Statistica Sinica **21** (2011), 1713-1733
doi:http://dx.doi.org/10.5705/ss.2009.070

SENSITIVITY ANALYSIS OF NONGAUSSIANITY BY PROJECTION PURSUIT

Yufen Huang[1], Ching-Ren Cheng[1] and Tai-Ho Wang[2]

[1]*National Chung Cheng University and* [2]*Baruch College, CUNY*

Abstract: From the information-theoretic point of view, the Gaussian distribution is the least structured. Therefore, the most non-Gaussian direction in which to explore the clustering structure of data is considered to be the most interesting linear projection direction when applying projection pursuit. Non-Gaussianity is often measured by kurtosis. However, kurtosis is well-known to be sensitive to influential points/outliers and so the projection direction can be unduly affected by abnormal points. In this paper, we focus on developing influence functions of projection directions in order to detect abnormal observations, especially on high-dimensional data. For multivariate data, a new technique is proposed for defining and developing influence functions of projection directions. In addition, a new influence function is suggested. Two simulated data examples and one concrete data example are provided for illustration.

Key words and phrases: Influence function, kurtosis, non-gaussianity, projection pursuit.

1. Introduction

Friedman and Tukey (1974) introduced the term "projection pursuit" for a technique designed to search for "interesting" linear projections of multivariate data. Structure can then be visualized, for example, by the distribution of data projections on one-dimensional subspaces, or two-dimensional planes defined by one or two of the projection pursuit directions, respectively. Projection pursuit is an extension of the classic method of using principal component analysis (PCA) for visualization, in which the distribution of the data is shown on the plane spanned by the two first principal components. However, a clustering structure is not always visible in the covariance or correlation matrix on which PCA is based. Hence, in this paper, we focus on finding the projection directions that reveal the clustering or other structure of the data when projection pursuit is used for exploratory data analysis and its corresponding influence analysis. The following example is for the illustration of finding certain structure underlying in a data set, say clustering, by projection pursuit. Figure 1 represents the scatter plot of the simulated data from a bivariate distribution. It is obvious

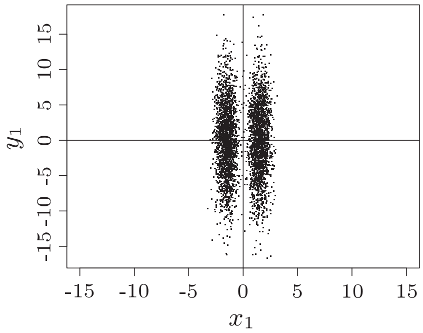

Figure 1. Plots the simulated data.

that projecting the original data onto the x-axis reveals the clustering of the data (the projected data aggregate into two clusters) while projecting onto y-axis does not. Therefore, projecting the original data onto the “right” subspace does help to discover the data structure under consideration, which is clustering in the example. Of course one can argue that the clustering is easily seen from the scatter plot in such an example, but for higher dimensional data, due to the lack of panoramic visualization, applying projection pursuit is at least a plausible option.

These issues were first addressed in Huang, Cheng, and Wang (2007) for the two-dimensional case. The idea there is that the most non-Gaussian direction to explore the clustering structure of the data is the most interesting linear projection direction found by applying projection pursuit. Kurtosis is often used as a measure of non-Gaussianity. The kurtosis of the standard Gaussian distribution is 0; the most interesting projection is the one such that the kurtosis of the projected variables is as far from 0 as possible. This is discussed for general dimensions in Section 2. It is well known that kurtosis is sensitive to influential points/outliers, and that the projection directions can well be affected by these unusual points. Hence the single-perturbation influence functions in Huang, Cheng, and Wang (2007) were developed to detect abnormal observations. Single-perturbation diagnostics can suffer from masking effects (see Riani and Atkinson (2001)). Identification of outliers in a multivariate points cloud is difficult, especially when there are several outliers. The classical detection method does not always find them, because it is based on the sample mean and covariance matrix, that are themselves affected by the outliers. To avoid the masking effect, Rousseeuw and van Zomeren (1990) proposed to compute distance, based on robust estimates of location and covariance, to detect outliers in a multivariate point cloud. However, Fung (1993) pointed out that the high-breakdown robust estimation method, and the least median of squares and minimum volume ellipsoid methods proposed by Rousseeuw and van Zomeren (1990) tend to declare too many observations as extreme. Hence, Fung (1993) proposed a stepwise analysis, using diagnostic measures to add back omitted observations, that performs well for confirming outliers. These methods rely on robustness ideas to uncover masked outliers, so one needs to address the issue of determining breakdown points. Huang, Cheng, and Wang (2007, 2008) and this paper devote themselves to the development of pair-perturbations influence functions that provide not only a sensible way to detect outliers but also a useful auxiliary scheme to uncover masked outliers. The robust method and the influence function methodology are complementary in our detection of multiple outliers.

Three types of influence functions in use are the empirical influence function (EIF), the deleted empirical influence function (DIF), and the sample influence function (SIF). They have been applied in various contexts in the literature. For instance, in principal component analysis, influence functions have been considered by Critchley (1985) and Tanaka (1988), and generalizations to pair-perturbations were discussed by Huang, Kao, and Wang (2007). In linear discriminant analysis, influence analysis has been discussed by Fung (1992, 1993, 1995, 1996), He and Fung (2000), Poon (2004) and Huang, Kao, and Wang (2007).

This paper extends previous work by Huang, Cheng, and Wang (2007, 2008) that discussed single-perturbation influence functions and pair-perturbation influence functions of projection direction in applying projection pursuit. Earlier work looked at the two-dimensional case. Here we go on to high-dimensional cases, see Remark 1 for more details. We also propose a new influence function, the averaged influence function (AIF), that averages the information obtained by the EIF, DIF and SIF functions. Examples in Section 4 show that EIF, DIF, and SIF do not always agree with one another on outliers. Averaging them after standardization can retain useful information from each while it dampens misleading noise variation that individual influence functions might trigger. These characteristics of AIF will be seen in Section 4.

The remainder of this paper is organized as follows. Our implementation of non-Gaussianity to search projection directions in multivariate data by projection pursuit is given in Section 2. In Section 3 we propose a new technique to develop a general framework for influence functions of the projection directions for multivariate data to detect abnormal points via single perturbation and pair-perturbation cases. The selection of cut points of influence functions is also discussed in Section 3. In Section 4, simulations and a data example are used to illustrate the application of these approaches. Conclusions, additional remarks, and a brief discussion are in Section 5.

2. Searching Non-Gaussian Projection Direction via Projection Pursuit for Multivariate Data

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a multivariate random variable with distribution function F. Let Y be the projection of $(X_1, \ldots, X_n)$ in the direction of a unit vector $\mathbf{a} = (a_1, \ldots, a_n)$, that is, $Y = \mathbf{a} \cdot \mathbf{X} = \sum_{i=1}^n a_i X_i$ is the projected random variable, and $\mathbf{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n$ is the direction vector with $\|\mathbf{a}\| = \sqrt{a_1^2 + \ldots + a_n^2} = 1$. We use the word "direction", instead of "vector", to emphasize that any unit vector $\mathbf{a}$ and its opposite, $-\mathbf{a}$, result in the same projection direction. This also applies to the intervals for polar angles, discussed next.

A convenient parametrization of the projection directions is provided by the use of polar coordinates. Let $\Theta = (\theta_1, \ldots, \theta_{n-1})$ be a vector of angles in polar coordinates with $-\pi/2 < \theta_i \leq \pi/2$ for $i = 1, \ldots n-1$. Projection directions can now be parametrized by their polar angles as

$$a_i = \begin{cases} \cos\theta_1 & \text{if} \quad i = 1, \\ \prod_{j=1}^{i-1} \sin\theta_j (\cos\theta_i)^{1-\delta_{in}} & \text{for} \quad i = 2, \ldots n, \end{cases}$$

where δ_{in} is the Kronecker delta, and Y can be regarded as a function of Θ, $Y(\Theta)$.

Let κ be the kurtosis of Y,

$$\kappa = M_4 - 3M_2^2 - 6M_1^4 + 12M_1^2 M_2 - 4M_1 M_3, \tag{2.1}$$

where M_i denotes the ith moment $\mathbb{E}[Y^i]$ of Y. Let $\mathbf{m} = (m_1, \ldots, m_n)$ be a multi-index, $\mathbf{a} = (a_1, \ldots, a_n)$ and $\mathbf{X} = (X_1, \ldots, X_n)$ be vectors. The following notation is used:

$$|\mathbf{m}| = \sum_{i=1}^n m_i, \qquad \mathbf{m}! = m_1! m_2! \ldots m_n! \;, \qquad \mathbf{a}^{\mathbf{m}} = \prod_{i=1}^n a_i^{m_i},$$

$$\mathbf{X}^{\mathbf{m}} = \prod_{i=1}^n X_i^{m_i}, \qquad c_{\mathbf{m}}^k = \frac{k!}{\mathbf{m}!}, \qquad \text{where } k \text{ is a positive integer.}$$

Note that the kth moment M_k of Y can be expressed in terms of the moments of $\mathbf{X}$ as

$$M_k = \mathbb{E}[Y^k] = \mathbb{E}\left[(\mathbf{a} \cdot \mathbf{X})^k\right] = \mathbb{E}\left[\sum_{|\mathbf{m}|=k} c_{\mathbf{m}}^k \mathbf{a}^{\mathbf{m}} \mathbf{X}^{\mathbf{m}}\right].$$

We remark that, since Y is a function of Θ, the kurtosis κ and the moments M_k of Y are also functions of Θ. We suppress the dependence of these quantities on Θ from time to time. The search for the projection direction that drives the projected data as far away from being Gaussian as possible is that of maximizing the absolute value of the kurtosis κ of Y as a function of Θ.

Definition 1. *A direction $\boldsymbol{a}^*$ is the most non-Gaussian direction if its corresponding vector of angles Θ^* satisfies*

$$|\kappa(\Theta^*)| = \max_{\Theta} |\kappa(\Theta)|. \tag{2.2}$$

The first order partial derivatives of the moments involved kurtosis with respect to the θ_k's are summarized in a lemma; the proof is by straightforward but tedious computations, and is omitted. We write $|\mathbf{m}^i| = \sum_{j=1}^{i-1} m_j$ and $\tan\Theta = (\tan\theta_1, \ldots, \tan\theta_{n-1})$.

Lemma 1. *The partial derivatives with respect to θ_k of the moments needed in defining the kurtosis κ of Y are*

$$\frac{\partial M_4}{\partial \theta_k} = \left(\prod_{i=1}^{n-1}\cos\theta_i\right)^4 h_1^k(\tan\Theta), \qquad \frac{\partial (M_2^2)}{\partial \theta_k} = \left(\prod_{i=1}^{n-1}\cos\theta_i\right)^4 h_2^k(\tan\Theta),$$
$$\frac{\partial (M_1^4)}{\partial \theta_k} = \left(\prod_{i=1}^{n-1}\cos\theta_i\right)^4 h_3^k(\tan\Theta), \quad \frac{\partial (M_1^2 M_2)}{\partial \theta_k} = \left(\prod_{i=1}^{n-1}\cos\theta_i\right)^4 h_4^k(\tan\Theta),$$
$$\frac{\partial (M_1 M_3)}{\partial \theta_k} = \left(\prod_{i=1}^{n-1}\cos\theta_i\right)^4 h_5^k(\tan\Theta), \tag{2.3}$$

where, with $\boldsymbol{T} = (\xi_1, \ldots, \xi_{n-1})$,

$$h_1^k(\boldsymbol{T}) = \phi_1^k(\boldsymbol{T},4), \quad h_2^k(\boldsymbol{T}) = 2\phi_2(\boldsymbol{T},2)\phi_1^k(\boldsymbol{T},2), \quad h_3^k(\boldsymbol{T}) = 4\rho_1^3(\boldsymbol{T})\rho_2^k(\boldsymbol{T}),$$
$$h_4^k(\boldsymbol{T}) = 2\rho_1(\boldsymbol{T})\rho_2^k(\boldsymbol{T})\phi_2(\boldsymbol{T},2) + \rho_1^2(\boldsymbol{T})\phi_1^k(\boldsymbol{T},2),$$
$$h_5^k(\boldsymbol{T}) = \rho_2^k(\boldsymbol{T})\phi_2(\boldsymbol{T},3) + \rho_1(\boldsymbol{T})\phi_1^k(\boldsymbol{T},3),$$

and

$$\phi_1^k(\boldsymbol{T},p) = \sum_{|\boldsymbol{q}|=p} \Bigg\{ c_{\boldsymbol{q}}^p \prod_{i\neq k} \xi_i^{p-|\boldsymbol{q}^{i+1}|} \prod_{i=1}^{n-1} \left(1+\xi_i^2\right)^{|\boldsymbol{q}^i|/2} \left((p-|\boldsymbol{q}^{k+1}|)\xi_k^{p-1-|\boldsymbol{q}^{k+1}|} - q_k \xi_k^{p+1-|\boldsymbol{q}^{k+1}|}\right)\mathbb{E}[\boldsymbol{X}^{\boldsymbol{q}}]\Bigg\}, \tag{2.4}$$

$$\phi_2(\boldsymbol{T},p) = \sum_{|\boldsymbol{q}|=p} \left\{ c_{\boldsymbol{q}}^p \prod_{i=1}^{n-1} \xi_i^{p-|\boldsymbol{q}^{i+1}|} \left(1+\xi_i^2\right)^{|\boldsymbol{q}^i|/2} \mathbb{E}[\boldsymbol{X}^{\boldsymbol{q}}]\right\}, \tag{2.5}$$

$$\rho_1(\boldsymbol{T}) = \sum_{i=1}^{n} \left(\prod_{l=1}^{i-1}\xi_l \prod_{l=i+1}^{n-1} \left(1+\xi_l^2\right)^{(1-\delta_{in})/2} \mathbb{E}X_i\right), \tag{2.6}$$

$$\rho_2^k(\boldsymbol{T}) = \sum_{i=1}^{n} \left(\prod_{l=1}^{i-1}\xi_l \prod_{l=i+1}^{n-1} \left(1+\xi_l^2\right)^{(1-\delta_{in})/2} \left[(1-\delta_{ik})\xi_k^2 - (1-\delta_{in})\delta_{ik}\right] \mathbb{E}X_i\right). \tag{2.7}$$

In obtaining the equations for the partial derivatives of the M_k's and their products in (2.3), we have removed the common factor $\prod_{i=1}^{n-1}\cos\theta_i$. We have taken $\cos\theta_i \neq 0$ for all $i = 1,\dots,n-1$, equivalently $\theta_i \neq \pi/2$. After $\prod_{i=1}^{n-1}\cos\theta_i$ has been factored out, the h_i's become algebraic functions of the ξ_k's only, where $\xi_k = \tan\theta_k$. In fact, the h_i are polynomials in ξ_k and $\sqrt{1+\xi_k^2}$ of degree at most four, for $k = 1,\dots,n-1$.

Lemma 2. *The first order conditions for the maximization problem* (2.2)*, obtained by setting the gradient of* κ *to zero, is equivalent to the system of equations*

$$\boldsymbol{H}(\boldsymbol{T}) = \boldsymbol{0}, \tag{2.8}$$

where

$$\boldsymbol{H}(\boldsymbol{T}) = (H_1(\boldsymbol{T}),\dots,H_{n-1}(\boldsymbol{T})), \tag{2.9}$$
$$H_k(\boldsymbol{T}) = h_1^k(\boldsymbol{T}) - 3h_2^k(\boldsymbol{T}) - 6h_3^k(\boldsymbol{T}) + 12h_4^k(\boldsymbol{T}) - 4h_5^k(\boldsymbol{T}). \tag{2.10}$$

Proof. In the notation of Lemma 1, the partial derivative of $\kappa(\Theta)$ with respect to θ_k can be written as

$$\frac{\partial\kappa(\Theta)}{\partial\theta_k} = \frac{\partial}{\partial\theta_k}\left[M_4 - 3M_2^2 - 6M_1^4 + 12M_1^2M_2 - 4M_1M_3\right] = \left(\prod_{i=1}^{n-1}\cos\theta_i\right)^4 \cdot H_k(\mathbf{T}),$$

where $H_k(\mathbf{T})$ is given as in (2.10) by letting $\mathbf{T} = \tan\Theta$. Then setting the gradient $\nabla\kappa(\Theta)$ of κ to zero comes to

$$\nabla\kappa(\Theta) = \left(\prod_{i=1}^{n-1}\cos\theta_i\right)^4 \mathbf{H}(\mathbf{T}) = \mathbf{0},$$

which, since $\cos\theta_i \neq 0$ for $i = 1,\dots,n-1$, is equivalent to (2.8).

The system (2.8) needs to be solved numerically, for example by Newton's method, in order to obtain the critical $\mathbf{T}$. Once we have $\mathbf{T}$, the critical Θ is obtained as $\Theta = \arctan(\mathbf{T})$. Finally, the most non-Gaussian direction is obtained by picking the one among the critical Θ's that maximizes the objective function. While we seek to maximize the *absolute value* of the kurtosis, the same maximum of the absolute value of kurtosis is obtained whether the sign of the kurtosis is positive or negative; the first order conditions for kurtosis provide the critical points for both cases.

To simplify the computations, one can center and whiten the observations before analyzing them, whitening makes the components of the observed vector uncorrelated and their variances equal to 1. A popular method here is to use the eigenvalue decomposition (EVD) of the covariance matrix that, after centering and whitening, is the identity matrix. Hence, the h functions h_3^k, h_4^k and h_5^k, which are involved in the derivation of M_1^4, $(M_1)^2M_2$ and M_1M_3 vanish, and $H_k(\mathbf{T}) = h_1^k(\mathbf{T}) - 3h_2^k(\mathbf{T})$.

3. Influence Functions for the Projection Directions for Multivariate Data

We next develop influence functions to detect abnormal points for the projection directions of multivariate data. Three influence functions for practical use are empirical, deleted empirical, and sample influence functions. We sometimes emphasize dependence on underlying distributions. For example, $\kappa(\Theta; F)$ denotes the kurtosis of the projected random variable in the direction determined by Θ, computed under the distribution function F.

Let F be the distribution function of an n-dimensional random vector and F_ε the perturbed distribution function $F_\varepsilon = (1-\varepsilon)F + \varepsilon\delta_{\mathbf{x}}$, where $\delta_{\mathbf{x}}$ denotes the distribution function (Dirac measure) of the point mass concentrated at $\mathbf{x} = (x_1, \ldots, x_n)$.

Definition 2. If T is a functional acting on an n-dimensional distribution function F by $T : F \to \mathbb{R}$, the influence function $\mathbb{I}(F, \mathbf{x})$ of F at $\mathbf{x}$ is

$$\mathbb{I}(F, \mathbf{x}) := \lim_{\varepsilon \to 0^+} \frac{T(F_\varepsilon) - T(F)}{\varepsilon} = \left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} T(F_\varepsilon).$$

When T is given by an expectation, say

$$T(F) = \mathbb{E}[g(\mathbf{X})] = \int_{\mathbb{R}^n} g(\mathbf{y})dF(\mathbf{y}), \tag{3.1}$$

where $g(\mathbf{y})$ is an integrable function, then

$$\mathbb{I}(F, \mathbf{x}) = \left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} \int_{\mathbb{R}^n} g(\mathbf{y})dF_\varepsilon(\mathbf{y}) = -\mathbb{E}[g(\mathbf{X})] + g(\mathbf{x}).$$

In particular, if g is the monomial $\mathbf{X}^{\mathbf{m}}$, then its influence function at $\mathbf{x}$ is $\mathbf{x}^{\mathbf{m}} - \mathbb{E}[\mathbf{X}^{\mathbf{m}}]$.

Here $\mathbf{T} = \tan\Theta$, which is implicitly defined by the solution of the system of equations at (2.8). Since $\mathbf{T}$ is implicitly defined, the computation of its influence function appeals to the technique of implicit differentiation; the result is summarized in Theorem 1.

Theorem 1. *The influence function $\mathbb{I}(\boldsymbol{T}^*, \boldsymbol{x}; F)$ of the most non-Gaussian projection direction statistic $\boldsymbol{T}^*$ is obtained by solving the linear system*

$$\boldsymbol{G}(\boldsymbol{T}^*; F) \cdot \mathbb{I}(\boldsymbol{T}^*, \boldsymbol{x}; F) = -\boldsymbol{t}(\boldsymbol{T}^*, \boldsymbol{x}; F), \tag{3.2}$$

where $\boldsymbol{G}(\boldsymbol{T}^{*};F)=\left[g_{j}^{k}(\boldsymbol{T}^{*})\right]_{(n-1)\times(n-1)}$, $\boldsymbol{t}(\boldsymbol{T}^{*},\boldsymbol{x};F)=[t_{k}(\boldsymbol{T}^{*},\boldsymbol{x})]_{(n-1)\times 1}$ *and, for* $j,k=1,\ldots,n-1$, $g_{j}^{k}(\boldsymbol{T}^{*})=\lambda_{1,j}^{k}-3\lambda_{2,j}^{k}-6\lambda_{3,j}^{k}+12\lambda_{4,j}^{k}-4\lambda_{5,j}^{k}$, *and* $t_{k}(\boldsymbol{T}^{*},\boldsymbol{x})=\pi_{1}^{k}-3\pi_{2}^{k}-6\pi_{3}^{k}+12\pi_{4}^{k}-4\pi_{5}^{k}$. *Exact expressions for the* λ *and* π *functions are given by* (1.7)−(1.11) *and* (1.12)−(1.16) *in the Supplement.*

Finally, since the influence function $\mathbb{I}(\boldsymbol{T}^{*},\boldsymbol{x};F)$ *is an* $n-1$ *dimensional vector; it is natural to use the vector norm* $|\mathbb{I}(\boldsymbol{T}^{*},\boldsymbol{x};F)|$ *as the magnitude of influence.*

Proof. We only sketch the proof of the theorem in the following because of the complexity of the expressions. Details of full expressions and derivations are in the Supplement. Given any $\varepsilon\geq 0$, let $\mathbf{T}_{\varepsilon}^{*}=(\xi_{1,\varepsilon}^{*},\ldots,\xi_{n-1,\varepsilon}^{*})$ be a solution of the system of

$$\mathbf{H}(\mathbf{T};F_{\varepsilon})=\mathbf{0}, \tag{3.3}$$

where $\mathbf{H}$ is defined at (2.9) and (2.10). This for every $\varepsilon\geq 0$,

$$\mathbf{H}(\mathbf{T}_{\varepsilon}^{*};F_{\varepsilon})=\mathbf{0}. \tag{3.4}$$

Write $\mathbf{T}_{0}^{*}$, the critical $\mathbf{T}$ under the unperturbed distribution F, as $\mathbf{T}^{*}=(\xi_{1}^{*},\ldots,\xi_{n-1}^{*})$. Let $I_{i}(\mathbf{T}^{*},\mathbf{x})=\frac{d}{d\varepsilon}\xi_{i,\varepsilon}^{*}\big|_{\varepsilon=0}$ be the influence function of ξ_{i}^{*} at $\mathbf{x}$, and let $\mathbb{I}(\mathbf{T}^{*},\mathbf{x})=(I_{1}(\mathbf{T}^{*},\mathbf{x}),\ldots,I_{n-1}(\mathbf{T}^{*},\mathbf{x}))$ be the influence function of $\mathbf{T}$ at $\mathbf{x}$. To derive the influence function of $\mathbb{I}(\mathbf{T}^{*},\mathbf{x})$, we implicitly differentiate (3.4) with respect to ε, then evaluate the resulting equation at $\varepsilon=0$. Via straightforward computations (for details see the Supplement) plus the use of (3.4), the resulting equation can be written as

$$\begin{aligned}\frac{d}{d\varepsilon}\mathbf{H}(\mathbf{T}_{\varepsilon}^{*},F_{\varepsilon})\bigg|_{\varepsilon=0}&=\mathbf{G}(\mathbf{T}^{*})\cdot\mathbb{I}(\mathbf{T}^{*},\mathbf{x})+\mathbf{t}(\mathbf{T}^{*},\mathbf{x})\\&\equiv\begin{bmatrix}g_{1}^{1}&\cdots g_{n-1}^{1}\\ \vdots&\ddots\ \vdots\\ g_{1}^{n-1}&\cdots g_{n-1}^{n-1}\end{bmatrix}\cdot\begin{bmatrix}I_{1}(\mathbf{T}^{*},\mathbf{x})\\ \vdots\\ I_{n-1}(\mathbf{T}^{*},\mathbf{x})\end{bmatrix}+\begin{bmatrix}t_{1}\\ \vdots\\ t_{n-1}\end{bmatrix}=\mathbf{0},\end{aligned}$$

where, for $j=1,\ldots,n-1$ and $k=1,\ldots,n-1$,

$$\begin{aligned}g_{j}^{k}(\mathbf{T}^{*})&=\lambda_{1,j}^{k}-3\lambda_{2,j}^{k}-6\lambda_{3,j}^{k}+12\lambda_{4,j}^{k}-4\lambda_{5,j}^{k},\\t_{k}(\mathbf{T}^{*},\mathbf{x})&=\pi_{1}^{k}-3\pi_{2}^{k}-6\pi_{3}^{k}+12\pi_{4}^{k}-4\pi_{5}^{k}.\end{aligned}$$

We remark that all the variables λ's and π's involved in the above expressions are functions of $\mathbf{T}^{*}$ and $\mathbf{x}$; we have suppressed the dependence for simplicity. The exact expressions for the λ and π functions are given by (1.7)−(1.11) and (1.12)−(1.16) in the Supplement. Then, the influence function $\mathbb{I}(\mathbf{T}^{*},\mathbf{x};F)$ can be obtained by solving the linear system in (3.2).

Remark 1. For the two-dimensional case, only a single θ is needed to parametrize the possible projection directions. To find the direction with maximal kurtosis, the derivative of kurtosis can be regarded as a polynomials of tanθ and set to zero; we end up with an explicit expression for the influence function of tanθ. In the multi-dimensional case, we need to consider the projection directions as a vector of angles $\Theta = (\theta_1, \ldots, \theta_{n-1})$ in polar coordinates. In finding the maximizer for kurtosis, the first order criterion, $\nabla\kappa(\Theta) = 0$, can only be solved numerically. Accordingly, the derivation of pair-perturbation influence functions for the projection directions is much more involved.

3.1. Empirical influence function

In practice, the exact distribution function F is unknown but can be estimated by an empirical distribution function $\hat{F}$ based on a random sample. The empirical influence function (EIF) is obtained by replacing $\mathbf{G}(\cdot; F)$ and $\mathbf{t}(\cdot; F)$ in (3.2) by the sample estimates $\mathbf{G}(\cdot; \hat{F})$ and $\mathbf{t}(\cdot; \hat{F})$, respectively. That is, $\hat{\mathbb{I}}$ solves the linear system

$$\mathbf{G}(\mathbf{T}^*; \hat{F}) \cdot \hat{\mathbb{I}}(\mathbf{T}^*, \mathbf{x}; \hat{F}) = -\mathbf{t}(\mathbf{T}^*, \mathbf{x}; \hat{F}).$$

3.2. Deleted empirical influence function

Let $\hat{F}_{(\mathbf{x})}$ be the cdf obtained by deleting $\mathbf{x}$ from the empirical cdf $\hat{F}$ and then substituting it into $\mathbf{G}(\cdot; F)$ and $\mathbf{t}(\cdot; F)$ in (3.2). The deleted empirical influence function is obtained by solving the linear system

$$\mathbf{G}(\mathbf{T}^*; \hat{F}_{(\mathbf{x})}) \cdot \mathring{\mathbb{I}}(\mathbf{T}^*, \mathbf{x}; \hat{F}_{(\mathbf{x})}) = -\mathbf{t}(\mathbf{T}^*, \mathbf{x}; \hat{F}_{(\mathbf{x})}).$$

3.3. Sample influence function

The sample influence function (SIF) was defined by Devlin, Gnanadesikan and Kettenring, (1975) as

$$SIF_{T,F}(z_i) = -(N-1)(T(\hat{F}_{(i)}) - T(\hat{F})), \tag{3.5}$$

where N is the size of a random sample. It directly measures the effect on the functional when the ith observation is removed from the empirical cdf $\hat{F}$. As the statistical functional is $\mathbf{T} = \tan\Theta$ in our case, the sample influence function of the projection direction is

$$\tilde{\mathbb{I}}(\mathbf{T}^*, \mathbf{x}; \hat{F}) = -(N-1)\left[\mathbf{T}^*(\hat{F}_{(\mathbf{x})}) - \mathbf{T}^*(F)\right],$$

where $\mathbf{T}^*(\hat{F}_{(\mathbf{x})})$ and $\mathbf{T}^*(F)$ are the solutions of the system (2.8) based on the deleted empirical cdf $\hat{F}_{(\mathbf{x})}$ and the empirical cdf $\hat{F}$.

3.4. Averaged influence function

In order to retain useful information and average out misleading noise variation from EIF, DIF, and SIF, an averaged influence function is proposed. The values of the three influence functions are first standardized by their corresponding means and standard deviations,

$$S(IF) = \frac{(IF - \overline{IF})}{\sqrt{\mathrm{Var}\,(IF)}}, \tag{3.6}$$

and then the averaged influence function is

$$AIF = \frac{1}{3}\{S(EIF) + S(DIF) + S(SIF)\}.$$

3.5. Pair-perturbation influence function for the projection direction

It is well known that single-perturbation diagnostics can suffer from a masking effect. Here we generalize the construction of the single-perturbation influence function of projection direction to the pair-perturbation influence function of projection direction to uncover masked influential points.

Let F_ε be the pair-perturbed cdf of F at $\mathbf{x}_1$ and $\mathbf{x}_2$, $F_\varepsilon = (1-2\varepsilon)F + \varepsilon(\delta_{\mathbf{x}_1} + \delta_{\mathbf{x}_2})$, and let T be a statistical functional acting on an n-dimensional distribution function F by $T : F \to \mathbb{R}$.

Definition 3. The pair-perturbation influence function $\mathbb{I}(F, \mathbf{x}_1, \mathbf{x}_2)$ of F at $(\mathbf{x}_1, \mathbf{x}_2)$ is

$$\mathbb{I}(F, \mathbf{x}_1, \mathbf{x}_2) := \lim_{\varepsilon \to 0^+} \frac{T(F_\varepsilon) - T(F)}{\varepsilon} = \frac{d}{d\varepsilon}\bigg|_{\varepsilon=0} T(F_\varepsilon).$$

Lemma 3. *The pair-perturbation influence function* $\mathbb{I}(F, \boldsymbol{x}_1, \boldsymbol{x}_2)$ *for the statistic in* (3.1) *is additive in the sense that* $\mathbb{I}(F, \boldsymbol{x}_1, \boldsymbol{x}_2) = \mathbb{I}(F, \boldsymbol{x}_1) + \mathbb{I}(F, \boldsymbol{x}_2)$.

Proof. Indeed, the pair-perturbed influence function $\mathbb{I}(F, \mathbf{x}_1, \mathbf{x}_2)$ of the statistic in (3.1) is determined by

$$\begin{aligned}\mathbb{I}(F, \mathbf{x}_1, \mathbf{x}_2) &= \frac{d}{d\varepsilon}\bigg|_{\varepsilon=0} \int_{\mathbb{R}^n} g(\mathbf{y}) dF_\varepsilon(\mathbf{y}) = -2\mathbb{E}[g(\mathbf{X})] + g(\mathbf{x}_1) + g(\mathbf{x}_2) \\ &= \mathbb{I}(F, \mathbf{x}_1) + \mathbb{I}(F, \mathbf{x}_2).\end{aligned}$$

For the monomial function $\mathbf{X}^{\mathbf{m}}$, one gets

$$\frac{d}{d\varepsilon}\mathbb{E}[\mathbf{X}^{\mathbf{m}}; F_\varepsilon]\bigg|_{\varepsilon=0} = -2\mathbb{E}[\mathbf{X}^{\mathbf{m}}] + \mathbf{x}_1^{\mathbf{m}} + \mathbf{x}_2^{\mathbf{m}}.$$

Moreover, by direct computation, one can also show that the ϕ functions and ρ functions share additivity in the form

$$\begin{aligned}
\frac{d}{d\varepsilon}\phi^k_{1,\varepsilon}(p)\bigg|_{\varepsilon=0} &= \sum_{j=1}^{n-1} I_j \xi^k_{1,j}(p) - 2\phi^k_1(p) + \phi^k_1(p,\mathbf{x}_1) + \phi^k_1(p,\mathbf{x}_2),\\
\frac{d}{d\varepsilon}\phi_{2,\varepsilon}(p)\bigg|_{\varepsilon=0} &= \sum_{j=1}^{n-1} I_j \xi_{2,j}(p) - 2\phi_2(p) + \phi_2(p,\mathbf{x}_1) + \phi_2(p,\mathbf{x}_2),\\
\frac{d}{d\varepsilon}\rho_{1,\varepsilon}\bigg|_{\varepsilon=0} &= \sum_{j=1}^{n-1} I_j \varrho_{1,j} - 2\rho_1 + \rho_1(\mathbf{x}_1) + \rho_1(\mathbf{x}_2),\\
\frac{d}{d\varepsilon}\rho^k_{2,\varepsilon}\bigg|_{\varepsilon=0} &= \sum_{j=1}^{n-1} I_j \varrho^k_{2,j} - 2\rho_2 + \rho^k_2(\mathbf{x}_1) + \rho^k_2(\mathbf{x}_2),
\end{aligned}$$

where I_j is the pair-perturbed influence function of ξ^*_j at $(\mathbf{x}_1, \mathbf{x}_2)$. The definitions of the functions $\phi^k_1(p,\cdot)$, $\phi_2(p,\cdot)$, $\rho_1(\cdot)$, and $\rho^k_2(\cdot)$ are given by (1.4), (1.3), (1.5), and (1.6), respectively, in the Supplement. (The dependence of the functions $\phi^k_1(p,\cdot)$, $\phi_2(p,\cdot)$, $\rho_1(\cdot)$, and $\rho^k_2(\cdot)$ on $\mathbf{T}$ is not shown for notational simplicity.) We conclude that

$$\begin{aligned}
\frac{d}{d\varepsilon}h^k_1(\mathbf{T}^*_\varepsilon; F_\varepsilon)\bigg|_{\varepsilon=0} &= \sum_{j=1}^{n-1} I_j \lambda^k_{1,j}(\mathbf{T}^*; F) - 2h^k_1(\mathbf{T}^*; F) + \pi_1^{\prime k}(\mathbf{T}^*, \mathbf{x}_1, \mathbf{x}_2; F),\\
\frac{d}{d\varepsilon}h^k_2(\mathbf{T}^*_\varepsilon; F_\varepsilon)\bigg|_{\varepsilon=0} &= \sum_{j=1}^{n-1} I_j \lambda^k_{2,j}(\mathbf{T}^*; F) - 2h^k_2(\mathbf{T}^*; F) + \pi_2^{\prime k}(\mathbf{T}^*, \mathbf{x}_1, \mathbf{x}_2; F),\\
\frac{d}{d\varepsilon}h^k_3(\mathbf{T}^*_\varepsilon; F_\varepsilon)\bigg|_{\varepsilon=0} &= \sum_{j=1}^{n-1} I_j \lambda^k_{3,j}(\mathbf{T}^*; F) - 2h^k_1(\mathbf{T}^*; F) + \pi_3^{\prime k}(\mathbf{T}^*, \mathbf{x}_1, \mathbf{x}_2; F),\\
\frac{d}{d\varepsilon}h^k_4(\mathbf{T}^*_\varepsilon; F_\varepsilon)\bigg|_{\varepsilon=0} &= \sum_{j=1}^{n-1} I_j \lambda^k_{4,j}(\mathbf{T}^*; F) - 2h^k_2(\mathbf{T}^*; F) + \pi_4^{\prime k}(\mathbf{T}^*, \mathbf{x}_1, \mathbf{x}_2; F),\\
\frac{d}{d\varepsilon}h^k_5(\mathbf{T}^*_\varepsilon; F_\varepsilon)\bigg|_{\varepsilon=0} &= \sum_{j=1}^{n-1} I_j \lambda^k_{5,j}(\mathbf{T}^*; F) - 2h^k_2(\mathbf{T}^*; F) + \pi_5^{\prime k}(\mathbf{T}^*, \mathbf{x}_1, \mathbf{x}_2; F),
\end{aligned}$$

where

$$\begin{aligned}
\pi_1^{\prime k}(\mathbf{T}^*, \mathbf{x}_1, \mathbf{x}_2; F) &= \phi^k_1(4,\mathbf{x}_1) + \phi^k_1(4,\mathbf{x}_2),\\
\pi_2^{\prime k}(\mathbf{T}^*, \mathbf{x}_1, \mathbf{x}_2; F) &= 2\big[\big(\phi_2(2,\mathbf{x}_1) + \phi_2(2,\mathbf{x}_2)\big)\phi^k_1(2) + \phi_2(2)\big(\phi^k_1(2,\mathbf{x}_1) + \phi^k_1(2,\mathbf{x}_2)\big)\\
&\qquad - 2\phi_2(2)\phi^k_1(2)\big],
\end{aligned}$$

$$
\begin{aligned}
\pi_3^{'k}(\mathbf{T}^*,\mathbf{x}_1,\mathbf{x}_2;F)=&4\big[3\rho_1^2\rho_2^k\big(-2\rho_1+\rho_1(\mathbf{x}_1)+\rho_1(\mathbf{x}_2)\big)+\rho_1^3\big(\rho_2^k(\mathbf{x}_1)+\rho_2^k(\mathbf{x}_2)\big)\big],\\
\pi_4^{'k}(\mathbf{T}^*,\mathbf{x}_1,\mathbf{x}_2;F)=&2\big(-2\rho_1+\rho_1(\mathbf{x}_1)+\rho_1(\mathbf{x}_2)\big)(\rho_2^k\phi_2(2)+\rho_1\phi_1^k(2))\\
&+2\rho_1\rho_2^k\big(\phi_2(2,\mathbf{x}_1)+\phi_2(2,\mathbf{x}_2)\big)+\rho_1^2\big(\phi_1^k(2,\mathbf{x}_1)+\phi_1^k(2,\mathbf{x}_2)\big)\\
&+2\rho_1\phi_2(2)\big(-2\rho_2^k+\rho_2^k(\mathbf{x}_1)+\rho_2^k(\mathbf{x}_2)\big),\\
\pi_5^{'k}(\mathbf{T}^*,\mathbf{x}_1,\mathbf{x}_2;F)=&\phi_2(3)\big(-2\rho_2^k+\rho_2^k(\mathbf{x}_1)+\rho_2^k(\mathbf{x}_2)\big)+\rho_2^k\big(\phi_2(3,\mathbf{x}_1)+\phi_2(3,\mathbf{x}_2)\big)\\
&+\phi_1^k(3)\big(-2\rho_1+\rho_1(\mathbf{x}_1)+\rho_1(\mathbf{x}_2)\big)+\rho_1\big(\phi_1^k(3,\mathbf{x}_1)+\phi_1^k(3,\mathbf{x}_2)\big).
\end{aligned}
$$

Applying the same trick as in the proof of Theorem 1 and using these computations, we obtain the pair-perturbation influence function for the project direction statistic $\mathbf{T}^*$.

Theorem 2. *The pair-perturbation influence function* $\mathbb{I}(\boldsymbol{T}^*,\boldsymbol{x}_1,\boldsymbol{x}_2;F)$ *of the most non-Gaussian projection direction statistic* $\boldsymbol{T}^*$ *is obtained by solving the linear system*

$$
\boldsymbol{G}(\boldsymbol{T}^*,F)\cdot\mathbb{I}(\boldsymbol{T}^*,\boldsymbol{x}_1,\boldsymbol{x}_2;F)=-\boldsymbol{t}^{'}(\boldsymbol{T}^*,\boldsymbol{x}_1,\boldsymbol{x}_2;F),\tag{3.7}
$$

where

$$
\begin{aligned}
\boldsymbol{t}^{'}(\boldsymbol{T}^*,\boldsymbol{x}_1,\boldsymbol{x}_2;F)&=(t_1^{'}(\boldsymbol{T}^*,\boldsymbol{x}_1,\boldsymbol{x}_2;F),\ldots,t_{n-1}^{'}(\boldsymbol{T}^*,\boldsymbol{x}_1,\boldsymbol{x}_2;F))^T,\\
t_k^{'}(\boldsymbol{T}^*,\boldsymbol{x}_1,\boldsymbol{x}_2;F)&=\pi_1^{'k}-3\pi_2^{'k}-6\pi_3^{'k}+12\pi_4^{'k}-4\pi_5^{'k}.
\end{aligned}
$$

Moreover, the pair-perturbation influence for the most non-Gaussian projection direction, defined implicitly by the solution of the system (2.8) and therefore not immediately an expectation functional, has the additivity property.

Corollary 1. *The pair-perturbation influence function* $\mathbb{I}(\boldsymbol{T}^*,\boldsymbol{x}_1,\boldsymbol{x}_2;F)$ *of the most non-Gaussian projection direction statistic* $\boldsymbol{T}^*$*, defined by the solution to the system* (3.7)*, is additive in the sense that* $\mathbb{I}(\boldsymbol{T}^*,\boldsymbol{x}_1,\boldsymbol{x}_2;F)=\mathbb{I}(\boldsymbol{T}^*,\boldsymbol{x}_1;F)+\mathbb{I}(\boldsymbol{T}^*,\boldsymbol{x}_2;F)$.

Proof. Referring to (1.12) to (1.16) for the π functions in the Supplement, we note that the π' functions also satisfy the additivity property, $\pi_a^k(\mathbf{T}^*,\mathbf{x}_1,\mathbf{x}_2;F)=\pi_a^k(\mathbf{T}^*,\mathbf{x}_1;F)+\pi_a^k(\mathbf{T}^*,\mathbf{x}_1;F)$, for $a=1,\ldots,5$, and therefore $\mathbf{t}'$ satisfies $\mathbf{t}^{'}(\mathbf{T}^*,\mathbf{x}_1,\mathbf{x}_2;F)=\mathbf{t}(\mathbf{T}^*,\mathbf{x}_2;F)+\mathbf{t}(\mathbf{T}^*,\mathbf{x}_2;F)$. The additivity of $\mathbf{t}'$ and (3.7) imply the additivity of the influence function $\mathbb{I}$.

3.6. Sample version of the pair-perturbation influence functions

3.6.1. Pair-perturbation empirical influence function

We replace F by $\hat{F}$ in (3.7) to obtain the pair-perturbation empirical influence function of projection direction by solving

$$
\mathbf{G}(\mathbf{T}^*;\hat{F})\cdot\hat{\mathbb{I}}(\mathbf{T}^*,\mathbf{x}_1,\mathbf{x}_2;\hat{F})=-\mathbf{t}^{'}(\mathbf{T}^*;\hat{F}).
$$

3.6.2. Pair-perturbation delete empirical influence function

Let $\hat{F}_{(\mathbf{x}_1,\mathbf{x}_2)}$ be the cdf obtained by deleting the $\mathbf{x}_1$ and $\mathbf{x}_2$ observations from the empirical cdf $\hat{F}$. Then the pair-perturbation delete empirical influence of projection direction is obtained by replacing F by $\hat{F}_{(\mathbf{x}_1,\mathbf{x}_2)}$ in (3.7) as

$$\mathbf{G}(\mathbf{T}^*; \hat{F}_{(\mathbf{x}_1,\mathbf{x}_2)}) \cdot \hat{\mathbb{I}}_{(\mathbf{x}_1,\mathbf{x}_2)}(\mathbf{T}^*, \mathbf{x}_1, \mathbf{x}_2; \hat{F}_{(\mathbf{x}_1,\mathbf{x}_2)}) = -\mathbf{t}^{'}(\mathbf{T}^*; \hat{F}_{(\mathbf{x}_1,\mathbf{x}_2)}).$$

3.6.3. Pair-perturbation sample influence function

The pair-perturbation sample influence function of projection direction is

$$\tilde{\mathbb{I}}(\mathbf{T}^*, \mathbf{x}_1, \mathbf{x}_2; \hat{F}) = -(N-2)(\mathbf{T}^*(\hat{F}_{(\mathbf{x}_1,\mathbf{x}_2)}) - \mathbf{T}^*(\hat{F})),$$

where $\mathbf{T}^*(\hat{F}_{(\mathbf{x}_1,\mathbf{x}_2)})$ and $\mathbf{T}^*(\hat{F})$ are the solutions of the system (2.8) based on the deleted empirical cdf $\hat{F}_{(\mathbf{x}_1,\mathbf{x}_2)}$ and the empirical cdf $\hat{F}$, respectively.

3.6.4. Pair-perturbation averaged influence function

To calculate the pair-perturbation averaged influence function, we first standardize the values of the influence functions by their corresponding means and standard deviations as in (3.6). Then the averaged pair-perturbation influence function is

$$AIF = \frac{\{SP(EIF) + SP(DIF) + SP(SIF)\}}{3},$$

where SP means standardized pair-perturbation influence function.

3.7. Cut points selection for influence functions

One can grasp the influential observations visually when the values of the influence function are exhibited in a diagram. However, such a figure can only present the relative magnitudes of the oscillation for the values of influence functions; a choice of cut points is required to determine the "influential points".

Write the interquartile range (IQR) as IQR $:= q_3 - q_1$, where q_1 is the lower (25^{th}) quantile and q_3 is the upper (75^{th}) quantile; the quantities LOF $=$ $q_1 - 3$(IQR) and UOF $= q_3 + 3$(IQR) are the lower outer fence and upper outer fence, respectively. Observations with an influence value outside the outer fences are referred to as influential points.

4. Examples

Three examples (a simulation study, a data set and a simulated data set) illustrate the application of the proposed techniques of projection pursuit and the use of influence functions of projection directions.

Table 1. The single-perturbation results of a simulated data with 500 runs.

Influential observation	EIF	DIF	SIF	AIF
61^{st}	84(16.8%)	159(13.8%)	436(87.2%)	236(47.2%)
62^{nd}	497(99.4%)	180(36%)	481(96.2%)	499(99.8%)

Table 2. The pair-perturbation results of a simulation experiment with 500 runs.

Influential paired observations	EIF	DIF	SIF	AIF
(61^{st}, 62^{nd})	492(98.4%)	468(93.6%)	468(93.6%)	499(99.8%)

4.1. A simulation study

We generated 60 observations from five independent uniforms on $(-0.5,\ 0.5)$. The data were mixed using the (randomly chosen) mixing matrix

$$\begin{bmatrix} 6 & 2 & 9 & 5 & 4 \\ 2 & 8 & 1 & 5 & 5 \\ 2 & 6 & 4 & 8 & 3 \\ 9 & 4 & 2 & 2 & 1 \\ 1 & 8 & 6 & 8 & 1 \end{bmatrix}$$

and then points $(0.2,\ 12,\ 0.7,\ 0.6,\ 9)$ and $(0.1,\ 15,\ 14,\ 0.5,\ 0.3)$, indexed by numbers 61 and 62, were added. We centered and whitened the observations and then found the solution to the polynomial (2.8) in $\tan\Theta$ with maximum kurtosis. Next, the single-perturbation influence function and pair-perturbation influence function for the projection direction were calculated. This whole procedure was implemented 500 times.

Table 1 summarizes, the frequencies of the 61^{st} and 62^{nd} artificial observations when detected as influential points. The results for single-perturbation show that the sample influence function (SIF) successfully detected both the 61^{st} and 62^{nd} observations as influential points with high percentages (87.2% and 96.2%) of the time. On the other hand, EIF and AIF detected the 62^{nd} point as influential with a high percentages (99.4%) and (99.8%), respectively; DIF detected the 61^{st} and the 62^{nd} points as influential with low percentages (13.8%) and (36%), respectively. In order to further investigate the pair-perturbation results, we computed the average frequencies of the influential pair(s) for each observation among 500 simulation runs, and here plot the average counts (frequencies) of the influential pair(s) versus the observation index number as a diagnostic plot in Figure 2. The results are also summarized in Table 2. Clearly, observations paired with the 61^{st} and 62^{nd} observations have larger average counts of influential pairs for all of the influence functions. These simulation results suggest that the DIF does not succeed in detecting any influential point by single-perturbation, but the pair-perturbation DIF successfully detects the 61^{st} and the

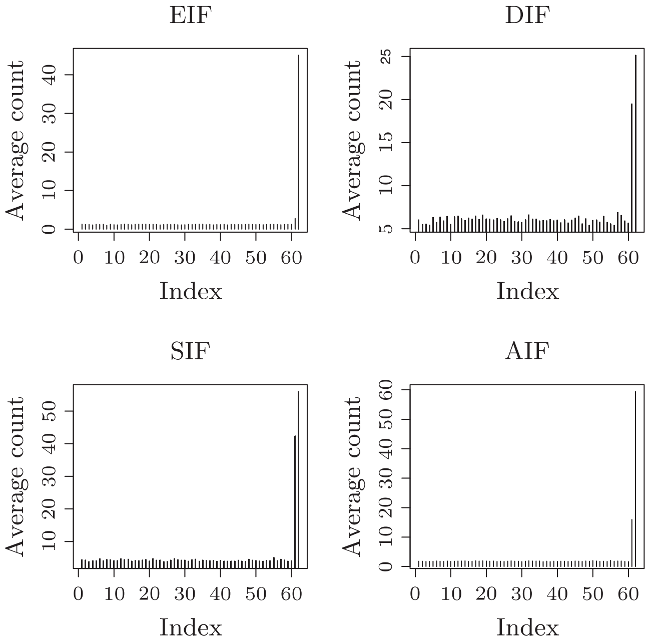

Figure 2. Plots of average counts of influential pair(s) versus the observation index number for a multivariate simulated data set with 500 replications.

62^{nd} as influential. Moreover, AIF, which averages all information from three influence functions, also successfully uncovered the masking 61^{st} point via the pair-perturbation scheme. These results suggest that the pair-perturbation influence functions provide a very useful auxiliary scheme to detect masked unusual points not detected by single perturbation influence functions.

4.2. Salinity data

Our second illustration uses the Salinity data set taken from Rousseeuw and Leory (1987, p.82). This comprises 28 measurements of water salinity and river discharge taken in North Carolina's Pamlico Sound and three explanatory variables: salinity, lagged by two weeks (x_1); the trend, that is the number of biweekly periods elapsed since the beginning of the spring season (x_2); and the volume of river discharge into the sound (x_3). Carroll and Ruppert (1985) described the physical background of the data and pointed out that cases 5 and 16 correspond to periods of very heavy discharge. Rousseeuw and van Zomeren (1990) concluded that three good leverage points (cases 5, 23, 24) and one bad leverage point (case 16) would be identified using larger cutoff values ±5 by applying the least median of squares (LMS) method. However, Fung (1993) proposed a stepwise confirmatory analysis and concluded that observation 16 is

Table 3. Influential cases of the single-perturbation influence functions for Salinity data.

Influence Function	influence cases detected
EIF	5, 16
DIF	3, 5, 13, 16
SIF	16
AIF	16

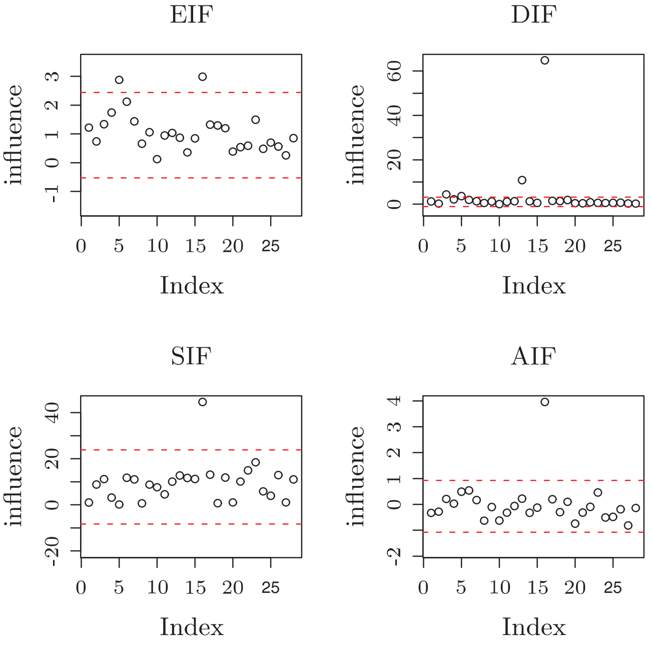

Figure 3. Plots of values of single-perturbation influence functions of $\tan\Theta^*$ versus the observation index number for a multivariate salinity data set.

the only outlier in the data set and that observations 5 and 23 are slightly high leverage points.

We first centered and whitened the observations. The solution to the polynomial (2.8) in $\tan\Theta$ with maximum kurtosis is 6.5527; thus the projection direction Θ^* is $(0.2466\pi, 0)$. Next, we computed the four influence functions of projection directions and here plot their values versus the observation index as a diagnostic plot in Figure 3. The cut points of the influence functions were selected based on the lower-upper inner fences (see Section 3). Table 3 summarizes the influential cases of single-perturbation influence functions. The results in Figure 3 and Table 3 show that EIF and DIF both detect the 5th and 16th observations as influential points. On the other hand, SIF and AIF detect only the 16th observation as influential.

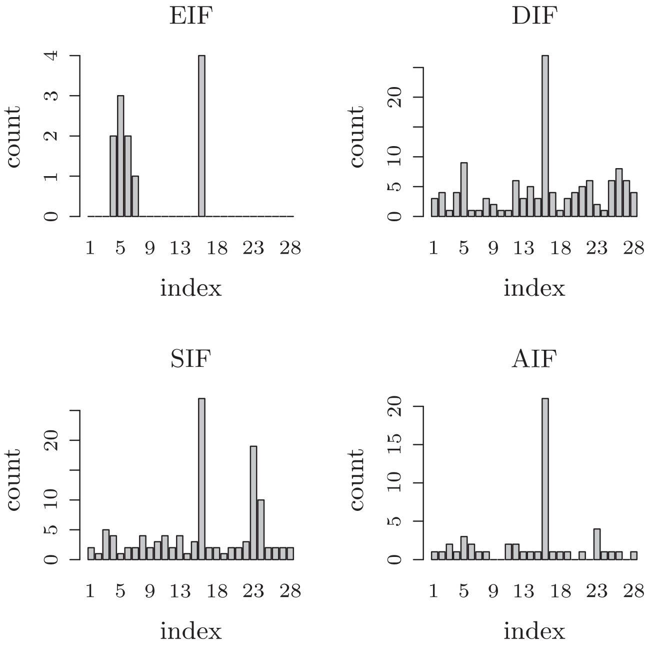

Figure 4. Plots of average counts of influential pair(s) versus the observation index number for a multivariate salinity data set.

We also considered the use of pair-perturbation influence functions to detect masking influential points. The values of the pair-perturbation influence functions versus the observation index are shown in the diagnostic plot of Figure 4. Note that observations paired with the 5th and 16th observations have clearly larger counts of influential pairs for EIF and DIF. This is not surprising for EIF because of its additivity property. On the other hand, the observations paired with the 16th and 23th observations have clearly larger counts of influential pairs for SIF. The results for this data set suggest that the EIF and DIF succeed in detecting the 5th and 16th influential points, both for the single-perturbation and the pair-perturbation cases. This agrees with the results of Carroll and Ruppert (1985). The SIF pair-perturbation influence function indicates that the 23th observation is a potential outlier/influential point. This is in agreement with the residual plot with the LMS fit in Rousseeuw and Leory (1987, p.84) and the confirmatory analysis of Fung (1993, p.519). Moreover, AIF successfully detects only the 16th observation as influential, both for single-perturbation and pair-perturbation cases. This also agrees with the confirmatory analysis of Fung (1993, p.519). In brief, our proposed outlier/influential observations detection method shows excellent performance.

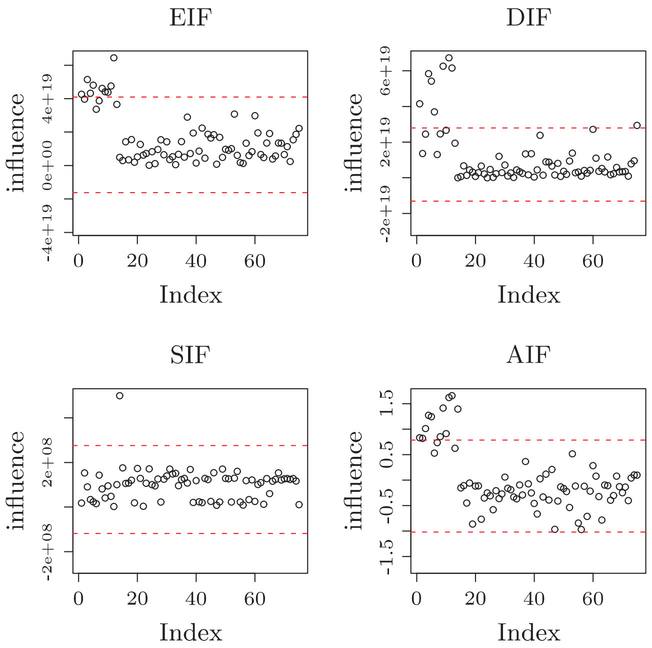

Figure 5. Plots of values of single-perturbation influence functions of $\tan \Theta^*$ versus the observation index number for the multivariate Hawkins-Bradu-Kass data set.

4.3. Hawkins-Bradu-Kass data

For the final illustration, we use the Hawkins-Bradu-Kass data generated artificially by Hawkins, Bradu, and Kass (1984, Table 4). This data set consists of 75 observations with three explanatory variables. The first ten observations are outliers as well as leverage points. They are called bad leverage points by Rousseeuw and van Zomeren (1990). The next four observations (11th $-$14th) are good leverage points (not outliers). Rousseeuw and van Zomeren (1990) correctly identified the first 14 extreme observations using the LMS and MVE methods. The first 14 observations are also confirmed as high leverage influential points via a stepwise analysis proposed by Fung (1993).

We first centered and whitened the observations. The solution to the polynomial (2.8) in $\tan\Theta$ with maximum kurtosis is 20.63586, thus the projection direction Θ^* is $(0, -0.5\pi)$. Next, we computed the influence functions of projection direction and here plot the values of the influence functions versus the observation index as a diagnostic plot in Figure 5. The cut points of influence functions were selected based on the lower-upper inner fences. Table 4 summarizes the influential cases detected by the single-perturbation influence functions. The results show that EIF detects observations 1, 3, 4, 5, 8, 9, 10, 11,

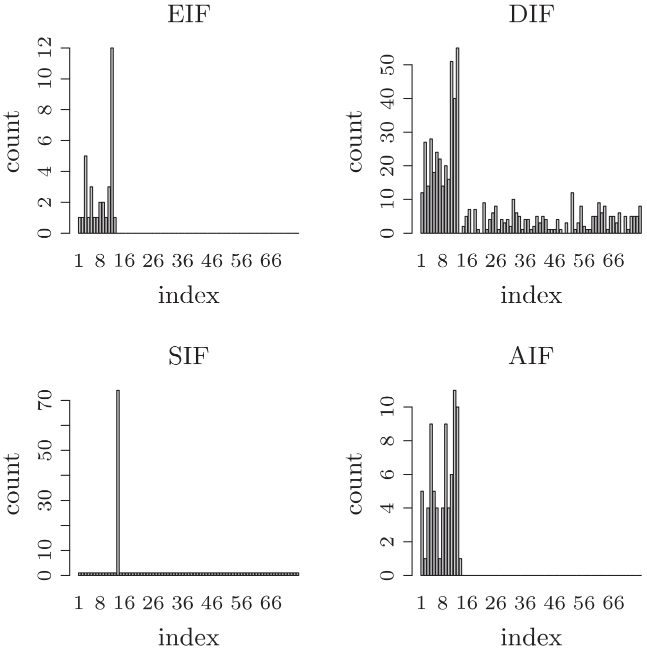

Figure 6. Plots of average counts of influential pair(s) versus the observation index number for the multivariate Hawkins-Bradu-Kass data set.

Table 4. Influential cases of the single-perturbation influence functions for Hawkins-Bradu-Kass data.

Influence Function	influence cases detected
EIF	1, 3, 4, 5, 8, 9, 10, 11, 12
DIF	1, 4, 5, 6, 9, 11, 12, 75
SIF	14
AIF	1, 2, 3, 4, 5, 8, 9, 10, 11, 12, 14

12 as influential; DIF detects observations 1, 4, 5, 6, 9, 11, 12, 75 as influential; SIF detects only the 14th observation as influential; AIF detects observations 1, 2, 3, 4, 5, 8, 9, 10, 11, 12, 14 as influential. We also use the pair-perturbation influence function to detect the masking influential points. A plot of the values of the pair-perturbation influence functions versus the corresponding observation index is shown in Figure 6. Note that the first 13 observations clearly have larger counts of influential pairs for EIF, DIF, and AIF; the 14th observation clearly has larger counts of influential pairs for SIF and AIF. The results suggest that the EIF and DIF pair-perturbation influence functions successfully uncover masking observations *except* for observation 14. The AIF successfully uncovers *all* masking observations not detected by single-perturbation. In fact, our proposed AIF method confirms the analyses of Rousseeuw and van Zomeren (1990) and of Fung

(1993).

5. Concluding Remarks

This paper extends previous work by Huang, Cheng, and Wang (2007, 2008) that discussed single-perturbation influence functions and pair-perturbation influence functions of projection direction by applying projection pursuit. The previous papers provided a way to detect influential observations/outliers for the two-dimensional data case. Here, we generalize the results to general higher dimensions. For multivariate high-dimensional data, a technique is proposed for defining and developing influence functions of projection directions. Also a new influence function is suggested, the averaged influence function (AIF), which averages the information obtained by the empirical influence function (EIF), the deleted empirical influence function (DIF), and the sample influence function (SIF). Three specific numerical examples (two simulated data, one real data) are discussed.

In our numerical examples, the proposed pair-perturbation method uncovers the case of one observation masked by another. The method can of course be extended to multiple perturbation schemes. However, in the case where the number of observation is large and/or the data dimensions are high, the number of combinations of cases needing to be analyzed becomes extremely large. An alternative solution then would be to apply the local influence functions proposed by Cook (1986), as discussed in the context of principal component analysis in Shi (1997) and Huang, Cheng, and Wang (2007).

Acknowledgements

The authors would like to thank an associate editor and an anonymous referee for detailed and thoughtful comments that aided our revision. The first author and corresponding author are partially supported by a grant from the National Science Council of Taiwan (NSC-95-2118-M-194-003).

References

Carroll, R. J. and Ruppert, D. (1985). Transformations in regression: a robust analysis, *Technometrics* **27**, 1-12.

Cook, R. D. (1986). Assessment of local influence. *J. Roy. Statist. Soc. Ser. B* **48**, 133-169.

Critchley, F.(1985). Influence in principal component analysis. *Biometrika* **72**, 627-636.

Devlin, S. J., Gnanadesikan, R. and Kettenring, J. R., (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika* **62**, 531-545.

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

Fung, W. K. (1992). Some diagnostic measures in discriminant analysis. *Statist. Probab. Lett.* **13**, 279-285.

Fung, W. K. (1993). Unmasking outliers and leverage points: a confirmation. *J. Amer. Statist. Assoc.* **88**, 518-519.

Fung, W. K. (1995). Diagnostics in linear discriminant analysis. *J. Amer. Statist. Assoc.* **90**, 952-956.

Fung, W. K. (1996). The influence of an observation on the misclassification probability in multiple discriminant analysis. *Comm. Statist. Theory Methods* **25**, 1917-1930.

Hawkins, D. M., Bradu, D., and Kass, G. V. (1984). Location of several outliers in multiple regression data using elemental sets. *Technometrics* **26**, 197-208.

He, X. and Fung, W. K. (2000). High breakdown estimation for multiple populations with applications to discriminant analysis. *J. Multivariate Anal.* **72**, 151-162.

Huang, Y., Cheng, C. R. and Wang, T. H. (2007). Influence analysis of nongaussianity by applying projection pursuit. *Statist. Probab. Lett.* **77**, 1515-1521.

Huang, Y., Cheng, C. R. and Wang, T. H. (2008). Pair-perturbation influence functions of Nongaussianity by Applying Projection Pursuit. *Comput. Statist. Data Anal.* **52**, 3971-3987.

Huang, Y., Kao, T. L. and Wang, T. H. (2007). Influence functions and local influence in linear discriminant analysis. *Comput. Statist. Data Anal.* **51**, 3844-3861.

Huang, Y., Kuo, M. L. and Wang, T. H. (2007). Pair-perturbation influence functions and local influence in PCA. *Comput. Statist. Data Anal.* **51**, 5886-5899.

Poon, W. Y. (2004). Identifying influential observations in discriminant analysis. *Statistical Methods in Medical Research* **13**, 291-308.

Riani, M. and Atkinson, A. C. (2001). A unified approach to outliers, influence, and transformations in discriminant analysis. *J. Comput. Graph. Statist.* **10**, 513-544.

Rousseeuw, P. J. and Leory, A. (1987). *Robust Regression and Outlier Detection.* Wiley, New York.

Rousseeuw, P. J. and van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *J. Amer. Statist. Assoc.* **85**, 633-651.

Shi, L. (1997). Local influence in principal component analysis. *Biometrika* **84**, 175-186.

Tanaka, Y. (1988). Sensitivity analysis in principal component analysis: influence on the subspace spanned by principal components. *Commun. Statist. A* **17**, 3157-3175.

Department of Mathematics, National Chung Cheng University, 168 University Rd., Min-Hsiung, Chia-Yi 621, Taiwan.

E-mail: yfhuang@math.ccu.edu.tw

Department of Mathematics, National Chung Cheng University, 168 University Rd., Min-Hsiung, Chia-Yi 621, Taiwan.

E-mail: iingcheng@yahoo.com.tw

Department of Mathematics, Baruch College, CUNY, One Bernard Baruch Way, New York, NY10010, USA.

E-mail: tai-ho.wang@baruch.cuny.edu

(Received March 2009; accepted August 2010)