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

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

2025-02-05

Let $Y = XA$ denote the projection of a p-dimensional dataset, $X \in \mathbb{R}^{n\times p}$, onto d-dimensional space, $Y \in \mathbb{R}^{n\times d}$, where the projection is defined by an orthonormal matrix $A \in \mathbb{R}^{p\times d}$. Projection pursuit aims to find the matrix A that maximizes an index function $f(\cdot)$, which measures interesting features of the projection, such as multi-modality, linear or non-linear relationship:

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

This appendix defines all indexes $f(\cdot)$ used in the paper, namely holes, MIC, TIC, dcor, loess, skinny, splines, and stringy). Since all the indexes are defined for 2D projections ($d = 2$), the projected data can also be written as $Y = (y_1, y_2)$.

The holes index

The holes index (Cook, Buja, and Cabrera 1993) defines as follows detects the presence of multi-modality in the projection:

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

The dcor index

The dcor index (Grimm 2016), sometimes called dcor2D, measures the **d**istance **cor**relation between the two projection axes (y_1, y_2). 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 $i, j = 1, 2, \ldots, n$
- 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:

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

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.

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

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

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

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

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 skinny and stringy index

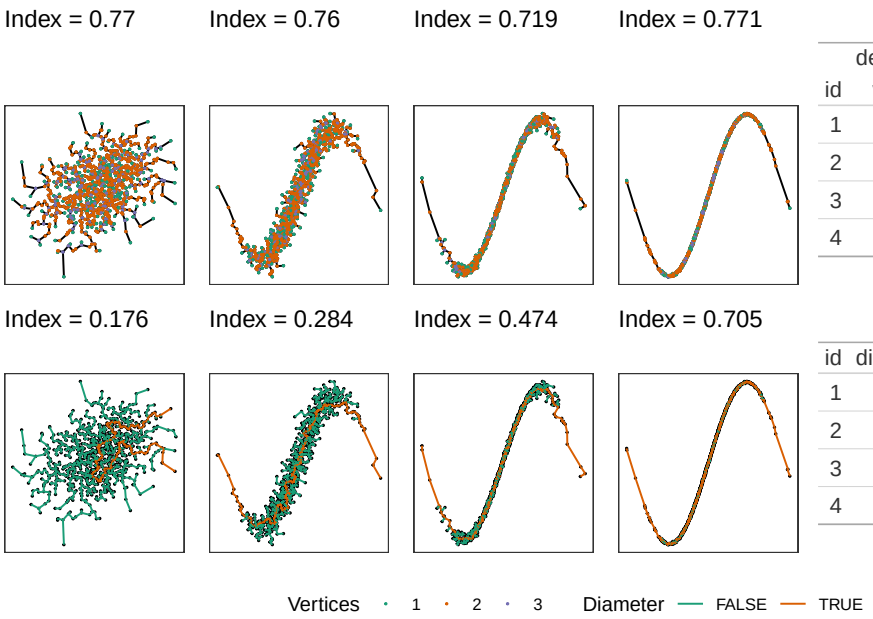
Both skinny and stringy indexes are non-smooth indexes based on scagnostics (Leland Wilkinson and Wills 2008; L. Wilkinson, Anand, and Grossman 2005). Scagnostics are defined based on graph concepts, specifically, the alpha hull for the skinny index and the minimum spanning tree for the stringy index. The skinny index is defined based on the alpha hull (AH) as:

$$I_{\text{skinny}} = 1 - \frac{\sqrt{4\pi \text{area(AH)}}}{\text{perimeter(AH)}}$$

The stringy index is defined based on the minimal spanning tree (T). We adopt the definition in L. Wilkinson, Anand, and Grossman (2005) based on the diameter and length of the tree. The diameter refers to the longest shortest path through T, and the length refers to the total number of edges in T.

$$I_{\text{stringy}} = \frac{\text{diameter}(T)}{\text{length}(T)}$$

An alternative definition of the stringy index is based on number of vertices, given by $\frac{|V^{(2)}|}{|V|-|V^{(1)}|}$, where $|V^{(2)}|$ represents the number of degree-2 vertices and $|V| - |V^{(1)}|$ is the overall number of vertices minus the number of degree-1 vertices. However, we prefer the original definition based on the diameter and length of the minimal spanning tree, as it is more robust to noise, especially as the number of observations increases. In Figure 1, the stringy index defined based on the number of vertices (row 1) fails differentiate the non-linear sine wave from the non-pattern (left to right), while the diameter/length-based definition (row 2) does.



id	degree 2 vertices	degree 2+ vertices	stringy
1	620	806	0.769
2	620	805	0.770
3	554	770	0.719
4	607	798	0.761

id	diameter	length	stringy
1	176	1000	0.176
2	284	1000	0.284
3	474	1000	0.474
4	705	1000	0.705

Figure 1: Projections and the stringy index values calculated based on the number of vertices definition (row 1) and the diameter/length definition (row 2) for a sine wave pattern with 1000 observations. The color represents the number of vertices in row 1 and highlighs the longest minimal path (the points forming the diameter) in row 2. The noise around the sine wave creates local structures in the minimal spanning tree, resulting in a high number of degree-2 and higher vertices when calculated based on the number of vertices. This causes stringy index based on the number of vertices fails to differentiate the non-linear sine wave from the non-pattern, while the diameter/length definition, which captures the dominant main structure, is able to differentiate the sine wave form the non-pattern.

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.