Informative Data Projections: A Framework and Two Examples

Tijl De Bie[1,2], Jefrey Lijffijt[1,2], Raúl Santos-Rodríguez[2], and Bo Kang[1] *

1- Data Science Lab - Ghent University

2- Dept. of Engineering Mathematics - University of Bristol

Abstract. Projection Pursuit aims to facilitate visual exploration of high-dimensional data by identifying interesting low-dimensional projections. A major challenge in Projection Pursuit is the design of a *projection index*—a suitable quality measure to maximise. We introduce a strategy for tackling this problem based on quantifying the amount of information a projection conveys, given a user's *prior beliefs* about the data. The resulting projection index is a subjective quantity, explicitly dependent on the intended user. As an illustration, we developed this principle for two kinds of prior beliefs; the first leads to PCA, the second leads to a novel projection index, which we call t-PCA, that can be regarded as a robust PCA-variant. We demonstrate t-PCA's usefulness in comparative experiments against PCA and FastICA, a popular PP method.

1 Introduction

The analysis of high-dimensional data often starts with dimensionality reduction, to facilitate initial visual exploration by a human user. Most analysts instinctively use Principal Component Analysis (PCA) [1]: it is widely available, computationally efficient, easy to interpret, and in the common situation where the data lies close to a low-dimensional subspace, PCA is effective in retrieving it. However, in user interactions with the PRIM-9 system for interactive data exploration [2], it was observed that users tended to prefer projections that reveal *some form of structure*, rather than *high variance* as preferred by PCA. Later, [3] gave theoretical arguments for why Normally-distributed projections are *least* interesting; they essentially reveal no structure in the data.

Quantifying the extent to which a projection *is* interesting is riddled with conceptual and practical difficulties. To the early Projection Pursuit (PP) research protagonists, it seemed obvious that a universally useful *projection index* (which formalizes the interestingness of a projection) cannot exist (see e.g. [3]). Therefore, lots of different projection indices were introduced. Most of these projection indices quantify the extent to which the projected data's distribution departs from the Normal distribution, and all strike a different balance between practical usefulness, computational complexity, and robustness against outliers (see, e.g. [2, 3]). Indeed, due to the elusive nature of the core question of what makes a projection interesting, the focus shifted towards secondary questions; robustness aspects and computational properties of projection indices.

*This work was supported by the European Union through the ERC Consolidator Grant FORSIED (project reference 615517).

Contributions. Our aim is to return focus on the user once again, and directly ask the question of how interesting a given data projection is *to a particular user*. Our work presents the first generic design strategy for projection indices that explicitly depend on the intended user. In Section 2, we introduce a strategy for quantifying the interestingness of a projection. In Sections 3 and 4, we then apply this strategy for two particular types of prior beliefs, leading to PCA in the first case, and a novel projection index in the second, which we call t-PCA and which could be considered a robust variant of PCA. We end with an empirical comparison that illustrates the benefits of t-PCA as compared to standard PCA and FastICA, a popular PP method also used for ICA [4]. A version of this manuscript with detailed derivations is available as a technical report online[1].

2 The subjective information content of a data projection

Our strategy to quantify the subjective information content of a data projection follows the FORSIED framework [5, 6]. To quantify the information content of a projection, or more generally any pattern, FORSIED relies on the availability of a user's prior belief state in the form of a probability density $p_{\mathbf{X}}$ over the set of possible values for the data $\mathbf{X}$—*in casu* over $\mathbb{R}^{n \times d}$. Given this so-called *background distribution*, one can then compute the marginal probability density function of a data projection $\mathbf{q_w} = \mathbf{Xw}$ defined by the weight vector $\mathbf{w} \in \mathbb{R}^d$.

We call a *projection pattern* a statement of the form $\mathbf{q_w} \in [\hat{\mathbf{X}}\mathbf{w}, \hat{\mathbf{X}}\mathbf{w} + \Delta\mathbf{1})$, specifying that the value $\mathbf{q_w}$ of the projected data lies within an interval of width Δ around $\hat{\mathbf{X}}\mathbf{w}$ (with $\hat{\mathbf{X}} \in \mathbb{R}^{n \times d}$ the empirical data). This is what is conveyed to a user through a scatter plot of the projections $\hat{\mathbf{X}}\mathbf{w}$, with plotting resolution Δ. Writing $p_{\mathbf{Xw}}$ to denote the marginal probability density function of a projection, the smaller the probability $\text{Prob}_{\mathbf{q_w} \sim p_{\mathbf{Xw}}} \Big(\mathbf{q_w} \in [\hat{\mathbf{X}}\mathbf{w}, \hat{\mathbf{X}}\mathbf{w} + \Delta\mathbf{1}) \Big)$, the more surprising and hence informative this pattern would be *to that particular user*. In [5], it is argued that the negative logarithm of this probability is shown to be a good measure of the Subjective Information Content (SIC):

$$\text{SIC}\left(\hat{\mathbf{X}}\mathbf{w}\right) = -\log\left(\text{Prob}_{\mathbf{q_w} \sim p_{\mathbf{Xw}}}\left(\mathbf{q_w} \in [\hat{\mathbf{X}}\mathbf{w}, \hat{\mathbf{X}}\mathbf{w} + \Delta\mathbf{1})\right)\right)$$

This is what we propose as a generic projection index, quantifying the interestingness of a projection.

An important question is how $p_{\mathbf{X}}$ and hence the marginals $p_{\mathbf{Xw}}$ can be obtained, without overburdening the user. [5] suggests that the user is often capable of specifying aspects of their belief state as constraints on expected values of specified statistics of the data. It is then argued that the Maximum Entropy (MaxEnt) distribution subject to these constraints is an attractive choice, given its unbiasedness and robustness, and in being an *exponential family* model, the inference of which is well understood and often computationally tractable.

[1] http://arxiv.org/abs/1511.08762

3 PCA: an information theoretic interpretation

Standard PCA can be derived using the above strategy as follows. A user not expecting any outliers can be assumed capable of expressing an expectation about the value of the average two-norm squared of the data points:

$$\mathbb{E}_{\mathbf{X}\sim p_{\mathbf{X}}}\left\{\frac{1}{n}\sum_{i}^{n}\mathbf{x}_i'\mathbf{x}_i\right\} = \sigma^2. \tag{1}$$

The MaxEnt distribution subject to this constraint is well known and equal to a product distribution of multivariate Normal distributions $\mathcal{N}(\mathbf{0}, \sigma\mathbf{I})$, with one factor for each data point $\mathbf{x}_i$. Given a Normal random vector $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \sigma\mathbf{I})$, a projection onto a weight vector $\mathbf{w}$ with $\mathbf{w}'\mathbf{w} = 1$ is also Normal: $\mathbf{x}'\mathbf{w} \sim \mathcal{N}(0, \sigma)$. Thus, the marginal probability density function $p_{\mathbf{X}\mathbf{w}}$ for the projection $\mathbf{q}_\mathbf{w} = \mathbf{X}\mathbf{w}$ of a dataset $\mathbf{X}$ sampled from the background distribution is given by:

$$p_{\mathbf{X}\mathbf{w}}(\mathbf{q}_\mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}^n}\exp\left(-\frac{\mathbf{q}_\mathbf{w}'\mathbf{q}_\mathbf{w}}{2\sigma^2}\right).$$

We can then compute the SIC of a projection pattern $\mathbf{q}_\mathbf{w} \in [\hat{\mathbf{X}}\mathbf{w}, \hat{\mathbf{X}}\mathbf{w} + \Delta\mathbf{1})$ as minus the logarithm of its probability under this marginal density function $p_{\mathbf{X}\mathbf{w}}$. Noting that for small enough Δ, $\text{Prob}_{\mathbf{q}_\mathbf{w}\sim p_{\mathbf{X}\mathbf{w}}}\left(\mathbf{q}_\mathbf{w} \in [\hat{\mathbf{X}}\mathbf{w}, \hat{\mathbf{X}}\mathbf{w} + \Delta\mathbf{1})\right) \approx \Delta^n \cdot p_{\mathbf{X}\mathbf{w}}(\hat{\mathbf{X}}\mathbf{w})$, this leads to:

$$\begin{aligned}\text{SIC}\left(\hat{\mathbf{X}}\mathbf{w}\right) &= -\log\left(p_{\mathbf{X}\mathbf{w}}(\hat{\mathbf{X}}\mathbf{w})\right) - n\log(\Delta) \\ &= \frac{n}{2}\log(2\pi\sigma^2) - n\log(\Delta) + \frac{1}{2\sigma^2}\mathbf{w}'\hat{\mathbf{X}}'\hat{\mathbf{X}}\mathbf{w}. \end{aligned}\tag{2}$$

Finally, for fixed Δ, maximizing the SIC from Eq. (2) is done by solving:

$$\max_{\mathbf{w}} \mathbf{w}'\hat{\mathbf{X}}'\hat{\mathbf{X}}\mathbf{w}, \text{s. t. } \mathbf{w}'\mathbf{w} = 1,$$

equivalent to the optimization problem for finding the first principal component in PCA. This can be extended straighforwardly to sets of components.

4 t-PCA: for users expecting a heavy tailed distribution

The previous section elucidates the assumptions on the user (prior belief on the average squared norm of the data points) and visualization approach (constant resolution) for PCA to be optimal. We develop an alternative to PCA for when the user's prior beliefs are altered to be more accommodating for outliers. More specifically, we propose the user's prior beliefs could have the following form:

$$\mathbb{E}_{\mathbf{X}\sim p_{\mathbf{X}}}\left\{\frac{1}{n}\sum_{i}^{n}\log\left(1 + \frac{1}{\rho}\mathbf{x}_i'\mathbf{x}_i\right)\right\} = c.$$

That is, rather than specifying an expectation on the spread of the data, for small values of ρ the user specifies an expectation on the *order of magnitude* of the spread. In other words, when the user expects outliers to be present, they may feel able to specify an expectation on the average *order of magnitude* of the 2-norms of the data points, rather than on the average of their 2-norms.

For notational convenience, let us introduce the function $\kappa(\nu) = \psi\left(\frac{\nu+d}{2}\right) - \psi\left(\frac{\nu}{2}\right)$, where ψ is the digamma function. In the sequel, the value of $\kappa^{-1}(c)$ will need to be used, denoted as ν for brevity. Writing Γ for the gamma function, the background distribution can be derived using [7], where it is shown that the MaxEnt distribution subject to the specified prior information is the product of independent multivariate t-distributions with density function $p_{\mathbf{x}}$:

$$p_{\mathbf{x}}(\mathbf{x}) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\sqrt{(\pi\rho)^d}\Gamma\left(\frac{\nu}{2}\right)} \cdot \frac{1}{\left(1 + \frac{1}{\rho}\mathbf{x}'\mathbf{x}\right)^{\frac{\nu+d}{2}}}.$$

Again, for each data point there is a factor in this product distribution.

Note that for $\rho, \nu \to \infty, \frac{\rho}{\nu} \to \sigma^2$ this density function tends to the multivariate Normal density function with mean $\mathbf{0}$ and covariance $\sigma^2\mathbf{I}$. For $\rho = \nu = 1$ it is a multivariate standard Cauchy distribution, which is so heavy-tailed that its mean is undefined and its second moment is infinitely large. Thus, this type of prior belief can model the expectation of outliers to varying degrees.

The marginals of a t-distribution with given correlation matrix are again a t-distribution with the same number of degrees of freedom, obtained by simply selecting the relevant part of the correlation matrix [8, 9]. This means that the marginal density function for the data projections $\mathbf{q}_{\mathbf{w}} = \mathbf{X}\mathbf{w}$ onto a vector $\mathbf{w}$ with $\mathbf{w}'\mathbf{w} = 1$ (and $q_{\mathbf{w},i} \triangleq \mathbf{x}_i'\mathbf{w}$) is:

$$p_{\mathbf{X}\mathbf{w}}(\mathbf{q}_{\mathbf{w}}) = \prod_i p_{\mathbf{x}'\mathbf{w}}(q_{\mathbf{w},i}), \text{where } p_{\mathbf{x}'\mathbf{w}}(q_{\mathbf{w},i}) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\rho}\Gamma\left(\frac{\nu}{2}\right)} \cdot \frac{1}{\left(1 + \frac{1}{\rho}q_{\mathbf{w},i}^2\right)^{\frac{\nu+1}{2}}}.$$

Thus the SIC of the projection pattern $\mathbf{q}_{\mathbf{w}} \in [\hat{\mathbf{X}}\mathbf{w}, \hat{\mathbf{X}}\mathbf{w} + \Delta\mathbf{1})$ is:

$$\text{SIC}\left(\hat{\mathbf{X}}\mathbf{w}\right) = \frac{\nu+1}{2}\sum_{i=1}^{n}\log\left(1 + \frac{1}{\rho}(\hat{\mathbf{x}}_i'\mathbf{w})^2\right) - n\log(\Delta) + \text{a constant}. \tag{3}$$

If we assume again that $\mathbf{\Delta}$ is constant, using $\mathbf{w}'\mathbf{w} = 1$, and ignoring some constant factors and terms, maximising the SIC is thus equivalent to solving:

$$\max_{\mathbf{w}} \sum_{i=1}^{n} \log\left(\rho + (\hat{\mathbf{x}}_i'\mathbf{w})^2\right), \text{s. t. } \mathbf{w}'\mathbf{w} = 1. \tag{4}$$

Clearly, the larger $\mathbf{w}'\mathbf{w}$, the larger the objective, so the constraint can be relaxed to $\mathbf{w}'\mathbf{w} \leq 1$. The optimization problem is more complex, but can be efficiently addressed with standard toolboxes; we used ManOpt (http://www.manopt.org).

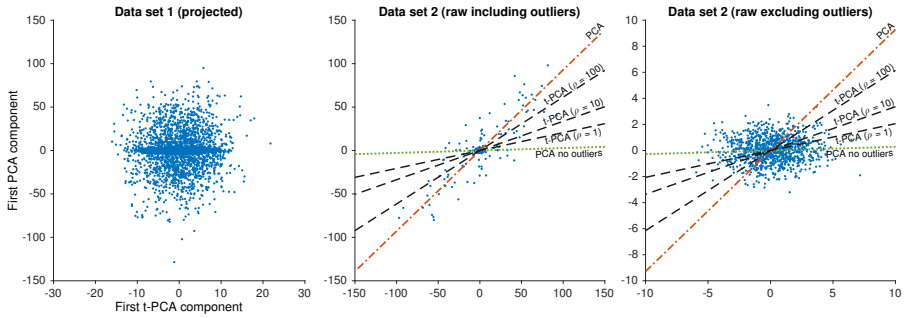

Fig. 1: Left: dominant PCA vs. t-PCA ($\rho = 0$) projections. Middle, right: data in the original space, visualised including and excluding outliers, with weight vectors of PCA, t-PCA ($\rho = 1, 10, 100$), and PCA fitted excluding the outliers.

5 Empirical evaluation

We first compared PCA and t-PCA on synthetic data. We generated a dataset with two populations, both sampled from a 100-dimensional multivariate Normal distribution with diagonal covariance: a population of 8000 points with a small spread, and a population of 2000 points with a large spread. Figure 1 (data set 1, left) shows that for this data the dominant PCA component is determined almost fully by the small population with large spread. In contrast, t-PCA offers an insight into the large population with lower spread as well.

To analyse the robustness of t-PCA, we generated a dataset consisting of two populations with different covariance structures: 1000 data points sampled from $\mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}\right)$, and 100 'outliers' from $\mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 16 & 12 \\ 12 & 13 \end{pmatrix}\right)$. Figure 1 (data set 2, middle and right) shows the weight vectors resulting from PCA, t-PCA, and PCA had there been no outliers. The middle plot shows that the PCA result is determined primarily by the outliers. The right plot shows the same weight vectors on top of a scatter plot without the 100 outliers, illustrating that t-PCA is hardly affected by outliers (the lower ρ, the less it is affected).

We also tested PCA, t-PCA[2], and FastICA[3] on the Shuttle data[4], and a reduced version[5] of the 20 NewsGroups data. Figure 2 shows that in both cases, t-PCA reveals more interesting structure in the data, although for 20 NewsGroups the structure is somewhat similar to PCA and does not appear to separate the classes. The FastICA projection for 20 NewsGroups is practically useless, because the weight vectors have all mass on a single dimension. The t-PCA weight vectors are also sparse, yet t-PCA gives the largest spread of data points in both visualizations.

[2] $\rho = 10^{-5}$ multiplied by a measure of the scale of the data equal to the square root of the average squared norm of all data points.

[3] With default parameters.

[4] `https://archive.ics.uci.edu/ml/datasets/Statlog+(Shuttle)`

[5] http://cs.nyu.edu/∼roweis/data.html

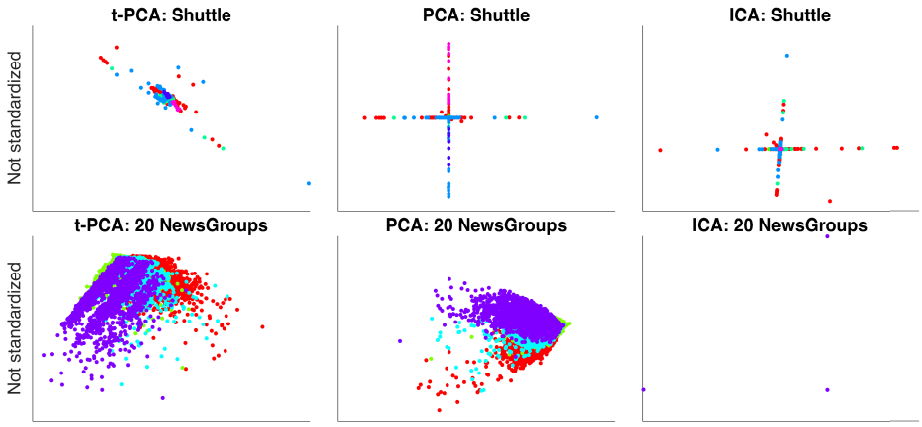

Fig. 2: Top two-component projections by t-PCA, PCA, and FastICA.

6 Conclusions

PCA is often a suboptimal choice for dimensionality reduction, e.g. in the presence of outliers. The Projection Pursuit literature addressed this by means of the introduction of numerous *projection indices* that quantify the interestingness of projections in different ways. Alternatively, various authors proposed principled *robust* versions of PCA recently, e.g. [10]. These lines of work are useful when the assumptions made are valid, but they do not address fundamentally how interesting a data projection is to a user. We introduced an approach to this elusive problem, explicitly recognizing the subjective nature of 'interestingness'.

References

[1] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[2] Jerome H Friedman and John W Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Tr. Comp.*, 100(23), 1974.

[3] Peter J Huber. Projection pursuit. *Ann. Stat.*, 13(2):435–475, 1985.

[4] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*. John Wiley & Sons, 2001.

[5] Tijl De Bie. An information theoretic framework for data mining. In *Proc. of KDD*, pages 564–572. ACM, 2011.

[6] Tijl De Bie. Subjective interestingness in exploratory data mining. In *Proc. of IDA*, pages 19–31. Springer, 2013.

[7] Konstantinos Zografos. On maximum entropy characterization of pearson's type II and VII multivariate distributions. *J. Multiv. Anal.*, 71(1):67–75, 1999.

[8] Samuel Kotz and Saralees Nadarajah. *Multivariate t distributions and their applications*. Cambridge University Press, 2004.

[9] Michael Roth. On the multivariate t distribution. Technical Report LiTH-ISY-R-3059, Department of Electrical Engineering, Linköping universitet, 2013.

[10] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3), 2011.