.i.d.}}{\sim} \text{a zero-mean sub-Gaussian distribution } F \tag{2.2}$$

is made in Remark 1 and Remark 3.

The assumption A3 on the dimension p growing with sample size n in various scenarios,

© 2022 The Authors. International Statistical Review published by John Wiley & Sons Ltd on behalf of International Statistical Institute.

17515823, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/insr.12517 by Chunming Zhang - University Of Wisconsin - Madison , Wiley Online Library on [04/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$$n \to \infty,\ p \to \infty \text{ with } p/n \to \gamma \in [0, +\infty] \begin{cases} = \infty, \\ \in (0, \infty), \\ = 0, \end{cases} \tag{2.3}$$

$$n \to \infty,\ p \to \infty \text{ with } p/n \to \gamma \in (0, \infty),\ \text{and} \|z\|_0 / n \to 0, \tag{2.4}$$

where z denotes the projection vector in (1.2) and $\|z\|_0$ denotes the number of non-zeros.

For subsequent use, Table 1 concisely summarizes the asymptotically feasible projections in Bickel *et al.* (2018) under settings (2.1)–(2.4), with additional notations gathered in Appendix A. There, a target distribution G is called 'feasible', if there exists a sequence of unit-length directions $z = z(\mathbf{X}, G)$ relying on both the target distribution G and the data matrix $\mathbf{X} = (\boldsymbol{X}_{\cdot,1}, \ldots, \boldsymbol{X}_{\cdot,n}) \in \mathbb{R}^{p \times n}$, such that the E.D.F. s $\hat{G}_z$ of projected data points, $\{z^T \boldsymbol{X}_{\cdot,i}\}_{i=1}^n$, converge uniformly to the distribution function G in probability as n tends to ∞, that is,

$$\|\hat{G}_z - G\|_\infty = \max_{-\infty < s < \infty} |\hat{G}_z(s) - G(s)| \xrightarrow{\text{P}} 0, \tag{2.5}$$

where $\xrightarrow{\text{P}}$ denotes converges in probability (van der Vaart, 1998); moreover, the distribution G is 'uniquely feasible', if the convergence (2.5) holds for all unit-length vectors z. Likewise, for an 'infeasible' distribution G, no sequence of projections has corresponding E.D.F.s $\hat{G}_z$ that converges to the distribution G and, consequently, the target distribution G can not be approximated by any projections onto unit-length vectors.

Accompanying with Table 1, more detailed descriptions of and self-contained statements on the feasible target distributions G in (Bickel *et al.*, 2018) are provided below.

1. The Case $p_n/n \to \infty$: Theorem 1 indicates that, under condition (2.1), any distribution G is feasible. The same conclusion holds in Remark 1 under more general assumptions (2.2). Remark 2 extends Theorem 1 to multi-dimensional projections.
2. The Case $p_n/n \to \gamma \in (1, \infty)$: Theorem 2(i) tells that G is feasible provided its second moment $\mu_2(G)$ is below $\gamma - 1$, while Theorem 2(ii) tells that $\mu_2(G)$ larger than $(\sqrt{\gamma} + 1)^2$ leads to an infeasible G. Remark 3 extends Theorem 2(i) to variables $X_1, \ldots, X_p$ following the zero-mean sub-Gaussian distribution in (2.2). Corollary 1 ensures that a distribution $G(\frac{s - u_0}{\sigma_0})$ with appropriate location parameter u_0 and scale parameter σ_0 is feasible whenever G has a finite second moment.
3. The Case $p_n/n \to \gamma \in (0, 1)$: Theorem 3 tells that the mixture distribution $(\gamma/L)G + (1 - \gamma/L)\Phi$ is feasible, if $\mu_2(G)$ is below $L - 1$, with a finite constant L greater than 1. Theorem 4 shows that a distribution G is infeasible if it is 'far' from the standard Gaussian distribution Φ, that is, the maximum discrepancy between G and Φ exceeds some value depending on the limiting constant γ.
4. The Case $p_n/n \to 1$: Theorem 3 continues to hold. Remark 4 points out that the feasibility of certain distributions, which are not a mixture of Gaussian, depends on the convergence rate of p_n/n to 1.
5. The Case $p_n/n \to 0$: Theorem 5 says that the standard Gaussian distribution Φ is uniquely feasible, and thus a non-Gaussian projection is rare.

© 2022 The Authors. International Statistical Review published by John Wiley & Sons Ltd on behalf of International Statistical Institute.

17515823, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/insr.12517 by Chunming Zhang - University Of Wisconsin - Madison , Wiley Online Library on [04/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

6. The Case of sparse projection with $p_n/n \to \gamma \in (0, \infty)$ and $\|z\|_0/n \to 0$: Theorem 6 states that, analogous to Theorem 5, the standard Gaussian distribution Φ is uniquely feasible, and thus a non-Gaussian projection is rare.

2.2 Ways of Assessing Feasible G in Feature Extraction

Following the previous review, the infeasibility of a distribution G in PP is verified by the departure of $\|\hat{G}_z - G\|_\infty$ from zero for any arbitrary direction of unit-length. In contrast, the feasibility of PP in extracting features of a distribution G hinges on both the 'existence' of certain unit-length direction z and the 'stochastic convergence' of $\|\hat{G}_z - G\|_\infty$ to zero in (2.5). In Sections 2.2.1 and 2.2.2, we develop computationally fast and effective methods for verifying the 'existence' and quantifying 'convergence'.

2.2.1 Verifying the 'existence' of a feasible unit-length vector z

For high-dimensional data with $p \geq n$, major steps for validating the existence of a feasible projection vector were described in Bickel *et al.* (2018) (p. 4., the proof of Theorem 1) and are outlined as follows:

1. Step 1: For the target distribution G, find an n-element source vector $\boldsymbol{a} = (a_1, \ldots, a_n)^T = \boldsymbol{a}(G)$ such that $\max_s |n^{-1}\sum_{i=1}^n \mathrm{I}(a_i \leq s) - G(s)|$ converges to zero with a high probability as n increases. Choices of a_i include $a_i = G^{-1}(\frac{i}{n+1})$, $i = 1, \ldots, n$, where $G^{-1}(\mathrm{p})$ denotes the quantile of the distribution G at the probability $\mathrm{p} \in (0, 1)$.
2. Step 2: For the data matrix $\mathbf{X} = (\boldsymbol{X}_{\cdot,1}, \ldots, \boldsymbol{X}_{\cdot,n}) \in \mathbb{R}^{p \times n}$, take an initial vector
$$z_0 = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{a} \in \mathbb{R}^p, \tag{2.6}$$
such that the length of z_0 is less than or equal to 1 with a high probability as n diverges. Modify z_0 as needed to obtain a feasible direction $z = z(\mathbf{X}, G)$ of unit length in asymptotic feasibility results, such that
$$(z^T\boldsymbol{X}_{\cdot,1}, \ldots, z^T\boldsymbol{X}_{\cdot,n}) = (a_1, \ldots, a_n), \text{ i.e., } \mathbf{X}^T z = \boldsymbol{a}. \tag{2.7}$$

On the other hand, for low-dimensional data with $p < n$, the vector z in (2.7) based on z_0 in (2.6) does not apply. Nonetheless, as seen from Table 1, the necessity of the existence of a feasible direction in asymptotic feasibility results comes solely from Theorem 3. Hence, in the setting of Theorem 3 with a finite constant L greater than 1 (appearing in Table 1), we set $n_1 = [p/L]$ which ensures $n_1 \leq p$, where $[x]$ denotes the largest integer less than or equal to x, and we can implement a modified version of the feasible direction,
$$z((\boldsymbol{X}_{\cdot,1}, \ldots, \boldsymbol{X}_{\cdot,n_1}), G) \text{ of unit length}, \tag{2.8}$$
by substituting the full data matrix $\mathbf{X}$ and the n-element source vector $\boldsymbol{a}$ in Steps 1–2 above with the subset $(\boldsymbol{X}_{\cdot,1}, \ldots, \boldsymbol{X}_{\cdot,n_1}) \in \mathbb{R}^{p \times n_1}$ and an n_1-element source vector with entries, for example, $G^{-1}(\frac{i}{n_1+1})$, $i = 1, \ldots, n_1$, respectively. Similar arguments for our extended Lemma E.2 in Appendix A.1 will ensure that (2.8) offers the desired feasible direction in PP.

© 2022 The Authors. International Statistical Review published by John Wiley & Sons Ltd on behalf of International Statistical Institute.

17515823, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/insr.12517 by Chunming Zhang - University Of Wisconsin - Madison , Wiley Online Library on [04/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2.2.2 *Quantifying 'convergence' of* $\|\hat{G}_z - G\|_\infty$ *to zero for a given direction* z

From the computational viewpoint, evaluating the closeness of the sup-norm $\|\hat{G}_z - G\|_\infty$ to zero in (2.5) is essential for visualizing the recovery of the distributional feature of the C.D.F. G from projection onto a vector z in PP. We now develop two types of non-parametric methods, which are computationally efficient in implementing the evaluations.

The distribution-based Method 1 is inspired by comparing distribution functions G and the E.D.F. $\hat{G}_z$ in (1.3) of the projected data $\{S_i\}_{i=1}^n$, where

$$S_i = z^T X_{\cdot, i}, \quad i = 1, \ldots, n,$$

onto a vector z, that is, $\hat{G}_z(s) = n^{-1} \sum_{i=1}^{n} \mathrm{I}(S_i \le s)$, for $s \in \mathbb{R}$. Then for distribution functions G, we utilize a result (Gibbons & Chakraborti, 2003, Theorem 4.3.1, p. 109) on the Kolmogorov–Smirnov (KS) distance to simplify the computation of $\|\hat{G}_z - G\|_\infty$ via an explicit expression,

$$\|\hat{G}_z - G\|_\infty = \max\left(\max_{1 \le i \le n} \left\{ \frac{i}{n} - G(S_{(i)}) \right\}, \max_{1 \le i \le n} \left\{ G(S_{(i)}) - \frac{i-1}{n} \right\}, 0 \right), \tag{2.9}$$

where $S_{(1)} \le \ldots \le S_{(n)}$ denote the order-statistics of $\{S_i\}_{i=1}^n$. Refer to van der Vaart (1998) and Groeneboom & Wellner (2001) for related distributional results of the maximal deviation.

In the same vein, for a distribution function $G(s_1, \ldots, s_K)$ in the K-variate setting with $K \ge 2$ together with the K-variate E.D.F. $\hat{G}_{z_1, \ldots, z_K}(s_1, \ldots, s_K)$ defined as

$$\hat{G}_{z_1, \ldots, z_K}(s_1, \ldots, s_K) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}(z_1^T X_{\cdot, i} \le s_1, \ldots, z_K^T X_{\cdot, i} \le s_K), \tag{2.10}$$

the maximal deviation

$$\|\hat{G}_{z_1, \ldots, z_K} - G\|_\infty = \max_{(s_1, \ldots, s_K) \in \mathbb{R}^K} |\hat{G}_{z_1, \ldots, z_K}(s_1, \ldots, s_K) - G(s_1, \ldots, s_K)| \tag{2.11}$$

generalizes the univariate analogue (2.5). Still, the numerical complications and challenges associated with this sup-norm are considerable, due to the lack of a closed-form expression as readily as (2.9); see Justel *et al.* (1997), Markatou & Sofikitou (2019) and references therein for discussions of the multivariate KS distance, and Perisic & Posse (2005) for empirical approximations for a bivariate E.D.F. and a bivariate KS distance. In the context of PP, we suggest approximating (2.11) via taking the greatest vertical distance between the two joint distribution functions across grid points of $(s_1, \ldots, s_K)$ in $\mathbb{R}^K$. For the bivariate case, computational efficiency and statistical guarantee are seen from the boxplot of $\|\hat{G}_{z_1, z_2} - G\|_\infty$ in Figure 11 and Figure 13 in Appendix S2, obtained using simulated data matrices.

Different from the distribution-based Method 1, the density-based Method 2 directly estimates the probability density function (p.d.f.) of the projected data points $S_1, \ldots, S_n$, and visualizes features in the estimated p.d.f.. Here, the histogram counts of n points $\{S_i\}_{i=1}^n$ give B non-overlapping (equally spaced) bins, for example, $B = 10$, associated with bin centres $c_1, \ldots, c_B$, and bin counts $\mathrm{n}_1, \ldots, \mathrm{n}_B$ satisfying $n = \mathrm{n}_1 + \ldots + \mathrm{n}_B$. The relative frequencies within bins are n_b/n, $b = 1, \ldots, B$, satisfying $\sum_{b=1}^{B} \mathrm{n}_b/n = 1$. To strike a balance between flexibility and interpretability, consider the estimated p.d.f. to be piecewise-constant within bins. It follows that we estimate the p.d.f. $g(c_b)$ at the bin centre c_b, by calibrating the 'empirical probability density function' (Epdf),

© 2022 The Authors. International Statistical Review published by John Wiley & Sons Ltd on behalf of International Statistical Institute.

17515823, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/insr.12517 by Chunming Zhang - University Of Wisconsin - Madison , Wiley Online Library on [04/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$$\hat{g}_z(c_b) = \frac{\mathrm{n}_b/n}{c_2 - c_1} = \frac{\text{relative frequency within the } bth \text{ bin}}{\text{width of the } bth \text{ bin}}, \quad b = 1, \ldots, B, \tag{2.12}$$

which ensures that the area under the estimated density function, $\hat{g}_z(\cdot)$, is equal to $\sum_{b=1}^{B} \hat{g}_z(c_b) \times (c_2 - c_1) = \sum_{b=1}^{B} (\mathrm{n}_b/n)/(c_2 - c_1) \times (c_2 - c_1) = \sum_{b=1}^{B} \mathrm{n}_b/n = 1$.

In the literature, the kernel density estimator (KDE) (Devroye & Gyorfi, 1985; Silverman, 1986) also serves as a non-parametric method for estimating a density function, via $\hat{f}_h(s) = (nh)^{-1} \sum_{i=1}^{n} \mathrm{K}((S_i - s)/h)$, where $h > 0$ is the bandwidth parameter and $\mathrm{K}(\cdot)$ is a non-negative symmetric kernel function. Meanwhile, the performance of KDE relies largely on the choice of h; data-driven methods of h include the normal-reference method, plug-in method, and cross-validation; see Silverman (1986) for details. Additionally, the conventional KDE is well-suited for estimating the unimodal, smooth or Gaussian-type of densities underlying i.i.d. data. See Hall *et al.* (2004) for related discussions on KDE for i.i.d. data.

Nevertheless, in the search for features underlying

$$\text{multi-modal, non-smooth, } \textit{or} \text{ non-Gaussian distributions} \tag{2.13}$$

$$\text{from correlated data points } \{\boldsymbol{z}^T \boldsymbol{X}_{\cdot,i}\}_{i=1}^{n} \tag{2.14}$$

onto the data-dependent feasible vector $\boldsymbol{z}$ (discussed in Section 2.2.1), the proposed Epdf $\hat{g}_z$ will better uncover latent structures from the sampling distribution of $\{\boldsymbol{z}^T \boldsymbol{X}_{\cdot,i}\}_{i=1}^{n}$ in a simpler way. As an illustration, consider three types of bimodal distribution functions: G in (3.5), the 'Type-G ' distribution G_{u_0,σ_0} in Section 3.1.5, and G^* in (3.8). Three panels in Figure 1 compare the finite-sample performance of the proposed Epdf $\hat{g}_z$ using 10 bins with the KDE (via the Matlab function ksdensity) applied to the same projected data $\{\boldsymbol{z}^T \boldsymbol{X}_{\cdot,i}\}_{i=1}^{n}$. The Epdf retains the shape of the bimodal mixture distribution, with the two modes more accurately caught. In comparison, the KDE tends to be oversmoothed in fitting distributions which have multiple peaks, bumps or spikes. Besides, the KDE in the left panel of Figure 1 is comparable to the example of KDE (indicated by the blue solid line) given in Bickel *et al.*, (2018) (on p. 5, Figure 1).

To overcome challenging issues (2.13) and (2.14), an adaptive choice of bandwidth parameter could be developed to improve the accuracy of KDE while preserving the smoothness of the estimates. The resulting procedure, in practice, will substantially escalate the computational complexity, and would be time consuming as well. For the exploratory PP, the Epdf eases the complexity and more closely resembles the shape (not necessarily smooth) of the sampling distribution of data points $\{\boldsymbol{z}^T \boldsymbol{X}_{\cdot,i}\}_{i=1}^{n}$, and thus works well for finding the structure or pattern from projected data, without a loss of sensitivity.

3 Graphically Illustrating Feasibility of PP in Feature Extraction

To illustrate the performance of the numerical tool in Section 2.2 used for assessing feasibility results of PP in Bickel *et al.* (2018) (reviewed in Section 2.1 and summarized in Table 1) and the extended results (listed in Appendix A.1), we conduct numerical experiments. In each simulation study, for each sample, we randomly generate a data matrix $\mathbf{X} = (\boldsymbol{X}_{\cdot,1}, \ldots, \boldsymbol{X}_{\cdot,n}) \in \mathbb{R}^{p \times n}$ consisting of data vectors $\boldsymbol{X}_{\cdot,1}, \ldots, \boldsymbol{X}_{\cdot,n} \overset{\text{i.i.d.}}{\sim} \boldsymbol{X}$, where variables $X_1, \ldots, X_p$ in $\boldsymbol{X}$ are i.i.d. $\mathcal{N}(0,1)$ according to (2.1) as discussed in Bickel *et al.* (2018), unless otherwise specified. To perform a visual assessment from a single data matrix $\mathbf{X}$, we utilize the

© 2022 The Authors. International Statistical Review published by John Wiley & Sons Ltd on behalf of International Statistical Institute.
17515823, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/insr.12517 by Chunming Zhang - University Of Wisconsin - Madison , Wiley Online Library on [04/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

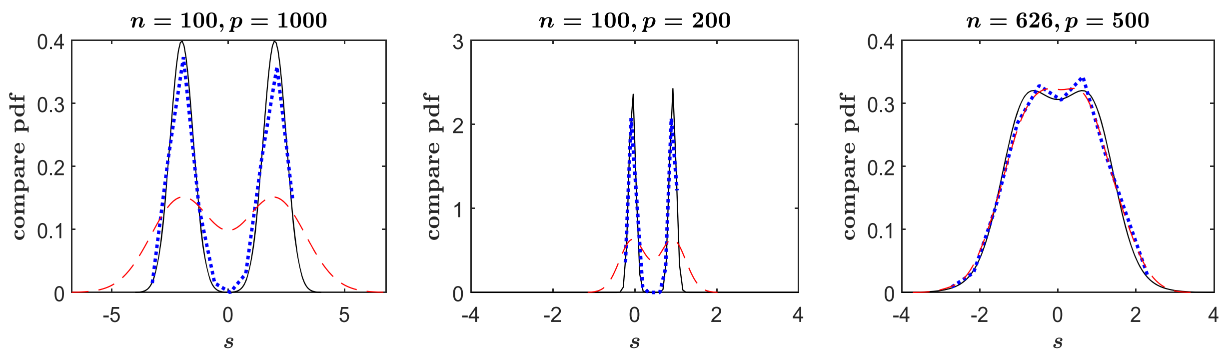

FIGURE 1. *Compare* Epdf *and* KDE *for estimating density functions. Left panel: for G in (3.5); middle panel: for* G_{u_0,σ_0} *in Section 3.1.5; right panel: for* G^* *in (3.8). True* p.d.f.*: solid line* $-$; Epdf*: dots ..;* KDE*: dashed line* $--$. *The online version of this figure is in colour*

© 2022 The Authors. International Statistical Review published by John Wiley & Sons Ltd on behalf of International Statistical Institute.

17515823, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/insr.12517 by Chunming Zhang - University Of Wisconsin - Madison , Wiley Online Library on [04/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Epdf developed in (2.12) with 10 bins. To more precisely characterize the KS distance, we present boxplots of $\|\hat{G}_z - G\|_\infty$ across 100 replicate samples calculated via either (2.9) for one-dimensional PPs in Section 3.1 or a grid approximation for two-dimensional PPs in Section 3.2.

For the sake of comparison, the following types of projection vectors are examined:

1. A data-dependent feasible vector which 'exists' in asymptotic feasibility results (abbreviated 'z (exist)' in what follows),

$$\begin{aligned} &z(\mathbf{X}^{o}, G) \text{of length 1, from} (2.6) - (2.7), \text{ using the data matrix} \mathbf{X}^{o} \in \mathbb{R}^{p \times n^{o}}, \\ &\text{and source vector } \boldsymbol{a}^{o} \in \mathbb{R}^{n^{o}} \text{with entries} G^{-1}(\frac{i}{n^{o}+1}), \ i = 1, \ldots, n^{o}, \end{aligned} \tag{3.1}$$

where

$$\mathbf{X}^{o} = \begin{cases} \mathbf{X}, & \text{if} p \geq n, \\ (\boldsymbol{X}_{\cdot,1}, \ldots, \boldsymbol{X}_{\cdot,n_1}), & \text{if} p < n, \end{cases} \text{ and } n^{o} = \begin{cases} n, & \text{if} p \geq n, \\ n_1, & \text{if} p < n, \end{cases}$$

with $n_1 = \lfloor p/L \rfloor$ discussed in Section 2.2.1 for the case $p < n$.

2. A data-dependent vector,

$$z = \frac{\boldsymbol{d}}{(\boldsymbol{d}^T\boldsymbol{d})^{1/2}}, \text{ where } \boldsymbol{d} = (d_1, \ldots, d_p)^T, \text{ with entries } d_j = \sum_{i=1}^{n} X_{j,i}/n. \tag{3.2}$$

3. A data-dependent vector,

$$z = \frac{\boldsymbol{d}}{(\boldsymbol{d}^T\boldsymbol{d})^{1/2}}, \text{ where } \boldsymbol{d} = (d_1, \ldots, d_p)^T \text{is independent of } (\boldsymbol{X}_{\cdot,1}, \ldots, \boldsymbol{X}_{\cdot,n}). \tag{3.3}$$

4. A data-dependent 'critical' vector,

$$z_{\text{crt}} \approx \arg \min_{z \in \mathbb{R}^p : z^T z = 1} \sum_{i=1}^{n} (z^T \boldsymbol{X}_{\cdot,i} - a_i)^2, \tag{3.4}$$

where a_i's are given in Step 1 of Section 2.2.1.

With regard to the choice between data-'independent' and data-'dependent' directions z for the originally independent data vectors $\{\boldsymbol{X}_{\cdot,i}\}_{i=1}^{n}$, the former choice preserves the independence among the projected data $\{z^T\boldsymbol{X}_{\cdot,i}\}_{i=1}^{n}$, whereas the latter one induces correlation among those points for which typical results of KDE may not hold. The feasible direction given in (3.1) follows the discussion of (2.6)–(2.8) in Section 2.2.1. In the absence of a feasible direction, an exhaustive search for all possible projection vectors is impractical. The candidate vectors in (3.2), (3.3) and (3.4) are for illustrative purposes only; alternative vectors can be added when necessary to the comparisons.

3.1 One-dimensional PP

3.1.1 Illustrate Theorem 1 *and* Theorem 2(i) *in Bickel et al. (2018) with* $p > n$*: feasible* G

We assess the feasibility of PP in retrieving the bimodal feature in the Gaussian mixture distribution,

© 2022 The Authors. International Statistical Review published by John Wiley & Sons Ltd on behalf of International Statistical Institute.

17515823, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/insr.12517 by Chunming Zhang - University Of Wisconsin - Madison , Wiley Online Library on [04/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$$G = \frac{1}{2}\mathbb{N}\left(-2, \frac{1}{2^2}\right) + \frac{1}{2}\mathbb{N}\left(2, \frac{1}{2^2}\right), \text{ with} \mu_2(G) = 4.25, \tag{3.5}$$

using simulated datasets of sample size $n = 100$ and dimension $p = 1000$, satisfying $\gamma = p/n = 10 > 1$ and $\mu_2(G) < \gamma - 1$.

In Figure 2, the left panel compares

(i) the true bimodal p.d.f. of the target distribution G in (3.5),
(ii) the developed Epdfs in (2.12) for the projected data $\{z^T \boldsymbol{X}_{\cdot,i}\}_{i=1}^n$ using four types of projection vectors: $z_1 = z(\mathbf{X}, G)$ in (3.1), z_2 in (3.2), z_3 in (3.3) with $\boldsymbol{d} \sim \mathbb{N}(\mathbf{0}, \mathbf{I}_p)$, where $\mathbf{I}_p$ denotes a $p \times p$ identity matrix, and z_{crt} in (3.4),

based on one simulated data matrix $\mathbf{X}$. Clearly, the direction 'z_1 (exist)' outperforms z_{crt} in extracting the true bimodal structure, whereas vectors z_2 and z_3 suggest unimodal distributions that do not reflect the true structure. The right panel compares boxplots of $\|\hat{G}_z - G\|_\infty$, exhibiting more pronounced distinction between choices of projection vectors. For the feasible direction z_1, $\|\hat{G}_{z_1} - G\|_\infty$ decreases to zero, and thus the data projection onto z_1 can well recover the two modes of the target distribution G; vectors z_2, z_3 and z_{crt} increase $\|\hat{G}_z - G\|_\infty$ up to 0.7, 0.4 and 0.26, respectively, substantially above zero, and hence are not eligible for the feasible projection direction even when the distribution G is feasible.

3.1.2 *Illustrate* Remark *1 and* Remark *3 in Bickel et al. (2018) with $p > n$: feasible G*

To illustrate Remark 1 and Remark 3 in Bickel *et al.* (2018), we adopt the same target distribution G as in (3.5), but assume that data are i.i.d. realizations of a non-Gaussian random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ with i.i.d. variables $X_1, \ldots, X_p$ following some non-Gaussian distribution F according to (2.2). Sample size $n = 100$ and dimension $p = 1000$ are used, with $\gamma = p/n = 10 > 1$ and $\mu_2(G) < (\sqrt{\gamma} - 1)^2$. Two types of mean-zero sub-Gaussian distributions F are considered,

$$F = \text{Uniform}(-3, 3), \text{ and } F = \text{C.D.F. of } 6(B_{1,2} - 1/3),$$

where $B_{1,2}$ denotes the random variable having a beta distribution with parameters 1 and 2. The boxplots of $\|\hat{G}_z - G\|_\infty$ over 100 samples are displayed in Figure 3. Similar to the choices of

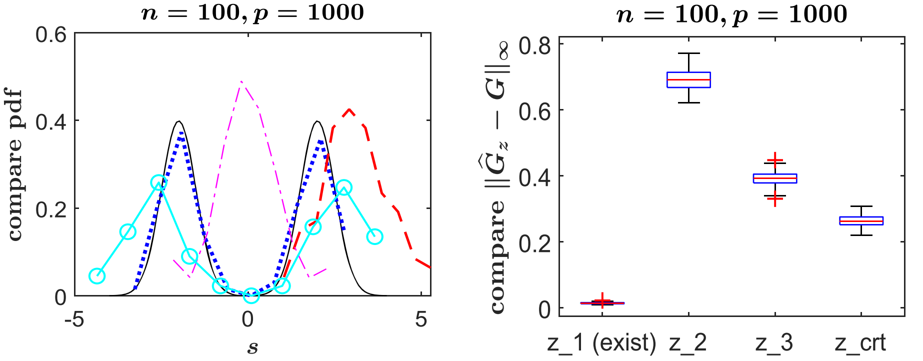

FIGURE 2. *Illustrate* Theorem *1 and* Theorem 2(i) *summarized in Table 1 for Bickel et al. (2018) with $p > n$: feasible distribution G in (3.5). Left panel: compare the* Epdfs *of $\{z^T \boldsymbol{X}_{\cdot,i}\}_{i=1}^n$ using vectors 'z_1(exist)' in (3.1) (dots ..), z_2 (dashed line −−), z_3 (dashed dotted line −·), z_{crt} (solid line with circles o–o), and the true* p.d.f. *(solid line −) of the* C.D.F. *G. Right panel: compare boxplots of $\|\hat{G}_z - G\|_\infty$ using vectors 'z_1 (exist)', z_2, z_3 and z_{crt}. The online version of this figure is in colour*

© 2022 The Authors. International Statistical Review published by John Wiley & Sons Ltd on behalf of International Statistical Institute.

17515823, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/insr.12517 by Chunming Zhang - University Of Wisconsin - Madison , Wiley Online Library on [04/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

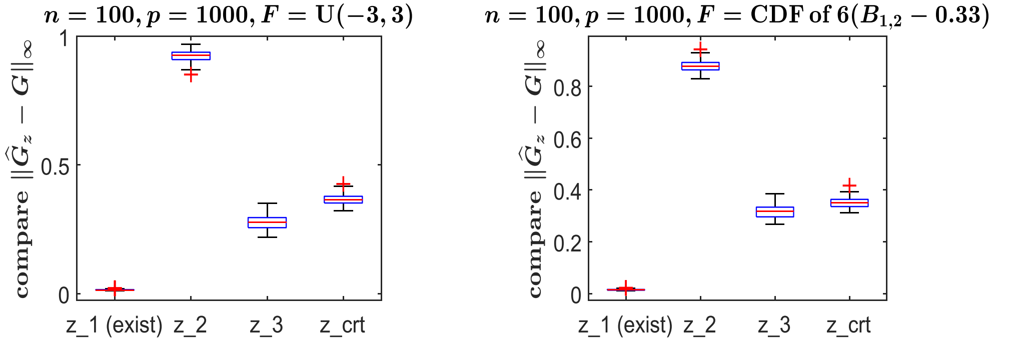

FIGURE 3. *Illustrate* Remark *1 and* Remark *3 summarized in Table 1 for Bickel et al. (2018) with $p > n$: feasible distribution G in (3.5). The caption is similar to the right panel of Figure 2. Left:* $X_1, \ldots, X_p \overset{\text{i.i.d.}}{\sim} F = \mathrm{Uniform}(-3, 3)$*; right:* $X_1, \ldots, X_p \overset{\text{i.i.d.}}{\sim} F =$C.D.F.*of* $6(B_{1,2} - 1/3)$*. The online version of this figure is in colour*

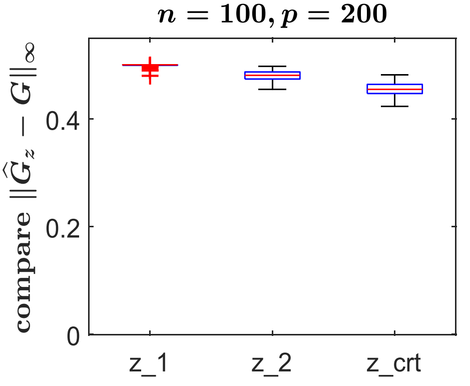

FIGURE 4. *Illustrate* Theorem 2(ii) *summarized in Table 1 for Bickel et al. (2018) with $p > n$: infeasible distribution G in (3.6). Compare boxplots of $\|\hat{G}_z - G\|_\infty$ using different projection vectors. The online version of this figure is in colour*

data-dependent and data-independent projection vectors in Figure 2, the bimodal feature of the distribution G can feasibly be reconstructed from projecting non-Gaussian datasets onto the data-dependent vector '$\boldsymbol{z}_1$ (exist)', while none of the other vectors achieves this goal.

3.1.3 *Illustrate* Theorem 2(ii) *in Bickel et al. (2018) with $p > n$: infeasible G*

In an attempt to visually inspect the infeasibility result of Theorem 2(ii) in Bickel et al. (2018), we consider the target distribution,

$$G = \frac{1}{2}\mathbb{N}\left(-3, \frac{1}{2^2}\right) + \frac{1}{2}\mathbb{N}\left(3, \frac{1}{2^2}\right), \quad \text{with}\,\mu_2(G) = 9.25, \tag{3.6}$$

and simulate samples of size $n = 100$, dimension $p = 200$, with $\gamma = p/n = 2$, satisfying $\mu_2(G) > (\sqrt{\gamma} + 1)^2$. Three types of projection vectors are compared: 'data-dependent' $\boldsymbol{z}_1$ in (3.2), 'data-dependent' $\boldsymbol{z}_2$ in (3.3) with $\boldsymbol{d} \sim \mathbb{N}(\boldsymbol{0}, \mathbf{I}_p)$, and $\boldsymbol{z}_{\mathrm{crt}}$ in (3.4).

The magnitudes of the KS distance $\|\hat{G}_z - G\|_\infty$ in boxplots of Figure 4 consistently exceed 0.4, across all these projection vectors, lending numerical support to the claim that PP is not

© 2022 The Authors. International Statistical Review published by John Wiley & Sons Ltd on behalf of International Statistical Institute.

17515823, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/insr.12517 by Chunming Zhang - University Of Wisconsin - Madison , Wiley Online Library on [04/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

capable of recovering the distribution G in (3.6) when $\gamma \in (1, \infty)$ and $\mu_2(G) > (\sqrt{\gamma}+1)^2$, no matter which vector of unit length is used.

3.1.4 *Illustrate the length of* z_0 *in* Theorem 2(i) *and* (ii) *in Bickel et al. (2018) with* $p > n$

Recall from Section 2.2.1 that the algorithmic simplicity in the search for the feasible projection vector z, in asymptotic feasibility results, is partly attributed to the achievability of (2.6)–(2.7). It is thus natural to check whether the underlying condition $z_0^T z_0 \le 1$ is fulfilled, in finite-samples, for the initial vector z_0 in (2.6). There, we consider the same source vector $\boldsymbol{a}$ as in Step 1 of Section 2.2.1, the same distribution G as in (3.6), sample size $n = 100$ and dimension $p = n \times \gamma$ with $1 < \gamma \le 20$. From Figure 5, the case of $\gamma > \mu_2(G)+1$ gives $z_0^T z_0 \le 1$, coinciding with the feasibility of the distribution G declared in Theorem 2(i). In contrast, the case of $\gamma < \{\sqrt{\mu_2(G)} - 1\}^2$ reflects $z_0^T z_0 > 1$, in accordance with the infeasibility of the distribution G claimed in Theorem 2(ii).

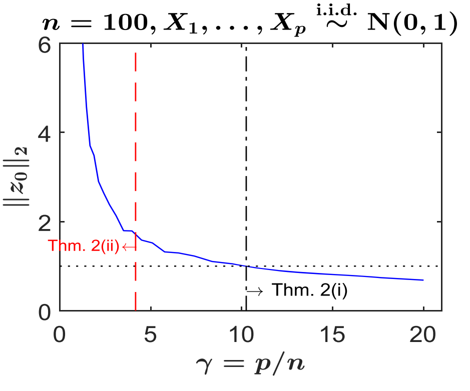

FIGURE 5. *Illustrate the length of* z_0 *in* Theorem 2(i) *and* (ii) *summarized in Table 1 for (Bickel et al., 2018). Plots of* $(z_0^T z_0)^{1/2}$ *(solid line –) as the ratio* $\gamma = p/n$ *increases. The case of* $\gamma < \{\sqrt{\mu_2(G)} - 1\}^2$ *(dashed line – –) corresponds to* Theorem 2(ii)*; the case of* $\gamma > \mu_2(G)+1$ *(dashed dotted line – ·) corresponds to* Theorem 2(i)*. The online version of this figure is in colour*

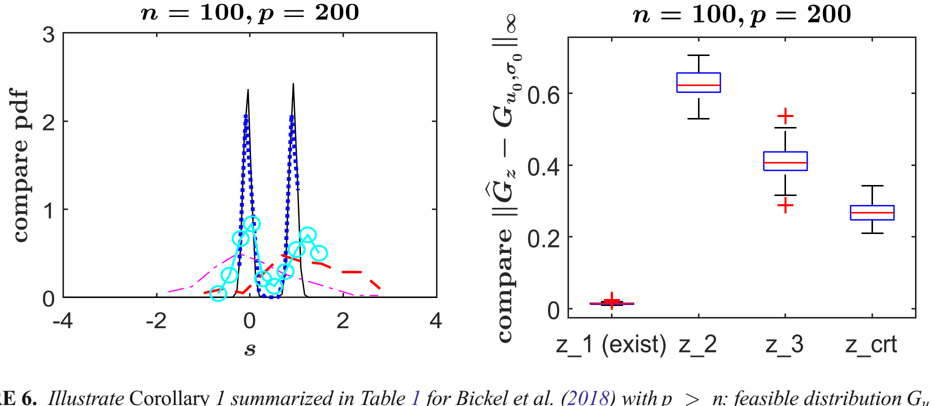

FIGURE 6. *Illustrate* Corollary *1 summarized in Table 1 for Bickel et al. (2018) with* $p > n$*: feasible distribution* G_{u_0,σ_0} *for* G *in (3.6). The caption is similar to that of Figure 2. The online version of this figure is in colour*

© 2022 The Authors. International Statistical Review published by John Wiley & Sons Ltd on behalf of International Statistical Institute.

17515823, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/insr.12517 by Chunming Zhang - University Of Wisconsin - Madison , Wiley Online Library on [04/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

3.1.5 *Illustrate* Corollary *1 in Bickel et al. (2018) with* $p > n$*: feasible* G_{u_0,σ_0}

In this case, we consider the same distribution G as in (3.6), together with sample size $n = 100$ and data dimension $p = 200$. The 'Type-G' distribution function (defined in Appendix A), $G_{u_0,\sigma_0}(s) = G(\frac{s - u_0}{\sigma_0})$, is chosen according to the location and scale constants,

$$\sigma_0 = \sqrt{(\gamma - 1)/\mu_2(G)}/2,\ u_0 = \{\max(u_1, 0)+u_2\}/2,$$

with $u_1 < u_2$ corresponding to two roots of the equation $u^2 + 2\{\sigma_0\mu_1(G)\}u - \{(\gamma - 1) - \sigma_0^2\mu_2(G)\} = 0$; the lengthy derivations of σ_0 and u_0 are omitted.

Unlike Figure 4, where the distribution G itself is infeasible, that is, not extractable by PP onto any projection vectors, the boxplots of $\|\hat{G}_z - G_{u_0,\sigma_0}\|_\infty$ in the right panel of Figure 6 clearly support the claim that the 'Type-G' distribution G_{u_0,σ_0} can be well approximated by projection onto the suitable data-based vector 'z_1 (exist)', that is, $z_1(\mathbf{X}, G_{u_0,\sigma_0})$ in (3.1). In contrast, vectors z_2, z_3 and z_{crt} as used in Figure 2 could not act as the proper projection directions. An empirical evidence is similarly found from the proposed Epdf in the left panel of Figure 6.

3.1.6 *Illustrate* Theorem *3 in Bickel et al. (2018) with* $p < n$*: feasible* G^*

To illustrate Theorem 3 in Bickel et al. (2018), we simulate datasets of sample size $n = 626$, dimension $p = 500$, with $\gamma = p/n = 0.7987$, and adopt the distribution,

$$G = \frac{1}{2}\mathbb{N}\left(-1, \frac{1}{2^2}\right) + \frac{1}{2}\mathbb{N}\left(1, \frac{1}{2^2}\right),\ \text{with}\,\mu_2(G) = 1.25, \tag{3.7}$$

which satisfies the condition $\mu_2(G) < L - 1$ with $L = 2.5$ in Theorem 3.

In Figure 7, the left panel compares the true density function of the target distribution,

$$G^* = (\gamma/L)G + (1 - \gamma/L)\Phi, \tag{3.8}$$

using the developed Epdfs in (2.12) along vectors: $z_1 = z(\mathbf{X}, G)$ described in (3.1), z_2 in (3.2), z_3 in (3.3) with entries of $\boldsymbol{d}$ i.i.d. from the Uniform$(0, 1)$ distribution, and z_{crt} in (3.4), based on one simulated data matrix $\mathbf{X}$. The multi-modal feature in (3.8) is better extracted from projection onto the data-dependent feasible projection vector 'z_1 (exist)' than z_{crt}, but was obscured from the other vectors. The boxplots of $\|\hat{G}_z - G^*\|_\infty$ in the right panel reveal the superiority of employing the vector 'z_1 (exist)' over others in recovering the true feature in distribution (3.8).

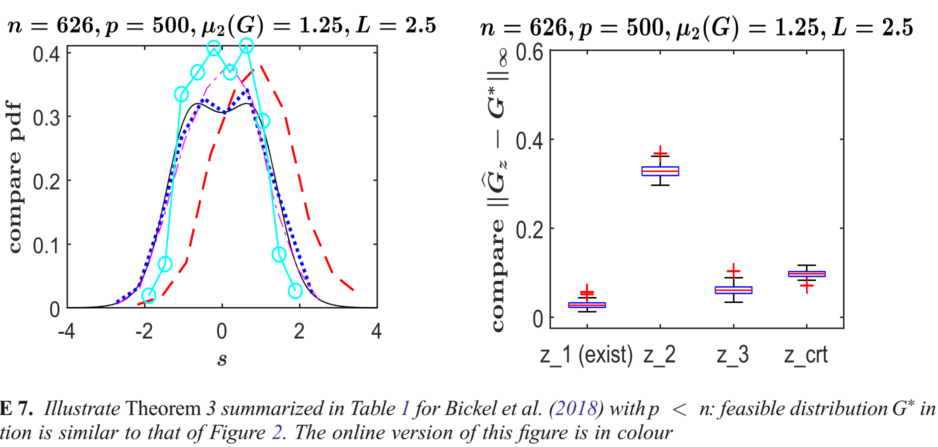

FIGURE 7. *Illustrate* Theorem *3 summarized in Table 1 for Bickel et al. (2018) with* $p < n$*: feasible distribution* G^* *in (3.8). The caption is similar to that of Figure 2. The online version of this figure is in colour*

© 2022 The Authors. International Statistical Review published by John Wiley & Sons Ltd on behalf of International Statistical Institute.

17515823, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/insr.12517 by Chunming Zhang - University Of Wisconsin - Madison , Wiley Online Library on [04/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Also, due to the nature of the specific direction z_3, we infer directly that $\{z_3^T \boldsymbol{X}_{\cdot,1}, \ldots, z_3^T \boldsymbol{X}_{\cdot,n}\} \overset{\text{i.i.d.}}{\sim} \mathbb{N}(0,1)$, and thus $\|\hat{G}_{z_3} - G^*\|_\infty$ is lower than $\|\hat{G}_{z_2} - G^*\|_\infty$ and $\|\hat{G}_{z_{\text{crt}}} - G^*\|_\infty$.

3.1.7 *Illustrate* Theorem *4 in Bickel et al. (2018) with $p < n$: infeasible G*

Recall that Theorem 4 in Bickel et al. (2018) conveys an infeasibility result: in the low-dimensional case, a target distribution which is far from the standard Gaussian distribution could not be extracted by PP onto any directions. To illustrate such impossibility result, we simulate $n = 626$ data vectors of dimension $p = 500$, with $\gamma = p/n = 0.7987$, and adopt the non-Gaussian distribution G as in (3.5). Five types of directions are inspected: z_1 is as in (3.2); $z_2 = \boldsymbol{X}_{\cdot,10}/(\boldsymbol{X}_{\cdot,10}^T \boldsymbol{X}_{\cdot,10})^{1/2}$; z_3 and z_4 are as in (3.3), where all entries in $\boldsymbol{d}$ are i.i.d. following the $\mathbb{N}(0,1)$ distribution in z_3 and the Uniform$(0,100)$ distribution in z_4; z_{crt} in (3.4). As noticed from Figure 8, the aberration of $\|\hat{G}_z - G\|_\infty$ from zero provides evidence to support Theorem 4.

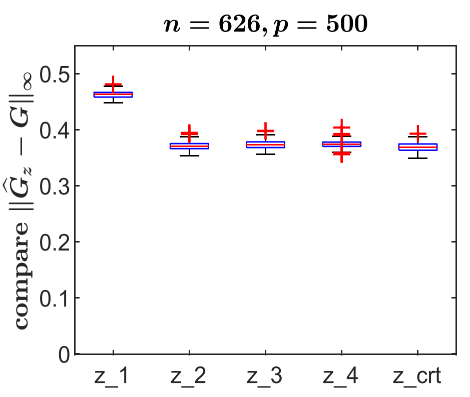

FIGURE 8. *Illustrate* Theorem *4 summarized in Table 1 for Bickel et al. (2018) with $p < n$: infeasible distribution G in (3.5). Compare boxplots of $\|\hat{G}_z - G\|_\infty$ using different projection vectors. The online version of this figure is in colour*

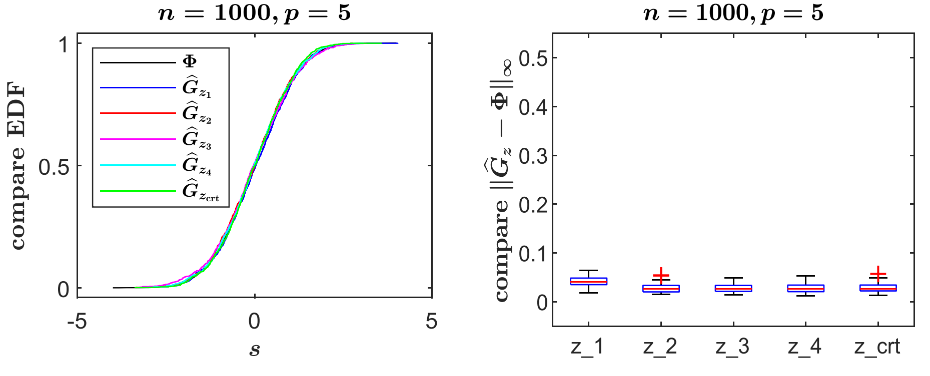

FIGURE 9. *Illustrate* Theorem *5 summarized in Table 1 for (Bickel et al., 2018) with $p << n$: uniquely feasible distribution Φ. Left: compare the* E.D.F.s *$\hat{G}_z$ using different projection vectors with the standard Gaussian* C.D.F. *Φ. Right: compare boxplots of $\|\hat{G}_z - \Phi\|_\infty$ using different projection vectors. The online version of this figure is in colour*

© 2022 The Authors. International Statistical Review published by John Wiley & Sons Ltd on behalf of International Statistical Institute.

17515823, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/insr.12517 by Chunming Zhang - University Of Wisconsin - Madison , Wiley Online Library on [04/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

3.1.8 *Illustrate* Theorem *5 in Bickel et al. (2018) with* $p \ll n$*: uniquely feasible* Φ

For the illustration of Theorem 5, we generate $n = 1000$ data vectors of dimension $p = 5$, that is, $\gamma = 0.005$, and consider the same types of projection vectors as used in Figure 8.

For all these vectors z_1, z_2, z_3, z_4 and z_{crt}, the left panel in Figure 9 plots the corresponding empirical distributions $\hat{G}_z$, in striking consistency with the Gaussian distribution function Φ. The boxplots in the right panel confirm that the KS distances $\|\hat{G}_z - \Phi\|_\infty$ closely approach zero, no matter whether projection directions are dependent on, weakly dependent on, or independent of data.

3.1.9 *Illustrate* Theorem *6 in Bickel et al. (2018) with sparse projections* $\|z\|_0 \ll n$*: uniquely feasible* Φ

To demonstrate Theorem 6 which states that sparse projection vectors can only recover the Gaussian distribution Φ, we take the sample size $n = 1000$, dimension $p \in \{1200;\ 120\}$ and $\mathrm{s} = 5$ in the vector z, covering both the high- and low-dimensional cases with sparse projection vectors. The following types of sparse vectors z are considered: z_1, z_2, z_3, z_4 and z_{crt} similar to those used in Figure 9, except that the first s coordinates are non-zero and the rest $(p - \mathrm{s})$ entries are zero.

In Figure 10, the boxplots indicate that for both the high-dimensional (corresponding to $p > n$ in the left panel) and low-dimensional (corresponding to $p < n$ in the right panel) data, the KS distances $\|\hat{G}_z - \Phi\|_\infty$ using all these sparse vectors, are in a small neighbourhood of zero, agreeing with Theorem 6 where Φ is uniquely feasible.

3.2 *Two-dimensional* PP

In practice, projection strategies are frequently used for projecting multivariate data down to one-, two-, or even three-dimensional space. An approach for three-dimensional PP was given in Nason (1995); see Klinke (1995) and Jee (2009) for further discussions on the dimension of the projection space. This subsection focuses on the two-dimensional PP for high-dimensional datasets to ease numerical work and graphical presentation.

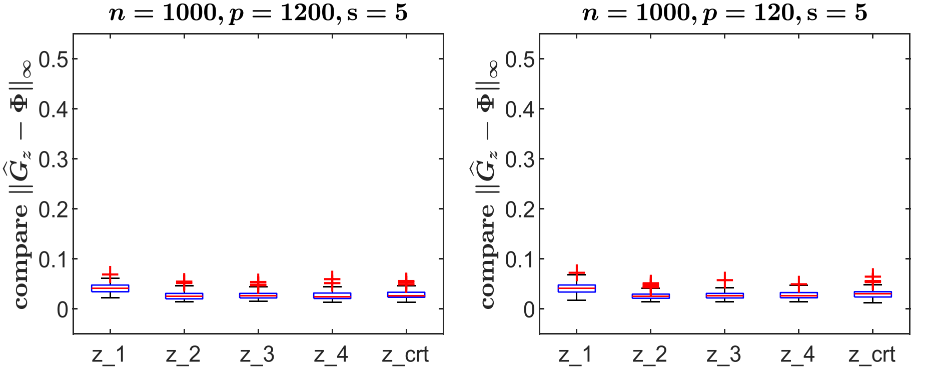

FIGURE 10. *Illustrate* Theorem *6 summarized in Table 1 for Bickel et al. (2018) with* $\|z\|_0 < < n$*: uniquely feasible distribution* Φ*. The caption is similar to that of Figure 8. Left: dimension* $p = 1200$*. Right: dimension* $p = 120$*. The online version of this figure is in colour*

© 2022 The Authors. International Statistical Review published by John Wiley & Sons Ltd on behalf of International Statistical Institute.

17515823, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/insr.12517 by Chunming Zhang - University Of Wisconsin - Madison , Wiley Online Library on [04/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

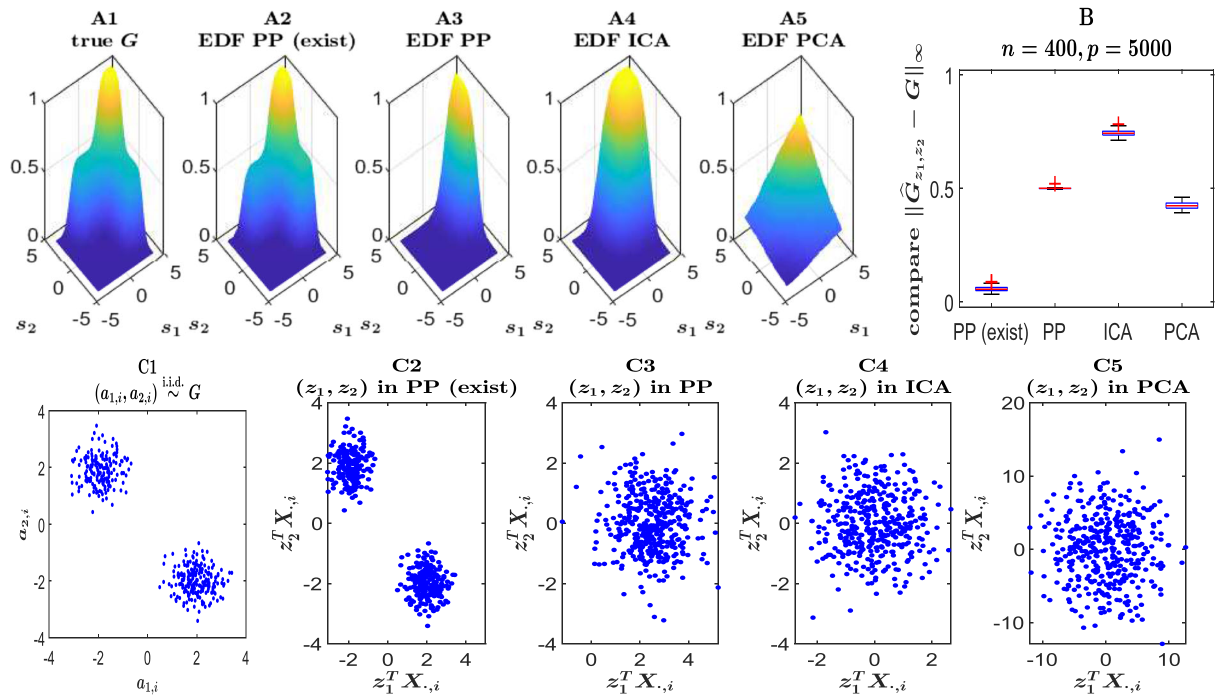

FIGURE 11. *Illustrate* Remark *2 summarized in Table 1 for Bickel et al. (2018) for two-dimensional* PP *and* $p > n$*: feasible distribution* G *in (3.9).* ***(A1)–(A5)*** *Compare true* C.D.F. $G(s_1, s_2)$ *in (3.9) with the bivariate* E.D.F.s *of* $\{(z_1^T X_{\cdot,i}, z_2^T X_{\cdot,i})\}_{i=1}^n$ *obtained from a* 'PP *(exist)' direction, a candidate* PP *direction,* ICA *and* PCA *methods.* ***(B)*** *Compare boxplots of* $\|\hat{G}_{z_1,z_2} - G\|_\infty$ *using the 4 types of projections.* ***(C1)–(C5)*** *Compare the scatterplot of sampled points* $\{(a_{1,i}, a_{2,i})\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} G$ *with score plots of* $\{(z_1^T X_{\cdot,i}, z_2^T X_{\cdot,i})\}_{i=1}^n$ *using the 4 types of projections. The online version of this figure is in colour*

© 2022 The Authors. International Statistical Review published by John Wiley & Sons Ltd on behalf of International Statistical Institute.

17515823, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/insr.12517 by Chunming Zhang - University Of Wisconsin - Madison , Wiley Online Library on [04/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

3.2.1 *Illustrate* Remark *2 in Bickel et al. (2018) for two-dimensional* PP *and* $p > n$*: feasible* G

To illustrate Remark 2 in Bickel et al. (2018) for the two-dimensional PP and to compare with ICA and PCA using two components, we consider the bivariate Gaussian mixture distribution plotted in Figure 11**A1**, with the joint distribution function,

$$G = \frac{1}{2}\mathbb{N}_2\left(\begin{pmatrix} -2 \\ 2 \end{pmatrix}, \begin{pmatrix} 1/2^2 & 0 \\ 0 & 1/2^2 \end{pmatrix}\right) + \frac{1}{2}\mathbb{N}_2\left(\begin{pmatrix} 2 \\ -2 \end{pmatrix}, \begin{pmatrix} 1/2^2 & 0 \\ 0 & 1/2^2 \end{pmatrix}\right), \tag{3.9}$$

together with its mean vector and covariance matrix equal to

$$\boldsymbol{\mu}_G = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ and } \Sigma_G = \begin{pmatrix} 4.25 & -4.00 \\ -4.00 & 4.25 \end{pmatrix}. \tag{3.10}$$

The simulated data sets include 400 data points in the 5000 dimensional space.

Figure 11 reflects the advantages of PP onto its feasible direction in uncovering the bimodal feature of the distribution G.

(i) Along the feasible bivariate projection direction (z_1, z_2) (abbreviated 'PP (exist)'), which is a bivariate extension of (3.1), the bivariate E.D.F. defined in (2.10) of $\{(z_1^T \boldsymbol{X}_{\cdot,i}, z_2^T \boldsymbol{X}_{\cdot,i})\}_{i=1}^n$, plotted in Figure 11**A2**, restores features underlying the target distribution G. Here, constructions of the data-dependent orthonormal directions (z_1, z_2) (in ' PP (exist)') are similar to those used in the proof of our extended Lemma E.1 in Appendix A.1, where source vectors $\boldsymbol{a}_1 = (a_{1,1}, \ldots, a_{1,n})^T$ and $\boldsymbol{a}_2 = (a_{2,1}, \ldots, a_{2,n})^T$ are chosen such that the bivariate random vectors $\{(a_{1,i}, a_{2,i})\}_{i=1}^n$ are i.i.d. following the bivariate distribution G. See the scatterplot of $a_{1,i}$ versus $a_{2,i}$ in Figure 11**C1**. Moreover, Figure 11**C2** easily identifies that the projected data $\{(z_1^T \boldsymbol{X}_{\cdot,i}, z_2^T \boldsymbol{X}_{\cdot,i})\}_{i=1}^n$ onto the feasible bivariate PP direction fall into two separated clusters, which go undetected by other three methods (PP onto a candidate direction, ICA and PCA) in Figure 11**C3–C5**, respectively.

(ii) For a candidate bivariate direction (z_1, z_2) in PP, where z_1 is in (3.2) and z_2 is in (3.3) with entries in $\boldsymbol{d}$ i.i.d. from the $\mathbb{N}(0,1)$ distribution, the E.D.F. in Figure 11**A3** is unable to find the bimodal distribution G.

(iii) The E.D.F.s (Figure 11**A4** and 11**A5**) of both the statistically independent components using the (standard linear) ICA method (via the FastICA algorithm (Hyvärinen & Oja, 2000), with the 'pow3' nonlinearity, 'symm' orthogonalization, and retaining 2 largest eigenvalues) and the principal components using PCA (via the singular value decomposition) bear little resemblance to the target distribution G.

Further comparisons are made in the boxplots in Figure 11**B**: values of the KS distances $\|\hat{G}_{z_1,z_2} - G\|_\infty$ are surrounded by 0.0 using the 'PP (exist)' direction, close to 0.5 using the candidate direction in PP, rising to 0.7 using ICA, and dropping to 0.4 using PCA, respectively. This is due to the fact:

(a) In view of PCA (Jolliffe & Cadima, 2016), it requires the principal components to be uncorrelated, while preserving as much 'variation' in a dataset as possible. As remarked in Guo *et al.* (2001), Bouveyron & Brunet-Saumard (2014), and Lever *et al.* (2017), PCA may not always find interesting data features, like clusters.

© 2022 The Authors. International Statistical Review published by John Wiley & Sons Ltd on behalf of International Statistical Institute.

17515823, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/insr.12517 by Chunming Zhang - University Of Wisconsin - Madison , Wiley Online Library on [04/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(b) The ICA method, as a useful extension of PCA, extracts hidden components as independent and non-Gaussian as possible, whereas the target distribution in (3.9) echoes the joint distribution of two dependent random variables with the covariance matrix in (3.10).

Apart from the bivariate Gaussian mixture distribution, Appendix B.1 presents additional simulation studies where the nonnormal features of a bivariate asymmetric distribution can feasibly be revealed by the PP method as opposed to other projection methods.

3.2.2 *Illustrate extended results in Appendix A.1 for two-dimensional* PP

To illustrate the feasibility result in our extended Result E.1 in Appendix A.1, we consider two distribution functions,

$$G_1 = \frac{1}{2}\mathbb{N}(-1,\, 1^2) \quad + \frac{1}{2}\mathbb{N}(1,\, 1^2), \qquad G_2 = \frac{1}{2}\mathbb{N}\left(-3,\, \frac{1}{2^2}\right) \quad + \frac{1}{2}\mathbb{N}\left(3,\, \frac{1}{2^2}\right).$$

The boxplots of $\|\hat{G}_{z_k;\,k} - G_{k;\,u_{0,k},\,\sigma_{0,k}}\|_\infty$, $k = 1, 2$, in Figure 12 with sample size $n = 100$ and data dimension $p = 400$ indicate that the pair of feasible projection directions, containing '$z_{1,1}$ (exist)' and '$z_{2,1}$ (exist)', that 'exist' in Result E.1, can be obtained using our algorithm, through source vectors $\boldsymbol{a}_{1,1} = \boldsymbol{a}_{1,1}(G_1)$ and $\boldsymbol{a}_{2,1} = \boldsymbol{a}_{2,1}(G_2)$ as in (3.1), and work better than other pairs of directions $(z_{1,2},\, z_{2,2})$ and $(z_{1,3},\, z_{2,3})$ in recovering features of the distributions $G_{1;\,u_{0,1},\,\sigma_{0,1}}(s) = G_1(\frac{s - u_{0,1}}{\sigma_{0,1}})$ and $G_{2;\,u_{0,2},\,\sigma_{0,2}}(s) = G_2(\frac{s - u_{0,2}}{\sigma_{0,2}})$, where constants $u_{0,1},\ \sigma_{0,1},\ u_{0,1},\ \sigma_{0,1}$ are determined in a way similar to that in Section 3.1.5.

Other results extended in Appendix A.1 can be illustrated via similar computational and graphical schemes, and we omit the details.

4 Discussion

Finding a suitable representation of multivariate data has wide applications in pattern recognition, blind source separation (Comon & Jutten, 2010), causal discovery, data summary, and many other scientific disciplines (Daszykowski, 2007). For computational simplicity, commonly used multivariate statistical methods, such as PP, PCA and ICA, often seek this

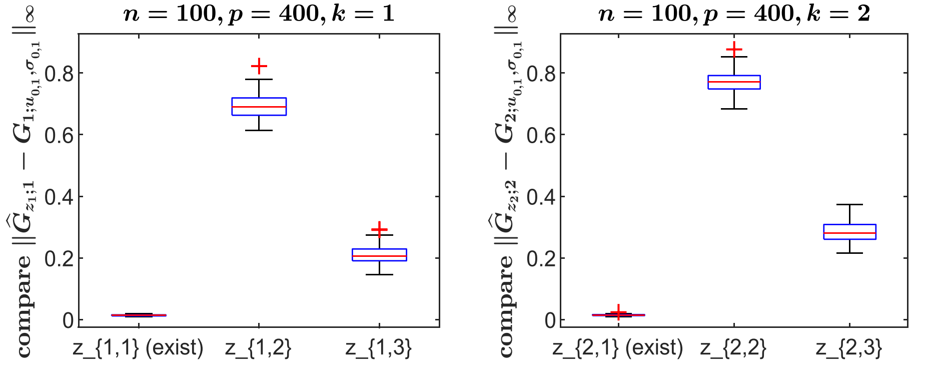

FIGURE 12. *Illustrate the extended Result E.1 in Appendix A.1 with two directions. Compare boxplots of* $\|\hat{G}_{z_k;\,k} - G_{k;\,u_{0,k},\,\sigma_{0,k}}\|_\infty$ *using 3 types of directions* z_k, '$z_{k,1}$ (exist)', $z_{k,2}$, $z_{k,3}$. *Left:* $k = 1$. *Right:* $k = 2$. *The online version of this figure is in colour*

© 2022 The Authors. International Statistical Review published by John Wiley & Sons Ltd on behalf of International Statistical Institute.

17515823, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/insr.12517 by Chunming Zhang - University Of Wisconsin - Madison , Wiley Online Library on [04/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

representation through a linear transformation in search of interesting structures of data. For large p small n data, (Lee & Cook, 2010) focuses on the projection pursuit index for classification; the recent work developed in Bickel *et al.*, (2018) offers new theoretical perspectives on the feasibility (in an asymptotic sense) of PP for high-dimensional data when both the dimension and sample size diverge.

For the exploratory data analysis, the computational and visualization tool we developed in the paper enables the asymptotic feasibility results of PP in high dimensions to be accessible to practitioners in a simplified way. On the practical side, the devised Epdf gains advantage over the traditional KDE in tracking multi-modal and/or non-Gaussian features from the sampling distribution of the projected data, which are correlated.

Although much effort in the literature on PP has been put into theoretical aspects and computational approaches for low-dimensional data, various open problems and challenging issues remain in high dimensions. These include (a) conjectures in Bickel *et al.*, (2018) and cases not yet covered for asymptotic studies, especially, observed data of correlated and/or non-Gaussian variables $X_1, \ldots, X_p$, with applications to financial time series satisfying stylized facts (De Luca & Loperfido, 2004; De Luca *et al.*, 2006); (b) handling practical data of discrete variables in discrete exploratory PP (Klinke, 1995), and analysis of discrete data arising from syllable patterns (Diaconis & Salzman, 2008), symbolic data or data with special structure in numerous disciplines; (c) nonlinear projections (Blanchard *et al.*, 2006; Guo *et al.*, 2020). There is more to be explored along the lines, driven by the need of new adaptations and methodological results, as well as real applications. Again, the numerical scheme and computational strategies will play an indispensable role and be beneficial to further expedite the exploration in the light of modern data.

ACKNOWLEDGEMENTS

The authors thank the Editor, Associate Editor and two reviewers for insightful comments. C. Zhang's work was partially supported by US National Science Foundation grants DMS-2013486 and DMS-1712418 and provided by the University of Wisconsin-Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation. J. Ye's research was supported by the Fundamental Research Funds for the Central Universities (No. QTZX22054). X. Wang's work was supported by the National Social Science Foundation of China, project number 18BGL286.

References

Bickel, P.J., Gil, K. & Boaz, N. (2018). Projection pursuit in high dimensions. *Proc. Natl. Acad. Sci.*, **115**(37), 9151–9156.

Blanchard, G., Kawanabe, M., Sugiyama, M., Spokoiny, V. & Muller, K.-R. (2006). In search of non-gaussian components of a high-dimensional distribution. *J. Mach. Learn. Res.*, **7**, 247–282.

Bouveyron, C. & Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: a review. *Comput. Statist. Data Anal.*, **71**, 52–78.

Comon, P. & Jutten, C. (2010). *Handbook of Blind Source Separation*, Independent Component Analysis and Applications. Academic Press.

Daszykowski, M. (2007). From projection pursuit to other unsupervised chemometric techniques. *J. Chemometr.*, **21**, 270–279.

De Bie, T., Lijffijt, J., Santos-Rodriguez, R. & Kang, B. (2016). Informative data projections: A framework and two examples. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN, Bruges, Belgium)*, pp. 635–640.

De Luca, G., Genton, M.G. & Loperfido, N. 2006. A Multivariate Skew-Garch Model. In *Advances in Econometrics: Econometric Analysis of Economic and Financial Time Series, Part A (Special volume in honor of Robert Engle and*

© 2022 The Authors. International Statistical Review published by John Wiley & Sons Ltd on behalf of International Statistical Institute.

17515823, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/insr.12517 by Chunming Zhang - University Of Wisconsin - Madison , Wiley Online Library on [04/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Clive Granger, the 2003 winners of the Nobel prize in Economics), Ed. Terrell, D., Vol. **20**, Elsevier: Oxford, UK, pp. 33–57.

De Luca, G. & Loperfido, N. 2004. A Skew-in-Mean GARCH Model for Financial Returns. In *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, pp. 205–222.

Devroye, L. & Gyorfi, L. (1985). *Nonparametric Density Estimation: the L_1 View*. Wiley: New York.

Diaconis, P. & Freedman, D. (1984). Asymptotics of graphical projection pursuit. *Ann. Stat.*, **12**, 793–815.

Diaconis, P. & Salzman, J. 2008. Projection pursuit for discrete data. In *IMS Collections, Probability and Statistics: Essays in Honor of David A. Freedman*, Vol. **2**, pp. 265-288.

Friedman, J.H. (1987). Exploratory projection pursuit. *J Amer. Statist. Assoc.*, **82**, 249–266.

Friedman, J.H. & Tukey, J. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.*, **23**, 881–889.

Geman, S. (1980). A limit theorem for the norm of random matrices. *Ann. Probab.*, **8**, 252–261.

Gibbons, J.D. & Chakraborti, S. (2003). *Nonparametric Statistical Inference*, 4th ed. Marcel Dekker, Inc.: New York, NY.

Groeneboom, P. & Wellner, J.A. (2001). Computing Chernoff's distribution. *J. Comput. Graph. Stat.*, **10**(2), 388–400.

Guo, Q., Wu, W., Massart, D.L., Boucon, C. & de Jong, S. (2001). Feature selection in sequential projection pursuit. *Anal. Chim. Acta*, **446**, 85–96.

Guo, R.S., Zhang, C.M. & Zhang, Z.J. (2020). Maximum independent component analysis with application to EEG data. *Stat. Sci.*, **35**(1), 145–157.

Hall, P., Minnotte, M.C. & Zhang, C.M. (2004). Bump hunting with non-Gaussian kernels. *Ann. Stat.*, **38**, 2124–2141.

Huber, P.J. (1985). Projection pursuit. *Ann. Stat.*, **13**, 435–475.

Hyvärinen, A., Karhunen, J. & Oja, E. (2001). *Independent Component Analysis*. John Wiley & Sons Inc.

Hyvärinen, A. & Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Netw.*, **13**(4-5), 411–430.

Jee, J.R. (2009). Projection pursuit. *Wiley Interdiscip. Rev.: Comput. Stat.*, **1**(2), 208–215.

Jolliffe, I.T. (2002). *Principal Component Analysis Springer Series in Statistics*. Springer-Verlag: New York.

Jolliffe, I.T. & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Phil. Trans. R. Soc. A*, **374**, 20150202.

Jones, M.C. & Sibson, R. (1987). What is projection pursuit? *J R Stat Soc Ser A (General)*, **150**(1), 1–37.

Justel, A., Pena, D. & Zamar, R. (1997). A multivariate Kolmogorov-Smirnov test of goodness of fit. *Stat. Probab. Lett.*, **35**(3), 251–259.

Klinke, S. 1995. Exploratory projection pursuit-the multivariate and discrete case. In *Proceedings of NTTS'95*, Eds. W. Kloesgen, P. Nanopoulos & A. Unwin, Bonn, pp. 247–262.

Kruskal, J.B. 1969. Toward a practical method which helps uncover the structure of a set of observations by finding the line transformation which optimizes a new "index of condensation". In *Statistical computation*, Eds. Milton, RC & Nelder, JA, Academic Press: New York, pp. 427–440.

Lee, E.K. & Cook, D. (2010). A projection pursuit index for large p small n data. *Stat. Comput.*, **20**, 381–392.

Lever, J., Krzywinski, M. & Altman, N. (2017). Points of significance: principal component analysis. *Nature Methods*, **14**(7), 641–642.

Loperfido, N. (2018). Skewness-based projection pursuit: a computational approach. *Comput. Stat. Data Anal.*, **120**, 42–57.

Markatou, M. & Sofikitou, E.M. (2019). Statistical distances and the construction of evidence functions for model adequacy. *Front. Ecol. Evol.*, **7**, 447.

Nason, G. (1995). Three-dimensional projection pursuit. *Appl. Stat.*, **44**, 1430.

Perisic, I. & Posse, C. (2005). Projection pursuit indices based on the empirical distribution function. *J. Comput. Graph. Stat.*, **14**, 700–715.

Posse, C. (1995). Projection Pursuit Exploratory Data Analysis. *Comput. Stat. Data Anal.*, **20**, 669–687.

Rudelson, M. & Vershynin, R. (2009). Smallest singular value of a random rectangular matrix. *Commun. Pure Appl. Math.*, **62**, 1707–1739.

Sasaki, H., Niu, G. & Sugiyama, M. (2016). Non-Gaussian component analysis with log-density gradient estimation. *PMLR*, **51**, 1177–1185.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall: London.

Sun, J. 2006. Projection Pursuit. In *In Encyclopedia of Statistical Sciences*, Vol. **10**, Wiley: New York.

Tyler, D., Critchley, F., Dmbgen, L. & Oja, H. (2009). Invariant co-ordinate selection (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **71**, 1–27.

van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press: Cambridge.

© 2022 The Authors. International Statistical Review published by John Wiley & Sons Ltd on behalf of International Statistical Institute.

17515823, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/insr.12517 by Chunming Zhang - University Of Wisconsin - Madison , Wiley Online Library on [04/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Virta, J., Nordhausen, K. & Oja, H. 2016. Projection pursuit for non-Gaussian independent components. arXiv:1612.05445.

[*Received December 2021; accepted July 2022*]

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

© 2022 The Authors. International Statistical Review published by John Wiley & Sons Ltd on behalf of International Statistical Institute.

17515823, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/insr.12517 by Chunming Zhang - University Of Wisconsin - Madison , Wiley Online Library on [04/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Supplementary Appendix

A Notations and symbols, and extended results

Notations and symbols. Let $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ be a random vector, and $\mathbf{X} = (X_{j,i})_{j=1,\ldots,p;\, i=1,\ldots,n} = (\boldsymbol{X}_{\bullet,1}, \ldots, \boldsymbol{X}_{\bullet,n}) \in \mathbb{R}^{p\times n}$ be the data matrix, where $\{\boldsymbol{X}_{\bullet,1}, \ldots, \boldsymbol{X}_{\bullet,n}\} \overset{\text{i.i.d.}}{\sim} \boldsymbol{X}$. For a vector $\boldsymbol{z}$, $\|\boldsymbol{z}\|_2 = \sqrt{\boldsymbol{z}^T\boldsymbol{z}}$ denotes the Euclidean norm, and $\|\boldsymbol{z}\|_0$ denotes the number of nonzeros in $\boldsymbol{z}$. Let $\mathbb{S}^{p-1} = \{\boldsymbol{z} = (z_1, \ldots, z_p)^T \in \mathbb{R}^p : \boldsymbol{z}^T\boldsymbol{z} = 1\}$ denote the unit sphere in $\mathbb{R}^p$, $\mathbf{1}_p = (1, \ldots, 1)^T$ denote a p-variate vector of ones, and $\mathbf{I}_p$ denote a $p \times p$ identity matrix. For a function $F(\boldsymbol{x}) : \mathbb{R}^k \mapsto \mathbb{R}$, we define the sup-norm of F by $\|F\|_\infty = \max_{\boldsymbol{x}\in\mathbb{R}^k} |F(\boldsymbol{x})|$. Use $\mathbb{N}(\boldsymbol{\mu}, \Sigma)$ for the multivariate Gaussian distribution with the mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. Denote by $\phi(\cdot)$ and $\Phi(\cdot)$ the density function and distribution function of a standard Gaussian distribution $\mathbb{N}(0, 1)$. For a C.D.F. G, denote its kth-moment by $\mu_k(G) = \int x^k \,\mathrm{d}G(x)$ for integers $k = 0, 1, 2, \ldots$; the family of "Type-G" C.D.F.s is defined as

$$\Big\{G_{u,\sigma}(s) = G\Big(\frac{s-u}{\sigma}\Big) : 0 \le u < \infty,\ \sigma \in (0, \infty)\Big\}. \tag{A.1}$$

For random quantities V_1 and V_2, $V_1 \perp\!\!\!\perp V_2$ denotes that V_1 and V_2 are independent. Let $[x]$ denote the largest integer less than or equal to x.

A.1 Extended results for multi-dimensional PP

To justify some results used in Sections 2.2.1 and 3.2.1 but not explicitly addressed in [2], we extend the asymptotic feasibility results of one-dimensional PP to K-dimensional orthonormal projections $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K$, with the number $K \ge 2$ being a finite integer. From the computational viewpoint, the orthogonal directions in the projection space are also used in PCA and ICA, which are graphically compared with PP in Section 3.2.1 and Appendix B.1. Further discussions on the orthogonal directions are given in [4, 6, 7, 5, 9, 8]. Developing non-orthogonal projections will be an interesting topic for future research.

Our extended results assume $p/n \to \gamma$ as $n \to \infty$, and are discussed separately according to scenarios $\gamma \in (K, \infty)$ and $\gamma \in (0, K)$. Regarding the case of $\gamma \in (K, \infty)$, we present two kinds of extensions, in Lemma E.1 and Result E.1 below, respectively.

Lemma E.1 ($K \ge 2$, $\gamma \in (K, \infty)$, $\mu_2(G_k) < \gamma/K - 1$) *Assume* (1.1)*, where* $\boldsymbol{X} \sim \mathbb{N}(\mathbf{0}, \mathbf{I}_p)$*,* $n \to \infty$*, and* $p \to \infty$*. Assume that* $K \ge 2$ *is an integer. Suppose* $p/n \to \gamma \in (K, \infty)$*. Let* G_k *be the* C.D.F.*s, associated with the* p.d.f.*s* g_k*, and the second-moments* $\mu_2(G_k) =$

$\int s^2 g_k(s)\,\mathrm{d}s$, $k = 1, \ldots, K$. *If* $\mu_2(G_k) < \gamma/K - 1$, $k = 1, \ldots, K$, *then there exist sequences of* K *orthonormal directions* $\boldsymbol{z}_k = \boldsymbol{z}_k(\mathbf{X}, G_k) \in \mathbb{S}^{p-1}$, $k = 1, \ldots, K$, *such that the* E.D.F.*s* $\widehat{G}_{\boldsymbol{z}_k;k}$ *of projected data points,* $\{\boldsymbol{z}_k^T \boldsymbol{X}_{\bullet,i}\}_{i=1}^n$, *converge to the* C.D.F. G_k, $k = 1, \ldots, K$.

As a comparison, Lemma E.1 extends Theorem 2(i) in [2] from $K = 1$ to $K \geq 2$. On the other hand, both results require knowing the upper bound $\gamma/K - 1$ for the second-moment of the distribution G_k. Such bound may not be available in advance or known in practice. To relax this constraint, Result E.1 below simply requires the finiteness of the second-moment of G_k. Interestingly, Result E.1 also extends Corollary 1 in [2] from $K = 1$ to $K \geq 2$.

Result E.1 ($K \geq 2$, $\gamma \in (K, \infty)$, $\mu_2(G_k) < \infty$) *Assume* (1.1), *where* $\boldsymbol{X} \sim \mathbb{N}(\mathbf{0}, \mathbf{I}_p)$, $n \to \infty$, *and* $p \to \infty$. *Assume that* $K \geq 2$ *is an integer. Suppose* $p/n \to \gamma \in (K, \infty)$. *Let* G_k *be the* C.D.F.*s, associated with* p.d.f.*s, and the second-moments* $\mu_2(G_k)$, $k = 1, \ldots, K$. *If* $\mu_2(G_k) < \infty$, $k = 1, \ldots, K$, *then there exist some "Type-G_k"* C.D.F. $G_{k;u_{0,k},\sigma_{0,k}}$, *for which there exist sequences of* K *orthonormal directions* $\boldsymbol{z}_k = \boldsymbol{z}_k(\mathbf{X}, G_k) \in \mathbb{S}^{p-1}$, *such that the* E.D.F.*s* $\widehat{G}_{\boldsymbol{z}_k;k}$ *of projected data points,* $\{\boldsymbol{z}_k^T \boldsymbol{X}_{\bullet,i}\}_{i=1}^n$, *converge to the* C.D.F. $G_{k;u_{0,k},\sigma_{0,k}}$, $k = 1, \ldots, K$.

For the case of $\gamma \in (0, K)$, the special case of $K = 1$ is given in Lemma E.2 below, which is further extended to $K \geq 2$ in Result E.2. Lemma E.2 shares a similar spirit with Theorem 3 in [2], in the sense that the target distributions in both results are a finite mixture distribution, though they differ in both the mixing weights and the upper bound for the second moment of the distribution G. Thus, Lemma E.2 and Result E.2 extend the scope of PP in applications.

Lemma E.2 ($K = 1$, $\gamma \in (0, 1)$, $\mu_2(G) < (1 - \gamma)/\gamma$) *Assume* (1.1), *where* $\boldsymbol{X} \sim \mathbb{N}(\mathbf{0}, \mathbf{I}_p)$, $n \to \infty$, *and* $p \to \infty$. *Suppose* $p/n \to \gamma \in (0, 1)$. *Let* G *be a* C.D.F. *with the second-moment* $\mu_2(G)$. *If* $\mu_2(G) < (1 - \gamma)/\gamma$, *then there exists a sequence of directions* $\boldsymbol{z} = \boldsymbol{z}(\mathbf{X}, G) \in \mathbb{S}^{p-1}$ *that rely on* $\mathbf{X}$ *and* G, *such that the* E.D.F.*s* $\widehat{G}_{\boldsymbol{z}}$ *of projected data points,* $\{\boldsymbol{z}^T \boldsymbol{X}_{\bullet,i}\}_{i=1}^n$, *converge to the* C.D.F. *of the mixture distribution,*

$$G^* \;=\; \gamma^2\, G + (1 - \gamma^2)\, \Phi. \tag{A.2}$$

Result E.2 ($K \geq 2$, $\gamma \in (0, K)$, $\mu_2(G_k) < (K - \gamma)/\gamma$) *Assume* (1.1), *where* $\boldsymbol{X} \sim \mathbb{N}(\mathbf{0}, \mathbf{I}_p)$, $n \to \infty$, *and* $p \to \infty$. *Assume that* $K \geq 2$ *is an integer. Suppose* $p/n \to \gamma \in (0, K)$. *Let* G_k *be the* C.D.F.*s, associated with* p.d.f.*s, and the second-moments* $\mu_2(G_k)$, $k = 1, \ldots, K$. *If* $\mu_2(G_k) < (K - \gamma)/\gamma$, $k = 1, \ldots, K$, *then there exist sequences of* K *orthonormal directions*

$\boldsymbol{z}_k = \boldsymbol{z}_k(\mathbf{X}, G_k) \in \mathbb{S}^{p-1}$, *such that the* E.D.F.*s* $\widehat{G}_{\boldsymbol{z}_k;k}$ *of projected data points,* $\{\boldsymbol{z}_k^T \boldsymbol{X}_{\bullet,i}\}_{i=1}^n$, *converge to the* C.D.F. *of the mixture distribution,*

$$G_k^* \;=\; (\gamma/K)^2\, G_k + \{1-(\gamma/K)^2\}\,\Phi, \quad k = 1,\dots,K. \tag{A.3}$$

A.2 Proofs of extended results in Appendix A.1

Proof of Lemma E.1. For the dimension p of $\boldsymbol{X}$, coordinates $\{1,\dots,p\}$ can be partitioned into $K+1$ disjoint index subsets $J_1,\dots,J_K, J_{K+1}$, of lengths

$$p_1 = \cdots = p_K = [p/K], \quad \text{and} \quad p_{K+1} = p - \sum_{k=1}^{K} p_k. \tag{A.4}$$

Accordingly, the random vector $\boldsymbol{X}$ and the data matrix $\mathbf{X} = (\boldsymbol{X}_{\bullet,1},\dots,\boldsymbol{X}_{\bullet,n})$ can be partitioned,

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_{J_1} \in \mathbb{R}^{p_1} \\ \vdots \\ \boldsymbol{X}_{J_K} \in \mathbb{R}^{p_K} \\ \boldsymbol{X}_{J_{K+1}} \in \mathbb{R}^{p_{K+1}} \end{pmatrix}, \; \mathbf{X} = \begin{pmatrix} \mathbf{X}(J_1,:) \in \mathbb{R}^{p_1\times n} \\ \vdots \\ \mathbf{X}(J_K,:) \in \mathbb{R}^{p_K\times n} \\ \mathbf{X}(J_{K+1},:) \in \mathbb{R}^{p_{K+1}\times n} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_{J_1,1} & \cdots & \boldsymbol{X}_{J_1,n} \\ \vdots & \cdots & \vdots \\ \boldsymbol{X}_{J_K,1} & \cdots & \boldsymbol{X}_{J_K,n} \\ \boldsymbol{X}_{J_{K+1},1} & \cdots & \boldsymbol{X}_{J_{K+1},n} \end{pmatrix}, \tag{A.5}$$

where the sub-vector $\boldsymbol{X}_{J_k} \in \mathbb{R}^{p_k}$ and the data sub-matrix $\mathbf{X}(J_k,:) \in \mathbb{R}^{p_k\times n}$ are formed by rows of $\boldsymbol{X}$ and $\mathbf{X}$, respectively, in J_k, $k = 1,\dots,K,K+1$. The assumption $\boldsymbol{X} \sim \mathbb{N}(\mathbf{0}, \mathbf{I}_p)$ together with (1.1) implies that

$$\text{all entries in } \mathbf{X} \text{ are i.i.d. } \mathbb{N}(0,1). \tag{A.6}$$

For the sake of notations in the rest of derivations, define $\mathrm{r} = \gamma/K$.

For the data sub-matrix $\mathbf{X}(J_1,:) \in \mathbb{R}^{p_1\times n}$, the condition $p/n \to \gamma \in (K,\infty)$ and (A.4) guarantee that $p_1/n = [p/K]/n = (p/n)/K + o(1) \to \gamma/K = \mathrm{r} \in (1,\infty)$. For $\mu_2(G_1) < \mathrm{r} - 1$, following the proof in [2] for Theorem 2(i) indicates that there exists a sequence of directions $\boldsymbol{z}_{J_1} = \boldsymbol{z}_{J_1}(\mathbf{X}(J_1,:), \boldsymbol{a}_1(G_1)) \in \mathbb{S}^{p_1-1}$ that rely on $\mathbf{X}(J_1,:)$ and $\boldsymbol{a}_1(G_1)$, where $\boldsymbol{a}_1(G_1) = (a_{1,1},\dots,a_{1,n})^T$ is a vector whose distribution relies on the C.D.F. G_1, e.g., $\{a_{1,1},\dots,a_{1,n}\} \overset{\text{i.i.d.}}{\sim} G_1$, and $\boldsymbol{a}_1(G_1) \perp\!\!\!\perp \mathbf{X}(J_1,:)$, such that the E.D.F.s $\widehat{G}_{\boldsymbol{z}_1;1}$ of projected data points in

$$\boldsymbol{z}_{J_1}^T \mathbf{X}(J_1,:) = (\boldsymbol{z}_{J_1}^T \boldsymbol{X}_{J_1,1},\dots,\boldsymbol{z}_{J_1}^T \boldsymbol{X}_{J_1,n}) = \boldsymbol{a}_1(G_1)^T \tag{A.7}$$

converge to the C.D.F. G_1, i.e., $\|\widehat{G}_{\boldsymbol{z}_1;1} - G_1\|_\infty \xrightarrow{\mathrm{P}} 0$.

Similarly, for the data sub-matrix $\mathbf{X}(J_k,:) \in \mathbb{R}^{p_k \times n}$, $k = 2, \dots, K$, note that $p_k/n = [p/K]/n = (p/n)/K + o(1) \to \gamma/K = \mathrm{r} \in (1, \infty)$. For $\mu_2(G_k) < \mathrm{r} - 1$, the proof in [2] for Theorem 2(i) indicates that there exists a sequence of directions,

$$\boldsymbol{z}_{J_k} = \boldsymbol{z}_{J_k}(\mathbf{X}(J_k,:), \boldsymbol{a}_k(G_k)) \in \mathbb{S}^{p_k - 1}, \tag{A.8}$$

that rely on $\mathbf{X}(J_k,:)$ and $\boldsymbol{a}_k(G_k)$, where $\boldsymbol{a}_k(G_k) = (a_{k,1}, \dots, a_{k,n})^T$ is a vector relying on the C.D.F. G_k, e.g.,

$$\{a_{k,1}, \dots, a_{k,n}\} \overset{\text{i.i.d.}}{\sim} G_k,\ \boldsymbol{a}_k(G_k) \perp\!\!\!\perp \mathbf{X}(J_k,:),\ \boldsymbol{a}_k(G_k) \perp\!\!\!\perp \{\mathbf{X}(J_1,:), \dots, \mathbf{X}(J_{k-1},:)\}, \tag{A.9}$$

such that the E.D.F.s $\widehat{G}_{\boldsymbol{z}_k;k}$ of projected data points in

$$\boldsymbol{z}_{J_k}^T \mathbf{X}(J_k,:) = (\boldsymbol{z}_{J_k}^T \boldsymbol{X}_{J_k,1}, \dots, \boldsymbol{z}_{J_k}^T \boldsymbol{X}_{J_k,n}) = \boldsymbol{a}_k(G_k)^T \tag{A.10}$$

converge to the C.D.F. G_k, i.e., $\|\widehat{G}_{\boldsymbol{z}_k;k} - G_k\|_\infty \xrightarrow{\mathrm{P}} 0$.

Now, take K directions in $\mathbb{R}^p$,

$$\boldsymbol{z}_1 = \begin{pmatrix} \boldsymbol{z}_{J_1} \in \mathbb{R}^{p_1} \\ \mathbf{0} \in \mathbb{R}^{p_2} \\ \mathbf{0} \in \mathbb{R}^{p-p_1-p_2} \end{pmatrix},\ \boldsymbol{z}_k = \begin{pmatrix} \mathbf{0} \in \mathbb{R}^{p_1+\cdots+p_{k-1}} \\ \boldsymbol{z}_{J_k} \in \mathbb{R}^{p_k} \\ \mathbf{0} \in \mathbb{R}^{p-p_1-\cdots-p_k} \end{pmatrix}, \dots, \boldsymbol{z}_K = \begin{pmatrix} \mathbf{0} \in \mathbb{R}^{p_1+\cdots+p_{K-1}} \\ \boldsymbol{z}_{J_K} \in \mathbb{R}^{p_K} \\ \mathbf{0} \in \mathbb{R}^{p-p_1-\cdots-p_K} \end{pmatrix} \tag{A.11}$$

Moreover, choose $\boldsymbol{a}_1(G_1), \dots, \boldsymbol{a}_K(G_K)$ to be mutually independent. This, together with (A.8) and (A.6), implies that $\boldsymbol{z}_1, \dots, \boldsymbol{z}_K$ are mutually independent. Also, from (A.11), we see that $\boldsymbol{z}_k \in \mathbb{S}^{p-1}$ due to $\|\boldsymbol{z}_k\|_2 = \|\boldsymbol{z}_{J_k}\|_2 = 1$, $k = 1, \dots, K$, and that $\boldsymbol{z}_{k_1}^T \boldsymbol{z}_{k_2} = 0$ for $1 \le k_1 \ne k_2 \le K$. Combining (A.7), (A.10) and (A.11) gives

$$\begin{array}{rcl} (\boldsymbol{z}_1^T \boldsymbol{X}_{\bullet,1}, \dots, \boldsymbol{z}_1^T \boldsymbol{X}_{\bullet,n}) = \boldsymbol{z}_1^T \mathbf{X} & = & \boldsymbol{z}_{J_1}^T \mathbf{X}(J_1,:) = \boldsymbol{a}_1(G_1)^T, \\ \cdots & = & \cdots \\ (\boldsymbol{z}_K^T \boldsymbol{X}_{\bullet,1}, \dots, \boldsymbol{z}_K^T \boldsymbol{X}_{\bullet,n}) = \boldsymbol{z}_K^T \mathbf{X} & = & \boldsymbol{z}_{J_K}^T \mathbf{X}(J_K,:) = \boldsymbol{a}_K(G_K)^T, \end{array} \tag{A.12}$$

and thus $\boldsymbol{z}_1^T \mathbf{X}$, ..., $\boldsymbol{z}_K^T \mathbf{X}$ are mutually independent. This completes the proof. ■

Proof of Result E.1. Similar to the proof of Lemma E.1, for the dimension p, indices $\{1, \dots, p\}$ can be partitioned into $K+1$ disjoint index subsets $J_1, \dots, J_K, J_{K+1}$, of lengths $p_1, \dots, p_K, p_{K+1}$ as in (A.4), and the random vector $\boldsymbol{X}$ and the data matrix $\mathbf{X}$ can be partitioned as in (A.5).

For the data sub-matrix $\mathbf{X}(J_1,:) \in \mathbb{R}^{p_1 \times n}$, the condition $p/n \to \gamma \in (K, \infty)$ implies that $p_1/n = [p/K]/n = (p/n)/K + o(1) \to \gamma/K = \mathrm{r} \in (1, \infty)$. For $\mu_2(G_1) < \infty$, applying Corollary 1 in [2] gives the existence of some "Type-G_1" C.D.F. $G_{1;u_{0,1},\sigma_{0,1}}$, for which there exists a sequence of directions $\boldsymbol{z}_{J_1} = \boldsymbol{z}_{J_1}(\mathbf{X}(J_1, :), \boldsymbol{a}_1(G_1)) \in \mathbb{S}^{p_1 - 1}$ that rely on $\mathbf{X}(J_1, :)$ and $\boldsymbol{a}_1(G_1)$, such that the E.D.F.s $\widehat{G}_{\boldsymbol{z}_1;1}$ of projected data points as defined in (A.7) converge to the C.D.F. $G_{1;u_{0,1},\sigma_{0,1}}$, i.e., $\|\widehat{G}_{\boldsymbol{z}_1;1} - G_{1;u_{0,1},\sigma_{0,1}}\|_\infty \xrightarrow{\mathrm{P}} 0$.

Similarly, for the data sub-matrix $\mathbf{X}(J_k, :) \in \mathbb{R}^{p_k \times n}$, $k = 2, \ldots, K$, note that $p_k/n = [p/K]/n = (p/n)/K + o(1) \to \gamma/K = \mathrm{r} \in (1, \infty)$. For $\mu_2(G_k) < \infty$, Corollary 1 in [2] indicates the existence of some "Type-G_k" C.D.F. $G_{k;u_{0,k},\sigma_{0,k}}$, for which there exists a sequence of directions $\boldsymbol{z}_{J_k} = \boldsymbol{z}_{J_k}(\mathbf{X}(J_k, :), \boldsymbol{a}_k(G_k)) \in \mathbb{S}^{p_k - 1}$ that rely on $\mathbf{X}(J_k, :)$ and $\boldsymbol{a}_k(G_k)$, where the vector $\boldsymbol{a}_k(G_k) = (a_{k,1}, \ldots, a_{k,n})^T$ relying on the C.D.F. G_k satisfies (A.9), such that the E.D.F.s $\widehat{G}_{\boldsymbol{z}_k;k}$ of projected data points as defined in (A.10) converge to the C.D.F. $G_{k;u_{0,k},\sigma_{0,k}}$, i.e., $\|\widehat{G}_{\boldsymbol{z}_k;k} - G_{k;u_{0,k},\sigma_{0,k}}\|_\infty \xrightarrow{\mathrm{P}} 0$.

Now define K directions, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K$ in $\mathbb{R}^p$, according to (A.11). Similar to (A.12) in the proof of Lemma E.1,

$$(\boldsymbol{z}_k^T \boldsymbol{X}_{\bullet,1}, \ldots, \boldsymbol{z}_k^T \boldsymbol{X}_{\bullet,n}) = \boldsymbol{z}_k^T \mathbf{X} = \boldsymbol{z}_{J_k}^T \mathbf{X}(J_k, :) = \boldsymbol{a}_k(G_k)^T, \quad k = 1, \ldots, K,$$

where $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K$ are orthonormal, and $\boldsymbol{z}_1^T \mathbf{X}$, ..., $\boldsymbol{z}_K^T \mathbf{X}$ are mutually independent. This completes the proof. ∎

Proof of Lemma E.2. The data matrix $\mathbf{X}$, partitioned according to columns, can be rewritten as

$$\mathbf{X} = (\boldsymbol{X}_{\bullet,1}, \ldots, \boldsymbol{X}_{\bullet,n_1}, \ \boldsymbol{X}_{\bullet,n_1+1}, \ldots, \boldsymbol{X}_{\bullet,n}) \equiv (\mathbf{X}(:, I_1), \ \ \mathbf{X}(:, I_2)),$$

with two sub-matrices $\mathbf{X}(:, I_1) = (\boldsymbol{X}_{\bullet,1}, \ldots, \boldsymbol{X}_{\bullet,n_1}) \in \mathbb{R}^{p \times n_1}$ and $\mathbf{X}(:, I_2) = (\boldsymbol{X}_{\bullet,n_1+1}, \ldots, \boldsymbol{X}_{\bullet,n}) \in \mathbb{R}^{p \times (n - n_1)}$.

Let $n_1 = [p\gamma]$. From the condition $p/n \to \gamma \in (0, 1)$, we observe that $n_1 < p < n$ as $n \to \infty$. Moreover, $p \to \infty$ implies $n_1 = [p\gamma] \to \infty$. Clearly, for the data sub-matrix $\mathbf{X}(:, I_1) \in \mathbb{R}^{p \times n_1}$, the ratio of the number p of rows to the number n_1 of columns of $\mathbf{X}(:, I_1)$ belongs to the setting of Theorem 2 in [2], i.e.,

$$n_1 \to \infty, \quad p \to \infty, \quad p/n_1 \to \gamma_1 = 1/\gamma \in (1, \infty).$$

If $\mu_2(G) < \gamma_1 - 1 = (1 - \gamma)/\gamma$, then the result of Theorem 2(i) in [2] indicates the existence of a sequence of directions $\boldsymbol{z} = \boldsymbol{z}(\mathbf{X}(:, I_1), \boldsymbol{a}_1(G)) \in \mathbb{S}^{p-1}$ that rely on $\mathbf{X}(:, I_1)$ and $\boldsymbol{a}_1(G)$, where $\boldsymbol{a}_1(G) = (a_1, \ldots, a_{n_1})^T$ is a vector relying on the C.D.F. G, e.g., $\{a_1, \ldots, a_{n_1}\} \overset{\text{i.i.d.}}{\sim} G$, $\boldsymbol{a}_1(G) \perp\!\!\!\perp \mathbf{X}(:, I_1)$ (and the proof in [2] also allows $\boldsymbol{a}_1(G) \perp\!\!\!\perp \mathbf{X}(:, I_2)$), such that the E.D.F.s of n_1 data points in $\boldsymbol{z}^T \boldsymbol{X}(:, I_1) = (\boldsymbol{z}^T \boldsymbol{X}_{\bullet,1}, \ldots, \boldsymbol{z}^T \boldsymbol{X}_{\bullet,n_1})$ converge to the C.D.F. G, i.e.,

$$\max_{s \in \mathbb{R}} \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \mathrm{I}(\boldsymbol{z}^T \boldsymbol{X}_{\bullet,i} \le s) - G(s) \right| \xrightarrow{\mathrm{P}} 0. \tag{A.13}$$

For the data sub-matrix $\mathbf{X}(:, I_2) \in \mathbb{R}^{p\times(n-n_1)}$, since the vector $\boldsymbol{z}$ only depends on $\{\mathbf{X}(:, I_1), \boldsymbol{a}_1(G)\}$, where $\boldsymbol{a}_1(G) \perp\!\!\!\perp \mathbf{X}$, we conclude $\boldsymbol{z} \perp\!\!\!\perp \mathbf{X}(:, I_2)$. Moreover, utilizing (A.6) and properties of the multivariate Gaussian distribution, we deduce that all $n - n_1$ entries $\boldsymbol{z}^T\boldsymbol{X}_{\bullet,n_1+1}, \dots, \boldsymbol{z}^T\boldsymbol{X}_{\bullet,n}$ of the vector $\boldsymbol{z}^T\boldsymbol{X}(:, I_2)$ are i.i.d. $\mathbb{N}(0, 1)$ variables, indicating that the E.D.F.s of $n - n_1$ data points in $\boldsymbol{z}^T\boldsymbol{X}(:, I_2) = (\boldsymbol{z}^T\boldsymbol{X}_{\bullet,n_1+1}, \dots, \boldsymbol{z}^T\boldsymbol{X}_{\bullet,n})$ converge to the C.D.F. Φ, i.e.,

$$\max_{s\in\mathbb{R}} \left| \frac{1}{n-n_1} \sum_{i=n_1+1}^{n} \mathrm{I}(\boldsymbol{z}^T\boldsymbol{X}_{\bullet,i} \le s) - \Phi(s) \right| \xrightarrow{\mathrm{P}} 0. \tag{A.14}$$

Also, $n_1/n = [p\,\gamma]/n \to \gamma^2 \in (0, 1)$. This, combined with (A.13) and (A.14), proves (A.2). ■

Proof of Result E.2. Recall that for the dimension p, indices $\{1, \dots, p\}$ can be partitioned into $K + 1$ parts in index sets $J_1, \dots, J_K, J_{K+1}$, of lengths $p_{_1}, \dots, p_{_K}, p_{_{K+1}}$ as in (A.4), and the random vector $\boldsymbol{X}$ and the data matrix $\mathbf{X}$ can be partitioned as in (A.5).

For the data sub-matrix $\mathbf{X}(J_1, :) \in \mathbb{R}^{p_1\times n}$, the condition $p/n \to \gamma \in (0, K)$ gives that $p_{_1}/n = [p/K]/n = (p/n)/K + o(1) \to \gamma/K = \mathrm{r} \in (0, 1)$. If $\mu_2(G_1) < (1 - \gamma/K)/(\gamma/K) = (K - \gamma)/\gamma$, then Lemma E.2 implies that there exists a sequence of directions $\boldsymbol{z}_{_{J_1}} = \boldsymbol{z}_{_{J_1}}(\mathbf{X}(J_1, I_{1;1}), \boldsymbol{a}_1(G_1)) \in \mathbb{S}^{p_1-1}$ that rely on $\mathbf{X}(J_1, I_{1;1})$ and $\boldsymbol{a}_1(G_1)$, where $I_{1;1} = \{1, \dots, n_{1;1}\}$ with $n_{1;1} = [p_{_1}\,\gamma/K]$, $\boldsymbol{a}_1(G_1) = (a_{1,1}, \dots, a_{1,n_{1;1}})^T$ relies on G_1 and $\boldsymbol{a}_1(G_1) \perp\!\!\!\perp \mathbf{X}(J_1, I_{1;1})$, such that the E.D.F.s of n projected data in $\boldsymbol{z}_{_{J_1}}^T\mathbf{X}(J_1, :) = (\boldsymbol{z}_{_{J_1}}^T\boldsymbol{X}_{J_1,1}, \dots, \boldsymbol{z}_{_{J_1}}^T\boldsymbol{X}_{J_1,n})$ converge to the C.D.F.,

$$G_1^* \;=\; (\gamma/K)^2\, G_1 + \{1 - (\gamma/K)^2\}\, \Phi,$$

in which $n_{1;1}/n \to (\gamma/K)^2 \in (0, 1)$.

Similarly, for the data sub-matrix $\mathbf{X}(J_k, :) \in \mathbb{R}^{p_k\times n}$, $k = 2, \dots, K$, note that $p_{_k}/n = [p/K]/n = (p/n)/K + o(1) \to \gamma/K = \mathrm{r} \in (0, 1)$. If $\mu_2(G_k) < (1 - \gamma/K)/(\gamma/K) = (K - \gamma)/\gamma$, then Lemma E.2 implies that there exists a sequence of directions $\boldsymbol{z}_{_{J_k}} = \boldsymbol{z}_{_{J_k}}(\mathbf{X}(J_k, I_{1;k}), \boldsymbol{a}_k(G_k)) \in \mathbb{S}^{p_k-1}$ that rely on $\mathbf{X}(J_k, I_{1;k})$ and $\boldsymbol{a}_k(G_k)$, where $I_{1;k} = \{1, \dots, n_{1;k}\}$ with $n_{1;k} = [p_{_k}\,\gamma/K]$, $\boldsymbol{a}_k(G_k) = (a_{k,1}, \dots, a_{k,n_{1;k}})^T$ relies on G_k, $\boldsymbol{a}_k(G_k) \perp\!\!\!\perp \mathbf{X}(J_k, I_{1;k})$ and $\boldsymbol{a}_k(G_k) \perp\!\!\!\perp \{\mathbf{X}(J_1, I_{1;k}), \dots, \mathbf{X}(J_{k-1}, I_{1;k})\}$, such that the E.D.F.s of projected data points in $\boldsymbol{z}_{_{J_k}}^T\mathbf{X}(J_k, :) = (\boldsymbol{z}_{_{J_k}}^T\boldsymbol{X}_{J_k,1}, \dots, \boldsymbol{z}_{_{J_k}}^T\boldsymbol{X}_{J_k,n})$ converge to the C.D.F.,

$$G_k^* \;=\; (\gamma/K)^2\, G_k + \{1 - (\gamma/K)^2\}\, \Phi,$$

in which $n_{1;k}/n \to (\gamma/K)^2 \in (0, 1)$.

Now, take K directions, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K$ in $\mathbb{R}^p$, as in (A.11). Then choices $\boldsymbol{a}_1(G_1), \ldots, \boldsymbol{a}_K(G_K)$ are mutually independent, and thus $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K$ are mutually independent and orthonormal, Also, we see from (A.12) that $\boldsymbol{z}_1^T\mathbf{X}$, …, $\boldsymbol{z}_K^T\mathbf{X}$ are mutually independent. This proves (A.3). ■

B Additional illustrations

B.1 Illustrate **Remark** 2 in [2] for 2-dimensional PP and $p > n$: another example of feasible G

In addition to specifying the bivariate mixture Gaussian distribution (3.9) in Section 3.2.1 for the bivariate target distribution G, we consider a bivariate asymmetric distribution ([1], p. 260) with the density function

$$g(s_1, s_2; \alpha) \quad = \quad 2\,\phi(s_1)\phi(s_2)\Phi(\alpha\, s_1 s_2), \tag{B.1}$$

for $-\infty < s_1, s_2 < +\infty$, and a shape parameter $\alpha = -10$. It is interesting to note that the univariate margins of the distribution are standard Gaussian, while conditional distributions are univariate skew-Normal.

The simulation study uses the setting similar to that of Figure 11 in Section 3.2.1, and displays the results in Figure 13. The projected data onto the feasible bivariate direction $(\boldsymbol{z}_1, \boldsymbol{z}_2)$ (in panel **C2**) clearly identify the twist, which is nonetheless not discoverable using either an alternative projection direction in panel **C3** or other projection methods ICA and PCA in panels **C4**–**C5**. These new graphical results are similar in spirit to those of the bivariate mixture Gaussian distribution in Figure 11, lending further support to feasible distributions G in Remark 2 of [2].

B.2 An exploratory study of multivariate t data with $p = 1000$ and $n = 100$, $p > n$

For an exploratory study, we randomly simulate $n = 100$ multivariate t data vectors $\boldsymbol{X}_{\cdot,1}, \ldots, \boldsymbol{X}_{\cdot,n}$, each of dimension $p = 1000$, coming from the joint distribution of a random vector $\boldsymbol{X}$. The construction of $\boldsymbol{X}$ follows from [3]:

$$\boldsymbol{X} \quad = \quad \left(\frac{v}{v-2}\mathrm{R}\right)^{-1/2} \sqrt{W_v}\, A\boldsymbol{Z}, \tag{B.2}$$

where the variable W_v has the distribution of v/χ_v^2 with a chi-squared variable χ_v^2 on degrees of freedom $v = 5$ and is dependent of the standard Gaussian vector $\boldsymbol{Z} \sim \mathbb{N}_p(\mathbf{0}, \mathbf{I}_p)$, together

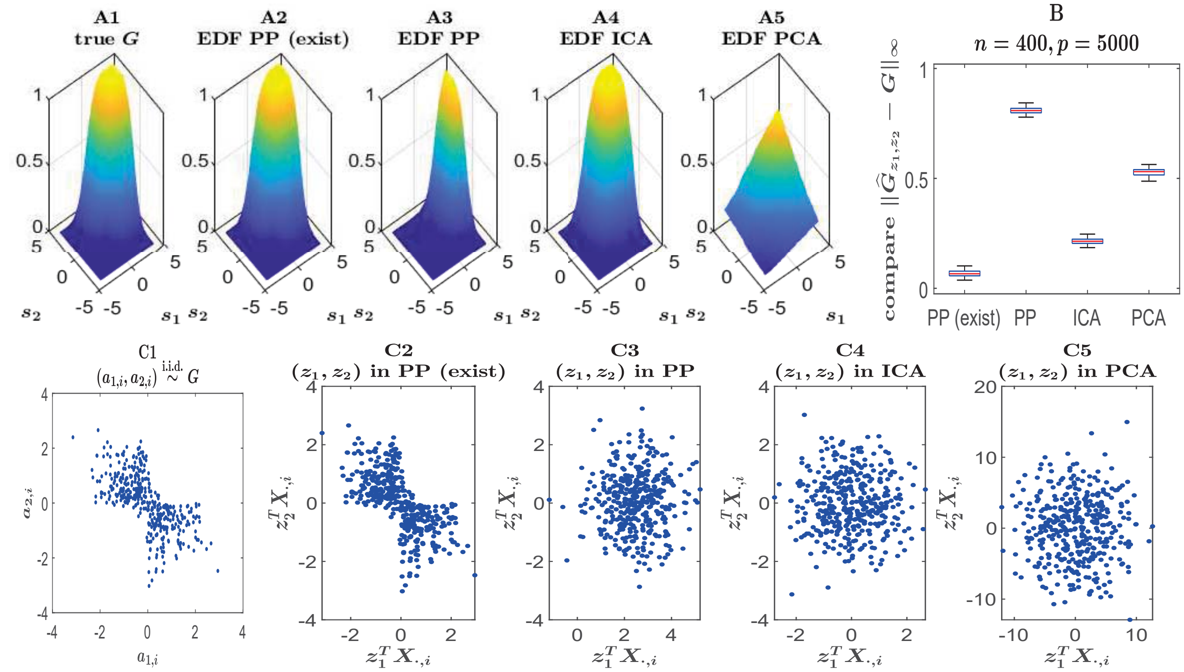

Figure 13: **Illustrate Remark 2 summarized in Table 1 for [2] with 2-dimensional PP and $p > n$: feasible distribution G with density function in (B.1).** *The caption is similar to that of Figure 11, except using the bivariate asymmetric target distribution G.*

with $p \times p$ matrices A and R satisfying $AA^T = \mathrm{R} = \rho \mathbf{1}_p \mathbf{1}_p^T + (1-\rho)\mathbf{I}_p$, with a p-variate vector $\mathbf{1}_p$ of ones and a constant $\rho = 0.8$. The joint distribution for $\boldsymbol{X}$ in (B.2) has a zero mean vector and an identity covariance matrix, and could better mimic some tail dependence in real data. Nonetheless, component variables $X_1, \ldots, X_p$ of $\boldsymbol{X}$ in (B.2) are dependent, thus violating the independence condition in either (2.1) or (2.2).

As in Section 3.1.1, we take the bimodal mixture distribution G in (3.5) as the target distribution. Interestingly, the boxplots of the KS distances in Figure 14 indicate that the target distribution G continues to be feasible, by means of the projection onto the direction "$\boldsymbol{z}_1$ (exist)". In the absence of theoretical guarantee yet, the numerical results do support that the interesting bimodal structure could be revealed from projecting multivariate t data. This empirical finding leads us to conjecture that Theorem 1 in [2] may generalize to the multivariate t data with an identity covariance. A further justification falls beyond the scope of the present paper and could be an interesting future work.

References

[1] Arnold, B., Castillo, E., and Sarabia, J. (2001). Conditionally specified distributions: an introduction (with discussion). *Statistical Science*, 16:249–274.

[2] Bickel, P.J., Kur, Gil, and Nadler, Boaz (2018). Projection pursuit in high dimensions. *Proceedings of the National Academy of Sciences*, **115** (37), 9151–9156.

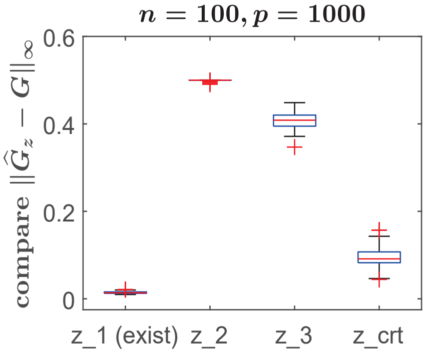

Figure 14: **An exploratory study of multivariate t data with $p > n$: target distribution G in (3.5).** *The caption is similar to that of the right panel of Figure 2, except using the data vector $\boldsymbol{X} \sim$ (B.2) with dependent variables $X_1, \ldots, X_p$.*

[3] Demarta, Stefano and McNeil, Alexander J. (2005). The t copula and related copulas. *International Statistical Review*, 73, 1, 111–129.

[4] Friedman, J.H., and Tukey, J. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, **23**:881–889.

[5] Friedman, J.H. (1987). Exploratory projection pursuit. *J Amer. Statist. Assoc.*, 82, pp. 249–266.

[6] Huber, P.J. (1985). Projection pursuit. *Annals of Statistics*, **13**:435–475.

[7] Jones, M.C. and Sibson, R. (1987). What is projection pursuit? *Journal of the Royal Statistical Society, Series A* (General), Vol. **150**, No. 1, pp. 1–37.

[8] Peña, D. and Prieto, F.J. (2001). Multivariate Outlier Detection and Robust Covariance Estimation (with discussion). *Technometrics* 43, 286–310.

[9] Posse, C. (1995). Projection Pursuit Exploratory Data Analysis. *Computational Statistics and Data Analysis*, 20, 669–687.