Appendix to " New Metrics for Assessing Projection Pursuit Indexes, and Guiding Optimisation Choices"

H. Sherry Zhang, Dianne Cook, Nicolas Langrené, Jessica Wai Yin Leung

2024-09-11

Given high-dimensional data $X \in \mathbf{R}^{n \times p}$ and the index function $f(\cdot)$, projection pursuit finds the orthonormal projection basis $A \in \mathbf{R}^{p \times d}$ by solving the following optimisation problem:

$$\max_{A} \quad f(XA) \quad \text{subject to} \quad A'A = I \tag{1}$$

This appendix presents definition of all indexes $f(\cdot)$ used in the paper. These indexes are defined in 2D cases ($d = 2$), where $Y \in \mathbf{R}^{n \times 2} = (y_1, y_2)$ represents the projected data.

The `holes` index

The `holes` index (Cook, Buja, and Cabrera 1993) is a smooth index for detecting the presence of multi-modality in the projection. The index is defined as:

$$I_{\text{holes}}(A) = \frac{1 - \frac{1}{n}\sum_{i=1}^{n} \exp(-\frac{1}{2y_i y_i'})}{1 - \exp(-\frac{d}{2})}$$

The `dcor2d` index

The `dcor2d` index (Grimm 2016) is a smooth index that measures the distance correlation between the two projection axes. The index uses pair-wise distance to define column sum, row sum, and overall sum and then calculates distance correlation from distance variance and covariance.

- pair-wised distance: $a_{ij} = \|y_1^i - y_1^j\|$ and $b_{ij} = \|y_2^i - y_2^j\|$ for $i, j = 1, 2, \dots, n$

- column sum, row sum, and overall sum:

$$a_{i\cdot} = \sum_{l=1}^{n} a_{il}, \quad a_{\cdot j} = \sum_{k=1}^{n} a_{kj}, \quad a_{\cdot\cdot} = \sum_{k,l=1}^{n} a_{kl}$$

$$b_{i\cdot} = \sum_{l=1}^{n} b_{il}, \quad b_{\cdot j} = \sum_{k=1}^{n} b_{kj}, \quad b_{\cdot\cdot} = \sum_{k,l=1}^{n} b_{kl}$$

- distance covariance (and variance defined similarly):

$$\mathrm{dCov}(a_{ij}, b_{ij}) = \frac{1}{n(n-3)} \sum_{i \neq j} a_{ij} b_{ij} - \frac{2 \sum_{i=1}^{n} a_{i\cdot} b_{i\cdot}}{n(n-2)(n-3)} + \frac{a_{\cdot\cdot} b_{\cdot\cdot}}{n(n-1)(n-2)(n-3)}$$

- distance correlation:

$$I_{\mathrm{dcor}}(A) = \mathrm{dCor}(a_{ij}, b_{ij}) = \frac{\mathrm{dCov}(a_{ij}, b_{ij})}{\sqrt{\mathrm{dVar}(a_{ij}) \mathrm{dVar}(b_{ij})}}$$

The `MIC` and `TIC` index

Both `MIC` and `TIC` are information-based indexes derived from mutual information. The indexes are defined based on a partition, $g(k,l)$, of the space (y_1, y_2) into $k \times l$ rectangles, For example, $g(2,3)$ divides the data space into 2 rectangles in the y_2 direction and 3 rectangles in the y_1 direction. Let G denotes all the possible partition.

MIC finds the **maximum** mutual information over G where k and l are bounded by the grid size: $k \times l < B(n) = n^{\alpha}$. $alpha = 0.3$ is used in our simulation.

$$I_{MIC}(A) = \max_{g \in G} \frac{I(y_1, y_2 | g)}{\log(\min(k^*, l^*))} = \max_{g \in G} \frac{\sum_{y1} \sum_{y2} P(y_1, y_2) \log \frac{P(y_1)}{P(y_1)P(y_2)}}{\log(\min(k^*, l^*))}$$

where k^* and l^* are the number of rectangles in the optimal partition g.

TIC calculates the **sum** of mutual information over G

$$I_{TIC}(A) = \sum_{g \in G} \frac{I(y_1, y_2 | g)}{\log(\min(k^*, l^*))} = \sum_{g \in G} \frac{\sum_{y1} \sum_{y2} P(y_1, y_2) \log \frac{P(y_1)}{P(y_1)P(y_2)}}{\log(\min(k^*, l^*))}$$

The `loess` and `splines` index

The `loess` and `splines` indexes detect non-linear structure in the projection, as captured by their respective model. These indexes are computed by regressing both axes against each other, and find the maximum variance from the residual that can be explained by the model. Let $e^{\text{model}}_{y_1 \sim y_2}$ denote the residual from the loess/splines model $y_1 \sim y_2$, and $e^{\text{model}}_{y_2 \sim y_1}$ denote the residual from the loess/spliens model $y_2 \sim y_1$. The two indexes are calculated as

$$I_{\text{loess}}(A) = \max\left(1 - \frac{var(e^{\text{loess}}_{y_1 \sim y_2})}{var(y_1)}, 1 - \frac{var(e^{\text{loess}}_{y_2 \sim y_1})}{var(y_2)}\right)$$

$$I_{\text{spline}}(A) = \max\left(1 - \frac{var(e^{\text{spline}}_{y_1 \sim y_2})}{var(y_1)}, 1 - \frac{var(e^{\text{spline}}_{y_2 \sim y_1})}{var(y_2)}\right)$$

The `stringy` index

The `stringy` (L. Wilkinson, Anand, and Grossman 2005; Leland Wilkinson and Wills 2008) index is a non-smooth index based on scagnostics. It measures the proportion of vertices with two edges in the minimum spanning tree (MST) of the projection to detect whether the projection forms a straight line. The index is calculated as

$$I_{stringy}(A) = \frac{\text{number of vertices with 2 edges}}{\text{number of total vertices with more than one edge}}$$

Reference

Cook, D., A. Buja, and J. Cabrera. 1993. "Projection Pursuit Indexes Based on Orthonormal Function Expansions." *Journal of Computational and Graphical Statistics* 2 (3): 225–50. https://doi.org/10.2307/1390644.

Grimm, Katrin. 2016. "Kennzahlenbasierte Grafikauswahl." Doctoral thesis, Universität Augsburg.

Wilkinson, L., A. Anand, and R. Grossman. 2005. "Graph-Theoretic Scagnostics." In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, 157–64. https://doi.org/10.1109/INFVIS.2005.1532142.

Wilkinson, Leland, and Graham Wills. 2008. "Scagnostics Distributions." *Journal of Computational and Graphical Statistics* 17 (2): 473–91. https://doi.org/10.1198/106186008X320465.