Robust projection indices

Guy P. Nason [1]
University of Bristol, UK

[Dec 28, 2000]

Summary. Loosely speaking a robust projection index is one that prefers projections involving true clusters over projections consisting of a cluster and an outlier. We introduce a mathematical definition of one-dimensional index robustness and describe a numerical experiment to measure it. We design five new indices based on measuring divergence from Student's t distribution which are intended to be especially robust: the experiment shows that they are more robust than several established indices. The experiment also reveals more generally that the robustness of moment indices depends on the number of approximation terms providing additional practical guidance for existing projection pursuit implementations. We investigate the theoretical properties of one new Student's t index and Hall's index and show that the new index automatically adapts its robustness to the degree of outlier contamination. We conclude by outlining the possibilities for extending our experiments both to higher dimensions and other new indices.

Keywords: Exploratory projection pursuit; Divergence from Student's t; Moment index; outlier contamination

1 Introduction

Exploratory projection pursuit (PP) is a technique for finding interesting low p-dimensional projections of high P-dimensional multivariate data, see Jones and Sibson (1987) for an introduction. Typically, PP uses a *projection index*, a functional computed on a projected density (or data set), to measure the "interestingness" of the current projection and then uses a numerical optimizer to move the projection direction to a more interesting position. What do we mean by a *robust* projection index? Generally speaking robust methods are those that perform well even when specific assumptions required for "normal operation" fail to hold or hold approximately. More specifically a robust method performs well in the presence of outliers. The field of robust statistics is vast however good general starting points are Huber (1981) or Hampel, Ronchetti, Rousseeuw and Stahel (1986). For robust approaches in multivariate analysis see Li and Chen (1985) or Ammann (1993) for example.

One aim of exploratory PP is to find clusters in high P-dimensional data. Generally, users may be disappointed when they obtain views that consist of one large cluster separated from one single outlier — we call such views *outlying projections*. In fact, outlying projections can occur all too often as Friedman (1987a) noted that any point can become a "pseudo-outlier" in the projection defined by the line joining the point to the origin.

[1] *Address for correspondence:* Department of Mathematics, University Walk, University of Bristol, Bristol, BS8 1TW, England
E-mail: G.P.Nason@bristol.ac.uk

Specifically Friedman noted that even for P-dimensional standard normally distributed data a significant proportion of points could become pseudo-outliers in their own projection. For $P = 5$ the χ^2_5 distribution shows that 5% of points lie a distance of 3.3 or greater from the origin, for $P = 10$ the distance is 4.3 and $P = 15$ the distance is 5.0. We can envisage practical applications in hyperspectral imaging where $P = 200$ is presently commonplace resulting in 5% of the observations being at a distance of 15.3 or greater. With advances in technology resulting in very large numbers of dimensions to analyse the problem can only get worse as Friedman predicted. On the other hand the detection, explanation and possible removal of real multivariate outliers is obviously important and an essential part of any sensible multivariate analysis which we strongly recommend but do not consider here (for further details see, for example, Hadi (1992) or Rocke and Woodruff (1996)).

Informally, we say that a projection index is more *robust* than another if it tends to select a smaller proportion of outlying projections on the average. This is a rather desirable but impractical definition because in actual implementations the index optimizer has a large influence on the projections that are selected, see Posse (1995) for example. For example, outlying projections may be obtained because a "bad" optimizer consistently finds one of the many sub-optima rather than the index being non-robust. Also, the informal definition is unsatisfactory because it permits many forms of non-robustness: for example, outliers and clusters in many different configurations.

To be precise we narrow our definition to one aspect of robustness and design an experiment to measure it in the next section. Our experiment measures an aspect of robustness that is independent of any optimizer and as such it provides a "test-bed" to evaluate the performance of and compare indices. It is important to stress that our experiment is *not* necessarily the *only* or *best* one. Our design could be extended to indices projecting into more than one dimension but we leave discussion of this possibility until Section 5.

Unless specified otherwise all integrals are from $-\infty$ to ∞. Denote the density and distribution functions of the standard normal by ϕ and Φ respectively.

2 An experiment to assess robustness

The diagram in Figure 1 shows a two-dimensional situation where there are two large "clusters" located on the horizontal axis and symmetric about the vertical axis and a movable outlier that is allowed to slide up the vertical axis from the origin. Now suppose PP was to be performed from $P = 2$ dimensions to $p = 1$ dimension. Clearly, the "ideal" projection of interest to users would be onto the horizontal axis which separates the two large clusters. The least desirable projection is the vertical axis which would give, at best, one large cluster and an outlier.

The projection selected by an index would typically depend on exactly where the outlier was located. If the outlier was located at the origin then ideally the horizontal axis would be chosen. However, as the outlier slides upwards the vertical axis projection becomes much more attractive (based on the value of the index) and eventually exceeds the attractiveness of the horizontal (although a human user would always prefer the horizontal axis projection

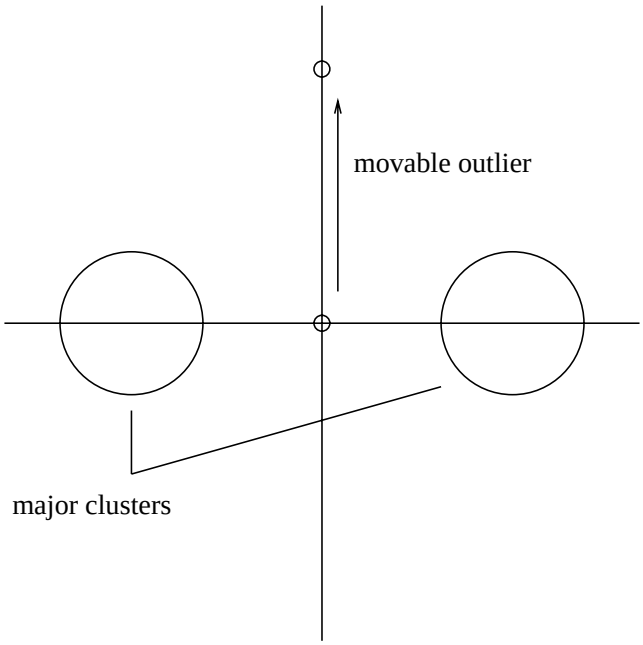

Figure 1: Robustness experiment data distribution

wherever the outlier was located). We name the point at which the vertical axis becomes more attractive to an index as follows.

Definition 1 (Switch point) *Let the location of the movable outlier on the vertical axis be denoted by η. For a given projection index I let the value of I on the horizontal and vertical axes projections be given by $I_H(\eta)$ and $I_V(\eta)$ respectively. The* switch point *is the point $\eta^*(I)$ such that*

$$I_H(\eta) > I_V(\eta) \quad \eta < \eta^*$$
$$I_H(\eta) < I_V(\eta) \quad \eta > \eta^*.$$

Here it is assumed that "interesting" projections are those that maximize the projection index (for those that minimize just change the sign). Note that it is possible that some projection indices may not have a switch point in that the horizontal axis might always be the most interesting projection — in this case we say that the index has a switch point $\eta^* = \infty$ (also, it would be possible to concoct a pathological index that always selected the vertical axis, or had multiple switch points or indeed no switch point at all). Our formal definition of robustness follows.

Definition 2 (Robustness) *Projection index I is more robust than index J if and only if*

$$\eta^*(I) > \eta^*(J).$$

In other words a robust(er) projection index prefers the horizontal axis projection over the vertical axis projection compared to another index as the outlier slides upwards. The reader

might note that our definition is *comparative* rather than absolute. In other words we have not defined what it means for an index to be robust except when comparing it to another index. One might borrow the concept of breakdown from robust statistics and call an index robust if its switch point is $\eta^* = \infty$ for some small $\epsilon > 0$. We develop this concept further in section 4.4 where robustness effectively depends on whether the index $I_V(\eta)$ decreases as η increases for small ϵ but not for large ϵ.

2.1 Experiment Implementation

The assessment of robustness according to definition 2 is performed using a numerical approach. A direct analytical approach would be difficult: some indices may be computed analytically on simple distributions but in general it would be hard, say, to compute them for the vertical projection. The other potential problem is that data are often sphered before applying PP which makes a general analytical solution challenging (sphering involves transforming the data set so it has zero mean and identity variance). In this article we define analytical densities and usually compute projection indices using numerical integration with the `integrate()` function in SPlus. However, for two of the indices which we study in more detail in Section 4.4 Nason (2000) derives explicit formulae on the following distributional setup. Unfortunately, the explicit formulae are too complicated to solve analytically to obtain switch points but they do provide a further check on our numerical results.

The two-dimensional situation illustrated by figure 1 can be represented by the following bivariate density for the random variable pair (X, Y):

$$f_{X,Y}(x, y) = \frac{1-\epsilon}{2}\phi_2(x+\mu, y) + \frac{1-\epsilon}{2}\phi_2(x-\mu, y) + \epsilon\phi_2(x, y-\eta), \qquad (1)$$

where $\phi_2(x, y)$ is the bivariate standard normal density and $x, y \in \mathbb{R}$. The density $f_{X,Y}(x, y)$ models two large clusters on the horizontal axis separated by 2μ with a movable outlier centred on $(0, \eta)$. The degree of outlier contamination in the density is specified by $0 \leq \epsilon \leq 1$ (although, of course, for large ϵ the outlier is no longer outlying).

The marginal density of X is symmetric and hence $\mathbb{E}X = 0$, further $\mathrm{var}(X) = (1-\epsilon)(1+\mu^2) + \epsilon := \tau_X^2(\epsilon, \mu)$. The mean of Y is $\mathbb{E}Y = \epsilon\eta$. The random variables (X, Y) are uncorrelated because:

$$\begin{aligned} \mathrm{cov}(X, Y) &= \mathbb{E}(XY) \\ &= \frac{1-\epsilon}{2}\int x\{\phi(x+\mu) + \phi(x-\mu)\}\,dx \int y\phi(y)\,dy \\ &+ \epsilon\int x\phi(x)\,dx \int y\phi(y-\eta)\,dy = 0. \end{aligned} \qquad (2)$$

To sphere $f_{X,Y}(x, y)$ we first centre: since $\mathbb{E}X = 0$ we only need centre Y by shifting its density down by $\epsilon\eta$ to obtain Y_C with new marginal density

$$f_{Y_C}(y) = (1-\epsilon)\phi(y+\epsilon\eta) + \epsilon\phi(y-(1-\epsilon)\eta)$$

with $\mathrm{var}(Y_C) = (1-\epsilon)(1+\epsilon^2\eta^2) + \epsilon\{1+(1-\epsilon)^2\eta^2\} := \tau_Y^2(\epsilon,\eta)$. Since the random variables are uncorrelated we only need to examine the effect of sphering on the marginals. The marginal densities of the sphered versions X°, Y° are given by

$$f_{X^\circ}(x) = \tau_X f_X(\tau_X x) \text{ and } f_{Y^\circ}(y) = \tau_Y f_{Y_C}(\tau_Y y). \tag{3}$$

In our experiment projection indices I will be computed on $f_{X^\circ}(x)$ and $f_{Y^\circ}(y)$ to obtain $I_X(\eta)$ and $I_Y(\eta)$ respectively. We are interested in finding the value of η when $D(\eta) = I_X(\eta) - I_Y(\eta) = 0$ (the switch point where the Y projection overtakes the X projection in interest). We numerically locate the zero of $D(\eta)$ by using the implementation of Brent's (1973) method found in Press *et al.* (1992).

The robustness experiment results appear in Section 4 after the next section describes some new indices designed to be robust.

3 Student's t is uninteresting!

One established method for dealing with the problem of outliers is to downweight the influence of the tails in a projection index by using a weight function. We adopt a different approach here and instead of searching for departures of the projected density from standard normality we look for departures from a standardized Student's t distribution. Historically, not measuring divergence from the standard normal density is not a crime: Friedman and Tukey's (1974) original index did not; Naito (1997) considers the possibility of measuring divergence from elliptically symmetric distributions in general (including the multivariate t and points out that one of the reasons Friedman (1987b) chose the normal from the class of elliptically symmetric distributions was because of computational tractability.) The heuristic behind the following indices is that Student's t distribution has heavier tails than the standard normal and so departures from it will also depart more frequently from mass in the tails, i.e. outliers. Later sections will verify that these new indices are indeed more robust than existing indices according to the definition given in Section 2. First let us present our uninteresting distribution!

Definition 3 (Multivariate sphered Student's density) *The p-dimensional* sphered Student's t-density *on $\nu \geq 3$ degrees of freedom is defined by $t_{\nu,p} : \mathbb{R}^p \to (0,\infty)$ such that*

$$t_{\nu,p}(\mathbf{x}) = \pi^{-p/2}(\nu-2)^{-p/2}\frac{\Gamma\left\{\frac{1}{2}(\nu+p)\right\}}{\Gamma\left\{\frac{1}{2}\nu\right\}}\left(1+\frac{\mathbf{x}^T\mathbf{x}}{\nu-2}\right)^{-(\nu+p)/2}. \tag{4}$$

Let the distribution function of $t_{\nu,p}$ be denoted by $T_{\nu,p}$.

The multivariate sphered Student's t-density is easily obtained by scaling Cornish's (1954) multivariate Student's t-density by the square-root of the reciprocal of the standard density's variance (which is $\nu/(\nu-2)I$, see Krzanowski and Marriott (1994)).

Next we introduce five new projection indices that measure divergence from the sphered Student's t-distribution.

3.1 Measures of divergence from $t_{\nu,p}$

Our first three indices are all weighted versions of the L^2 divergences from $t_{\nu,p}$ for $\nu \geq 3$. Given a sphered density $f(\mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^p$ our first three indices are given by

$$I^{\text{TL2}}_{\nu,\alpha} = \int \{f(\mathbf{x}) - t_{\nu,p}(\mathbf{x})\}^2 \, t^{\alpha}_{\nu,p}(\mathbf{x}) \, d\mathbf{x} \tag{5}$$

for $\alpha = 0, \frac{1}{2}, 1$. The choice of $\alpha = \frac{1}{2}$ may seem a little perverse at the moment but it is a good one for developing an orthogonal expansion approximation as we will demonstrate shortly. For the distributional setup given in Section 2.1 we have derived explicit formulae for the $\alpha = 0$ index in Nason (2000).

The next index is not derived from an L^2 measure but still measures departures from Student's t. It was initially developed by inverting a calculus of variations problem for finding a functional that could be minimised by $t_{\nu,p}(x)$ over sphered densities (and indeed calculus of variations arguments can be used to prove Theorem 1 although we actually use simpler methods of proof in Appendix A).

Definition 4 (Student's t-index) *For p-dimensional sphered densities f the* Student's t-index *(on $\nu \geq 3$ degrees of freedom) is defined by:*

$$I^{TI}_{\nu}(f) = -\int f(\mathbf{x})^{1-2/(\nu+p)} \, d\mathbf{x}. \tag{6}$$

Unlike the Student's L^2-indices it is not obvious that $I^{\text{TI}}_{\nu}(f)$ is minimized by anything. In fact, the Student's t-index is minimized over all sphered densities by $t_{\nu,p}(\mathbf{x})$ (both on ν degrees of freedom) as shown by the next theorem.

Theorem 1 *Let $C_{\nu,p}$ be the constant depending only on ν and p as defined by (14). The Student's t-index satisfies the following inequality*

$$I^{TI}_{\nu}(f) \geq C_{\nu,p}$$

for all sphered densities f with equality if and only if $f = t_{\nu,p}$ almost everywhere.

The proof of Theorem 1 is in the appendix.

If the above projection indices were used in practice then an estimate of the projected density would be required. Accurately estimating the projected density for each new projection direction in projection pursuit is computationally expensive. As a computationally efficient alternative several workers have circumvented the density estimation step by using index approximations built from orthogonal polynomial expansions (moment indices). It might seem that by using moment indices one also gains by not having to select a smoothing parameter for estimating the projected density. However, the number of terms in a moment index approximation has to be selected and acts as a surrogate smoothing parameter.

We now concentrate on seeking departures from $t_{3,1}(x) = \frac{2}{\pi}(1+x^2)^{-2}$ since the t-distribution on $\nu = 3$ degrees of freedom has the heaviest tails of the series with a finite mean and variance. The results from the experiments in the next section show that the $I^{\text{TL2}}_{\nu,\alpha}$ indices are most robust and of these the one with $\alpha = \frac{1}{2}$ is the most straightforward to develop an orthogonal expansion for. Unlike previous indices in the literature there is no obvious orthogonal *polynomial* expansion however with an appropriate transformation we can supply a natural *Fourier* expansion (as did Morton (1989) in the development of a different projection index.)

3.2 An orthogonal expansion index

Using the transformation $x = \tan(\theta)$ the $I^{\text{TL2}}_{3,\frac{1}{2}}$ index can be written as

$$I^{\text{TL2}}_{3,\frac{1}{2}} = (2\pi^{-1})^{\frac{1}{2}} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \left\{ g_\Theta(\theta) - \frac{2}{\pi}\cos^4(\theta) \right\}^2 d\theta,$$

where g_Θ is the density of the transformed projected data X. Using the Fourier series expansion of g_Θ on $[-\pi/2, \pi/2]$

$$g_\Theta(\theta) = \tfrac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos(2n\theta) + b_n \sin(2n\theta),$$

where

$$a_n = 2\pi^{-1} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} g_\Theta(\theta)\cos(2n\theta)\, d\theta$$

and similarly for b_n. We can write $I^{\text{TL2}}_{3,\frac{1}{2}}$ as

$$I^{\text{TL2}}_{3,\frac{1}{2}} = \sqrt{\pi/2} \left\{ \tfrac{1}{2}(a_0 - \tfrac{3}{2\pi})^2 + (a_1 - \tfrac{1}{\pi})^2 + (a_2 - \tfrac{1}{4\pi})^2 + \sum_{n=3}^{\infty} a_n^2 + \sum_{n=1}^{\infty} b_n^2 \right\}. \qquad (7)$$

An orthogonal expansion approximation to $I^{\text{TL2}}_{3,\frac{1}{2}}$ can be obtained by truncating the infinite sum(s) in (7) to J terms for $J \geq 1$ although one must remember that two terms are included per increment of J. Computation of the index in a sample situation could be easily performed by replacing a_n by the usual empirical estimates: $\hat{a}_n = N^{-1}\sum_{i=1}^{N}\cos(2n\Theta_i)$ where Θ_i, for $i = 1, \ldots, N$ is the transformed projected data and similarly for the b_n.

4 Projection index comparison

4.1 List of projection indices

To be self-contained we list the projection indices that we compare in the next section. To be brief we only provide minimal descriptions of each index. We urge the reader to consult the original references for more details on the motivation behind each index and other associated and important ideas such as, for example, structure removal, rotational invariance, varimax rotation and p-value computation.

The index of Friedman and Tukey (1974) can, as noted by Jones and Sibson (1987), essentially be represented by

$$I^{\text{FT}}(f) = \int f^2(x)\,dx. \tag{8}$$

Jones and Sibson suggested that the Friedman-Tukey index looks for departures from parabolic form rather than specifically look for clusters. Huber (1985) and Jones and Sibson (1987) both suggested the (order-1) entropy measure

$$I^{\text{ENT}}(f) = \int f(x)\log f(x)\,dx, \tag{9}$$

as a suitable projection index that measures departures of f from standard normality. Friedman (1987b) introduced a projection index based upon transforming data to mitigate the effect of outliers. Friedman's index can also be written as measuring a departure from normality:

$$I^{\text{FRI}}(f) = \int \phi(x)^{-1} f(x)^2\,dx.$$

Unfortunately there are some technical problems with this index as noted by Hall (1989, p. 591) who showed that densities that decay slower than $\exp(-x^2/4)$ have infinite I^{FRI} (this is the case for both our densities given in (3) so we do not analyse I^{FRI} in our experiments based on distributions in Section 4.2. However, a moment approximation to I^{FRI} has been found to be a useful projection index so we will analyse it in Section 4.3).

Hall (1989) proposed and analysed the L^2 metric

$$I^{\text{HAL}}(f) = \int \{f(x) - \phi(x)\}^2\,dx,$$

which clearly measures departures from standard normality. Nason (2000) derives explicit formulae for this index on the distributional setup given in Section 2.1. Cook, Buja and Cabrera (1993) developed the transformation ideas of Friedman (1987b) and obtained a family of transformed indices that included both I^{FRI} and I^{HAL} indices as well as their *natural hermite index*

$$I^{\text{NHI}}(f) = \int \phi(x)\{f(x) - \phi(x)\}^2\,dx.$$

Clearly, apart from I^{FT}, all these indices measure departures from standard normality. To

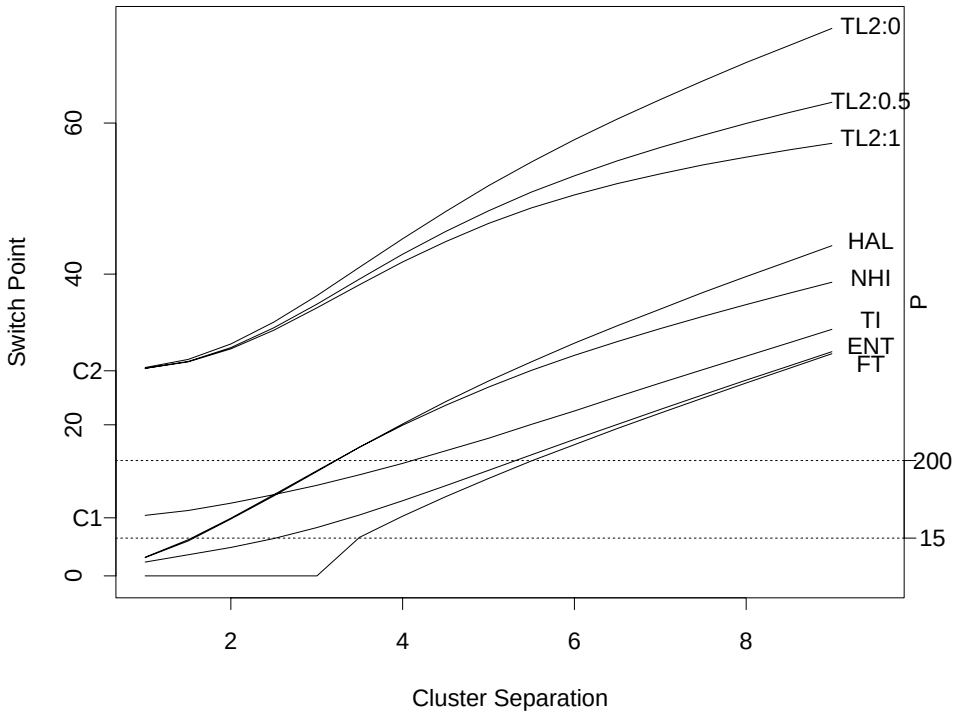

Figure 2: Switch points, η^*, for a range of cluster separations (2μ ranging from 1 to 9 in steps of 0.5) with $\epsilon = 1/201$. All indices tend to 0 as $\mu \to 0$ apart from those based on divergence from Student's t which tend to $C1 \approx 7.72$ or $C2 \approx 27.2$ as marked on the plot.

downweight the influence of outliers I^{NHI} includes the extra $\phi(x)$ which reduces the value of the index when f differs from ϕ in the tails.

We compare the above indices in Section 4.2 by direct numerical integration using the densities specified by (3) and compare four moment based indices in Section 4.3.

4.2 Distributional index comparison

For the results in this section we set $\epsilon = 1/201$ which closely models the two main clusters on the x-axis and the main cluster with outlier on the y-axis as in figure 1. Figure 2 shows the switch points for the projection indices given above for a range of cluster separations (the figure shows the discrete points joined as lines for ease of interpretation).

The horizontal dotted lines in Figure 2 at $P = 15$ and $P = 200$ (marked on the right hand axis) indicate the point, η, at which 5% of the points from a P-dimensional standard normal distribution lie at a distance greater than η, as mentioned in section 1. The way to interpret Figure 2 is to pick a cluster separation, 2μ and then draw an imaginary vertical line. If the switch point line of a projection index is below a horizontal dotted line (e.g. for $P = 15$) then for a reasonable number of points (at least 5%) pseudo-outliers will be preferred to large clusters at the given separation. It can be seen that the t based indices do

well in this *particular* test.

Figure 2 shows that the $I_{3,\alpha}^{\mathrm{TL2}}$ indices are most robust over all computed cluster separations. The next most robust indices are Hall's and the natural Hermite index for cluster separations $2\mu > 2.5$ and (slightly disappointingly) the Student's t index I_3^{TI} otherwise.

The line for the I^{FT} index is a little strange in that the switch points appear to be zero for cluster separations of 1 to 3. This happens because the I^{FT} index decreases on the sphered x-density, $f_{X^\circ}(x)$ for $2\mu = 1$ to $2\mu = 2.0$ and then monotonically increases from $2\mu = 2.5$ but only exceeding the $\mu = 0$ no separation case at $2\mu = 3.5$. In other words for the I^{FT} index rates the zero separation (one cluster) set to be more attractive than two slightly separated clusters which is undesirable behaviour and we conjecture this is due to I^{FT} measuring departures from parabolic form as was mentioned by Jones and Sibson (1987).

It is also interesting to note that for the indices based around L^2 divergence ($I_{3,\alpha}^{\mathrm{TL2}}$, I^{HAL} and I^{NHI}) the $I_{3,0}^{\mathrm{TL2}}$ and I^{HAL} indices are more robust than those that downweight the tails by the normal or Student's t density. *A priori* one might have expected the weighted indices (I^{NHI}, $I_{3,\frac{1}{2}}^{\mathrm{TL2}}$, $I_{3,1}^{\mathrm{TL2}}$) to be more robust.

One might feel a little suspicious at this point as the experimental distributions in (1) are all Gaussian and therefore it is no surprise that the switch points of the "Gaussian departure" indices tend to zero as their separation $\mu \to 0$. The t based indices tend to $C1$ or $C2$ in Figure 2 as $\mu \to 0$. One might argue that it is the Gaussian nature of the experimental distributions that causes the t based indices to appear more robust. To explore this argument we repeated the robustness experiment for Hall's index and the $I_{3,0}^{\mathrm{TL2}}$ index but this time using a distributional configuration similar to that in (1) but using Student's t-distributions on 3 d.f. instead of normal distributions. The $I_{3,0}^{\mathrm{TL2}}$ index was still significantly more robust than Hall's index. Moreover, with this distributional configuration Hall's index preferred a single cluster (y-axis projection with $\eta = 0$) to a slightly separated double cluster (x-axis projection with $\mu < 1.07$ approx). In other words with this "heavy-tailed" experimental configuration the Hall index exhibits potentially undesirable behaviour as it prefers a unimodal heavy tailed distribution rather than a very clear bimodal one. We do not believe the Hall index is unique in this respect and further work would be needed to check the other projection indices.

4.3 Moment index comparison

This section compares the robustness of Friedman's moment based index from Friedman (1987b), Hall's moment index from Hall (1989), the natural Hermite (NHI) moment index from Cook, Buja and Cabrera (1993) and our orthogonal expansion index given by (7). Figures 3 to 6 show perspective plots for their switch points against cluster separation and the number of terms, J, in the index (truncation point). Tables containing the actual values of the switch points that form the basis of the perspective plots can be obtained from web site of the author or the *Journal*.

For large truncation points Friedman's moment index is more robust than the Hall and

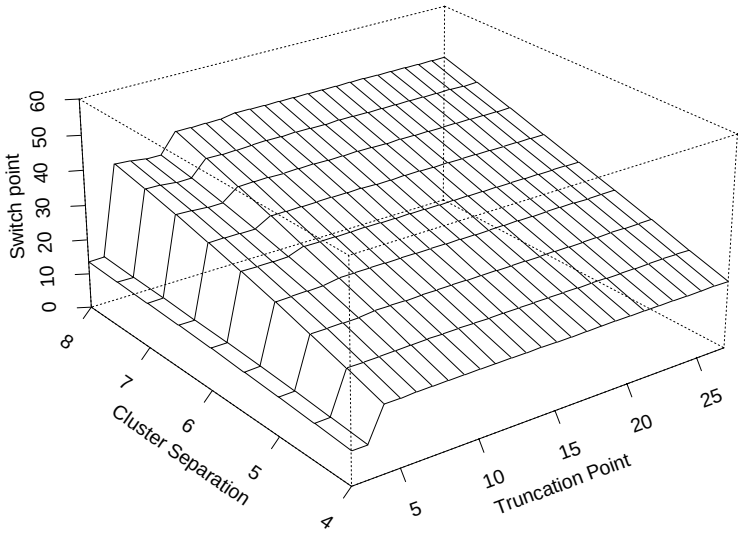

Figure 3: Switch points for Friedman's moment index against cluster separation and truncation point (ranging from 2 to 27).

NHI moment indices for all but the smallest cluster separations. For truncation points of $J = 2, 3$ Friedman's moment index is least robust for all cluster separations, and for $J = 4, 5, 6, 7$ the index is less robust for larger cluster separations when compared to truncation points greater than 7.

Next, figure 4 shows the switch points for Hall's moment index for truncation points ranging from 1 to 40 and the same cluster separations. As the number of terms in Hall's moment index gets large the robustness approaches that of Hall's distributional index (which is to be expected since I^{HAL} is finite, unlike I^{FRI}). However, the robustness is not a monotonically increasing function of the truncation point as there is a slight dip at $J = 7$ terms, and again at around $J = 11$ terms (especially for larger cluster separations). The index is most robust for about $J \geq 28$ terms (although it reaches its peak robustness for $J \geq 8$ terms for a cluster separation of 4).

A perspective plot for the NHI moment index appears in figure 5. In terms of overall robustness the NHI moment index appears to be of the same order of robustness as Hall's moment index (similar to conclusions for the distributional indices above).

For small numbers of terms both Hall's and the NHI moment indices exhibit a pairing effect which causes the "stepping" effect in the perspective plots. Moment indices with 2 and 3 terms, 4 and 5 terms, 6 and 7 terms in pairs exhibit similar robustness values — due no doubt to odd order polynomial terms in the expansions for the densities contributing little. This behaviour was noticed by Cook, Buja and Cabrera (1993).

Finally figure 6 shows the switch points for our Student's t moment index from (7). Remember for this index each increase in truncation point includes two extra Fourier coefficients. It is easily seen that our new moment index is superior in terms of robustness both for small and large cluster separation even if the number of terms in the expansion is

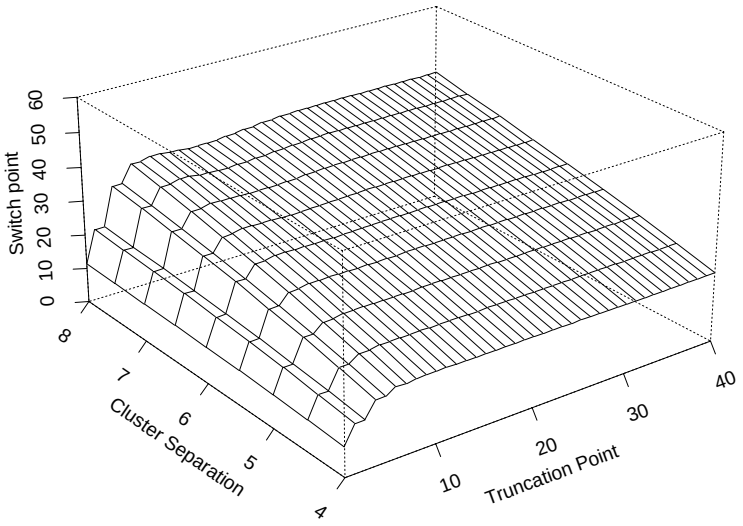

Figure 4: Switch points for Hall's moment index against cluster separation and truncation point (ranging from 1 to 40).

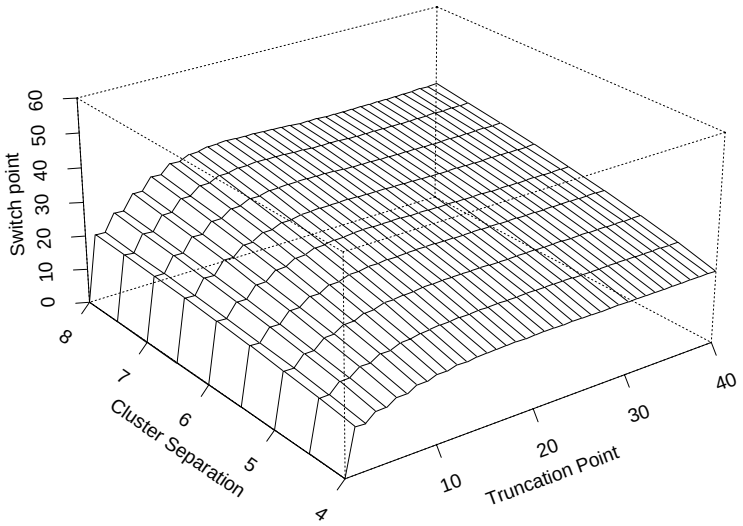

Figure 5: Switch points for natural Hermite moment index against cluster separation and truncation point (ranging from 1 to 40).

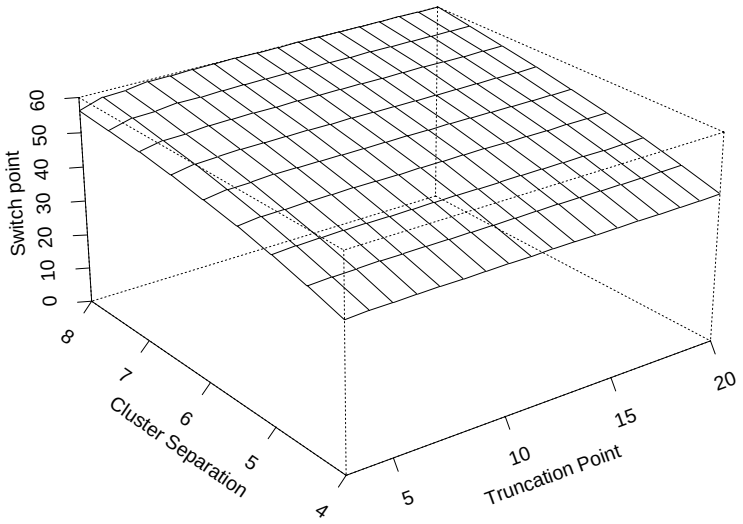

Figure 6: Switch points for Student's t moment index against cluster separation and truncation point (ranging from 3 to 20).

small.

4.4 Student's t indices are affected by degree of outlier contamination

In this section we only consider the Student's t index based on an unweighted L^2 measure: $I^{\mathrm{TL2}}_{\nu,0}$ and compare it to the equivalent L^2 distance from a sphered density to the standard normal (or I^{HAL}, Hall's index). For the following we are only interested in how the projection indices behave on the y-axis projection as a function of the movable outlier location, η, and assume the x-axis projection fixed.

For $\eta = 0$ Hall's index is always zero (since $f_{Y^\circ}(y) = \phi(y)$ in this case) and strictly greater than zero for values of $\eta > 0$. More to the point, in this situation Hall's index can never be negative. However, figure 7 shows that $I^{\mathrm{TL2}}_{\nu,0}$ can be negative for certain values of η. The behaviour in figure 7 is very interesting as it indicates that Hall's index is a monotone increasing function of η for all ϵ (not mathematically proven). However, the behaviour of $I^{\mathrm{TL2}}_{\nu,0}$ *depends* very much on the outlier contamination. Indeed, for small values of ϵ the $I^{\mathrm{TL2}}_{\nu,0}$ index initially decreases as η increases but then the index eventually increases. For small ϵ (when the movable point behaves like an outlier) the $I^{\mathrm{TL2}}_{\nu,0}$ index is robust. For larger ϵ (when the movable point behaves like a true cluster) the $I^{\mathrm{TL2}}_{\nu,0}$ index behaves similarly to Hall's index. *In other words the $I^{TL2}_{\nu,0}$ index repels outliers (decreasing index as a function of η for small ϵ) but is interested in large clusters (increasing index as a function of η for larger ϵ).*

Numerical experiments indicated that the point at which the $I^{\mathrm{TL2}}_{\nu,0}$ switches from being "robust" to "normal" (i.e. when $I^{\mathrm{TL2}}_{\nu,0}$ becomes an increasing function of η) occured at about $\epsilon \approx 0.2113$. Using the explicit projection index formulae given in Nason (2000) and the

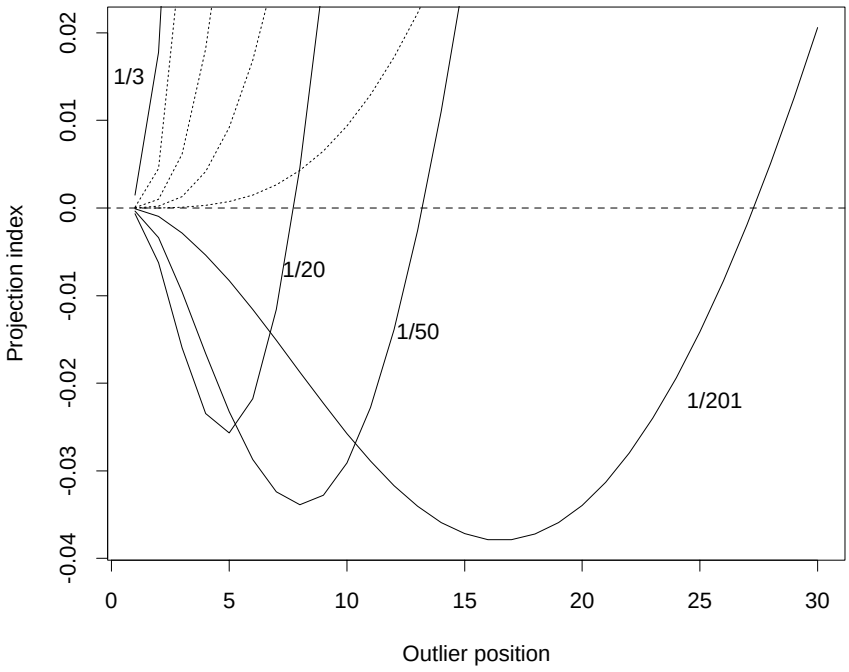

Figure 7: Response of projection index, $I_Y(\eta)$, versus outlier location, η, for two different projection indices and four outlier contaminations, ϵ. The solid lines correspond to the $I^{\mathrm{TL2}}_{\nu,0}$ index and the dotted line to Hall's I^{HAL} index. The outlier contaminations are indicated as fractions for the solid lines, the same contaminations apply to the dotted lines in the same order.

Table 1: Critical contamination point ϵ^* for $I^{\mathrm{TL2}}_{\nu,0}$ index for varying degrees of freedom ν.

ν	ϵ^*	ν	ϵ^*
3	0.2113	160	0.21
4	0.2111	1936	0.1
5	0.2111	4504	0.05
10	0.2108	23962	0.01

computer algebra package MAPLE (`www.maplesoft.com`) we derived a Taylor series of $I^{\mathrm{TL2}}_{\nu,0}$ near $\eta = 0$ and obtained

$$I^{\mathrm{TL2}}_{\nu,0}(f_{Y^\circ}) - I^{\mathrm{TL2}}_{\nu,0}(f_{X^\circ}) = F(\epsilon)\eta^4 + \mathcal{O}(\eta^6)$$

where

$$F(\epsilon) = K\epsilon(\epsilon - 1)(6\epsilon^2 - 6\epsilon + 1) \tag{10}$$

and $K = (96\sqrt{\pi} - 3\,\pi + 64\,\sqrt{2}e^{1/2}\pi\{\sqrt{\pi}\Phi(\sqrt{2}/2)/2 - 1\})/96\pi^{3/2} > 0$. The polynomial in ϵ in (10) governs whether $I^{\mathrm{TL2}}_{\nu,0}$ is increasing or decreasing as η increases from 0. If ϵ is greater than $\epsilon^* = (1 - 1/\sqrt{3})/2 \approx 0.2113$ then the $I^{\mathrm{TL2}}_{\nu,0}$ index behaves like Hall's index however for $\epsilon < \epsilon^*$ the index behaves robustly. (Indeed, more can be said. A referee kindly pointed out that the kurtosis parameter of the normal mixture on the vertical axis is positive if and only if $\epsilon < \epsilon^*$ so the $I^{\mathrm{TL2}}_{\nu,0}$ index is behaves robustly when the *normal* mixture is leptokurtic. Although what happens for other non-normal reference distributions is not known).

It is well-known that as $\nu \to \infty$ the Student's t-distribution tends to the standard normal distribution. Thus, under the same limiting regime we must have $I^{\mathrm{TL2}}_{\nu,0} \to I^{\mathrm{HAL}}$ and consequently the ϵ^* point must decrease as a function of ν (and the index becomes "less" robust) as shown by numerical experiments in Table 1. So, for example, a user of the $I^{\mathrm{TL2}}_{\nu,0}$ index in practice could adapt the behaviour of the index from being very robust to acting like Hall's index with Table 1 indicating the degree of robustness.

5 Conclusions and further work

This article has discussed the concept of comparing projection indices independently of the index optimizers and in particular has introduced one particular aspect of index behaviour: the robustness of an index which measures how sensitive it is to (pseudo-)outliers as compared to clusters. We designed four new projection indices based on measuring divergence from a sphered Student's t-distribution and further developed an orthogonal expansion approximation index for one of them. We designed an experiment to quantify robustness and compared a selection of established indices, our new indices and four moment/approximation indices. Through both numerical calculation and explicit analytical formulae we found that our new Student's t-indices were generally more robust and that indices based on L^2 divergences were also the most robust in their class. A detailed

analytical exploration of one of the new Student's t-indices ($I^{\mathrm{TL}2}_{\nu,0}$) shows that it acts robustly when outliers diverge from a main cluster, but acts like a standard projection index when two clusters diverge: i.e. its behaviour automatically changes depending on the degree of outlier contamination. The degree of outlier sensitivity can be reduced by increasing the degrees of freedom, ν, of the $I^{\mathrm{TL}2}_{\nu,0}$ index to make it behave increasingly like Hall's index as $\nu \to \infty$.

Posse (2000) has suggested that an alternative robust projection index might be constructed by comparing a sphered projected density to a normal density with a variance larger than one (i.e. with heavier tails like the t densities here). Such an index will almost certainly be more analytically tractable than the t-indices given here. Further, Posse has suggested that "squint-angle" plots (i.e. the *angle* at which the outlier becomes more attractive, rather than just x and y projections) might be an additional technique to evaluate the robustness of indices. Both of these ideas are beyond the current scope of this article but reserved for further work.

All of the projection indices in this article have been defined for projected data in p dimensions (apart from the orthogonal expansion index given in Section 3.2. A two-dimensional version of this would not be hard to develop but tedious). However, our robustness experiment, results and further analysis have only been carried out in one dimension. It would be straightforward to perform similar robustness experiments for projections into two and three dimensions. The problem with more dimensions is not that the methodology is hard but the question is which particular configuration should be chosen? There are many ways in which the robustness concept could be generalized to the multi-dimensional situation. For example, the two symmetric clusters in our current experiment could be generalized to three clusters each situated at the apex of an equilateral triangle. The moveable outlier could then initially be positioned at the centre of the triangle in the plane of the triangle and then η could represent the height of the outlier as it moves up the line through the centre of the triangle and perpendicular to the plane of the triangle. A problem with such a configuration might be that there were several interesting "real" views rather than the clear cut distinction between the x and y axes here. There are several different other experiments that one might try and the number of possible "reasonable" experiments that one could devise would increase with p. However, a simple ranking of the established projection indices might not be possible. Although the experiments in this article are one-dimensional we expect that at least some of the multi-dimensional versions would behave somewhat similarly. However, the possibility of interesting structures such as holes in more dimensions means that the experiments here can only be viewed as a first step to evaluating and comparing projection indices and further work in more dimensions would be interesting.

It almost goes without saying that a further line of development would be to actually use some of the indices proposed here in practice as part of a full PP implementation. Robustness experiments could be carried out by measuring the frequency that an index with optimizer finds outlying projections as a proportion of found projections. This too is beyond the scope of the current paper but would be an interesting avenue for further exploration.

Acknowledgments

Part of this article appeared in the unpublished PhD thesis of Nason (1992), helpful conversations with Robin Sibson and a studentship from the Science and Engineering Research Council are gratefully acknowledged. I would also like to thank Christian Posse for extremely useful enthusiastic discussions, very helpful suggestions, and persuading me to move away from my original simulation method to using a direct numerical approach. I would like to thank the referees and Associate Editor for very helpful questions and comments.

A Minimization of the Student's t-index

Here we show that the Student's t-index I_ν^{TI} in (6) is minimized over all sphered densities by the sphered Student's t-density using the theory of F-divergence. We first briefly review the essential theory of F-divergence: a class of dissimilarity measures for densities (or probability measures).

A.1 F-divergence

A formal (measure-theoretic) definition of F-divergence can be found in Vajda (1989). However, this article makes do with the following, slightly simpler, definitions.

Definition 5 ($\mathcal{K}$) *Define $\mathcal{K}$ to be the set of all functions $F : [0, \infty) \to \mathbb{R}$ that are continuous and convex on $[0, \infty)$, finite on $(0, \infty)$, and strictly convex at some point $0 < x < \infty$.*

Definition 6 (F-divergence) *Let f and g be probability density functions defined on $\mathbb{R}^p$. The F-divergence from f to g is given by*

$$\mathcal{F}_F(f|g) = \int f(\mathbf{x}) F\left\{f(\mathbf{x})/g(\mathbf{x})\right\}\, d\mathbf{x},$$

where $F \in \mathcal{K}$.

The reason why F-divergences are useful for designing projection indices is because they satisfy the following inequality due to Csiszár (1967).

Theorem 2 (Csiszár) *For densities f and g, the F-divergence from f to g satisfies*

$$\mathcal{F}_F(f|g) \geq F(1),$$

with equality if and only if $f = g$.

The utility of F-divergence for the design of projection indices becomes apparent when one substitutes (e.g.) the standard normal density ϕ for g in Theorem 2 and then the

F-divergence measures departures from standard normality. The following useful property of F-divergence is used later: define $\tilde{F}$ by

$$\tilde{F}(u) = uF(1/u) \text{ for all } u \in (0, \infty),$$

then Csiszar (1967) shows that

$$F \in \mathcal{K} \text{ if and only if } \tilde{F} \in \mathcal{K},$$

and this leads to the following F-divergence asymmetry property

$$\mathcal{F}_{\tilde{F}}(f|g) = \mathcal{F}_F(g|f). \tag{11}$$

It is very easy to put some well-known projection indices into F-divergence form (for example I^{ENT} with $F(u) = u \log u$, $q = \phi$ and f a sphered projection density).

A.2 Representation of Student's t-index as an F-divergence

Before the main result is proven we demonstrate that the t-index can be written as the sum of two F-divergences after we define the following constants.

Definition 7 (constants) *The following quantities $D^{\star}_{\nu,p}, D_{\nu,p}$ and $C_{\nu,p}$ are all constants dependent only on ν and p defined by*

$$D^{\star}_{\nu,p} = (\nu-2)^{-p/2}\pi^{-p/2}\left[\Gamma\{\tfrac{1}{2}(\nu+p)\}/\Gamma(\tfrac{1}{2}\nu)\right]. \tag{12}$$

Define

$$D_{\nu,p} = {D^{\star}_{\nu,p}}^{-2/(\nu+p)}. \tag{13}$$

Define

$$C_{\nu,p} = -\left(\frac{\nu+p-2}{\nu-2}\right) D_{n,p}. \tag{14}$$

Lemma 1 (F-divergence representation) *The Student's t-index, I^{TI}_{ν}, can be represented as the sum of two F-divergences (multiplied by a constant) as follows*

$$I^{TI}_{\nu}(f) - I^{TI}_{\nu}(t_{\nu,p}) = D_{\nu,p}\left\{\mathcal{F}_{\tilde{F}^*}(f|t_{\nu,p}) + (\nu-2)^{-1}\mathcal{F}_{\tilde{F}^*}(\mathbf{x}^T\mathbf{x}f|\mathbf{x}^T\mathbf{x}t_{\nu,p})\right\}, \tag{15}$$

where f is a sphered density, $D_{\nu,p}$ is a known constant depending only on ν and p and $F^*(u) \in \mathcal{K}$ defined by

$$F^*(u) = 1 - u^{2/(\nu+p)}.$$

Both of the terms on the right-hand side of (15) are F-divergences. Note how the second term measures divergence of $\mathbf{x}^T\mathbf{x}f$ from $\mathbf{x}^T\mathbf{x}t_{\nu,p}$. The extra $\mathbf{x}^T\mathbf{x}$ weight causes I^{TI}_{ν} to be large whenever f differs from $t_{\nu,p}$ in the tails (both $\mathbf{x}^T\mathbf{x}t_{\nu,p}$ and $\mathbf{x}^T\mathbf{x}f$ are densities because $t_{\nu,p}$ and f are sphered). Moreover, as ν increases this term is progressively down-weighted

and its influence wanes. This weighting concurs exactly with our aim of finding a projection index that has a large value for densities that differ from t in the tail.

Proof: (of Lemma 1).

First we establish some notation. Write the form of the sphered t-density (4) as

$$t_{\nu,p}(\mathbf{x}) = D^{\star}_{\nu,p}\left\{1 + (\nu-2)^{-1}\mathbf{x}^T\mathbf{x}\right\}^{-(\nu+p)/2},$$

where $D^{\star}_{\nu,p}$ was defined by (12). It is convenient to note that

$$t_{\nu,p}(\mathbf{x})^{-2/(\nu+p)} = D_{\nu,p}\left\{1 + (\nu-2)^{-1}\mathbf{x}^T\mathbf{x}\right\}, \tag{16}$$

is a simple quadratic form with no linear term and where $D_{\nu,p}$ was defined by (13).

We now move directly on to the representation of the Student's t-index as the sum of two F-divergences. Using definition 4 of $I^{\mathrm{TI}}_{\nu}$the difference we must examine is

$$I^{\mathrm{TI}}_{\nu}(f) - I^{\mathrm{TI}}_{\nu}(t_{\nu,p}) = -\int f f^{-2/(\nu+p)} + \int t_{\nu,p} t_{\nu,p}^{-2/(\nu+p)}.$$

We now introduce two new equal terms to this and obtain

$$\begin{aligned} I^{\mathrm{TI}}_{\nu}(f) - I^{\mathrm{TI}}_{\nu}(t_{\nu,p}) &= -\left\{\int f f^{-2/(\nu+p)} - f t_{\nu,p}^{-2/(\nu+p)} + f t_{\nu,p}^{-2/(\nu+p)} - t_{\nu,p} t_{\nu,p}^{-2/(\nu+p)}\right\} \\ &= -\left[\int f\left\{f^{-2/(\nu+p)} - t_{\nu,p}^{-2/(\nu+p)}\right\} + \int (f - t_{\nu,p}) t_{\nu,p}^{-2/(\nu+p)}\right]. \end{aligned}$$

The second of these integrals is zero because $t_{\nu,p}^{-2/(\nu+p)}$ is a quadratic form with no linear term by (16) and f and $t_{\nu,p}$ are sphered. Therefore

$$\begin{aligned} I^{\mathrm{TI}}_{\nu}(f) - I^{\mathrm{TI}}_{\nu}(t_{\nu,p}) &= \int f t_{\nu,p}^{-2/(\nu+p)}\left\{1 - (t_{\nu,p}/f)^{2/(\nu+p)}\right\} \\ &= \int f D_{\nu,p}\left\{1 + (\nu-2)^{-1}\mathbf{x}^T\mathbf{x}\right\} F^{\star}(t_{\nu,p}/f) d\mathbf{x}, \end{aligned}$$

from (16) and where

$$F^{\star}(u) = 1 - u^{2/(\nu+p)}$$

is a continuous, strictly convex, and finite function on $[0,\infty)$ and hence $F^{\star} \in \mathcal{K}$. Thus using the definition of F-divergence we have

$$I^{\mathrm{TI}}_{\nu}(f) - I^{\mathrm{TI}}_{\nu}(t_{\nu,p}) = D_{\nu,p}\left\{\mathcal{F}_{F^{\star}}(t_{\nu,p}|f) + (\nu-2)^{-1}\mathcal{F}_{F^{\star}}(\mathbf{x}^T\mathbf{x}t_n|\mathbf{x}^T\mathbf{x}f)\right\}. \tag{17}$$

and with $\tilde{F}^{\star}$ defined as in the statement of the lemma and using the asymmetry property (11) of F-divergence we obtain the result we require. □

Proof: (of Theorem 1). From Csiszár's Theorem 2 we know that

$$\mathcal{F}_{F^\star}(\mathbf{x}^T\mathbf{x}t_{\nu,p}|\mathbf{x}^T\mathbf{x}f) \geq \tilde{F^\star}(1) = F^\star(1) = 0,$$

and similarly

$$\mathcal{F}_{F^\star}(t_{\nu,p}|f) \geq 0,$$

with equality in both cases if and only if $f = t_{\nu,p}$ almost everywhere. Therefore using lemma 1 we have

$$I_\nu^{\mathrm{TI}}(f) - I_\nu^{\mathrm{TI}}(t_{\nu,p}) \geq 0$$

with equality if and only if $f = t_{\nu,p}$ almost everywhere. It is easy to show that $I_\nu^{\mathrm{TI}}(t_{\nu,p})$ is equal to $C_{\nu,p}$ defined in (14) and hence the theorem is proved. □

References

Ammann, L.P. (1993) Robust singular value decompositions: a new approach to project pursuit. *J. Am. Statist. Ass.*, **88**, 505–514.

Brent, R.P. (1973) *Algorithms for minimization without derivatives*. Englewood Cliffs: Prentice-Hall.

Cook, D., Buja, A. and Cabrera, J. (1993) Projection pursuit indices based on expansions with orthonormal functions. *J. Comput. Graph. Statist.*, **2**, 225–250.

Cornish, E.A. (1954) The multivariate t-distribution associated with a set of normal sample deviates. *Austral. J. Physics*, **7**, 531–542.

Csiszár, I. (1967) Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, **2**, 299–317.

Friedman, J. H. and Tukey, J. W. (1974) A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.*, **23**, 881–890.

Friedman, J.H. (1987a) Contribution to the discussion Jones and Sibson (1987). *J. R. Statist. Soc. A*, **150**, 26–27.

Friedman, J.H. (1987b) Exploratory projection pursuit. *J. Am. Statist. Ass.*, **82**, 249–266.

Hadi, A.S. (1992) Identifying multiple outliers in multivariate data. *J. R. Statist. Soc. B*, **54**, 761–771.

Hall, P. (1989) On polynomial-based projection indices for exploratory projection pursuit. *Ann. Statist.*, **17**, 589–605.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986) *Robust statistics*. New York: Wiley.

Huber, P. J. (1985) Projection pursuit (with discussion). *Ann. Statist.*, **13**, 435–525.

Huber, P.J. (1981) *Robust statistics*. New York: Wiley.

Jones, M. C. and Sibson, R. (1987) What is projection pursuit? (with discussion). *J. R. Statist. Soc. A*, **150**, 1–36.

Krzanowski, W.J. and Marriott, F.H.C. (1994) *Multivariate analysis*. London: Arnold.

Li, G. and Chen, Z. (1985) Projection-pursuit approach to robut dispersion matrices and principal components: primary theory and Monte Carlo. *J. Am. Statist. Ass.*, **80**, 759–766.

Morton, S.C. (1989) *Interpretable Projection Pursuit*. Ph.D. thesis, Department of Statistics, Stanford University.

Naito, K. (1997) A generalized projection pursuit procedure and its significance level. *Hiroshima Math. J.*, **27**, 513–554.

Nason, G.P. (1992) *Design and choice of projection indices*. Ph.D. thesis, University of Bath, Bath, U.K.

Nason, G.P. (2000) *Analytic formulae for projection indices in a robustness experiment*. Tech. rept. 00:06. Department of Mathematics, University of Bristol.

Posse, C. (1995) Tools for two-dimensional exploratory projection pursuit. *J. Comput. Graph. Statist.*, **4**, 83–100.

Posse, C. (2000) *Personal communication*.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. (1992) *Numerical recipes in C*. Cambridge: Cambridge University Press.

Rocke, D.M. and Woodruff, D.L. (1996) Identification of outliers in multivariate data. *J. Am. Statist. Ass.*, **91**, 1047–1061.

Vajda, I. (1989) *Theory of Statistical Inference and Information*. Theory and Decision Library. Series B: Mathematical and Statistical Methods. Dordrecht, Boston, London.: Kluwer Academic Publishers.