Noname manuscript No.
(will be inserted by the editor)

Using tours to visually investigate properties of new projection pursuit indexes with application to problems in physics – Appendix

Ursula Laa · **Dianne Cook** ·

Received: date / Accepted: date

Abstract Appendix material for the paper "Using tours to visually investigate properties of new projection pursuit indexes with application to problems in physics".

Keywords

Appendix

A Re-scaling of holes index

Holes and cmass indexes are derived from I_0^N of (Cook, Buja, and Cabrera 1993). As noted in proposition 1 of that paper the index takes local maxima for the minimum and maximum of a_0 which are achieved by central hole and central mass distributions respectively. Cook and Swayne (2007) then gives explicit index functions defined for sphered data (zero mean, identity variance-covariance matrix). They are defined such that each one is maximized for central holes or central mass type distributions, with maximum=1, and cmass=1-holes. It follows that for either index both large and small values signal deviation from the normal distribution, and given a normal distribution we expect to find "average" index values rather than values close to zero.

We can estimate the values found for normal distribution by comparing values of a_{00} of Sec 7.1 in (Cook, Buja, and Cabrera 1993). The maximum value is $1/(2\pi)$ found for cmass type distributions, the minimum is $1/(2\pi e)$ found for hole type distributions. Evaluating for normal distributions gives $1/(4\pi)$, rescaling such that the index values range from 0 to 1 then puts the value for normal distributions at approximately 0.2. This is consistent with

Ursula Laa
School of Physics and Astronomy, Monash University
E-mail: ursula.laa@monash.edu

Dianne Cook
Department of Econometrics and Business Statistics, Monash University
E-mail: dicook@monash.edu

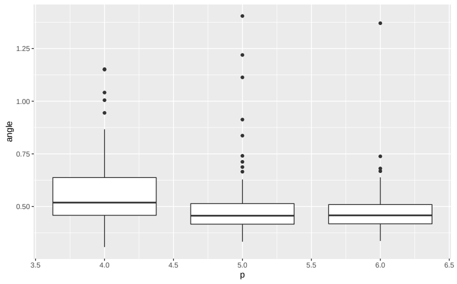

Fig. 1 Estimated squint angles for the Spiral dataset with 1000 datapoints, with p = 4, 5, 6, containing estimates evaluated for 100 randomly selected directions each.

the results we found, i.e. for normal distributions the cmass index is about 0.2, and the holes index (=1-cmass) is about 0.8.

We therefore rescale as follows: first have cut-off at the respective value for the normal distribution, i.e. any value below 0.2 (0.8) is set to zero for cmass (holes) index, and we rescale the remaining range to be between zero and one.

B Estimating the squint angle of the spiral

We estimate the squint angle of the spiral (for p = 4, 5, 6) as follows. First pick a random starting plane and generate a tour path from the starting plane to the ideal plane containing the spiral. Using skinny as the reference index we fix a lower index value that is attributed to indicate squintability at 0.6, and we move along the tour path towards the ideal plane until this value is reached. The distance between the thus identified plane and the ideal plane is used as an estimate of the squint angle in this direction. Since this will strongly depend on the random starting plane, i.e. the considered direction, we repeat the estimation 100 times and present the results in the form of a box plot in Figure 1. The result shows a large drop in squint angle when going from p = 4 to higher dimensions, and generally a large spread of squint angles depending on the direction.

C Computational performance

Computational time is important for using the PPIs with the guided tour, online. Figure 2 summarizes performance for each PPI. For simplicity, data with sample sizes ranging from 100-10000 are drawn from a 6-d solid sphere, using the geozoo package (Schloerke 2016). The time to compute the PPIs over 100 interpolated grand tour projections is recorded. The scagnostics PPI are computed as a bundle, since this is the code base, and that major computational constraint is common to all the scagnostics. There are two versions of the MIC and TIC algorithm, labelled MINE and MINE E, the second being a newer algorithm which improves their computational performance.

The results are interesting. The scagnostic indexes and splines2d are very fast regardless of sample size. MIC, TIC (both versions) and dcor2d slow rapidly as sample size increases.

D Effect of parameter choice in index value

Some parameters must be provided for some PPIs. This can be advantageous, allowing the index to more flexibly work for different types of structure, controlling trade-offs between noise and fine structure detection, and affecting computing time and precision.

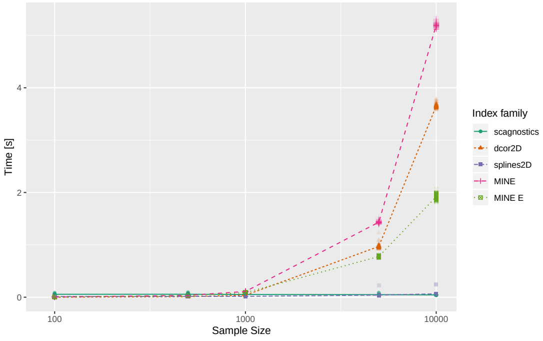

Fig. 2 Computational perfoamce for PPIs, using sample sizes 100-10000. Colour indicates the PPI. Because the scagnostics calculation is bundled together, the values are the same for all these indexes, and they are really fast to compute. MINE includes the MIC and TIC indexes, and MINE E are computationally more efficient algorithms for these. These, along with dcor2D, are slower with larger sample sizes.

- Binning:
  - Scagnostics: the number of bins can be controlled by the user, note however that internally the implementation will reduce the number of bins if too many non-empty bins are found (more than 250).
  - MINE: the maximum number of bins considered is fixed by the user as a function of the number of data points. The default is chosen as a trade-off between resolution and noise dependence, but it may be tuned based on requirements dictated by specific datasets. Apart from sensitivity to noise computing time may also be a consideration here.
- Spline knots: for the splines2d measure we need to fix the number of knots. By default it is fixed to be 10 (or lower if appropriate based on the data values). In our examples we find the number to be appropriate to identify functional dependence while rejecting noise, but some distributions may require tuning of this parameter.

The bins argument for the scagnostics might be reasonably expected to affect the smoothness of the index: a small number of bins should provide a smooth index function, but may affect its ability to detect fine structure. Figure 3 examines this. Scagnostics PPIs are computed for the spiral1000 data on a tour path between x_1-x_2 to x_5-x_6, for number of bins equal to 10, 20, 50. The interesting observation to make is that even with small bin size the functions are all relatively noisy. The problem with the small bin size is that the spiral becomes invisible to the PPI.

D.1 Binning sensitivity of MIC index

To examine the sensitivity of binning in the MIC PPI, the classic RANDU data (Marsaglia 1968), available in R, is used. Binning is defined by δ, where $B(n) = n^{\delta}$. Figure 4 shows the best projection, index value and computing time obtained when optimizing the MIC index with values $\delta = 0.6, 0.7, 0.8$. With small δ, less bins, the structure isn't visible, and with larger δ the structure is confounded with noise. It does appear that this parameter affects the performance of the MIC index.

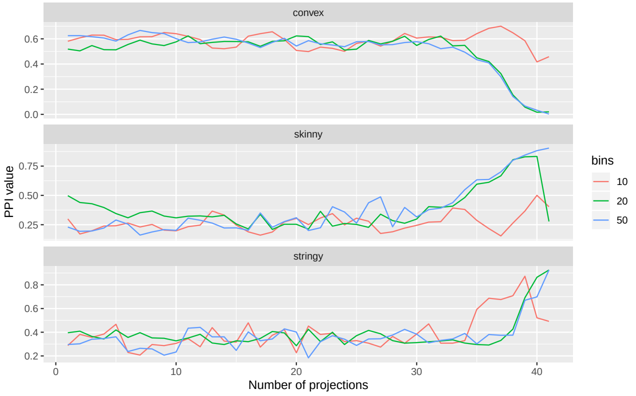

Fig. 3 Comparing the traces of the three scagnostics indices when changing the binning via the bins parameter set to 10, 20 and 50 in this example.

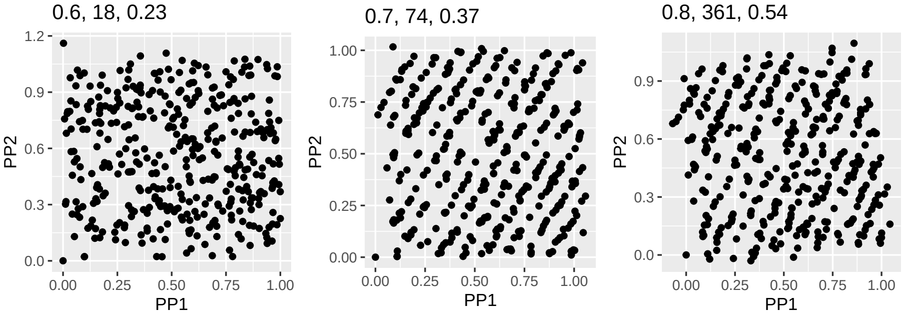

Fig. 4 Best projection obtained by optimizing the MIC index on the RANDU data, using different number of bins, defined by δ. The smaller the value the fewer bins. Above each plot is written the value of δ, time required to optimize (seconds) and the MIC index value. The best $\delta = 0.7$, and the result indicates that this parameter does affect MIC performance.

E Ways to refine the PPIs

The biggest issues revealed by the investigation into the new PPIs are a lack of smoothness, particularly for the scagnostics indexes, and the rotation invariance of Grimm's indexes. To fix the smoothness of an index function, it is possible to calculate the PPI for a small neighborhood of projections and average the value, or alternatively average the PPI for several jittered projections. This is investigated in Fig. 5. Rotation invariance is more difficult to fix, but an alternative tour interpolation method could be useful. The geodesic interpolation transitions between planes, and it ignores the basis defining the plane, creating a problem with rotation invariant indices. Alternative interpolations based on Givens or Householder rotations could be implemented to transition between bases, which should alleviate the need for rotationally invariant indices.

Two different methods are considered for smoothing the index values:

- Jittering points: using the jitter function we move each point by a random amount drawn from a uniform distribution between $\pm\beta$.
- Jittering angles: using the tourr implementation we can draw a random plane and move some small amount ϵ in that direction.

The mean value from a sample of projections is recorded as the PPI value. This could be robsitufied by dropping the most extreme values.

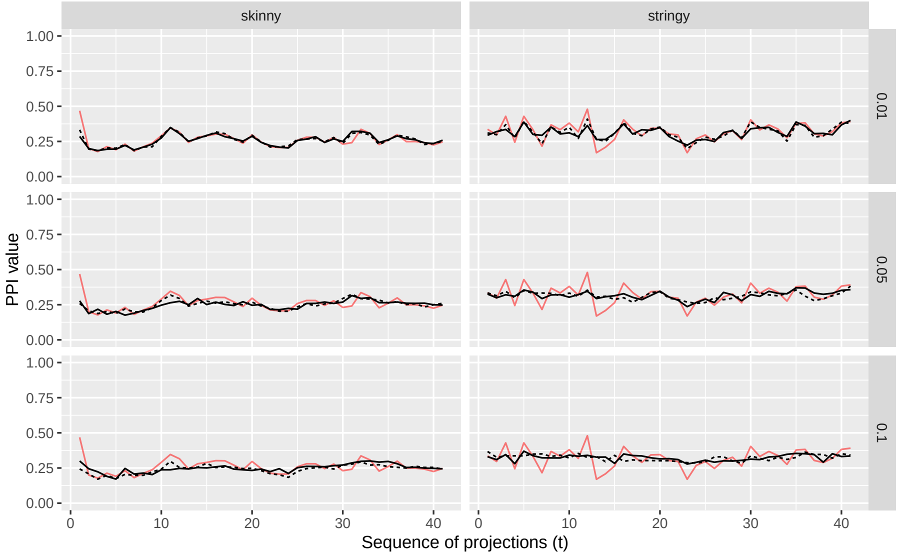

Fig. 5 Comparing the traces of the scagnostics indexes skinny and stringy when smoothing the index values, either by averaging over the index value after jittering the projection by some angle ϵ (full line) or after jittering the projected datapoints with some amount β (dashed line). For comparison the red line in the background shows the trace without any smoothing applied.

This is particularly interesting for the scagnostics indexes skinny and stringy which we found to be most noisy among the indexes considered. Figure 5 studies the potential of these two smoothing approaches, using the tour path between noise variables of the spiral1000 dataset and different ϵ and β values. Both methods appear to be promising in smoothing the function. Because the scagnostics are fast to compute, either of these methods is feasible. For this example we have smoothed over 10 randomly selected jittered views, computing time increases linearly with the number of randomly jittered views, as this is mostly determined by the time needed to evaluate the scagnostics indexes which is done separately for each view.

F Additional Figures

F.1 Final projection of pipe guided tour

Figure 6 shows the projection returned during the scouting phase (left) and the refinement phase (right). It was important to start the method 3 optimizer at the best projection returned during the scouting phase, to smoothly converge more closely to the ideal projection.

F.2 Dataset overview scatter plot matrices

Figures 7 and 8 show the scatter plot matrices of the gravitational wave datasets considered, see Section **??** for details.

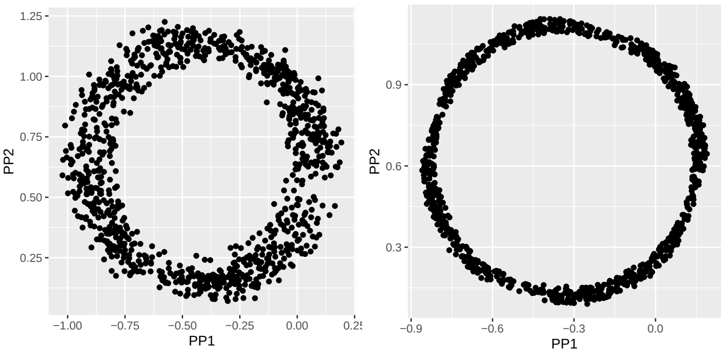

Fig. 6 Projections returned by TIC optimization: by the scouting phase (left) and refined by optimization method 3 (right), starting from the scouting phase projection.

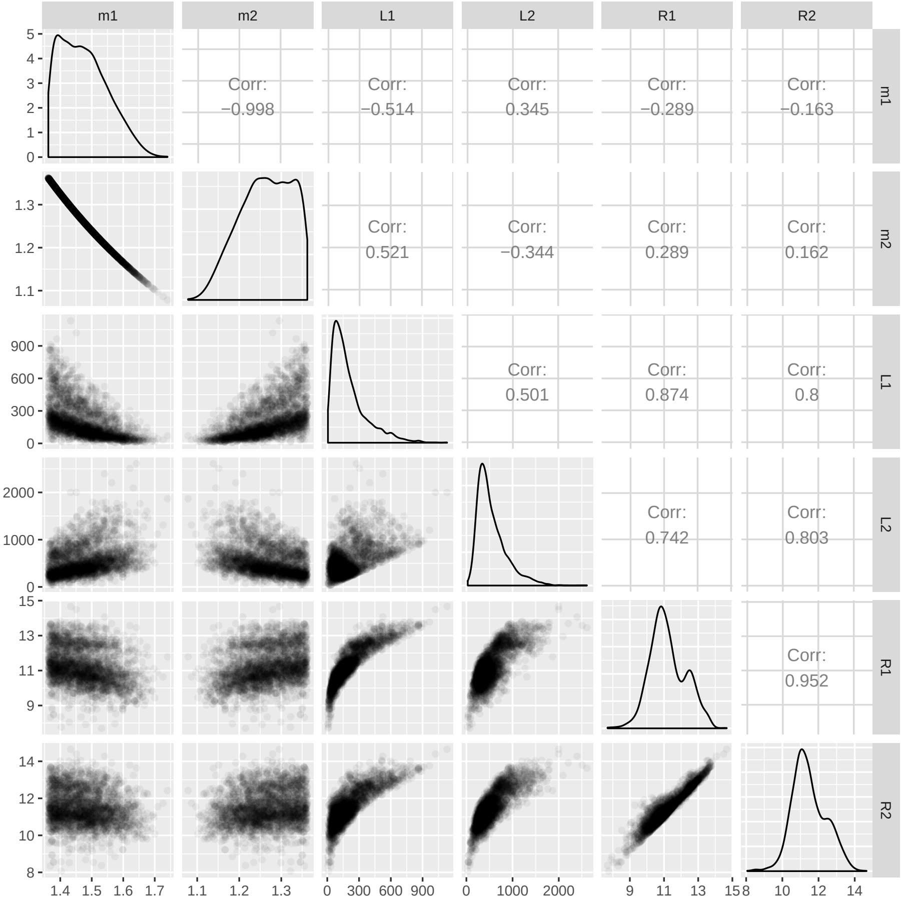

Fig. 7 Scatter plot matrix of the neutron star dataset, darker regions represent higher marginalised posterior densities.

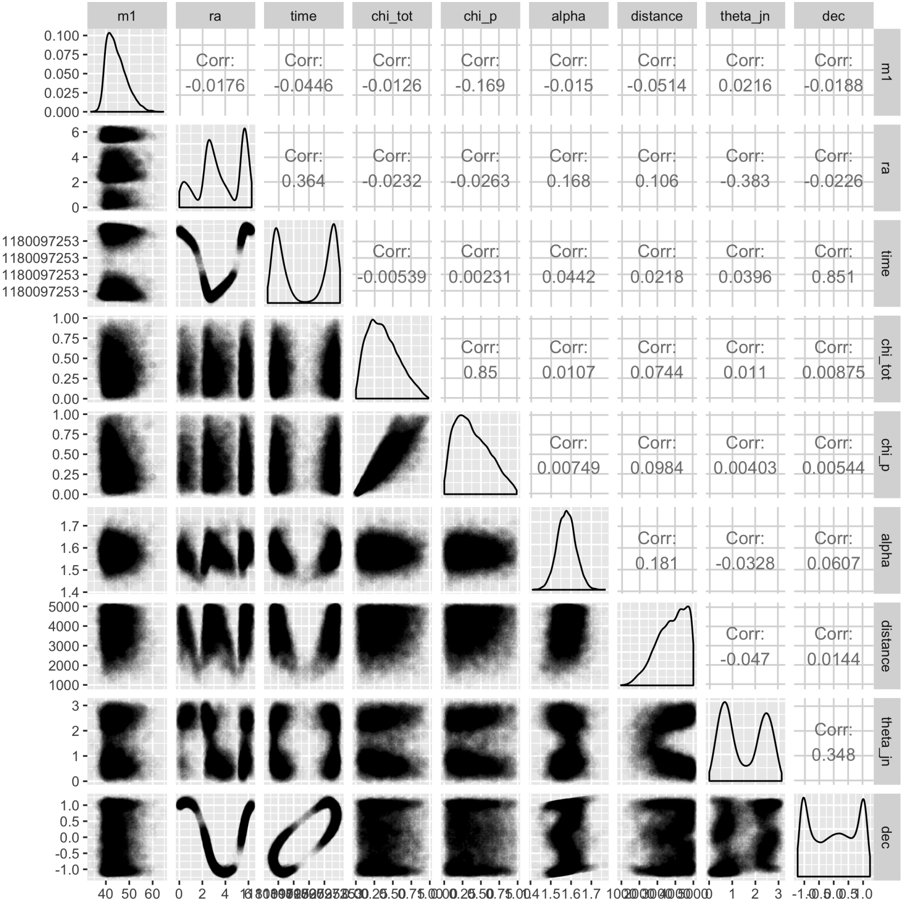

Fig. 8 Scatter plot matrix showing most of the variables included in the BBH dataset. Strong correlation between the parameters time, dec and ra can be observed.

References

Cook, Dianne, Andreas Buja, and Javier Cabrera. 1993. “Projection Pursuit Indexes Based on Orthonormal Function Expansions.” *Journal of Computational and Graphical Statistics* 2 (3): 225–50. `http://www.jstor.org/stable/1390644`.

Cook, Dianne, and Deborah F. Swayne. 2007. *Interactive and Dynamic Graphics for Data Analysis with R and Ggobi.* 1st ed. Springer Publishing Company, Incorporated.

Marsaglia, G. 1968. “Random Numbers Fall Mainly in the Planes.” *Proceedings of the National Academy of Science.*

Schloerke, Barret. 2016. *Geozoo: Zoo of Geometric Objects.* `https://CRAN.R-project.org/package=geozoo`.